
Methods of Multivariate Analysis

Second Edition

Abdel C. Rencher



WILEY SERIES IN
PROBABILITY AND STATISTICS

Methods of Multivariate Analysis

Second Edition

ALVIN C. RENCHER

Edgemon Young University

 **WILEY-
INTERSCIENCE**

A JOHN WILEY & SONS, INC. PUBLICATION

Methods of Multivariate Analysis

Second Edition

Methods of Multivariate Analysis

Second Edition

ALVIN C. RENCHER

Belgium Young University

 **WILEY-
INTERSCIENCE**

A JOHN WILEY & SONS, INC. PUBLICATION

This book is printed on acid-free paper. 

Copyright © 2002 by John Wiley & Sons, Inc. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0001, (212) 850-6000, fax (212) 850-6008, E-mail: PERMREQ@WILEY.COM.

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging-in-Publication Data

Bretz, John C., 1954-

Methods of multivariate analysis / John C. Bretz. — 2nd ed.

p. cm. — (Wiley series in probability and mathematical statistics)

1. Wiley-Interscience publication.

Includes bibliographical references and index.

ISBN 0-471-01897-2 (pbk.)

I. Multivariate analysis. II. Title. III. Series.

QA278.B83B264

519.3F5—dc22

2001m01897

Printed in the United States of America

0 4 7 1 0 1 8 9 7 2

Contents

1. Introduction	1
1.1 Why Multivariate Analysis?	1
1.2 Prerequisites	3
1.3 Objectives	3
1.4 Basic Types of Data and Analysis	3
2. Matrix Algebra	5
2.1 Introduction	5
2.2 Notation and Basic Definitions	5
2.2.1 Matrices, Vectors, and Scalars	5
2.2.2 Equality of Vectors and Matrices	7
2.2.3 Transpose and Symmetric Matrices	7
2.2.4 Special Matrices	8
2.3 Operations	9
2.3.1 Summation and Product Notation	9
2.3.2 Addition of Matrices and Vectors	10
2.3.3 Multiplication of Matrices and Vectors	11
2.4 Partitioned Matrices	20
2.5 Rank	23
2.6 Inverse	25
2.7 Positive Definite Matrices	25
2.8 Determinants	26
2.9 Trace	30
2.10 Orthogonal Vectors and Matrices	31
2.11 Eigenvalues and Eigenvectors	32
2.11.1 Definition	32
2.11.2 $\mathbf{I} + \mathbf{A}$ and $\mathbf{I} - \mathbf{A}$	33
2.11.3 $\sigma(\mathbf{A})$ and $ \mathbf{A} $	34
2.11.4 Positive Definite and Semidefinite Matrices	34
2.11.5 The Product $\mathbf{A}\mathbf{B}$	35
2.11.6 Symmetric Matrix	35

- 2.11.7 Spectral Decomposition, 35
- 2.11.8 Square-Root Matrix, 36
- 2.11.9 Square Matrices and Inverse Matrices, 36
- 2.11.10 Singular Value Decompositions, 38

3. Characterizing and Displaying Multivariate Data 43

- 3.1 Mean and Variance of a Univariate Random Variable, 43
- 3.2 Covariance and Correlation of Bivariate Random Variables, 49
 - 3.2.1 Covariance, 49
 - 3.2.2 Correlation, 49
- 3.3 Scatter Plots of Bivariate Samples, 50
- 3.4 Graphical Displays for Multivariate Samples, 52
- 3.5 Mean Vectors, 53
- 3.6 Covariance Matrices, 57
- 3.7 Correlation Matrices, 60
- 3.8 Mean Vectors and Covariance Matrices for Subsets of Variables, 62
 - 3.8.1 Two Subsets, 62
 - 3.8.2 Three or More Subsets, 64
- 3.9 Linear Combinations of Variables, 66
 - 3.9.1 Sample Properties, 68
 - 3.9.2 Population Properties, 72
- 3.10 Measures of Overall Variability, 75
- 3.11 Estimation of Missing Values, 74
- 3.12 Distances between Vectors, 76

4. The Multivariate Normal Distribution 82

- 4.1 Multivariate Normal Density Function, 82
 - 4.1.1 Univariate Normal Density, 82
 - 4.1.2 Multivariate Normal Density, 83
 - 4.1.3 Generalized Population Variance, 83
 - 4.1.4 Diversity of Applications of the Multivariate Normal, 83
- 4.2 Properties of Multivariate Normal Random Variables, 85
- 4.3 Estimation in the Multivariate Normal, 90
 - 4.3.1 Maximum Likelihood Estimation, 90
 - 4.3.2 Distribution of $\bar{Y}$ and S , 91
- 4.4 Assessing Multivariate Normality, 92
 - 4.4.1 Investigating Univariate Normality, 93
 - 4.4.2 Investigating Multivariate Normality, 95

4.5 Outliers, 99

4.5.1 Outliers in Univariate Samples, 100

4.5.2 Outliers in Multivariate Samples, 101

B. Tests on One or Two Mean Vectors, 112

112

5.1 Multivariate versus Univariate Tests, 112

5.2 Tests on μ with Σ Known, 1135.2.1 Review of Univariate Test for $H_0: \mu = \mu_0$ with σ Known, 1135.2.2 Multivariate Test for $H_0: \mu = \mu_0$ with Σ Known, 1145.3 Tests on μ When Σ Is Unknown, 1175.3.1 Review of Univariate t-Test for $H_0: \mu = \mu_0$ with σ Unknown, 1175.3.2 Hotelling's T^2 -Test for $H_0: \mu = \mu_0$ with Σ Unknown, 117

5.4 Comparing Two Mean Vectors, 121

5.4.1 Review of Univariate Two-Sample t-Test, 121

5.4.2 Multivariate Two-Sample T^2 -Test, 121

5.4.3 Likelihood Ratio Tests, 126

5.5 Tests on Individual Variables-Conditional on Rejection of M_0 by the T^2 -Test, 1265.6 Computation of T^2 , 1305.6.1 Obtaining T^2 from a SAS/PCNA Program, 1305.6.2 Obtaining T^2 from Multiple Regression, 130

5.7 Paired-Observations Test, 132

5.7.1 Univariate Case, 132

5.7.2 Multivariate Case, 134

5.8 Test for Additional Information, 136

5.9 Profile Analysis, 139

5.9.1 One-Sample Profile Analysis, 139

5.9.2 Two-Sample Profile Analysis, 140

6. Multivariate Analysis of Variance, 156

156

6.1 One-Way Models, 156

6.1.1 Univariate One-Way Analysis of Variance (ANOVA), 156

6.1.2 Multivariate One-Way Analysis of Variance Model (MANOVA), 156

6.1.3 "Wilks" Test Statistic, 161

6.1.4 Roy's Test, 164

6.1.5 Pillai and Lawley-Hotelling Tests, 166

- 6.1.6 Unbalanced One-Way MANOVA, 168
- 6.1.7 Summary of the Four Tests and Relationship to F^* , 168
- 6.1.8 Measures of Multivariate Association, 173
- 6.2 Comparisons of the Four MANOVA Test Statistics, 176
- 6.3 Contrasts, 178
 - 6.3.1 Univariate Contrasts, 178
 - 6.3.2 Multivariate Contrasts, 180
- 6.4 Tests on Individual Variables Following Rejection of H_0 by the Overall MANOVA Test, 183
- 6.5 Two-Way Classification, 186
 - 6.5.1 Review of Univariate Two-Way ANOVA, 186
 - 6.5.2 Multivariate Two-Way MANOVA, 188
- 6.6 Other Models, 189
 - 6.6.1 Higher Order Fixed Effects, 189
 - 6.6.2 Mixed Models, 196
- 6.7 Checking on the Assumptions, 198
- 6.8 Profile Analysis, 199
- 6.9 Repeated Measures Designs, 204
 - 6.9.1 Multivariate vs. Univariate Approach, 208
 - 6.9.2 One-Sample Repeated Measures Model, 208
 - 6.9.3 k-Sample Repeated Measures Model, 211
 - 6.9.4 Computation of Repeated Measures Tests, 212
 - 6.9.5 Repeated Measures with Two Within-Subjects Factors and One Between-Subjects Factor, 213
 - 6.9.6 Repeated Measures with Two Within-Subjects Factors and Two Between-Subjects Factors, 216
 - 6.9.7 Additional Topics, 221
- 6.10 Growth Curves, 221
 - 6.10.1 Growth Curve for One Sample, 221
 - 6.10.2 Growth Curves for Several Samples, 229
 - 6.10.3 Additional Topics, 230
- 6.11 Tests on a Selection, 231
 - 6.11.1 Test for Additional Information, 231
 - 6.11.2 Stepwise Selection of Variables, 233

7. Tests on Covariance Matrices

248

- 7.1 Introduction, 248
- 7.2 Testing a Specified Pattern for Σ , 248
 - 7.2.1 Testing $A\mathbf{b} : \Sigma = \Sigma_0$, 248

3.1.2	Testing Sphericity, 250	
3.1.3	Testing $H_0: \Sigma = \sigma^2 I (I = p(I + \rho J))$, 251	
3.3	Tests Comparing Covariance Matrices, 256	
3.3.1	Univariate Tests of Equality of Variances, 254	
3.3.2	Multivariate Tests of Equality of Covariance Matrices, 255	
3.4	Tests of Independence, 259	
3.4.1	Independence of Two Subvectors, 259	
3.4.2	Independence of Several Subvectors, 261	
3.4.3	Test for Independence of All Variables, 263	
8.	Discriminant Analysis: Description of Group Separation	270
8.1	Introduction, 270	
8.2	The Discriminant Function for Two Groups, 271	
8.3	Relationship between Two-Group Discriminant Analysis and Multiple Regression, 279	
8.4	Discriminant Analysis for Several Groups, 277	
8.4.1	Discriminant Functions, 277	
8.4.2	A Measure of Association for Discriminant Functions, 282	
8.5	Standardized Discriminant Functions, 282	
8.6	Tests of Significance, 284	
8.6.1	Tests for the Two-Group Case, 284	
8.6.2	Tests for the Several-Group Case, 285	
8.7	Interpretation of Discriminant Functions, 288	
8.7.1	Standardized Coefficients, 289	
8.7.2	Partial F -Values, 290	
8.7.3	Correlations between Variables and Discriminant Functions, 291	
8.7.4	Rotation, 291	
8.8	Scatter Plots, 291	
8.9	Stepwise Selection of Variables, 293	
9.	Classification Analysis: Allocation of Observations to Groups	299
9.1	Introduction, 299	
9.2	Classification into Two Groups, 300	
9.3	Classification into Several Groups, 304	
9.3.1	Equal Population Covariance Matrices: Linear Classification Functions, 304	
9.3.2	Unequal Population Covariance Matrices: Quadratic Classification Functions, 306	

8.4	Estimating Misclassification Rates, 307	
8.5	Improved Estimates of Error Rates, 309	
8.5.1	Partitioning the Sample, 310	
8.5.2	Holdout Method, 310	
8.6	Subset Selection, 311	
8.7	Nonparametric Procedures, 314	
8.7.1	Multinomial Data, 314	
8.7.2	Classification Based on Density Estimators, 315	
8.7.3	Nearest Neighbor Classification Rule, 318	
IX.	Multivariate Regression	321
10.1	Introduction, 322	
10.2	Multiple Regression: Fixed x 's, 323	
10.2.1	Model for Fixed x 's, 323	
10.2.2	Least Squares Estimation in the Fixed- x Model, 326	
10.2.3	An Estimator for σ^2 , 326	
10.2.4	The Model Corrected for Means, 327	
10.2.5	Hypothesis Tests, 329	
10.2.6	R^2 in Fixed- x Regression, 332	
10.2.7	Subset Selection, 333	
10.3	Multiple Regression: Random x 's, 335	
10.4	Multivariate Multiple Regression: Estimation, 337	
10.4.1	The Multivariate Linear Model, 337	
10.4.2	Least Squares Estimation in the Multivariate Model, 339	
10.4.3	Properties of Least Squares Estimators $\hat{B}$, 341	
10.4.4	An Estimator for Σ , 342	
10.4.5	Model-Corrected for Means, 342	
10.5	Multivariate Multiple Regression: Hypothesis Tests, 343	
10.5.1	Test of Overall Regression, 343	
10.5.2	Test on a Subset of the x 's, 347	
10.6	Measures of Association between the y 's and the x 's, 349	
10.7	Subset Selection, 351	
10.7.1	Stepwise Procedures, 351	
10.7.2	All Possible Subsets, 359	
10.8	Multivariate Regression: Random x 's, 358	
X.	Canonical Correlation	361
11.1	Introduction, 361	
11.2	Canonical Correlations and Canonical Variates, 363	

11.3	Properties of Canonical Correlations, 368	
11.4	Tests of Significance, 367	
11.4.1	Tests of No Relationship between the y 's and the x 's, 367	
11.4.2	Test of Significance of Successive Canonical Correlations after the First, 369	
11.5	Interpretation, 371	
11.5.1	Standardized Coefficients, 371	
11.5.2	Correlations between Variables and Canonical Variables, 373	
11.5.3	Rotation, 375	
11.5.4	Redundancy Analysis, 375	
11.6	Relationships of Canonical Correlation Analysis to Other Multivariate Techniques, 374	
11.6.1	Regression, 374	
11.6.2	MANCOV and Discriminant Analysis, 376	
12.	Principal Component Analysis	380
12.1	Introduction, 380	
12.2	Geometric and Algebraic Views of Principal Components, 381	
12.2.1	Geometric Approach, 381	
12.2.2	Algebraic Approach, 385	
12.3	Principal Components and Perpendicular Regression, 387	
12.4	Plotting of Principal Components, 389	
12.5	Principal Components from the Correlation Matrix, 393	
12.6	Deciding How Many Components to Retain, 397	
12.7	Information in the Last Few Principal Components, 403	
12.8	Interpretation of Principal Components, 401	
12.8.1	Special Patterns in S or R , 402	
12.8.2	Rotation, 403	
12.8.3	Correlations between Variables and Principal Components, 405	
12.9	Selection of Variables, 404	
13.	Factor Analysis	408
13.1	Introduction, 408	
13.2	Orthogonal Factor Model, 409	
13.2.1	Model Definition and Assumptions, 409	
13.2.2	Interpretations of Factor Loadings, 414	
13.3	Estimation of Loadings and Communalities, 415	
13.3.1	Principal Component Method, 415	
13.3.2	Principal Factor Method, 421	

13.3.3	Iterated Principal Factor Method, 428	
13.3.4	Maximum Likelihood Method, 425	
13.4	Choosing the Number of Factors, <i>in</i> , 426	
13.5	Rotation, 429	
13.5.1	Introduction, 430	
13.5.2	Orthogonal Rotation, 431	
13.5.3	Oblique Rotation, 435	
13.5.4	Interpretation, 438	
13.6	Factor Scores, 438	
13.7	Validity of the Factor Analysis Model, 440	
13.8	The Relationship of Factor Analysis to Principal Component Analysis, 447	
14.	Cluster Analysis	451
14.1	Introduction, 454	
14.2	Measures of Similarity or Dissimilarity, 457	
14.3	Hierarchical Clustering, 459	
14.3.1	Introduction, 459	
14.3.2	Single Linkage (Nearest Neighbor), 461	
14.3.3	Complete Linkage (Furthest Neighbor), 468	
14.3.4	Average Linkage, 467	
14.3.5	Centroid, 463	
14.3.6	Median, 465	
14.3.7	Ward's Method, 466	
14.3.8	Flexible Beta Method, 468	
14.3.9	Properties of Hierarchical Methods, 471	
14.3.10	Divisive Methods, 478	
14.4	Nonhierarchical Methods, 481	
14.4.1	Partitioning, 481	
14.4.2	Other Methods, 490	
14.5	Choosing the Number of Clusters, 484	
14.6	Cluster Validity, 498	
14.7	Clustering Variables, 497	
15.	Graphical Procedures	504
15.1	Multidimensional Scaling, 504	
15.1.1	Introduction, 504	
15.1.2	Metric Multidimensional Scaling, 509	
15.1.3	Nonmetric Multidimensional Scaling, 508	

15.2	Correspondence Analysis, 314	
15.2.1	Introduction, 314	
15.2.2	Row and Column Profiles, 315	
15.2.3	Testing Independence, 319	
15.2.4	Coordinates for Plotting Row and Column Profiles, 321	
15.2.5	Multiple Correspondence Analysis, 326	
15.3	Epistata, 331	
15.3.1	Introduction, 331	
15.3.2	Principal Component Plots, 331	
15.3.3	Singular Value Decomposition Plots, 331	
15.3.4	Coordinates, 333	
15.3.5	Other Methods, 335	
A.	Tables	349
B.	Answers and Hints to Problems	351
C.	Data Sets and SAS Files	379
	References	381
	Index	393

Preface

I have long been fascinated by the interplay of variables in multivariate data and by the challenge of unraveling the effect of each variable. My continuing objective in the second edition has been to present the power and utility of multivariate analysis in a highly readable format.

Practitioners and researchers in all applied disciplines often measure several variables on each subject or experimental unit. In some cases, it may be productive to isolate each variable in a system and study it separately. Typically, however, the variables are not only correlated with each other, but each variable is influenced by the other variables as it affects a test statistic or descriptive statistic. Thus, in many instances, the variables are intertwined in such a way that when analyzed individually they yield little information about the system. Using multivariate analysis, the variables can be examined simultaneously in order to assess the key features of the process that produced them. The multivariate approach enables us to (1) explore the joint performance of the variables and (2) determine the effect of each variable in the presence of the others.

Multivariate analysis provides both descriptive and inferential procedures—we can search for patterns in the data or test hypotheses about patterns of a priori interest. With multivariate descriptive techniques, we can peer beneath the tangled web of variables on the surface and extract the essence of the system. Multivariate inferential procedures include hypothesis tests that (1) process any number of variables without inflating the Type I error rate and (2) allow for whatever intercorrelations the variables possess. A wide variety of multivariate descriptive and inferential procedures is readily accessible in statistical software packages.

My selection of topics for this volume reflects many years of consulting with researchers in many fields of inquiry. A brief overview of multivariate analysis is given in Chapter 1. Chapter 2 reviews the fundamentals of matrix algebra. Chapters 3 and 4 give an introduction to sampling from multivariate populations. Chapters 5, 6, 7, 10, and 11 extend univariate procedures with one dependent variable (including *t*-tests, analysis of variance, tests on variances, multiple regression, and multiple correlation) to analogous multivariate techniques involving several dependent variables. A review of each univariate procedure is presented before covering the multivariate counterpart. These reviews may provide key insights for the student trained in previous courses.

Chapters 8, 9, 12, 13, 14, and 15 describe multivariate techniques that are not extensions of univariate procedures. In Chapters 8 and 9, we find functions of the variables that discriminate among groups in the data. In Chapters 12 and 13, we

find functions of the variables that reveal the basic dimensionality and characteristic patterns of the data, and we discuss procedures for finding the underlying latent variables of a system. In Chapters 14 and 15 (new in the second edition), we give methods for searching for groups in the data, and we provide plotting techniques that show relationships in a reduced dimensionality for various kinds of data.

In Appendix A, tables are provided for many multivariate distributions and tests. These enable the reader to conduct an exact test in many cases for which software packages provide only approximate tests. Appendix B gives answers and hints for most of the problems in the book.

Appendix C describes an ftp site that contains (1) all data sets and (2) SAS command files for all examples in the text. These command files can be adapted for use in teaching problems or in analyzing data sets encountered in applications.

To illustrate multivariate applications, I have provided many examples and exercises based on 50 real data sets from a wide variety of disciplines. A practitioner or consultant in multivariate analysis gains insights and assumes from long experience in working with data. It is not expected that a student can achieve this kind of expertise in a one-semester class. However, the examples provide a good start, and further development is gained by working problems with the data sets. For example, in Chapters 12 and 13, the exercises cover several typical patterns in the covariance or correlation matrix. The student's intuition is expanded by associating these covariance patterns with the resulting configuration of the principal components or factors.

Although this is a methods book, I have included a few derivations. For some students, an occasional proof provides insights obtainable in no other way. I hope that instructors who do not wish to use proofs will not be deterred by their presence. The proofs can be disregarded easily when reading the book.

My objective has been to make the book accessible to readers who have taken as few as two statistical methods courses. The students in my classes in multivariate analysis include majors in statistics and majors from other departments. With the applied researcher in mind, I have provided careful intuitive explanations of the concepts and have included many insights typically available only in journal articles or in the minds of practitioners.

My overriding goal in preparation of this book has been clarity of exposition. I hope that students and instructors alike will find this multivariate text more comfortable than most. In the final stages of development of both the first and second editions, I asked my students for written reports on their initial reaction as they read each day's assignment. They made many comments that led to improvements in the manuscript. I will be very grateful if readers will take the time to notify me of errors or of other suggestions they might have for improvements.

I have tried to use standard mathematical and statistical notation as far as possible and to maintain consistency of notation throughout the book. I have refrained from the use of abbreviations and mnemonic devices. These save space when one is reading a book page by page, but they are annoying in those using a book as a reference.

Equations are numbered sequentially throughout a chapter; for example, (3.79) indicates the 79th numbered equation in Chapter 3. Tables and figures are also num-

based sequentially throughout a chapter in the form “Table 3.8” or “Figure 3.1.” Examples are not numbered sequentially; each example is identified by the same number as the section in which it appears and is placed at the end of the section.

When citing references in the text, I have used the standard format involving the year of publication. For a journal article, the year alone suffices, for example, Fisher (1970). But for books, I have usually included a page number, as in Fisher (1984, p. 118).

This is the first volume of a two-volume set on multivariate analysis. The second volume is entitled *Multivariate Statistical Inference and Applications* (Wiley, 1998). The two volumes are not necessarily sequential; they can be read independently. I adopted the two-volume format in order to (1) provide broader coverage than would be possible in a single volume and (2) offer the reader a choice of approach.

The second volume includes proofs of many techniques covered in the first 15 chapters of the present volume and also introduces additional topics. The present volume includes many examples and problems using actual data sets, and there are fewer algebraic problems. The second volume emphasizes derivations of the results and contains fewer examples and problems with real data. The present volume has fewer references to the literature than the other volume, which includes a careful review of the latest developments and a more comprehensive bibliography. In this second edition, I have occasionally referred the reader to Bencher (1998) to note that added coverage of a certain subject is available in the second volume.

I am indebted to many individuals in the preparation of the first edition. My initial exposure to multivariate analysis came in courses taught by Rolf Hargmann at the University of Georgia and B. K. Jensen at Virginia Tech. Additional impetus to probe the subtleties of this field came from research conducted with Bruce Boewer at BYU. I wish to thank Bruce Boewer, Deane Brainerd, Del Scott, Robert Smith, and Ingram Olkin for reading various versions of the manuscript and making valuable suggestions. I am grateful to the following students at BYU who helped with computations and typing: Mitchell Tollard, Tawnia Newton, Marianne Marie Mahr, Gregg Littlefield, Suzanne Embert, Wendy Nielsen, Tiffany Nordgren, David Whiting, Katie Winkler, and Rachel Jones.

SECOND EDITION

For the second edition, I have added Chapters 14 and 15, covering cluster analysis, multidimensional scaling, correspondence analysis, and biploids. I also made numerous corrections and revisions (almost every page) in the first 13 chapters, in an effort to improve composition, readability, and clarity. Many of the first 13 chapters now have additional problems.

I have listed the data sets and SAS files on the Wiley Web site rather than on a diskette, as in the first edition. I have made improvements in labeling of these files.

I am grateful to the many readers who have pointed out errors or made suggestions for improvements. The book is better for their caring and their efforts.

I thank Loraine Stockard and Candace B. McNaughton for typing and J. D. Williams for computer support. As with my other books, I dedicate this volume to my wife, Lillian, who has supplied much needed support and encouragement.

ALVIN C. KIRSCHEN

Acknowledgments

I thank the authors, editors, and owners of copyrights for permission to reproduce the following materials:

- Figure 3.4 and Table 3.1, Krinner and Hartigan (1981), Reprinted by permission of *Journal of the American Statistical Association*
- Table 3.3, Krinner and Jensen (1999a), Reprinted by permission of *Journal of Quality Technology*
- Table 3.4, Rowers and Miller (1979), Reprinted by permission of *Diabetologia*
- Table 3.5, Tison (1979), Reprinted by permission of Elsevier North-Holland Publishing Company
- Table 3.6, Elston and Chisale (1992), Reprinted by permission of *Diabetica*
- Table 3.7, Fests (1921), Reprinted by permission of *Genetica*
- Table 3.8, O'Sullivan and Mahan (1985), Reprinted by permission of *American Journal of Clinical Nutrition*
- Table 4.1, Royston (1983), Reprinted by permission of *Applied Statistics*
- Table 5.1, Beall (1945), Reprinted by permission of *Psychometrika*
- Table 5.2, Hummel and Slips (1971), Reprinted by permission of *Psychological Bulletin*
- Table 5.3, Krinner and Jensen (1999b), Reprinted by permission of *Journal of Quality Technology*
- Table 5.5, Leitchow (1981), Reprinted by permission of *Biometrika*
- Table 5.6, Travers (1993), Reprinted by permission of *Psychometrika*
- Table 5.7, Andrews and Hensberg (1983), Reprinted by permission of Springer-Verlag
- Table 5.8, Thirner (1948), Reprinted by permission of *Journal of the American Statistical Association*
- Table 5.9, Krinner (1972), Reprinted by permission of the author
- Table 5.10, Cameron and Farling (1978), Reprinted by permission of National Academy of Science

- Table S.2, Andrews and Hunsberg (1983), Reprinted by permission of Springer-Verlag
- Table S.3, Raftery and Scott (1990), Reprinted by permission of *Communications in Statistics: Simulation and Computation*
- Table S.4, Paden (1905), Reprinted by permission of the author
- Table S.5, Crowder and Hand (1990, pp. 21–29), Reprinted by permission of Routledge Chapman and Hall
- Table S.12, Cochran and Cox (1957), Tinnis (1980), Reprinted by permission of John Wiley and Sons and Elsevier North-Holland Publishing Company
- Table S.14, Tinnis (1980), Reprinted by permission of Elsevier North-Holland Publishing Company
- Table S.15, Pothoff and Hey (1964), Reprinted by permission of Biometrika Truston
- Table S.17, Bates, Tick, and Rader (1958), Reprinted by permission of Quality Progress
- Table S.18, Katis et al. (1984), Reprinted by permission of *Journal of Statistical Software*
- Table S.19, Basilek (1975), Reprinted by permission of the author
- Table S.20, Ben (1950), Reprinted by permission of *Biometrics*
- Table S.21, Rao (1948), Reprinted by permission of *Biometrika Truston*
- Table S.22, Cameron and Pauling (1978), Reprinted by permission of National Academy of Science
- Table S.23, Williams and Lemons (1989), Reprinted by permission of Colorado State University
- Table S.24, Stanchak and Host (1974), Reprinted by permission of *Journal of Statistical Computation and Simulation*
- Table S.25, Ben (1950), Reprinted by permission of *Biometrics*
- Table S.26, Orlick and Allen (1955), Reprinted by permission of *Biometrics*
- Table S.27, Capon et al. (1985), Reprinted by permission of *Biometrics*
- Table S.28, Zerbe (1979a), Reprinted by permission of *Journal of the American Statistical Association*
- Table S.29, Tinnis (1980), Reprinted by permission of Elsevier North-Holland Publishing Company
- Table 7.1, Sarkar et al. (1983), Reprinted by permission of the Institute of Statistical Mathematics

- Table 7.1, Reprinted by permission of R. J. Freund
- Table 8.1, Krumer and Jonson (1959a), Reprinted by permission of *Journal of Quality Technology*
- Table 8.3, Reprinted by permission of G. R. Boyce and R. M. Statler
- Table 10.1, Box and Jenkins (1955), Reprinted by permission of *Biometrika*
- Tables 12.1, 12.3, and 12.4, Jeffries (1967), Reprinted by permission of *Applied Statistics*
- Table 13.1, Brown et al. (1984), Reprinted by permission of the *Journal of Plant, Soil, and Weather*
- Correlation matrix in Example 13.6, Brown, Strong, and Bancher (1973), Reprinted by permission of *The Journal of the Agricultural Society of America*
- Table 14.1, Hartigan (1975), Reprinted by permission of John Wiley and Sons
- Table 14.3, Durkin (1987), Reprinted by permission of *The American Statistician*
- Table 14.7, Hand et al. (1994), Reprinted by permission of D. J. Hand
- Table 14.12, Finkel and Rohlf (1981), Reprinted by permission of W. H. Freeman and Co.
- Table 14.13, Hand et al. (1994), Reprinted by permission of D. J. Hand
- Table 15.1, Kruskal and Wish (1978), Reprinted by permission of Sage Publications
- Tables 15.2 and 15.3, Hand et al. (1994), Reprinted by permission of D. J. Hand
- Table 15.13, Edwards and Kravitz (1983), Reprinted by permission of *Biometrika*
- Table 15.15, Hand et al. (1994), Reprinted by permission of D. J. Hand
- Table 15.16, Green (1987), Reprinted by permission of the author
- Table 15.17, Andrews and Herzberg (1985), Reprinted by permission of Springer Verlag
- Table 15.18, Clausen (1988), Reprinted by permission of Sage Publications
- Table 15.19, Andrews and Herzberg (1985), Reprinted by permission of Springer Verlag
- Table A.1, Møller-Hand (1977), Reprinted by permission of *Biometrika Trustee*
- Table A.2, D'Agostino and Pearson (1973), Reprinted by permission of *Biometrika Trustee*
- Table A.3, D'Agostino and Tibshirani (1971), Reprinted by permission of *Biometrika Trustee*

- Table A.4, D'Agostino (1971), Reprinted by permission of Biometrika Trustee
- Table A.5, Murda (1970, 1974), Reprinted by permission of Biometrika Trustee
- Table A.6, Barnett and Lewis (1978), Reprinted by permission of John Wiley and Sons
- Table A.7, Krumer and Jensen (1984), Reprinted by permission of *Journal of Quality Technology*
- Table A.8, Bailey (1977), Reprinted by permission of *Journal of the American Statistical Association*
- Table A.9, Wall (1963), Reprinted by permission of the author, *Albuquerque, NM*
- Table A.10, Pearson and Hartley (1972) and Pillai (1964, 1965), Reprinted by permission of Biometrika Trustee
- Table A.11, Schmeers et al. (1975), Reprinted by permission of *Journal of Statistical Computation and Simulation*
- Table A.12, Davis (1970a,b, 1980), Reprinted by permission of Biometrika Trustee
- Table A.13, Krishnan, Rappin, and Muller (1983), Reprinted by permission of PWS-KENT Publishing Company
- Table A.14, Lee et al. (1977), Reprinted by permission of Elsevier North-Holland Publishing Company
- Table A.15, Matal and Kariyar (1979), Reprinted by permission of Biometrika Trustee

CHAPTER 1

Introduction

1.1 WHY MULTIVARIATE ANALYSES?

Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. We will refer to the measurements as variables and to the individuals or objects as units (research units, sampling units, or experimental units) or observations. In practice, multivariate data sets are common, although they are not always analyzed as such, but the exclusive use of univariate procedures with such data is no longer reasonable, given the availability of multivariate techniques and inexpensive computing power to carry them out.

Historically, the bulk of applications of multivariate techniques have been in the behavioral and biological sciences. However, interest in multivariate methods has now spread to numerous other fields of investigation. For example, I have collaborated on multivariate problems with researchers in education, chemistry, physics, geology, engineering, law, business, literature, religion, public broadcasting, mining, mining, linguistics, biology, psychology, and many other fields. Table 1.1 shows some examples of multivariate observations.

The major distinction that in some cases all the variables are measured in the same scale (see 1 and 2 in Table 1.1). In other cases, measurements are in different scales (see 3 in Table 1.1). In a few techniques, such as profile analysis (Sections 5.9 and 6.8), the variables must be commensurate, that is, similar in scale of measurement; however, most multivariate methods do not require this.

Ordinarily the variables are measured simultaneously on each sampling unit. Typically, these variables are correlated. If this were not so, there would be little use for many of the techniques of multivariate analysis. We need to wrangle the overlapping information provided by correlated variables and peer beneath the surface to see the underlying structure. Thus the goal of many multivariate approaches is simplification. We seek to express what is going on in terms of a reduced set of dimensions. Such multivariate techniques are exploratory; they essentially generate hypotheses rather than test them.

On the other hand, if our goal is a formal hypothesis test, we need a technique that will (1) allow several variables to be tested and still preserve the significance level

Table 1.1. Examples of Multivariate Data

Units	Variables
1. Students	Several exam scores less single exams
2. Students	Grades in mathematics, history, music, art, physics
3. People	Height, weight, percentage of body fat, resting heart rate
4. Adults	Length, width, dental capacity
5. Companies	Expenditures for advertising, labor, raw materials
6. Manufactured items	Various measurements to check on compliance with specifications
7. Applicants for bank loans	Income, education level, length of residence, savings account, current debt load
8. Segments of literature	Sentence length, frequency of usage of certain words and of stylistic ornamentation
9. Human faces	Composition of features, distance
10. Birds	Length of various bones

and (2) do this for any intercorrelation structure of the variables. Many such tests are available.

As the two preceding paragraphs imply, multivariate analysis is concerned generally with two areas, descriptive and inferential statistics. In the descriptive realm, we often obtain optimal linear combinations of variables. The optimality criterion varies from one technique to another, depending on the goal in each case. Although linear combinations may seem too simple to reveal the underlying structure, we use them for two obvious reasons: (1) they have mathematical tractability (linear approximations are used throughout all science for the same reason) and (2) they often perform well in practice. These linear functions may also be useful as a follow-up to inferential procedures. When we have a statistically significant test result that compares several groups, for example, we can find the linear combination (or combinations) of variables that led to rejection of the hypothesis. Then the contribution of each variable to these linear combinations is of interest.

In the inferential area, many multivariate techniques are extensions of univariate procedures. In such cases, we follow the univariate procedure before presenting the analogous multivariate approach.

Multivariate inference is especially useful in curbing the researcher's natural tendency to read too much into the data. Total control is provided for experimenters: error rate, that is, no matter how many variables are tested simultaneously, the value of α (the significance level) remains at the level set by the researcher.

Some authors were against applying the common multivariate techniques to data for which the measurement scale is not interval or ratio. It has been found, however, that many multivariate techniques give reliable results when applied to ordinal data.

For many years the applications lagged behind the theory because the computations were beyond the power of the available desktop calculations. However, with modern computers, virtually any analysis one desires, no matter how many variables

or observations are involved, can be quickly and easily carried out. Perhaps it is not premature to say that multivariate analysis has come of age.

1.2 PREREQUISITES

The mathematical prerequisites for reading this book is matrix algebra. Calculus is not used (with a brief exception in equation (4.19)). But the basic tools of matrix algebra are essential, and the presentation in Chapter 2 is intended to be as thoroughly complete as that the reader with no previous experience can master matrix manipulation up to the level required in this book.

The statistical prerequisites are basic familiarity with the normal distribution, t -tests, confidence intervals, multiple regression, and analysis of variance. These techniques are reviewed as such is extended to the analogous multivariate procedures.

This is a multivariate methods text. Most of the results are given without proof. In a few cases proofs are provided, but the major emphasis is on heuristic explanations. Our goal is an intuitive grasp of multivariate analysis, in the same mode as other statistical methods courses. Some problems are algebraic in nature, but the majority involve data sets to be analyzed.

1.3 OBJECTIVES

I have formulated three objectives that I hope this book will achieve for the reader. These objectives are based on long experience teaching a course in multivariate methods, consulting on multivariate problems with researchers in many fields, and guiding statistics graduate students as they come faced with similar clients.

The first objective is to gain a thorough understanding of the details of various multivariate techniques, their purposes, their assumptions, their limitations, and so on. Many of these techniques are related, yet they differ in some essential ways. We emphasize these similarities and differences.

The second objective is to be able to select one or more appropriate techniques for a given multivariate data set. Recognizing the essential nature of a multivariate data set is the first step in a meaningful analysis. We introduce basic types of multivariate data in Section 1.4.

The third objective is to be able to interpret the results of a computer analysis of a multivariate data set. Reading the manual for a particular program package is not enough to make an intelligent appraisal of the output. Achievement of the first objective and practice on data sets in the text should help achieve the third objective.

1.4 BASIC TYPES OF DATA AND ANALYSIS

We will list four basic types of (continuous) multivariate data and then briefly describe some possible analyses. Some writers would consider this an oversimpli-

location and might prefer elaborate tree diagrams of data structure. However, many data sets can fit into one of these categories, and the simplicity of this structure makes it easier to remember. The four basic data types are as follows:

1. A single sample with several variables measured on each sampling unit (subject or object).
2. A single sample with two sets of variables measured on each unit.
3. Two samples with several variables measured on each unit.
4. Three or more samples with several variables measured on each unit.

Each data type has extensions, and various combinations of the four are possible. A few examples of analyses for each case are as follows:

1. A single sample with several variables measured on each sampling unit:
 - (a) Test the hypothesis that the means of the variables have specified values.
 - (b) Test the hypothesis that the variables are uncorrelated and have a common variance.
 - (c) Find a small set of linear combinations of the original variables that summarizes most of the variation in the data (principal-components).
 - (d) Express the original variables as linear functions of a smaller set of underlying variables that account for the original variables and their interrelations (factor analysis).
2. A single sample with two sets of variables measured on each unit:
 - (a) Determine the number, the size, and the nature of relationships between the two sets of variables (canonical correlation). For example, you may wish to relate a set of interest variables to a set of achievement variables. How much overall correlation is there between these two sets?
 - (b) Find a model to predict one set of variables from the other set (multivariate multiple regression).
3. Two samples with several variables measured on each unit:
 - (a) Compare the means of the variables across the two samples (Hotelling's T^2 -test).
 - (b) Find a linear combination of the variables that best separates the two samples (discriminant analysis).
 - (c) Find a function of the variables that accurately allocates the units into the two groups (classification analysis).
4. Three or more samples with several variables measured on each unit:
 - (a) Compare the means of the variables across the groups (multivariate analysis of variance).
 - (b) Extension of MAN to more than two groups.
 - (c) Extension of MCA to more than two groups.

CHAPTER 2

Matrix Algebra

2.1 INTRODUCTION

This chapter introduces the basic elements of matrix algebra used in the remainder of this book. It is essentially a review of the requisite matrix tools and is not intended to be a complete development. However, it is sufficiently self-contained so that those with no previous exposure to the subject should need no other reference. Anyone unfamiliar with matrix algebra should plan to work most of the problems containing numerical illustrations. It would also be helpful to explore some of the problems involving general matrix manipulation.

With the exception of a few derivations that seemed instructive, most of the results are given without proof. Some additional proofs are requested in the problems. For the remaining proofs, see any general text on matrix theory or one of the specialized matrix texts oriented to statistics, such as Graybill (1961), Searls (1982), or Harville (1987).

2.2 NOTATION AND BASIC DEFINITIONS

2.2.1 Matrices, Vectors, and Scalars

A matrix is a rectangular or square array of numbers or variables arranged in rows and columns. We use uppercase/boldface letters to represent matrices. All entries in matrices will be real numbers or variables representing real numbers. The elements of a matrix are displayed in brackets. For example, the ACT score and GPA for three students can be conveniently listed in the following matrix:

$$\mathbf{A} = \begin{pmatrix} 25 & 3.54 \\ 29 & 3.81 \\ 18 & 2.79 \end{pmatrix}. \quad (2.1)$$

The elements of $\mathbf{A}$ can also be variables, representing possible values of ACT and GPA for three students.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}. \quad (2.2)$$

In this double-subscript notation for the elements of a matrix, the first subscript indicates the row; the second identifies the column. The matrix $\mathbf{A}$ in (2.2) can also be expressed as

$$\mathbf{A} = (a_{ij}), \quad (2.3)$$

where a_{ij} is a general element.

With three rows and two columns, the matrix $\mathbf{A}$ in (2.1) or (2.2) is said to be 3×2 . In general, if a matrix $\mathbf{A}$ has m rows and p columns, it is said to be $m \times p$. Alternatively, we say the *size* of $\mathbf{A}$ is $m \times p$.

A vector is a matrix with a single column or row. The following could be the test scores of a student in a course in multivariate analysis:

$$\mathbf{x} = \begin{pmatrix} 88 \\ 95 \\ 93 \\ 87 \end{pmatrix}. \quad (2.4)$$

Variable elements in a vector can be identified by a single subscript:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}. \quad (2.5)$$

We use lowercase boldface letters for column vectors. Row vectors are expressed as

$$\mathbf{x}' = (x_1, x_2, x_3, x_4) \quad \text{or} \quad \mathbf{x}' = (x_1 \quad x_2 \quad x_3 \quad x_4),$$

where $\mathbf{x}'$ indicates the transpose of $\mathbf{x}$. The transpose operation is defined in Section 2.1.3.

Geometrically, a vector with p elements identifies a point in a p -dimensional space. The elements in the vector are the coordinates of the point. In (2.5) in Section 1.1.3, we define the distance from the origin to the point. In Section 3.12, we define the distance between two vectors. In some cases, we will be interested in a directed line segment or arrow from the origin to the point.

A single real number is called a *scalar*; to distinguish it from a vector or matrix. Thus 2, -4 , and 1.25 are scalars. A variable representing a scalar is usually denoted by a lowercase unbolded letter, such as $a = 2$. A product involving vectors and matrices may reduce to a matrix of size 1×1 , which then becomes a scalar.

2.1.2 Equality of Vectors and Matrices

Two matrices are equal if they are the same size and the elements in corresponding positions are equal. Thus if $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$, then $\mathbf{A} = \mathbf{B}$ if $a_{ij} = b_{ij}$ for all i and j . For example, let

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} 3 & -2 & 4 \\ 1 & 5 & 7 \end{pmatrix}, & \mathbf{B} &= \begin{pmatrix} 3 & 1 \\ -2 & 5 \\ 4 & 7 \end{pmatrix}, \\ \mathbf{C} &= \begin{pmatrix} 3 & -2 & 4 \\ 1 & 5 & 7 \end{pmatrix}, & \mathbf{D} &= \begin{pmatrix} 3 & -2 & 4 \\ 1 & 3 & 8 \end{pmatrix}.\end{aligned}$$

Then $\mathbf{A} = \mathbf{C}$. But even though $\mathbf{A}$ and $\mathbf{B}$ have the same elements, $\mathbf{A} \neq \mathbf{B}$ because the two matrices are not the same size. Likewise, $\mathbf{A} \neq \mathbf{D}$ because $a_{23} \neq d_{23}$. Thus two matrices of the same size are unequal if they differ in a single position.

2.1.3 Transpose and Symmetric Matrices

The transpose of a matrix $\mathbf{A}$, denoted by $\mathbf{A}'$, is obtained from $\mathbf{A}$ by interchanging rows and columns. Thus the columns of $\mathbf{A}'$ are the rows of $\mathbf{A}$, and the rows of $\mathbf{A}'$ are the columns of $\mathbf{A}$. The following examples illustrate the transpose of a matrix or vector:

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} -5 & 2 & 4 \\ 3 & 6 & -2 \end{pmatrix}, & \mathbf{A}' &= \begin{pmatrix} -5 & 3 \\ 2 & 6 \\ 4 & -2 \end{pmatrix}, \\ \mathbf{B} &= \begin{pmatrix} 3 & -3 \\ 4 & 1 \end{pmatrix}, & \mathbf{B}' &= \begin{pmatrix} 3 & 4 \\ -3 & 1 \end{pmatrix}, \\ \mathbf{a} &= \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}, & \mathbf{a}' &= (2, \quad -3, \quad 1).\end{aligned}$$

The transpose operation does not change a scalar, since it has only one row and one column.

If the transpose operator is applied twice to any matrix, the result is the original matrix:

$$(\mathbf{A}')' = \mathbf{A}. \quad (2.6)$$

If the transpose of a matrix is the same as the original matrix, the matrix is said to be symmetric; that is, $\mathbf{A}$ is symmetric if $\mathbf{A} = \mathbf{A}'$. For example,

$$\mathbf{A} = \begin{pmatrix} 3 & -2 & 4 \\ -2 & 10 & -7 \\ 4 & -7 & 9 \end{pmatrix}, \quad \mathbf{A}' = \begin{pmatrix} 3 & -2 & 4 \\ -2 & 10 & -7 \\ 4 & -7 & 9 \end{pmatrix}.$$

Clearly, all symmetric matrices are square.

2.1.4 Special Matrices

The diagonal of a $p \times p$ square matrix $\mathbf{A}$ consists of the elements $a_{11}, a_{22}, \dots, a_{pp}$. For example, in the matrix

$$\mathbf{A} = \begin{pmatrix} 5 & -2 & 4 \\ 7 & 8 & 3 \\ -6 & 8 & 1 \end{pmatrix},$$

the elements 5, 8, and 1 lie on the diagonal. If a matrix contains zeros in all off-diagonal positions, it is said to be a diagonal matrix. An example of a diagonal matrix is

$$\mathbf{D} = \begin{pmatrix} 18 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7 \end{pmatrix}.$$

This matrix can also be denoted as

$$\mathbf{D} = \text{diag}(18, -3, 0, 7). \quad (2.7)$$

A diagonal matrix can be formed from any square matrix by replacing off-diagonal elements by 0's. This is denoted by $\text{diag}(\mathbf{A})$. Thus for the preceding matrix $\mathbf{A}$, we have

$$\text{diag}(\mathbf{A}) = \text{diag} \left(\begin{pmatrix} 5 & -2 & 4 \\ 7 & 8 & 3 \\ -6 & 8 & 1 \end{pmatrix} \right) = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (2.8)$$

A diagonal matrix with a 1 in each diagonal position is called an identity matrix and is denoted by $\mathbf{I}$. For example, a 3×3 identity matrix is given by

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (2.9)$$

An upper triangular matrix is a square matrix with zeros below the diagonal, such as

$$\mathbf{T} = \begin{pmatrix} 8 & 3 & 4 & 7 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 6 \end{pmatrix}. \quad (2.10)$$

A lower triangular matrix is defined similarly.

A vector of 1 's is denoted by $\mathbf{j}$:

$$\mathbf{j} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \quad (2.1.1)$$

A square matrix of 1 's is denoted by $\mathbf{J}$. For example, a 3×3 matrix $\mathbf{J}$ is given by

$$\mathbf{J} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}. \quad (2.1.2)$$

Finally, we denote a vector of zeros by $\mathbf{0}$ and a matrix of zeros by $\mathbf{O}$. For example,

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{O} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (2.1.3)$$

2.2 OPERATIONS

2.2.1 Summation and Product Notation

For completeness, we review the standard mathematical notation for sums and products. The sum of a sequence of numbers $a_1, a_2, \dots, a_n$ is indicated by

$$\sum_{i=1}^n a_i = a_1 + a_2 + \dots + a_n.$$

If the a numbers are all the same, that $\sum_{i=1}^n a = a + a + \dots + a = na$. The sum of all the numbers in an array with double subscripts, such as

$$\begin{array}{ccc} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{array}$$

is indicated by

$$\sum_{i=1}^2 \sum_{j=1}^3 a_{ij} = a_{11} + a_{12} + a_{13} + a_{21} + a_{22} + a_{23}.$$

This is sometimes abbreviated to

$$\sum_{i=1}^2 \sum_{j=1}^3 a_{ij} = \sum_{ij} a_{ij}.$$

The product of a sequence of numbers $a_1, a_2, \dots, a_n$ is indicated by

$$\prod_{i=1}^n a_i = (a_1)(a_2)\cdots(a_n).$$

If the n numbers are all equal, the product becomes $\prod_{i=1}^n a = (a)(a)\cdots(a) = a^n$.

2.3.2 Addition of Matrices and Vectors

If two matrices (or two vectors) are the same size, their sum is found by adding corresponding elements; that is, if $\mathbf{A}$ is $n \times p$ and $\mathbf{B}$ is $n \times p$, then $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is also $n \times p$ and is found as $(c_{ij}) = (a_{ij} + b_{ij})$. For example,

$$\begin{pmatrix} -2 & 5 \\ 3 & 1 \\ 7 & -6 \end{pmatrix} + \begin{pmatrix} 3 & -2 \\ 4 & 3 \\ 10 & -3 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 7 & 4 \\ 17 & -9 \end{pmatrix},$$

$$\begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix} + \begin{pmatrix} 3 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 10 \end{pmatrix}.$$

Similarly, the difference between two matrices or two vectors of the same size is found by subtracting corresponding elements. Thus $\mathbf{C} = \mathbf{A} - \mathbf{B}$ is found as $(c_{ij}) = (a_{ij} - b_{ij})$. For example,

$$\begin{pmatrix} 3 & 7 & -4 \end{pmatrix} - \begin{pmatrix} 5 & -4 & 2 \end{pmatrix} = \begin{pmatrix} -2 & 11 & -6 \end{pmatrix}.$$

If two matrices are identical, their difference is a zero-matrix; that is, $\mathbf{A} = \mathbf{B}$ implies $\mathbf{A} - \mathbf{B} = \mathbf{0}$. For example,

$$\begin{pmatrix} 3 & -2 & 4 \\ 6 & 7 & 3 \end{pmatrix} - \begin{pmatrix} 3 & -2 & 4 \\ 6 & 7 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Matrix addition is commutative:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}. \quad (2.14)$$

The transpose of the sum (difference) of two matrices is the sum (difference) of the transposes:

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}', \quad (2.15)$$

$$(\mathbf{A} - \mathbf{B})' = \mathbf{A}' - \mathbf{B}', \quad (2.16)$$

$$(\mathbf{x} + \mathbf{y})' = \mathbf{x}' + \mathbf{y}', \quad (2.17)$$

$$(\mathbf{x} - \mathbf{y})' = \mathbf{x}' - \mathbf{y}'. \quad (2.18)$$

2.3.3 Multiplication of Matrices and Vectors

In order for the product $\mathbf{AB}$ to be defined, the number of columns in $\mathbf{A}$ must be the same as the number of rows in $\mathbf{B}$, in which case $\mathbf{A}$ and $\mathbf{B}$ are said to be *conformable*. Then the (i, j) element of $\mathbf{C} = \mathbf{AB}$ is

$$c_{ij} = \sum_{k=1}^p a_{ik}b_{kj}. \quad (2.18)$$

Thus c_{ij} is the sum of products of the i th row of $\mathbf{A}$ and the j th column of $\mathbf{B}$. We therefore multiply each row of $\mathbf{A}$ by each column of $\mathbf{B}$, and the size of $\mathbf{AB}$ consists of the number of rows of $\mathbf{A}$ and the number of columns of $\mathbf{B}$. Thus, if $\mathbf{A}$ is $n \times m$ and $\mathbf{B}$ is $m \times p$, then $\mathbf{C} = \mathbf{AB}$ is $n \times p$. For example, if

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 3 \\ 4 & 6 & 5 \\ 7 & 2 & 5 \\ 1 & 3 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 3 & 4 \\ 2 & 6 \\ 5 & 8 \end{pmatrix},$$

then

$$\begin{aligned} \mathbf{C} = \mathbf{AB} &= \begin{pmatrix} 2 \cdot 3 + 1 \cdot 2 + 3 \cdot 5 & 2 \cdot 4 + 1 \cdot 6 + 3 \cdot 8 \\ 4 \cdot 3 + 6 \cdot 2 + 5 \cdot 5 & 4 \cdot 4 + 6 \cdot 6 + 5 \cdot 8 \\ 7 \cdot 3 + 2 \cdot 2 + 5 \cdot 5 & 7 \cdot 4 + 2 \cdot 6 + 5 \cdot 8 \\ 1 \cdot 3 + 3 \cdot 2 + 2 \cdot 5 & 1 \cdot 4 + 3 \cdot 6 + 2 \cdot 8 \end{pmatrix} \\ &= \begin{pmatrix} 19 & 38 \\ 54 & 62 \\ 50 & 68 \\ 19 & 38 \end{pmatrix}. \end{aligned}$$

Note that $\mathbf{A}$ is 4×3 , $\mathbf{B}$ is 3×2 , and $\mathbf{AB}$ is 4×2 . In this case, $\mathbf{AB}$ is of a different size than either $\mathbf{A}$ or $\mathbf{B}$.

If $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$, then $\mathbf{AB}$ is also $n \times n$. Clearly, $\mathbf{A}^2$ is defined only if $\mathbf{A}$ is a square matrix.

In some cases $\mathbf{AB}$ is defined, but $\mathbf{BA}$ is not defined. In the preceding example, $\mathbf{BA}$ cannot be found because $\mathbf{B}$ is 3×2 and $\mathbf{A}$ is 4×3 and a row of $\mathbf{B}$ cannot be multiplied by a column of $\mathbf{A}$. Sometimes $\mathbf{AB}$ and $\mathbf{BA}$ are both defined but are different in size. For example, if $\mathbf{A}$ is 2×4 and $\mathbf{B}$ is 4×2 , then $\mathbf{AB}$ is 2×2 and $\mathbf{BA}$ is 4×4 . If $\mathbf{A}$ and $\mathbf{B}$ are square and the same size, then $\mathbf{AB}$ and $\mathbf{BA}$ are both defined. However,

$$\mathbf{AB} \neq \mathbf{BA}, \quad (2.20)$$

except for a few special cases. For example, let

$$\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & -2 \\ 3 & 5 \end{pmatrix}.$$

Then

$$AB = \begin{pmatrix} 18 & 18 \\ 14 & 18 \end{pmatrix}, \quad BA = \begin{pmatrix} -3 & -3 \\ 15 & 29 \end{pmatrix}.$$

Thus we must be careful to specify the order of multiplication. If we wish to multiply both sides of a matrix equation by a matrix, we must multiply *on the left* or *on the right* and be consistent on both sides of the equation.

Multiplication is distributive over addition or subtraction:

$$A(B + C) = AB + AC, \quad (2.20)$$

$$A(B - C) = AB - AC, \quad (2.21)$$

$$(A + B)C = AC + BC, \quad (2.22)$$

$$(A - B)C = AC - BC. \quad (2.23)$$

Note that, in general, because of (2.20),

$$A(B + C) \neq BA + CA. \quad (2.24)$$

Using the distributive law, we can expand products such as $(A - B)(C + D)$ to obtain

$$\begin{aligned} (A - B)(C + D) &= (A - B)C + (A - B)D && \text{[by (2.22)]} \\ &= AC - BC + AD - BD && \text{[by (2.23)]}. \end{aligned} \quad (2.25)$$

The transpose of a product is the product of the transposes in reverse order:

$$(AB)^T = B^T A^T. \quad (2.26)$$

Note that (2.26) holds as long as A and B are conformable. They need not be square.

Multiplication involving vectors follows the same rules as for matrices. Suppose A is $m \times p$, u is $p \times 1$, v is $p \times 1$, w is $p \times 1$, and x is $m \times 1$. Then some possible products are Av , $v^T A$, $u^T B$, $B^T u$, and Aw . For example, let

$$A = \begin{pmatrix} 3 & -2 & 4 \\ 1 & 3 & 2 \end{pmatrix}, \quad u = \begin{pmatrix} 3 \\ -2 \\ 2 \end{pmatrix}, \quad v = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}, \quad x = \begin{pmatrix} 2 \\ 3 \\ -5 \end{pmatrix}.$$

Then

$$Av = \begin{pmatrix} 3 & -2 & 4 \\ 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 18 \\ 20 \end{pmatrix}.$$

$$c'A = (2 \quad -2) \begin{pmatrix} 2 & -2 & 4 \\ 1 & 1 & 3 \end{pmatrix} = (1 \quad -18 \quad -17),$$

$$c'Ab = (2 \quad -2) \begin{pmatrix} 2 & -2 & 4 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = (2 \quad -2) \begin{pmatrix} 36 \\ 31 \end{pmatrix} = -123,$$

$$a'b = (2 \quad -2) \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = 8,$$

$$b'a = (2 \quad 3 \quad 4) \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} = 8,$$

$$ab' = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} (2 \quad 3 \quad 4) = \begin{pmatrix} 2 & 3 & 4 \\ -4 & -6 & -8 \\ 6 & 9 & 12 \end{pmatrix},$$

$$a'b' = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} (2 \quad -2) = \begin{pmatrix} 2 & -2 \\ -4 & 10 \\ 6 & -18 \end{pmatrix}.$$

Note that Ab is a column vector, $c'A$ is a row vector, $c'Ab$ is a scalar, and $a'b = b'a$. The triple product $c'Ab$ was obtained as $c'(Ab)$. The same result would be obtained if we multiplied in the order $(c'A)b$:

$$(c'A)b = (1 \quad -12 \quad -17) \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = -123.$$

This is true in general for a triple product:

$$A(BC) = (AB)C. \quad (2.28)$$

Thus multiplication of three matrices can be defined in terms of the product of two matrices, since (fortunately) it does not matter which two are multiplied first. Note that A and B must be conformable for multiplication, and B and C must be conformable. For example, if A is $n \times p$, B is $p \times q$, and C is $q \times m$, then both multiplications are possible and the product $A(BC)$ is $n \times m$.

We can sometimes factor a sum of triple products on both the right and left sides. For example,

$$A(BC) + A(B'C) = A(B + B')C. \quad (2.29)$$

As another illustration, let X be $n \times p$ and A be $n \times m$. Then

$$X'X - X'AX = X'(X - AX) = X'(I - A'X). \quad (2.30)$$

If $\mathbf{a}$ and $\mathbf{b}$ are both $n \times 1$, then

$$\mathbf{a}^T \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \quad (2.30)$$

is a sum of products and is a scalar. On the other hand, $\mathbf{ab}^T$ is defined for any size $\mathbf{a}$ and $\mathbf{b}$ and is a matrix, either rectangular or square:

$$\mathbf{ab}^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} (b_1 \ b_2 \ \cdots \ b_p) = \begin{pmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_p \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_p \\ \vdots & \vdots & \cdots & \vdots \\ a_n b_1 & a_n b_2 & \cdots & a_n b_p \end{pmatrix}. \quad (2.32)$$

Similarly,

$$\mathbf{a}^T \mathbf{a} = a_1^2 + a_2^2 + \cdots + a_n^2, \quad (2.33)$$

$$\mathbf{aa}^T = \begin{pmatrix} a_1^2 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & a_2^2 & \cdots & a_2 a_n \\ \vdots & \vdots & \cdots & \vdots \\ a_n a_1 & a_n a_2 & \cdots & a_n^2 \end{pmatrix}. \quad (2.34)$$

Thus $\mathbf{a}^T \mathbf{a}$ is a sum of squares, and $\mathbf{aa}^T$ is a square (symmetric) matrix. The products $\mathbf{a}^T \mathbf{a}$ and $\mathbf{aa}^T$ are sometimes referred to as the *dot product* and *matrix product*, respectively. The square root of the sum of squares of the elements of $\mathbf{a}$ is the distance from the origin to the point $\mathbf{a}$ and is also referred to as the *length* of $\mathbf{a}$:

$$\text{Length of } \mathbf{a} = \sqrt{\mathbf{a}^T \mathbf{a}} = \sqrt{\sum_{i=1}^n a_i^2}. \quad (2.35)$$

As a special case of (2.33) and (2.34), note that if $\mathbf{j}$ is $n \times 1$, then

$$\mathbf{j} \mathbf{j}^T = \mathbf{J}, \quad \mathbf{j} \mathbf{j}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \mathbf{J}, \quad (2.36)$$

where $\mathbf{j}$ and $\mathbf{J}$ were defined in (2.11) and (2.12). If $\mathbf{a}$ is $n \times 1$ and $\mathbf{J}$ is $n \times p$, then

$$\mathbf{a} \mathbf{J} = \mathbf{J}^T \mathbf{a} = \sum_{i=1}^n a_i \mathbf{e}_i, \quad (2.37)$$

$$iA = \left(\sum_i a_{i1}, \sum_i a_{i2}, \dots, \sum_i a_{ip} \right), \quad A_j = \begin{pmatrix} \sum_i a_{i1} \\ \sum_i a_{i2} \\ \vdots \\ \sum_i a_{ij} \end{pmatrix}. \quad (2.38)$$

Thus a_j^i is the sum of the elements in a_i , jA contains the column sums of A , and A_j contains the row sums of A . In a_j^i , the vector j is $a \leq 1$, in jA , the vector j is $a \leq 1$, and in A_j , the vector j is $p \leq 1$.

Since $a^i b$ is a scalar, it is equal to its transpose:

$$a^i b = (a^i b)' = b' (a^i)' = b' a. \quad (2.39)$$

This allows us to write $(a^i b)^2$ in the form

$$(a^i b)^2 = (a^i b)(a^i b) = (a^i b)(b' a) = a' (bb') a. \quad (2.40)$$

From (2.18), (2.26), and (2.39) we obtain

$$(x - y)'(x - y) = x'x - (x'y) + y'y. \quad (2.41)$$

Note that in analogous expressions with matrices, however, the two middle terms cannot be combined:

$$\begin{aligned} (A - B)'(A - B) &= A'A - A'B - B'A + B'B, \\ (A - B)^2 &= (A - B)(A - B) = A^2 - AB - BA + B^2. \end{aligned}$$

If a and $x_1, x_2, \dots, x_n$ are all $p \times 1$, and A is $p \times p$, we obtain the following factorizing results as consequences of (2.23) and (2.28):

$$\sum_{i=1}^n a' x_i = a' \sum_{i=1}^n x_i, \quad (2.42)$$

$$\sum_{i=1}^n A x_i = A \sum_{i=1}^n x_i, \quad (2.43)$$

$$\sum_{i=1}^n (a' x_i)^2 = a' \left(\sum_{i=1}^n x_i x_i' \right) a \quad \text{[by (2.40)]}, \quad (2.44)$$

$$\sum_{i=1}^n A x_i (A x_i)' = A \left(\sum_{i=1}^n x_i x_i' \right) A'. \quad (2.45)$$

We can express matrix multiplication in terms of row vectors and column vectors. If a_j^i is the i th row of A , and b_j is the j th column of B , then the (i, j) th element of AB

is $a_j^i b_{ij}$. For example, if $\mathbf{A}$ has three rows and $\mathbf{B}$ has two columns,

$$\mathbf{A} = \begin{pmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \end{pmatrix}, \quad \mathbf{B} = (b_{11}, b_{12}),$$

then the product $\mathbf{AB}$ can be written as

$$\mathbf{AB} = \begin{pmatrix} a_1^1 b_{11} & a_1^1 b_{12} \\ a_2^1 b_{11} & a_2^1 b_{12} \\ a_3^1 b_{11} & a_3^1 b_{12} \end{pmatrix}. \quad (2.46)$$

This can be expressed in terms of the rows of $\mathbf{A}$:

$$\mathbf{AB} = \begin{pmatrix} a_1^1(b_{11}, b_{12}) \\ a_2^1(b_{11}, b_{12}) \\ a_3^1(b_{11}, b_{12}) \end{pmatrix} = \begin{pmatrix} a_1^1 \mathbf{B} \\ a_2^1 \mathbf{B} \\ a_3^1 \mathbf{B} \end{pmatrix} = \begin{pmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \end{pmatrix} \mathbf{B}. \quad (2.47)$$

Note that the first column of $\mathbf{AB}$ in (2.46) is

$$\begin{pmatrix} a_1^1 b_{11} \\ a_2^1 b_{11} \\ a_3^1 b_{11} \end{pmatrix} = \begin{pmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \end{pmatrix} b_{11} = \mathbf{A} b_{11},$$

and likewise the second column is $\mathbf{A} b_{12}$. Thus $\mathbf{AB}$ can be written in the form

$$\mathbf{AB} = \mathbf{A}(b_{11}, b_{12}) = (\mathbf{A} b_{11}, \mathbf{A} b_{12}).$$

This result holds in general:

$$\mathbf{AB} = \mathbf{A}(b_{11}, b_{12}, \dots, b_{1p}) = (\mathbf{A} b_{11}, \mathbf{A} b_{12}, \dots, \mathbf{A} b_{1p}). \quad (2.48)$$

To further illustrate matrix multiplication in terms of rows and columns, let $\mathbf{A} = \begin{pmatrix} a_{ij} \end{pmatrix}$ be a $2 \times p$ matrix, $\mathbf{x} = \begin{pmatrix} x_i \end{pmatrix}$ be a $p \times 1$ vector, and $\mathbf{B} = \begin{pmatrix} b_{ij} \end{pmatrix}$ be a $p \times p$ matrix. Then

$$\mathbf{Ax} = \begin{pmatrix} a_{1j} \\ a_{2j} \end{pmatrix} \mathbf{x} = \begin{pmatrix} a_{1j} x_j \\ a_{2j} x_j \end{pmatrix}, \quad (2.49)$$

$$\mathbf{AxB} = \begin{pmatrix} a_{1j} b_{jk} & a_{1j} b_{j2} \\ a_{2j} b_{jk} & a_{2j} b_{j2} \end{pmatrix}. \quad (2.50)$$

Any matrix can be multiplied by its transpose. If $\mathbf{A}$ is $m \times p$, then

$\mathbf{AA}'$ is $m \times m$ and is obtained as products of rows of $\mathbf{A}$ (see 12.524).

Similarly,

AK is $p \times p$ and is obtained as products of columns of A [see (1.56)].

From (1.6) and (1.17), it is clear that both AA' and $A'A$ are symmetric.

In the preceding illustration for AB in terms of row and column vectors, the rows of A were denoted by $\mathbf{a}_i'$ and the columns of B by $\mathbf{b}_j$. If both rows and columns of a matrix A are under discussion, as in AA' and $A'A$, we will use the notation $\mathbf{a}_i'$ for rows and $\mathbf{a}_{ij}$ for columns. To illustrate, if A is 5×4 , we have

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{51} & a_{52} & a_{53} & a_{54} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \mathbf{a}_3' \\ \mathbf{a}_4' \\ \mathbf{a}_5' \end{pmatrix} = (\mathbf{a}_{11}, \mathbf{a}_{12}, \mathbf{a}_{13}, \mathbf{a}_{14}),$$

where, for example,

$$\mathbf{a}_1' = (a_{11} \ a_{12} \ a_{13} \ a_{14}),$$

$$\mathbf{a}_{23} = \begin{pmatrix} a_{23} \\ a_{33} \\ a_{43} \\ a_{53} \end{pmatrix}.$$

With this notation for rows and columns of A , we can express the elements of AA' or of $A'A$ as products of the rows of A or of the columns of A . Thus if we write A in terms of the rows as

$$A = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix},$$

then we have

$$AA' = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p) \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix} = \sum_{i=1}^p \mathbf{a}_i \mathbf{a}_i'. \quad (2.51)$$

$$A'A' = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix} (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p) = \begin{pmatrix} \mathbf{a}_1' \mathbf{a}_1 & \mathbf{a}_1' \mathbf{a}_2 & \dots & \mathbf{a}_1' \mathbf{a}_p \\ \mathbf{a}_2' \mathbf{a}_1 & \mathbf{a}_2' \mathbf{a}_2 & \dots & \mathbf{a}_2' \mathbf{a}_p \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{a}_p' \mathbf{a}_1 & \mathbf{a}_p' \mathbf{a}_2 & \dots & \mathbf{a}_p' \mathbf{a}_p \end{pmatrix}. \quad (2.52)$$

Similarly, if we express A in terms of the columns as

$$A = (\mathbf{a}_{11}, \mathbf{a}_{12}, \dots, \mathbf{a}_{1p}),$$

then

$$A\mathbf{R} = (R_{12}, R_{13}, \dots, R_{1p}) \begin{pmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1p} \end{pmatrix} = \sum_{j=1}^p R_{1j} a_{1j}, \quad (1.53)$$

$$\begin{aligned} \mathbf{R}^T A &= \begin{pmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1p} \end{pmatrix} (R_{12}, R_{13}, \dots, R_{1p}) \\ &= \begin{pmatrix} a_{11}R_{12} & a_{11}R_{13} & \cdots & a_{11}R_{1p} \\ a_{12}R_{12} & a_{12}R_{13} & \cdots & a_{12}R_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p}R_{12} & a_{1p}R_{13} & \cdots & a_{1p}R_{1p} \end{pmatrix}. \end{aligned} \quad (1.54)$$

Let $A = (a_{ij})$ be an $n \times n$ matrix and $\mathbf{D}$ be a diagonal matrix, $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_n)$. Then, in the product $\mathbf{D}A$, the i th row of A is multiplied by d_i , and in $A\mathbf{D}$, the j th column of A is multiplied by d_j . For example, if $n = 3$, we have

$$\begin{aligned} \mathbf{D}A &= \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\ &= \begin{pmatrix} d_1 a_{11} & d_1 a_{12} & d_1 a_{13} \\ d_2 a_{21} & d_2 a_{22} & d_2 a_{23} \\ d_3 a_{31} & d_3 a_{32} & d_3 a_{33} \end{pmatrix}, \end{aligned} \quad (1.55)$$

$$\begin{aligned} A\mathbf{D} &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} \\ &= \begin{pmatrix} d_1 a_{11} & d_2 a_{12} & d_3 a_{13} \\ d_1 a_{21} & d_2 a_{22} & d_3 a_{23} \\ d_1 a_{31} & d_2 a_{32} & d_3 a_{33} \end{pmatrix}, \end{aligned} \quad (1.56)$$

$$\mathbf{D}A\mathbf{D} = \begin{pmatrix} d_1^2 a_{11} & d_1 d_2 a_{12} & d_1 d_3 a_{13} \\ d_2 d_1 a_{21} & d_2^2 a_{22} & d_2 d_3 a_{23} \\ d_3 d_1 a_{31} & d_3 d_2 a_{32} & d_3^2 a_{33} \end{pmatrix}. \quad (1.57)$$

In the special case where the diagonal matrix is the identity, we have

$$\mathbf{I}A = A\mathbf{I} = A. \quad (1.58)$$

If A is rectangular, (2.58) still holds, but the two identities are of different sizes.

The product of a scalar and a matrix is obtained by multiplying each element of the matrix by the scalar.

$$cA = (c a_{ij}) = \begin{pmatrix} c a_{11} & c a_{12} & \cdots & c a_{1n} \\ c a_{21} & c a_{22} & \cdots & c a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c a_{m1} & c a_{m2} & \cdots & c a_{mn} \end{pmatrix}. \quad (2.59)$$

For example,

$$cI = \begin{pmatrix} c & 0 & \cdots & 0 \\ 0 & c & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c \end{pmatrix}, \quad (2.60)$$

$$c\mathbf{a} = \begin{pmatrix} c a_1 \\ c a_2 \\ \vdots \\ c a_n \end{pmatrix}. \quad (2.61)$$

Since $c a_{ij} = a_{ij} c$, the product of a scalar and a matrix is commutative:

$$cA = Ac. \quad (2.62)$$

Multiplication of vectors or matrices by scalars permits the use of linear combinations, such as

$$\sum_{i=1}^n a_i \mathbf{b}_i = a_1 \mathbf{b}_1 + a_2 \mathbf{b}_2 + \cdots + a_n \mathbf{b}_n,$$

$$\sum_{i=1}^n a_i \mathbf{B}_i = a_1 \mathbf{B}_1 + a_2 \mathbf{B}_2 + \cdots + a_n \mathbf{B}_n.$$

If A is a symmetric matrix, and $\mathbf{x}$ and $\mathbf{y}$ are vectors, the product

$$\mathbf{x}' A \mathbf{y} = \sum_i a_{ii} x_i y_i + \sum_{i < j} a_{ij} (x_i y_j + x_j y_i) \quad (2.63)$$

is called a *quadratic form*, whereas

$$\mathbf{x}' A \mathbf{x} = \sum_{i,j} a_{ij} x_i x_j \quad (2.64)$$

is called a *bilinear form*. Either of $\mathbf{f}^T \mathbf{A} \mathbf{f}$ or, of course, $\mathbf{f}^T \mathbf{A}^T \mathbf{f}$ can be treated in such. Expressions such as $\mathbf{f}^T \mathbf{A} \mathbf{g}$ or $\mathbf{g}^T \mathbf{A} \mathbf{f}$ are permissible (assuming $\mathbf{A}$ is positive definite) see Section 2.7).

2.4 PARTITIONED-MATRIX

It is sometimes convenient to partition a matrix into submatrices. For example, a partitioning of a matrix $\mathbf{A}$ into four submatrices could be indicated symbolically as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}.$$

For example, a 4×5 matrix $\mathbf{A}$ can be partitioned as

$$\mathbf{A} = \left(\begin{array}{ccc|cc} 2 & 1 & 3 & 8 & 4 \\ -3 & 4 & 0 & 2 & 7 \\ 9 & 3 & 6 & 5 & -2 \\ 4 & 8 & 3 & 1 & 6 \end{array} \right) = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where

$$\begin{aligned} \mathbf{A}_{11} &= \begin{pmatrix} 2 & 1 & 3 \\ -3 & 4 & 0 \\ 9 & 3 & 6 \end{pmatrix}, & \mathbf{A}_{12} &= \begin{pmatrix} 8 & 4 \\ 2 & 7 \\ 5 & -2 \end{pmatrix}, \\ \mathbf{A}_{21} &= [4 \quad 8 \quad 3], & \mathbf{A}_{22} &= [1 \quad 6]. \end{aligned}$$

If two matrices $\mathbf{A}$ and $\mathbf{B}$ are conformable and $\mathbf{A}$ and $\mathbf{B}$ are partitioned so that the submatrices are appropriately conformable, then the product $\mathbf{AB}$ can be found by following the usual row-by-column pattern of multiplication on the submatrices as if they were single elements; for example,

$$\begin{aligned} \mathbf{AB} &= \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix}. \end{aligned} \quad (2.66)$$

It can be seen that this formulation is equivalent to the usual row-by-column definition of matrix multiplication. For example, the $(1, 1)$ element of $\mathbf{AB}$ is the product of the first row of $\mathbf{A}$ and the first column of $\mathbf{B}$. In the $(1, 1)$ element of $\mathbf{A}_{11}\mathbf{B}_{11}$ we have the sum of products of part of the first row of $\mathbf{A}$ and part of the first column of $\mathbf{B}$. In the $(1, 1)$ element of $\mathbf{A}_{12}\mathbf{B}_{21}$ we have the sum of products of the rest of the first row of $\mathbf{A}$ and the remainder of the first column of $\mathbf{B}$.

Multiplication of a matrix and a vector can also be carried out in partitioned form. For example,

$$A\mathbf{b} = (A_1, A_2) \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = A_1\mathbf{b}_1 + A_2\mathbf{b}_2, \quad (2.66)$$

where the partitioning of the columns of A corresponds to the partitioning of the elements of $\mathbf{b}$. Note that the partitioning of A into two sets of columns is indicated by a comma, $A = (A_1, A_2)$.

The partitioned multiplication in (2.66) can be extended to individual columns of A and individual elements of $\mathbf{b}$:

$$\begin{aligned} A\mathbf{b} &= (a_1, a_2, \dots, a_p) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} \\ &= b_1a_1 + b_2a_2 + \dots + b_pa_p. \end{aligned} \quad (2.67)$$

Thus $A\mathbf{b}$ is expressible as a linear combination of the columns of A , the coefficients being elements of $\mathbf{b}$. For example, let

$$A = \begin{pmatrix} 3 & -2 & 1 \\ 2 & 1 & 0 \\ 4 & 3 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 4 \\ 2 \\ 3 \end{pmatrix}.$$

Then

$$A\mathbf{b} = \begin{pmatrix} 11 \\ 10 \\ 28 \end{pmatrix}.$$

Using a linear combination of columns of A as in (2.67), we obtain

$$\begin{aligned} A\mathbf{b} &= b_1a_1 + b_2a_2 + b_3a_3 \\ &= 4 \begin{pmatrix} 3 \\ 2 \\ 4 \end{pmatrix} + 2 \begin{pmatrix} -2 \\ 1 \\ 3 \end{pmatrix} + 3 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \\ &= \begin{pmatrix} 12 \\ 8 \\ 16 \end{pmatrix} + \begin{pmatrix} -4 \\ 2 \\ 6 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \\ 6 \end{pmatrix} = \begin{pmatrix} 11 \\ 10 \\ 28 \end{pmatrix}. \end{aligned}$$

We note that if $\mathbf{A}$ is partitioned as in (2.65), $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$, the transpose is not equal to $(\mathbf{A}_1', \mathbf{A}_2')$, but rather

$$\mathbf{A}' = (\mathbf{A}_1, \mathbf{A}_2)' = \begin{pmatrix} \mathbf{A}_1' \\ \mathbf{A}_2' \end{pmatrix}. \quad (2.68)$$

2.5 RANK

Before defining the rank of a matrix, we first introduce the notions of linear independence and dependence. A set of vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ is said to be linearly dependent if constants $c_1, c_2, \dots, c_n$ (not all zero) can be found such that

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \dots + c_n\mathbf{a}_n = \mathbf{0}. \quad (2.69)$$

If no constants $c_1, c_2, \dots, c_n$ can be found satisfying (2.69), the set of vectors is said to be linearly independent.

If (2.69) holds, then at least one of the vectors $\mathbf{a}_i$ can be expressed as a linear combination of the other vectors in the set. Thus linear dependence of a set of vectors implies redundancy in the set. Among linearly independent vectors there is no redundancy of this type.

The rank of any square or rectangular matrix $\mathbf{A}$ is defined as

$$\begin{aligned} \text{rank}(\mathbf{A}) &= \text{number of linearly independent rows of } \mathbf{A} \\ &= \text{number of linearly independent columns of } \mathbf{A}. \end{aligned}$$

It can be shown that the number of linearly independent rows of a matrix is always equal to the number of linearly independent columns.

If $\mathbf{A}$ is $n \times p$, the maximum possible rank of $\mathbf{A}$ is the smaller of n and p , to which case $\mathbf{A}$ is said to be of full rank (sometimes said full row rank or full column rank). For example,

$$\mathbf{A} = \begin{pmatrix} 1 & -2 & 3 \\ 2 & 1 & 4 \end{pmatrix}$$

has rank 2 because the two rows are linearly independent (neither row is a multiple of the other). However, even though $\mathbf{A}$ is full rank, the columns are linearly dependent because rank 2 implies there are only two linearly independent columns. Thus, by (2.69), there exist constants c_1, c_2 , and c_3 such that

$$c_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + c_2 \begin{pmatrix} -2 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (2.70)$$

By (2.67), we can write (2.70) in the form

$$\begin{pmatrix} 1 & -2 & 3 \\ 5 & 2 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

or

$$Ax = 0. \quad (2.71)$$

A solution vector to (2.70) or (2.71) is given by any multiple of $x = (14, -11, -12)'$. Hence we have the interesting result that a product of a matrix A and a vector x is equal to 0 , even though $A \neq 0$ and $x \neq 0$. This is a direct consequence of the linear dependence of the column vectors of A .

Another consequence of the linear dependence of rows or columns of a matrix is the possibility of expressions such as $AB = CB$, where $A \neq C$. For example, let

$$A = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & 1 & 1 \\ 3 & -5 & -4 \end{pmatrix}.$$

Then

$$AB = CB = \begin{pmatrix} 3 & 5 \\ 1 & 4 \end{pmatrix}.$$

All three matrices A , B , and C are full rank, but being rectangular, they have a rank deficiency in either rows or columns, which permits us to construct $AB = CB$ with $A \neq C$. Thus in a matrix equation, we cannot, in general, cancel matrices from both sides of the equation.

There are two exceptions to this rule. One exception involves nonsingular matrices as defined in Section 1.6. The other special case occurs when the expression holds for all possible values of the matrix common to both sides of the equation. For example,

$$T(Ax) = Bx \text{ for all possible values of } x, \text{ then } A = B. \quad (2.72)$$

To see this, let $x = (1, 0, \dots, 0)'$. Then the first column of A equals the first column of B . Now let $x = (0, 1, 0, \dots, 0)'$, and the second column of A equals the second column of B . Continuing in this fashion, we obtain $A = B$.

Suppose a rectangular matrix A is $m \times p$ of rank p , where $p < m$. We typically describe this statement as " A is $m \times p$ of rank p ."

2.6 INVERSE

If a matrix A is square and of full rank, then A is said to be nonsingular, and A has a unique inverse, denoted by A^{-1} , with the property that

$$AA^{-1} = A^{-1}A = I. \quad (2.73)$$

For example, let

$$\mathbf{A} = \begin{pmatrix} 2 & 4 \\ 1 & 5 \end{pmatrix}.$$

Then

$$\begin{aligned}\mathbf{A}^{-1} &= \begin{pmatrix} .5 & -.4 \\ -.2 & .3 \end{pmatrix}, \\ \mathbf{A}\mathbf{A}^{-1} &= \begin{pmatrix} 2 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} .5 & -.4 \\ -.2 & .3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.\end{aligned}$$

If $\mathbf{A}$ is square and of less than full rank, then an inverse does not exist, and $\mathbf{A}$ is said to be singular. Note that rectangular matrices do not have inverses as in (2.75), even if they are full rank.

If $\mathbf{A}$ and $\mathbf{B}$ are the same size and nonsingular, then the inverse of their product is the product of their inverses in reverse order,

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \quad (2.74)$$

Note that (2.74) holds only for nonsingular matrices. Thus, for example, if $\mathbf{A}$ is $n \times p$ of rank $p < n$, then $\mathbf{A}^T\mathbf{A}$ has an inverse, but $(\mathbf{A}^T\mathbf{A})^{-1}$ is not equal to $\mathbf{A}^{-1}(\mathbf{A}^T)^{-1}$ because $\mathbf{A}$ is rectangular and does not have an inverse.

If a matrix is nonsingular, it can be canceled from both sides of an equation, provided it appears on the left (or right) on both sides. For example, if $\mathbf{B}$ is nonsingular, then

$$\mathbf{A}\mathbf{B} = \mathbf{C}\mathbf{B} \quad \text{implies} \quad \mathbf{A} = \mathbf{C},$$

since we can multiply on the right by $\mathbf{B}^{-1}$ to obtain

$$\begin{aligned}\mathbf{A}\mathbf{B}\mathbf{B}^{-1} &= \mathbf{C}\mathbf{B}\mathbf{B}^{-1}, \\ \mathbf{A}\mathbf{I} &= \mathbf{C}\mathbf{I}, \\ \mathbf{A} &= \mathbf{C}.\end{aligned}$$

Otherwise, if $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are rectangular or square and singular, it is easy to construct $\mathbf{A}\mathbf{B} = \mathbf{C}\mathbf{B}$, with $\mathbf{A} \neq \mathbf{C}$, as illustrated near the end of Section 2.5.

The inverse of the transpose of a nonsingular matrix is given by the transpose of the inverse:

$$(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T. \quad (2.75)$$

If the symmetric nonsingular matrix $\mathbf{A}$ is partitioned in the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{21}^T & a_{22} \end{pmatrix},$$

therefore the inverse is given by

$$A^{-1} = \frac{1}{b} \begin{pmatrix} bA_{11}^{-1} + A_{11}^{-1}a_{12}a_{12}^tA_{11}^{-1} & -A_{11}^{-1}a_{12} \\ -a_{12}^tA_{11}^{-1} & 1 \end{pmatrix}, \quad (2.76)$$

where $b = a_{22} - a_{12}^tA_{11}^{-1}a_{12}$. A nonsingular matrix of the form $B + \alpha\alpha^t$, where B is nonsingular, has as its inverse

$$(B + \alpha\alpha^t)^{-1} = B^{-1} - \frac{B^{-1}\alpha\alpha^tB^{-1}}{1 + \alpha^tB^{-1}\alpha}. \quad (2.77)$$

2.7 POSITIVE DEFINITE MATRICES

The symmetric matrix A is said to be *positive definite* if $x^tAx > 0$ for all possible vectors x (except $x = 0$). Similarly, A is *positive semidefinite* if $x^tAx \geq 0$ for all $x \neq 0$. [A quadratic form x^tAx was defined in (2.63).] The diagonal elements a_{ii} of a positive definite matrix are positive. To see this, let $x^t = (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n)$ with $x_i \neq 0$ in the i th position. Then $x^tAx = a_{ii}x_i^2 > 0$. Similarly, for a positive semidefinite matrix A , $a_{ii} \geq 0$ for all i .

One way to obtain a positive definite matrix is as follows:

$$\text{If } A = B^tB, \text{ where } B \text{ is } n \times p \text{ of rank } p = n, \text{ then } B^tB \text{ is positive definite.} \quad (2.78)$$

This is easily shown:

$$x^tAx = x^tB^tBx = (Bx)^t(Bx) = x^t\alpha,$$

where $\alpha = Bx$. Thus $x^tAx = \sum_{i=1}^n \alpha_i^2$, which is positive (Bx cannot be 0 unless $x = 0$, because B is full rank). If B is less than full rank, then by a similar argument, B^tB is positive semidefinite.

Note that $A = B^tB$ is analogous to $a = b^2$ in real numbers, where the square of any number (including negative numbers) is positive.

In another analogy to positive real numbers, a positive definite matrix can be factored into a "square root" in two ways. We give one method in (2.79) and the other in Section 2.11.4.

A positive definite matrix A can be factored into

$$A = TT^t, \quad (2.79)$$

where T is a nonsingular upper triangular matrix. One way to obtain T is the Cholesky decomposition, which can be carried out in the following steps.

Let $A = (a_{ij})$ and $T = (t_{ij})$ be $n \times n$. Then the elements of T are found as follows:

$$\begin{aligned}
 a_i &= \sqrt{b_i}, & a_j &= \frac{a_i}{a_n} & 1 \leq j \leq n, \\
 c_k &= \sqrt{a_n - \sum_{i=1}^{k-1} c_i^2} & 1 \leq k \leq n, \\
 k_j &= \frac{a_j - \sum_{i=1}^{j-1} k_i c_i}{c_j} & 1 \leq i < j \leq n, \\
 k_j &= 0 & 1 \leq j = n \leq n.
 \end{aligned}$$

For example, let

$$\mathbf{A} = \begin{pmatrix} 3 & 0 & -3 \\ 0 & 6 & 3 \\ -3 & 3 & 6 \end{pmatrix}.$$

Then by the Cholesky method, we obtain

$$\begin{aligned}
 \mathbf{T} &= \begin{pmatrix} \sqrt{3} & 0 & -\sqrt{3} \\ 0 & \sqrt{6} & \sqrt{13} \\ 0 & 0 & \sqrt{13} \end{pmatrix}, \\
 \mathbf{T}\mathbf{T} &= \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{6} & 0 \\ -\sqrt{3} & \sqrt{13} & \sqrt{13} \end{pmatrix} \begin{pmatrix} \sqrt{3} & 0 & -\sqrt{3} \\ 0 & \sqrt{6} & \sqrt{13} \\ 0 & 0 & \sqrt{13} \end{pmatrix} \\
 &= \begin{pmatrix} 3 & 0 & -3 \\ 0 & 6 & 3 \\ -3 & 3 & 6 \end{pmatrix} = \mathbf{A}.
 \end{aligned}$$

2.8 DETERMINANTS

The determinant of an $n \times n$ matrix $\mathbf{A}$ is defined as the sum of all $n!$ possible products of n elements such that

1. each product contains one element from every row and every column, and
2. the factors in each product are written so that the column subscripts appear in order of magnitude and each product is then preceded by a plus or minus sign according to whether the number of inversions in the row subscripts is even or odd.

An *inversion* occurs whenever a larger number precedes a smaller one. The symbol $n!$ is defined as

$$n! = n(n-1)(n-2) \cdots 2 \cdot 1. \quad (2.86)$$

The determinant of $\mathbf{A}$ is a scalar denoted by $|\mathbf{A}|$ or by $\det(\mathbf{A})$. The preceding definition is not useful in evaluating determinants, except in the case of 2×2 or 3×3 matrices. For larger matrices, other methods are available for manual computations, but determinants are typically evaluated by computer. For a 2×2 matrix, the determinant is found by

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}. \quad (2.81)$$

For a 3×3 matrix, the determinant is given by

$$|\mathbf{A}| = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{31}a_{22}a_{13} - a_{32}a_{23}a_{11} - a_{33}a_{21}a_{12}. \quad (2.82)$$

This can be found by the following scheme. The three positive terms are obtained by



and the three negative terms, by



The determinant of a diagonal matrix is the product of the diagonal elements; that is, if $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_n)$, then

$$|\mathbf{D}| = \prod_{i=1}^n d_i. \quad (2.83)$$

As a special case of (2.83), suppose all-diagonal elements are equal, say,

$$\mathbf{D} = \text{diag}(c, c, \dots, c) = c\mathbf{I}.$$

Then

$$|\mathbf{E}| = |\mathbf{e}| = \prod_{i=1}^n e_i = e^n. \quad (2.84)$$

The extension of (2.84) to any square matrix $\mathbf{A}$ is

$$|\mathbf{A}| = e^{\text{tr } \mathbf{A}}. \quad (2.85)$$

Since the determinant is a scalar, we can carry out operations such as

$$|\mathbf{A}|^2, \quad |\mathbf{A}|^{1/2}, \quad \frac{1}{|\mathbf{A}|},$$

provided that $|\mathbf{A}| > 0$ for $|\mathbf{A}|^{1/2}$ and that $|\mathbf{A}| \neq 0$ for $1/|\mathbf{A}|$.

If the square matrix $\mathbf{A}$ is singular, its determinant is 0:

$$|\mathbf{A}| = 0 \text{ if } \mathbf{A} \text{ is singular.} \quad (2.86)$$

If $\mathbf{A}$ is near singular, then there exists a linear combination of the columns that is close to 0, and $|\mathbf{A}|$ is also close to 0. If $\mathbf{A}$ is nonsingular, its determinant is nonzero:

$$|\mathbf{A}| \neq 0 \text{ if } \mathbf{A} \text{ is nonsingular.} \quad (2.87)$$

If $\mathbf{A}$ is positive definite, its determinant is positive:

$$|\mathbf{A}| > 0 \text{ if } \mathbf{A} \text{ is positive definite.} \quad (2.88)$$

If $\mathbf{A}$ and $\mathbf{B}$ are square and the same size, the determinant of the product is the product of the determinants:

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|. \quad (2.89)$$

For example, let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -3 & 5 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}.$$

Then

$$\mathbf{AB} = \begin{pmatrix} 6 & 8 \\ -7 & 9 \end{pmatrix}, \quad |\mathbf{AB}| = 118,$$

$$|\mathbf{A}| = 11, \quad |\mathbf{B}| = 10, \quad |\mathbf{A}||\mathbf{B}| = 118.$$

The determinant of the transpose of a matrix is the same as the determinant of the matrix, and the determinant of the the inverse of a matrix is the reciprocal of the

determinant

$$|\mathbf{A}'| = |\mathbf{A}|, \quad (2.88)$$

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} = |\mathbf{A}|^{-1}. \quad (2.89)$$

If a partitioned matrix has the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{O} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square but not necessarily the same size, then

$$|\mathbf{A}| = \begin{vmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{O} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}| |\mathbf{A}_{22}|. \quad (2.92)$$

For a general partitioned matrix,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square and nonsingular (not necessarily the same size), the determinant is given by either of the following two expressions:

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}| |\mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}| \quad (2.93)$$

$$= |\mathbf{A}_{22}| |\mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}|. \quad (2.94)$$

Note the analogy of (2.93) and (2.94) to the case of the determinant of a 2×2 matrix as given by (2.81):

$$\begin{aligned} \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} &= a_{11}a_{22} - a_{21}a_{12} \\ &= a_{11} \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}} \right) \\ &= a_{22} \left(a_{11} - \frac{a_{12}a_{21}}{a_{22}} \right). \end{aligned}$$

If $\mathbf{B}$ is nonsingular and x is a vector, then

$$|\mathbf{B} + xx'| = |\mathbf{B}| (1 + x' \mathbf{B}^{-1} x). \quad (2.95)$$

2.5 TRACE

A simple function of an $n \times n$ matrix A is the *trace*, denoted by $\text{tr}(A)$ and defined as the sum of the diagonal elements of A ; that is, $\text{tr}(A) = \sum_{i=1}^n a_{ii}$. The trace is, of course, a scalar. For example, suppose

$$A = \begin{pmatrix} 2 & 4 & 1 \\ 1 & -3 & 1 \\ 4 & 7 & 2 \end{pmatrix}.$$

Then

$$\text{tr}(A) = 2 + (-3) + 2 = 1.$$

The trace of the sum of two square matrices is the sum of the traces of the two matrices:

$$\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B). \quad (2.56)$$

An important result for the product of two matrices is

$$\text{tr}(AB) = \text{tr}(BA). \quad (2.57)$$

This result holds for any matrices A and B where AB and BA are both defined. It is not necessary that A and B be square or that AB equal BA . For example, let

$$A = \begin{pmatrix} 1 & 2 \\ 2 & -1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & -2 & 1 \\ 1 & 4 & 2 \end{pmatrix}.$$

Then

$$AB = \begin{pmatrix} 9 & 10 & 18 \\ 4 & -8 & -3 \\ 20 & 16 & 24 \end{pmatrix}, \quad BA = \begin{pmatrix} 3 & 17 \\ 20 & 32 \end{pmatrix},$$

$$\text{tr}(AB) = 9 - 8 + 24 = 25, \quad \text{tr}(BA) = 3 + 32 = 35.$$

From (2.52) and (2.54), we obtain

$$\text{tr}(A^T A) = \text{tr}(AA^T) = \sum_{i=1}^n \sum_{j=1}^p a_{ij}^2, \quad (2.58)$$

where the a_{ij} 's are elements of the $n \times p$ matrix A .

2.10 ORTHOGONAL VECTORS AND MATRICES

Two vectors $\mathbf{a}$ and $\mathbf{b}$ of the same size are said to be orthogonal if

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_nb_n = 0. \quad (2.101)$$

Geometrically, orthogonal vectors are perpendicular [see (2.14)] and the converse also follows (2.14). If $\mathbf{a} \cdot \mathbf{a} = 1$, the vector $\mathbf{a}$ is said to be normalized. The vector $\mathbf{a}$ can always be normalized by dividing by its length, $\sqrt{\mathbf{a} \cdot \mathbf{a}}$. Thus

$$\mathbf{e} = \frac{\mathbf{a}}{\sqrt{\mathbf{a} \cdot \mathbf{a}}} \quad (2.102)$$

is normalized so that $\mathbf{e} \cdot \mathbf{e} = 1$.

A matrix $\mathbf{C} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ whose columns are normalized and mutually orthogonal is called an orthogonal matrix. Since the elements of $\mathbf{C}^T \mathbf{C}$ are products of columns of $\mathbf{C}$ [see (2.54)], which have the properties $\mathbf{e}_i \cdot \mathbf{e}_i = 1$ for all i and $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ for all $i \neq j$, we have

$$\mathbf{C}^T \mathbf{C} = \mathbf{I}_n \quad (2.103)$$

if $\mathbf{C}$ satisfies (2.101). It necessarily follows that

$$\mathbf{C} \mathbf{C}^T = \mathbf{I}_n \quad (2.104)$$

from which we see that the rows of $\mathbf{C}$ are also normalized and mutually orthogonal. It is clear from (2.103) and (2.102) that $\mathbf{C}^{-1} = \mathbf{C}^T$ for an orthogonal matrix $\mathbf{C}$.

We illustrate the creation of an orthogonal matrix by starting with

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \end{pmatrix},$$

whose columns are mutually orthogonal. To normalize the three columns, we divide by the respective lengths, $\sqrt{3}$, $\sqrt{2}$, and $\sqrt{5}$, to obtain

$$\mathbf{C} = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{5} \\ 1/\sqrt{3} & 1/\sqrt{2} & -1/\sqrt{5} \\ 1/\sqrt{3} & -2/\sqrt{2} & 0 \end{pmatrix}.$$

Note that the rows also become normalized and mutually orthogonal so that $\mathbf{C}$ satisfies both (2.103) and (2.102).

Multiplication by an orthogonal matrix has the effect of rotating, since, that is, if a point $\mathbf{x}$ is transformed to $\mathbf{x}' = \mathbf{C}\mathbf{x}$, where $\mathbf{C}$ is orthogonal, then

$$\|x\| = \|Cx\|/\|Cx\| = \|C^TC\|/C\|x\| = \|C\|/\|x\|, \quad (2.103)$$

and the distance from the origin to x is the same as the distance to Cx .

2.11 EIGENVALUES AND EIGENVECTORS

2.11.1 Definition

For every square matrix A , a scalar λ and a nonzero vector x can be found such that

$$Ax = \lambda x. \quad (2.104)$$

In (2.104), λ is called an *eigenvalue* of A , and x is an *eigenvector* of A corresponding to λ . To find λ and x , we write (2.104) as

$$(A - \lambda I)x = 0, \quad (2.105)$$

If $(A - \lambda I) \neq 0$, then $(A - \lambda I)$ has an inverse and $x = 0$ is the only solution. Hence, in order to obtain nontrivial solutions, we set $(A - \lambda I) = 0$ to find values of λ that can be substituted into (2.105) to find corresponding values of x . Alternatively, (2.105) and (2.11) require that the columns of $A - \lambda I$ be linearly dependent. Thus in $(A - \lambda I)x = 0$, the matrix $A - \lambda I$ must be singular in order to find a solution vector x that is not 0.

The equation $(A - \lambda I) = 0$ is called the *characteristic equation*. If A is $n \times n$, the characteristic equation will have n roots; that is, A will have n eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$. The λ 's will not necessarily all be distinct or all nonzero. However, if A arises from computations on real (continuous) data and is nonsingular, the λ 's will all be distinct (with probability 1). After finding $\lambda_1, \lambda_2, \dots, \lambda_n$, the accompanying eigenvectors $x_1, x_2, \dots, x_n$ can be found using (2.105).

If we multiply both sides of (2.105) by a scalar β and note by (2.12) that β and $A - \lambda I$ commute, we obtain

$$(A - \lambda I)\beta x = \beta(A - \lambda I)x = 0. \quad (2.106)$$

Thus if x is an eigenvector of A , βx is also an eigenvector, and eigenvectors are unique only up to multiplication by a scalar. Hence we can adjust the length of x , but the direction from the origin is unique; that is, the relative values of ratios of the components of $x = (x_1, x_2, \dots, x_n)^T$ are unique. Typically, the eigenvector x is scaled so that $x^T x = 1$.

To illustrate, we will find the eigenvalues and eigenvectors for the matrix

$$A = \begin{pmatrix} 1 & 2 \\ -2 & 4 \end{pmatrix}.$$

The characteristic equation is

$$\begin{aligned} (A - \lambda I) &= \begin{vmatrix} 1-\lambda & 2 \\ -1 & 4-\lambda \end{vmatrix} = (1-\lambda)(4-\lambda) + 2 = 0, \\ \lambda^2 - 5\lambda + 6 &= (\lambda-3)(\lambda-2) = 0, \end{aligned}$$

from which $\lambda_1 = 3$ and $\lambda_2 = 2$. To find the eigenvector corresponding to $\lambda_1 = 3$, we use (2.10'),

$$\begin{aligned} (A - \lambda I)x &= 0, \\ \begin{pmatrix} 1-3 & 2 \\ -1 & 4-3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ -2x_1 + 2x_2 &= 0 \\ -x_1 + x_2 &= 0. \end{aligned}$$

As expected, either equation is redundant in the presence of the other, and there remains a single equation with two unknowns, $x_1 = x_2$. The solution vector can be written with an arbitrary constant,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = v \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

If v is set equal to $1/\sqrt{2}$ to normalize the eigenvector, we obtain

$$u_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.$$

Similarly, corresponding to $\lambda_2 = 2$, we have

$$u_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

2.11.2 $I + A$ and $I - A$

If λ is an eigenvalue of A and x is the corresponding eigenvector, then $1 + \lambda$ is an eigenvalue of $I + A$ and $1 - \lambda$ is an eigenvalue of $I - A$. In either case, x is the corresponding eigenvector.

We demonstrate this for $I + A$:

$$\begin{aligned} Ax &= \lambda x, \\ x + Ax &= x + \lambda x, \\ (I + A)x &= (1 + \lambda)x. \end{aligned}$$

2.11.3 $\text{tr}(\mathbf{A})$ and $|\mathbf{A}|$

For any square matrix $\mathbf{A}$ with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, we have

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i \quad (2.107)$$

$$|\mathbf{A}| = \prod_{i=1}^n \lambda_i \quad (2.108)$$

Note that by the definitions in Section 2.9, $\text{tr}(\mathbf{A})$ is also equal to $\sum_{i=1}^n a_{ii}$, but $a_{ii} \neq \lambda_i$.

We illustrate (2.107) and (2.108) using the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 4 \end{pmatrix}$$

from the illustration in Section 2.11.1, for which $\lambda_1 = 3$ and $\lambda_2 = 2$. Using (2.107), we obtain

$$\text{tr}(\mathbf{A}) = \lambda_1 + \lambda_2 = 3 + 2 = 5,$$

and from (2.108), we have

$$|\mathbf{A}| = \lambda_1 \lambda_2 = \ln(2) = 6.$$

By definition, we obtain

$$\text{tr}(\mathbf{A}) = 1 + 4 = 5 \quad \text{and} \quad |\mathbf{A}| = (1)(4) - (-1)(2) = 6.$$

2.11.4 Positive Definite and Semidefinite Matrices

The eigenvalues and eigenvectors of positive definite and positive semidefinite matrices have the following properties:

1. The eigenvalues of a positive definite matrix are all positive.
2. The eigenvalues of a positive semidefinite matrix are positive or zero, with the number of positive eigenvalues equal to the rank of the matrix.

It is customary to list the eigenvalues of a positive definite matrix in descending order: $\lambda_1 > \lambda_2 > \dots > \lambda_p$. The eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ are listed in the same order: $\mathbf{u}_1$ corresponds to λ_1 , $\mathbf{u}_2$ corresponds to λ_2 , and so on.

The following result, known as the Perron–Frobenius theorem, is of interest in Chapter 1.2. If all elements of the positive definite matrix $\mathbf{A}$ are positive, then all elements of the first eigenvector are positive. (The first eigenvector is the one associated with the first eigenvalue, λ_1 .)

2.11.5 The Product AB

If A and B are square and the same size, the eigenvalues of AB are the same as those of BA , although the eigenvectors are usually different. (This result also holds if AB and BA are both square but of different sizes, so when A is $n \times p$ and B is $p \times n$. (In this case, the nonzero eigenvalues of AB and BA will be the same.)

2.11.6 Symmetric Matrix

The eigenvalues of an $n \times n$ symmetric matrix A are mutually orthogonal. It follows that if the n eigenvectors of A are normalized and inserted as columns of a matrix $C = (c_1, c_2, \dots, c_n)$, then C is orthogonal.

2.11.7 Spectral Decomposition

It was noted in Section 2.11.6 that if the matrix $C = (c_1, c_2, \dots, c_n)$ contains the normalized eigenvectors of an $n \times n$ symmetric matrix A , then C is orthogonal. Therefore, by (2.103), $C^T C = CC^T = I$, which we can multiply by A to obtain

$$A = AC^T C.$$

We now substitute $C = (c_1, c_2, \dots, c_n)$:

$$\begin{aligned} A &= A(c_1, c_2, \dots, c_n)C^T \\ &= (Ac_1, Ac_2, \dots, Ac_n)C^T && \text{[by (2.102)]} \\ &= (\lambda_1 c_1, \lambda_2 c_2, \dots, \lambda_n c_n)C^T && \text{[by (2.104)]} \\ &= C\Lambda C^T && \text{[by (2.104)]} \end{aligned} \quad (2.105)$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}. \quad (2.110)$$

The expression $A = C\Lambda C^T$ in (2.105) for a symmetric matrix A in terms of its eigenvalues and eigenvectors is known as the *spectral decomposition* of A .

Since C is orthogonal and $C^T C = CC^T = I$, we can multiply (2.105) on the left by C^T and on the right by C to obtain

$$C^T A C = \Lambda. \quad (2.112)$$

Thus a symmetric matrix A can be diagonalized by an orthogonal matrix containing normalized eigenvectors of A , and by (2.112) the resulting diagonal matrix contains eigenvalues of A .

2.11.8 Square Root Matrix

If $\mathbf{A}$ is positive definite, the spectral decomposition of $\mathbf{A}$ in (2.112) can be modified by taking the square roots of the eigenvalues to produce a square root matrix,

$$\mathbf{A}^{1/2} = \mathbf{C}\mathbf{D}^{1/2}\mathbf{C}^T, \quad (2.112)$$

where

$$\mathbf{D}^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{pmatrix}. \quad (2.113)$$

The square root matrix $\mathbf{A}^{1/2}$ is symmetric and serves as the square root of $\mathbf{A}$,

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = (\mathbf{A}^{1/2})^2 = \mathbf{A}. \quad (2.114)$$

2.11.9 Square Matrices and Inverse Matrices

Other functions of $\mathbf{A}$ have spectral decompositions analogous to (2.112). Two of these are the square and inverse of $\mathbf{A}$. If the square matrix $\mathbf{A}$ has eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ and corresponding eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, then $\mathbf{A}^2$ has eigenvalues $\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2$ and eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. If $\mathbf{A}$ is nonsingular, then $\mathbf{A}^{-1}$ has eigenvalues $1/\lambda_1, 1/\lambda_2, \dots, 1/\lambda_n$ and eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. If $\mathbf{A}$ is also symmetric, then

$$\mathbf{A}^2 = \mathbf{C}\mathbf{D}^2\mathbf{C}^T, \quad (2.115)$$

$$\mathbf{A}^{-1} = \mathbf{C}\mathbf{D}^{-1}\mathbf{C}^T, \quad (2.116)$$

where $\mathbf{C} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ has as columns the normalized eigenvectors of $\mathbf{A}$ (and of $\mathbf{A}^2$ and $\mathbf{A}^{-1}$), $\mathbf{D}^2 = \text{diag}(\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2)$, and $\mathbf{D}^{-1} = \text{diag}(1/\lambda_1, 1/\lambda_2, \dots, 1/\lambda_n)$.

2.11.10 Singular Value Decomposition

In (2.109) in Section 2.11.2, we expressed a symmetric matrix $\mathbf{A}$ in terms of its eigenvalues and eigenvectors in the spectral decomposition $\mathbf{A} = \mathbf{C}\mathbf{D}\mathbf{C}^T$. In a similar manner, we can express any real matrix $\mathbf{A}$ in terms of eigenvalues and eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. Let $\mathbf{A}$ be an $m \times p$ matrix of rank k . Then the singular value decomposition of $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \quad (2.117)$$

where $\mathbf{U}$ is $m \times k$, $\mathbf{D}$ is $k \times k$, and $\mathbf{V}$ is $p \times k$. The diagonal elements of the nonsingular diagonal matrix $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$ are the positive square roots of

$\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2$, which are the nonzero eigenvalues of $\mathbf{A}^t\mathbf{A}$ or of $\mathbf{A}\mathbf{A}^t$. The values $\lambda_1, \lambda_2, \dots, \lambda_n$ are called the singular values of $\mathbf{A}$. The i columns of $\mathbf{U}$ are the normalized eigenvectors of $\mathbf{A}\mathbf{A}^t$ corresponding to the eigenvalues $\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2$. The i columns of $\mathbf{V}$ are the normalized eigenvectors of $\mathbf{A}^t\mathbf{A}$ corresponding to the eigenvalues $\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2$. Since the columns of $\mathbf{U}$ and of $\mathbf{V}$ are (normalized) eigenvectors of symmetric matrices, they are mutually orthogonal (see Section 2.1.4(i)), and we have $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}$.

PROBLEMS

2.1 Let

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & 3 \\ 7 & 5 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 3 & -2 & 4 \\ 6 & 9 & -3 \end{pmatrix}.$$

- (a) Find $\mathbf{A} + \mathbf{B}$ and $\mathbf{A} - \mathbf{B}$.
- (b) Find $\mathbf{A}^t\mathbf{A}$ and $\mathbf{A}\mathbf{A}^t$.

2.2 Use the matrices $\mathbf{A}$ and $\mathbf{B}$ in Problem 2.1:

- (a) Find $(\mathbf{A} + \mathbf{B})^t$ and $\mathbf{A}^t + \mathbf{B}^t$ and compare them, thus illustrating (2.17).
- (b) Show that $(\mathbf{A}^t)^t = \mathbf{A}$, thus illustrating (2.8).

2.3 Let

$$\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 2 & 0 \\ 1 & 5 \end{pmatrix}.$$

- (a) Find $\mathbf{A}\mathbf{B}$ and $\mathbf{B}\mathbf{A}$.
- (b) Find $(\mathbf{A}\mathbf{B})^t$, $(\mathbf{A})^t(\mathbf{B})^t$, and $(\mathbf{B}\mathbf{A})^t$ and verify that (2.49) holds in this case.

2.4 Use the matrices $\mathbf{A}$ and $\mathbf{B}$ in Problem 2.3:

- (a) Find $\mathbf{A} + \mathbf{B}$ and $\text{tr}(\mathbf{A} + \mathbf{B})$.
- (b) Find $\text{tr}(\mathbf{A})$ and $\text{tr}(\mathbf{B})$ and show that (2.56) holds for these matrices.

2.5 Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & -1 & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 3 & -2 \\ 2 & 0 \\ -1 & 1 \end{pmatrix}.$$

- (a) Find $\mathbf{A}\mathbf{B}$ and $\mathbf{B}\mathbf{A}$.
- (b) Compare $\text{tr}(\mathbf{A}\mathbf{B})$ and $\text{tr}(\mathbf{B}\mathbf{A})$, and conclude that (2.67) holds here.

2.6 Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 6 \\ 3 & 10 & 15 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} -1 & 1 & -2 \\ -1 & 1 & -2 \\ 1 & -1 & 1 \end{pmatrix}.$$

- (iii) Show that $A\mathbf{B} = \mathbf{O}$.
 (iv) Find a vector $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{b}$.
 (v) Show that $|\mathbf{A}| = 0$.

3.7 Let

$$A = \begin{pmatrix} 1 & -1 & 4 \\ -1 & 1 & 3 \\ 4 & 3 & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 3 & -2 & 4 \\ 2 & 1 & 0 \\ 2 & 3 & 5 \end{pmatrix},$$

$$\mathbf{x} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}.$$

Find the following:

- | | | |
|------------------------------|---------------------------------------|-----------------------------|
| (a) $\mathbf{B}\mathbf{x}$ | (d) $\mathbf{x}'\mathbf{B}\mathbf{y}$ | (g) $\mathbf{x}\mathbf{x}'$ |
| (b) $\mathbf{y}'\mathbf{B}$ | (e) $\mathbf{x}'\mathbf{x}$ | (h) $\mathbf{y}\mathbf{y}'$ |
| (c) $\mathbf{x}'A\mathbf{x}$ | (f) $\mathbf{x}'\mathbf{y}$ | (i) $\mathbf{B}'\mathbf{B}$ |

3.8 Use $\mathbf{x}$, $\mathbf{y}$, and A as defined in Problem 3.7:

- (i) Find $\mathbf{x} + \mathbf{y}$ and $\mathbf{x} - \mathbf{y}$.
 (ii) Find $(\mathbf{x} - \mathbf{y})'A(\mathbf{x} - \mathbf{y})$.

3.9 Using $\mathbf{B}$ and $\mathbf{x}$ in Problem 3.7, find $\mathbf{B}\mathbf{x}$ as a linear combination of columns of $\mathbf{B}$, as in (2.47) and compare with $\mathbf{B}\mathbf{x}$ found in Problem 3.7(a).

3.10 Let

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 4 & 2 \\ 5 & 0 & 3 \end{pmatrix}, \quad \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

- (i) Show that $(A\mathbf{B})' = \mathbf{B}'A'$ as in (2.37).
 (ii) Show that $A\mathbf{I} = A$ and that $\mathbf{I}\mathbf{B} = \mathbf{B}$.
 (iii) Find $\det A$.

3.11 Let

$$\mathbf{a} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}.$$

- (i) Find $\mathbf{a}\mathbf{b}$ and $(\mathbf{a}\mathbf{b})^T$.
 (ii) Find $(\mathbf{b}\mathbf{b})'$ and $\mathbf{a}'(\mathbf{b}\mathbf{b})'\mathbf{a}$.
 (iii) Compare $\mathbf{a}'\mathbf{b}^T$ with $\mathbf{a}'(\mathbf{b}\mathbf{b})'\mathbf{a}$ and then illustrate (2.43).

2.12 Let

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \quad B = \begin{pmatrix} a & b & c \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}.$$

Find DA , AB , and BD .

2.13 Let the matrices A and B be partitioned as follows:

$$A = \left(\begin{array}{cc|c} 2 & 1 & 1 \\ 3 & 2 & 0 \\ \hline 1 & 0 & 1 \end{array} \right), \quad B = \left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 2 & 1 & 1 & 2 \\ \hline 2 & 3 & 1 & 2 \end{array} \right).$$

- Find AB as in (2.65) using the indicated partitioning.
- Check by finding AB in the usual way, ignoring the partitioning.

2.14 Let

$$A = \begin{pmatrix} 1 & 5 & 2 \\ 2 & 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 & 1 \\ 3 & -6 & -4 \end{pmatrix}.$$

Find AB and CB . Are they equal? What is the rank of A , B , and C ?

2.15 Let

$$A = \begin{pmatrix} 5 & 4 & 4 \\ 2 & -3 & 1 \\ 3 & 7 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 2 & 3 \end{pmatrix}.$$

- Find $\text{tr}(A)$ and $\text{tr}(B)$.
- Find $A + B$ and $\text{tr}(A + B)$. Is $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$?
- Find $|A|$ and $|B|$.
- Find AB and (AB) . Is $(AB) = (A)(B)$?

2.16 Let

$$A = \begin{pmatrix} 2 & 4 & 3 \\ 4 & 8 & 6 \\ 3 & 6 & 9 \end{pmatrix}.$$

- Show that $|A| = 0$.
- Using the Cholesky decomposition in Section 2.7, find an upper triangular matrix T such that $A = T^T T$.

2.17 Let

$$\mathbf{A} = \begin{pmatrix} 2 & 3 & -5 & -1 \\ -2 & 13 & 0 & \\ -1 & 0 & 1 & \end{pmatrix}.$$

- Show that $|\mathbf{A}| = 5$.
- Using the Cholesky decomposition in Section 2.7, find an upper triangular matrix $\mathbf{T}$ such that $\mathbf{A} = \mathbf{T}\mathbf{T}^T$.

2.18 The columns of the following matrix are mutually orthogonal:

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 1 \\ 2 & 1 & 0 \\ 1 & -1 & -1 \end{pmatrix}.$$

- Normalize the columns of $\mathbf{A}$ by dividing each column by its length, denote the resulting matrix by $\mathbf{C}$.
- Show that $\mathbf{C}$ is an orthogonal matrix, that is, $\mathbf{C}\mathbf{C}^T = \mathbf{C}^T\mathbf{C} = \mathbf{I}$.

2.19 Let

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & -2 \\ -1 & 2 & 1 \\ 2 & 1 & -1 \end{pmatrix}.$$

- Find the eigenvalues and associated normalised eigenvectors.
- Find $\text{tr}(\mathbf{A})$ and $|\mathbf{A}|$ and show that $\text{tr}(\mathbf{A}) = \sum_{i=1}^3 \lambda_i$ and $|\mathbf{A}| = \prod_{i=1}^3 \lambda_i$.

2.20 Let

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}.$$

- The eigenvalues of $\mathbf{A}$ are 1, 4, -2 . Find the normalised eigenvectors and use them as columns in an orthogonal matrix $\mathbf{C}$.
- Show that $\mathbf{C}^T\mathbf{A}\mathbf{C} = \mathbf{B}$ as in (2.117), where $\mathbf{B}$ is diagonal with the eigenvalues of $\mathbf{A}$ on the diagonal.
- Show that $\mathbf{A} = \mathbf{C}\mathbf{B}\mathbf{C}^T$ as in (2.120).

2.21 For the positive definite matrix

$$\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix},$$

calculate the eigenvalues and eigenvectors and find the square root matrix $\mathbf{A}^{1/2}$ as in (2.131). Check by showing that $(\mathbf{A}^{1/2})^2 = \mathbf{A}$.

2.21 Let

$$\mathbf{A} = \begin{pmatrix} 2 & 5 & -1 \\ 8 & 9 & 4 \\ -3 & 4 & 3 \end{pmatrix}.$$

- Find the spectral decomposition of $\mathbf{A}$ as in (2.107).
- Find the spectral decomposition of $\mathbf{A}^2$ and show that the diagonal matrix of eigenvalues is equal to the square of the matrix $\mathbf{D}$ found in part (a), thus illustrating (2.115).
- Find the spectral decomposition of $\mathbf{A}^{-1}$ and show that the diagonal matrix of eigenvalues is equal to the inverse of the matrix $\mathbf{D}$ found in part (a), thus illustrating (2.116).

2.22 Find the singular value decomposition of $\mathbf{A}$ as in (2.117), where

$$\mathbf{A} = \begin{pmatrix} 4 & -5 & -1 \\ 7 & -2 & 3 \\ -1 & 4 & -5 \\ 6 & 2 & 6 \end{pmatrix}.$$

2.23 If $\mathbf{j}$ is a vector of 1's, as defined in (2.11), show that the following hold:

- $\mathbf{j}\mathbf{A} = \mathbf{A}\mathbf{j} = \sum_{i=1}^n a_i$ as in (2.37).
- $\mathbf{j}\mathbf{A}$ is a row vector whose elements are the column sums of $\mathbf{A}$ as in (2.38).
- $\mathbf{A}\mathbf{j}$ is a column vector whose elements are the row sums of $\mathbf{A}$ as in (2.38).

2.24 Verify (2.41); that is, show that $(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y}) = \mathbf{x}'\mathbf{x} - 2\mathbf{x}'\mathbf{y} + \mathbf{y}'\mathbf{y}$.

2.25 Show that $\mathbf{A}'\mathbf{A}$ is symmetric, where $\mathbf{A}$ is $n \times p$.

2.26 If $\mathbf{a}$ and $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ are all $p \times 1$ and $\mathbf{A}$ is $p \times p$, show that (2.42)–(2.45) hold:

- $\sum_{i=1}^p \mathbf{a}'\mathbf{u}_i = \mathbf{a}' \sum_{i=1}^p \mathbf{u}_i$.
- $\sum_{i=1}^p \mathbf{A}\mathbf{u}_i = \mathbf{A} \sum_{i=1}^p \mathbf{u}_i$.
- $\sum_{i=1}^p (\mathbf{a}'\mathbf{u}_i)^2 = \mathbf{a}'(\sum_{i=1}^p \mathbf{u}_i\mathbf{u}_i')\mathbf{a}$.
- $\sum_{i=1}^p \mathbf{A}\mathbf{u}_i(\mathbf{A}\mathbf{u}_i)' = \mathbf{A}(\sum_{i=1}^p \mathbf{u}_i\mathbf{u}_i')\mathbf{A}'$.

2.27 Assume that $\mathbf{A} = \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}$ is $2 \times p$, $\mathbf{x}$ is $p \times 1$, and $\mathbf{B}$ is $p \times p$.

- Show that

$$\mathbf{A}\mathbf{x} = \begin{pmatrix} \mathbf{a}'\mathbf{x} \\ \mathbf{b}'\mathbf{x} \end{pmatrix},$$

as in (2.46).

(b) Show that

$$AA^* = \begin{pmatrix} \sum_{j=1}^n a_{1j}a_{1j} & \sum_{j=1}^n a_{1j}a_{2j} \\ \sum_{j=1}^n a_{2j}a_{1j} & \sum_{j=1}^n a_{2j}a_{2j} \end{pmatrix},$$

as in (2.50).

2.29 (a) If the rows of A are denoted by a_i , show that $A^*A = \sum_{i=1}^n aa_i^*$ as in (2.51).

(b) If the columns of A are denoted by a_{ij} , show that $AA^* = \sum_{j=1}^n a_{ij}a_{ij}^*$ as in (2.50).

2.30 Show that $(A^*)^{-1} = (A^{-1})^*$ as in (2.73).

2.31 Show that the inverse of the partitioned matrix given in (2.76) is correct by multiplying by

$$\begin{pmatrix} B_{11} & a_1 \\ a_1^* & cca \end{pmatrix}$$

to obtain an identity.

2.32 Show that the inverse of $B + aa^*$ given in (2.77) is correct by multiplying by $B + aa^*$ to obtain an identity.

2.33 Show that $\gamma(A) = \gamma^2(A)$ as in (2.85).

2.34 Show that $(A^{-1})^* = 1/\gamma(A)$ as in (2.81).

2.35 If B is nonsingular and a is a vector, show that $(B + aa^*)^{-1} = (B)^{-1} + a^*(B^{-1})^{-1}a$ as in (2.93).

2.36 Show that $\text{tr}(A^*A) = \text{tr}(AA^*) = \sum_{i,j} a_{ij}^2$ as in (2.93).

2.37 Show that $CC^* = I$ in (2.102) follows from $C^*C = I$ in (2.101).

2.38 Show that the eigenvalues of AB are the same as those of BA , as noted in Section 2.11.3.

2.39 If $A^{1/2}$ is the square root matrix defined in (2.111), show that

(a) $(A^{1/2})^2 = A$ as in (2.114),

(b) $(A^{1/2})^2 = |A|$,

(c) $|A^{1/2}| = |A|^{1/2}$.

CHAPTER 3

Characterizing and Displaying Multivariate Data

We review some univariate and bivariate procedures in Sections 3.1, 3.2, and 3.3 and then extend them to vectors of higher dimension in the remainder of the chapter.

3.1 MEAN AND VARIANCE OF A UNIVARIATE RANDOM VARIABLE

Informally, a *random variable* may be defined as a variable whose value depends on the outcome of a chance experiment. Generally, we will consider only *continuous* random variables. Some types of multivariate data are only approximations to this ideal, such as test scores or a series-point semantic differential (Likert) scale consisting of ordered responses ranging from strongly disagree to strongly agree. Special techniques have been developed for such data, but in many cases, the usual methods designed for continuous data work almost as well.

The *density function* $f(x)$ indicates the relative frequency of occurrence of the random variable x . (We do not use F to denote the random variable for reasons given at the beginning of Section 3.5.) Thus, if $f(x_1) > f(x_2)$, then points in the neighborhood of x_1 are more likely to occur than points in the neighborhood of x_2 .

The *population mean* of a random variable y is defined informally as the mean of all possible values of y and is denoted by μ . The mean is also related to us the *expected value* of y , or $E(y)$. If the density $f(y)$ is known, the mean can sometimes be found using methods of calculus, but we will not use those techniques in this text.

If $f(y)$ is unknown, the population mean μ will naturally vary with time or chance if it has been established from extensive past experience with a stable population. If a large random sample from the population represented by $f(y)$ is available, it is highly probable that the mean of the sample is close to μ .

The *sample mean* of a random sample of n observations $y_1, y_2, \dots, y_n$ is given by the ordinary arithmetic average

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (3.1)$$

Generally, $\bar{y}$ will never be equal to μ ; by this we mean that the probability is zero that a sample will ever arise in which $\bar{y}$ is exactly equal to μ . However, $\bar{y}$ is considerably good estimator for μ because $E(\bar{y}) = \mu$ and $\text{var}(\bar{y}) = \sigma^2/n$, where σ^2 is the variance of y . In other words, $\bar{y}$ is an unbiased estimator of μ and has a smaller variance than a single observation y . The variance σ^2 is defined shortly. The notation $E(\bar{y})$ indicates the mean of all possible values of $\bar{y}$ that is, conceptually, every possible sample is obtained from the population, the mean of each is found, and the average of all these sample means is calculated.

If every y in the population is multiplied by a constant a , the expected value is also multiplied by a :

$$E(a\bar{y}) = aE(\bar{y}) = a\mu. \quad (3.2)$$

The sample mean has a similar property. If $x_i = ay_i$ for $i = 1, 2, \dots, n$, then

$$\bar{x} = a\bar{y}. \quad (3.3)$$

The variance of the population is defined as $\text{var}(y) = \sigma^2 = E(y - \mu)^2$. This is the average squared deviation from the mean and is thus an indication of the extent to which the values of y are spread or scattered. It can be shown that $\sigma^2 = E(y^2) - \mu^2$.

The sample variance is defined as

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}, \quad (3.4)$$

which can be shown to be equal to

$$s^2 = \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{n-1}. \quad (3.5)$$

The sample variance s^2 is generally never equal to the population variance σ^2 (the probability of such an occurrence is zero), but it is an unbiased estimator for σ^2 ; that is, $E(s^2) = \sigma^2$. Again the notation $E(s^2)$ indicates the mean of all possible sample variances. The square root of either the population variance or sample variance is called the standard deviation.

If each y is multiplied by a constant a , the (population) variance is multiplied by a^2 , that is, $\text{var}(ay) = a^2\sigma^2$. Similarly, if $x_i = ay_i$, $i = 1, 2, \dots, n$, then the sample variance of x is given by

$$s_x^2 = a^2s^2. \quad (3.6)$$

3.2 COVARIANCE AND CORRELATION OF BIVARIATE RANDOM VARIABLES

3.2.1 Covariance

If two variables x and y are measured on each research unit (object or subject), we have a bivariate random variable (x, y) . Often, x and y will tend to covary; if one is above its mean, the other is more likely to be above its mean, and vice versa. For example, height and weight were observed for a sample of 20 college-age males. The data are given in Table 3.1.

The values of height x and weight y from Table 3.1 are both plotted in the vertical direction in Figure 3.1. The tendency for x and y to stay on the same side of the mean

Table 3.1. Height and Weight for a Sample of 20 College-age Males

Person	Height x	Weight y	Person	Height x	Weight y
1	69	153	11	72	148
2	74	175	12	78	205
3	68	155	13	74	185
4	70	159	14	67	112
5	72	172	15	66	148
6	67	188	16	71	188
7	66	139	17	74	169
8	70	167	18	73	185
9	76	208	19	73	218
20	68	146	20	76	228

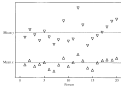


Figure 3.1. Two variables with a tendency to covary.

is clear in Figure 3.1. This illustrates positive covariance. With negative covariance the points would tend to deviate simultaneously to opposite sides of the mean.

The population covariance is defined as $\text{cov}(x, y) = \sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$, where μ_x and μ_y are the means of x and y , respectively. Thus if x and y are usually both above their means or both below their means, the product $(x - \mu_x)(y - \mu_y)$ will typically be positive, and the average value of the product will be positive. Conversely, if x and y tend to fall on opposite sides of their respective means, the product will usually be negative and the average product will be negative. It can be shown that $\sigma_{xy} = E(xy) - \mu_x\mu_y$.

If the two random variables x and y in a bivariate random variable are added or multiplied, a new random variable is obtained. The mean of $x + y$ or of xy is as follows:

$$E(x + y) = E(x) + E(y) \quad (3.7)$$

$$E(xy) = E(x)E(y) \quad \text{if } x \text{ and } y \text{ are independent.} \quad (3.8)$$

Formally, x and y are independent if their joint density factors into the product of their individual densities: $f(x, y) = g(x)h(y)$. Informally, x and y are independent if the random behavior of either of the variables is not affected by the behavior of the other. Note that (3.7) is true whether or not x and y are independent, but (3.8) holds only for x and y independently distributed.

The notion of independence of x and y is more general than that of zero covariance. The covariance σ_{xy} measures linear relationship only, whereas if two random variables are independent, they are not related either linearly or nonlinearly. Independence implies $\sigma_{xy} = 0$, but $\sigma_{xy} = 0$ does not imply independence. It is easy to show that if x and y are independent, then $\sigma_{xy} = 0$:

$$\begin{aligned} \sigma_{xy} &= E(xy) - \mu_x\mu_y \\ &= E(x)E(y) - \mu_x\mu_y \quad \text{[by (3.8)]} \\ &= \mu_x\mu_y - \mu_x\mu_y = 0. \end{aligned}$$

One way to demonstrate that the converse is not true is to construct examples of bivariate x and y that have zero covariance and yet are related in a nonlinear way (the relationship will have zero slope). This is illustrated in Figure 3.2.

If x and y have a bivariate normal distribution (see Chapter 4), then zero covariance implies independence. This is because (1) the covariance measures only linear relationships and (2) in the bivariate normal case, the means of y given x (or x given y) is a straight line.

The sample covariance is defined as

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}, \quad (3.9)$$

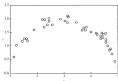


Figure 3.1. A sample from a population where x and y have zero covariance and yet are dependent.

It can be shown that

$$s_{xy} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{n - 1}. \quad (3.1.6)$$

Note that s_{xy} is essentially never equal to s_{yx} (for continuous data); that is, the probability is zero that s_{xy} will equal s_{yx} . It is true, however, that s_{xy} is an unbiased estimator for σ_{xy} , that is, $E(s_{xy}) = \sigma_{xy}$.

Since $s_{xy} \neq s_{yx}$ in any given sample, this is also true when $\sigma_{xy} = 0$. Thus when the population covariance is zero, no random sample from the population will have zero covariance. The only way a sample from a continuous bivariate distribution will have zero covariance is for the experimenter to choose the values of x and y so that $s_{xy} = 0$. (Such a sample would not be a random sample.) One way to achieve this is to place the values in the form of a grid. This is illustrated in Figure 3.2.

The sample covariance measures only linear relationships. If the points in a bivariate sample follow a curved trend, as, for example, in Figure 3.1, the sample covariance will not measure the strength of the relationship. To see that s_{xy} measures only linear relationships, note that the slope of a simple linear regression line is

$$\hat{A}_1 = \frac{\sum_{i=1}^n x_i (y_i - \bar{y}) / (y_i - \bar{y})}{\sum_{i=1}^n x_i^2 / (y_i - \bar{y})} = \frac{s_{xy}}{s_x^2}. \quad (3.1.7)$$

Thus s_{xy} is proportional to the slope, which shows only the linear relationship between y and x .

Variables with zero sample covariance can be said to be orthogonal. By (2.6F), two sets of variables $x_1, x_2, \dots, x_k$ and $y_1, y_2, \dots, y_k$ are orthogonal if $\sum_{i=1}^k x_i y_i =$

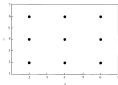


Figure 3.3. A sample of (x, y) values with zero correlation.

4. This is true for the centered variables $x_i - \bar{x}$ and $y_i - \bar{y}$ where the sample covariance is zero, that is, $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 0$.

Example 3.3.1. To obtain the sample variances for the height and weight data in Table 3.1, we first calculate $\bar{x}$, $\bar{y}$, and $\sum_{i=1}^n x_i y_i$, where x is height and y is weight:

$$\bar{x} = \frac{68 + 74 + \cdots + 78}{20} = 71.45,$$

$$\bar{y} = \frac{153 + 179 + \cdots + 220}{20} = 164.7,$$

$$\sum_{i=1}^{20} x_i y_i = (68)(153) + (74)(179) + \cdots + (76)(220) = 283,805.$$

Now, by (3.16), we have

$$\begin{aligned} s_{xy} &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n-1} \\ &= \frac{283,805 - (20)(71.45)(164.7)}{19} = 126.88. \end{aligned}$$

By itself, the sample covariance 126.88 is not very meaningful. We are not sure if this represents a small, moderate, or large amount of relationship between y and x . A method of standardizing the covariance is given in the next section. $\square$

3.1.2 Correlation

Since the covariance depends on the scale of measurement of x and y , it is difficult to compare covariances between different pairs of variables. For example, if we change a measurement from inches to centimeters, the covariance will change. To find a measure of linear relationship that is invariant to changes of scale, we can standardize the covariance by dividing by the standard deviations of the two variables. This standardized covariance is called a *correlation*. The *population correlation* of two random variables x and y is

$$\rho_{xy} = \text{corr}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{E\{(x - \mu_x)(y - \mu_y)\}}{\sqrt{E\{x - \mu_x\}^2} \sqrt{E\{y - \mu_y\}^2}}, \quad (3.1.2)$$

and the *sample correlation* is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (3.1.3)$$

Either of these correlations will range between -1 and 1 .

The sample correlation r_{xy} is related to the cosine of the angle between two vectors. Let θ be the angle between vectors $\mathbf{a}$ and $\mathbf{b}$ in Figure 3.4. The vector from the terminal point of $\mathbf{a}$ to the terminal point of $\mathbf{b}$ can be represented as $\mathbf{c} = \mathbf{b} - \mathbf{a}$. Then the law of cosines can be stated in vector form as

$$\cos \theta = \frac{\mathbf{a} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} - (\mathbf{b} - \mathbf{a}) \cdot (\mathbf{b} - \mathbf{a})}{2\sqrt{\mathbf{a} \cdot \mathbf{a}} \sqrt{\mathbf{b} \cdot \mathbf{b}}}$$

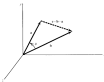


Figure 3.4. Vectors $\mathbf{a}$ and $\mathbf{b}$ in 2-space.

$$\begin{aligned}
&= \frac{\mathbf{a}'\mathbf{a} + \mathbf{b}'\mathbf{b} - (\mathbf{b}'\mathbf{b} + \mathbf{a}'\mathbf{a} - 2\mathbf{a}'\mathbf{b})}{2\sqrt{(\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b})}} \\
&= \frac{\mathbf{a}'\mathbf{b}}{\sqrt{(\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b})}}.
\end{aligned} \tag{3.14}$$

Since $\cos(90^\circ) = 0$, we see from (3.14) that $\mathbf{a}'\mathbf{b} = 0$ when $\theta = 90^\circ$. Thus $\mathbf{a}$ and $\mathbf{b}$ are perpendicular when $\mathbf{a}'\mathbf{b} = 0$. By (2.89), two vectors $\mathbf{a}$ and $\mathbf{b}$, such that $\mathbf{a}'\mathbf{b} = 0$, are also said to be orthogonal. Hence orthogonal vectors are perpendicular in a geometric sense.

To express the correlation in the form given in (3.14), let the n observation vectors $(x_1, y_1), (x_2, y_1), \dots, (x_n, y_1)$ be two dimensional (or represented as two vectors $\mathbf{x}' = (x_1, x_2, \dots, x_n)$ and $\mathbf{y}' = (y_1, y_2, \dots, y_n)$ in n dimensions, and let $\bar{x}$ and $\bar{y}$ be centered as $\mathbf{x} - \bar{x}\mathbf{j}$ and $\mathbf{y} - \bar{y}\mathbf{j}$. Then the cosine of the angle θ between them [see (3.14)] is equal to the sample correlation between x and y :

$$\begin{aligned}
\cos \theta &= \frac{(\mathbf{x} - \bar{x}\mathbf{j})'(\mathbf{y} - \bar{y}\mathbf{j})}{\sqrt{[(\mathbf{x} - \bar{x}\mathbf{j})'(\mathbf{x} - \bar{x}\mathbf{j})][(\mathbf{y} - \bar{y}\mathbf{j})'(\mathbf{y} - \bar{y}\mathbf{j})]}} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \\
&= r_{xy}.
\end{aligned} \tag{3.15}$$

Thus if the angle θ between the two centered vectors $\mathbf{x} - \bar{x}\mathbf{j}$ and $\mathbf{y} - \bar{y}\mathbf{j}$ is small so that $\cos \theta$ is near 1, r_{xy} will be close to 1. If the two vectors are perpendicular, $\cos \theta$ and r_{xy} will be zero. If the two vectors have nearly opposite directions, r_{xy} will be close to -1 .

Example 3.2.1. To obtain the correlation for the height and weight data of Table 3.1, we first calculate the sample variance of x :

$$s_x^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{302,379 - (20)(71.45)^2}{19} = 14.378.$$

Then $s_x = \sqrt{14.378} = 3.8179$ and, similarly, $s_y = 77.964$. By (3.15), we have

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{126.68}{(3.8179)(77.964)} = .899. \quad \square$$

3.3 SCATTER PLOTS OF BIVARIATE SAMPLES

Figures 3.1 and 3.2 are examples of scatter plots of bivariate samples. In Figure 3.1, the two variables x and y were plotted separately for the data in Table 3.1. Figure 3.2 shows a bivariate scatter plot of the same data.

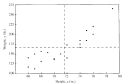


Figure 3.5. Rotated scatter plot of the data in Figure 3.1.

If the origin is shifted to $(\bar{x}, \bar{y})$, as indicated by the dashed lines, then the first and third quadrants contain most of the points. Scatter plots for correlated data typically show a substantial positive or negative slope.

A hypothetical sample of the uncorrelated variables height and IQ is shown in Figure 3.6. We could change the shape of the swarm of points by altering the scale on either axis. But, because of the independence assumed for these variables, each quadrant is likely to have as many points as any other quadrant. A tall person is as likely to have a high IQ as a low IQ. A person of low IQ is as likely to be short as to be tall.

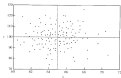


Figure 3.6. A sample of data from a population where x and y are uncorrelated.

3.4 GRAPHICAL DISPLAYS FOR MULTIVARIATE SAMPLES

It is a relatively simple procedure to plot bivariate samples as in Section 3.3. The position of a point shows at once the value of both variables. However, for three- or more variables it is a challenge to show graphically the values of all the variables in an observation vector $\mathbf{y}$. On a two-dimensional plot, the value of a third variable could be indicated by color or intensity or size of the plotted point. Four dimensions might be represented by starting with a two-dimensional scatter plot and adding two additional dimensions as line segments at right angles, as in Figure 3.7. The "corner point" represents y_1 and y_2 , whereas y_3 and y_4 are given by the lengths of the two line segments.

We will now describe various methods proposed for representing p dimensions in a plot of an observation vector, where $p > 2$.

Profiles represent each point by p vertical bars, with the heights of the bars depicting the values of the variables. Sometimes the profile is outlined by a polygonal line rather than bars.

Stars portray the value of each (normalized) variable as a point along a ray from the center to the outside of a circle. The points on the rays are usually joined to form a polygon.

Clype (Anderson 1960) are circles of fixed size with rays whose lengths represent the values of the variables. Anderson suggested using only three lengths of rays, thus rounding the variable values to three levels.

Icons (Chernoff 1973) depict each variable as a feature on a face, such as length of nose, size of eyes, shape of eyes, and so on. Flury and Rindwayl (1980) suggested using any metric faces, thus increasing the number of representable variables.



Figure 3.7. Four-dimensional plot.

Boxes (Hartigan 1975) show each variable as the length of a dimension of a box. For more than three variables, the dimensions are partitioned into segments.

Among these five methods, Chambers and Kleiner (1982) prefer the star plot because they "combine a reasonably distinctive appearance with computational simplicity and ease of interpretation." Commenting on the other methods, they state, "Profiles are not so easy to compare as a general shape. Faces are memorable, but they are more complex to draw, and are most be useful in assigning variables to parameters and in choosing parameter ranges. Faces to some extent disguise the data in the sense that individual data values may not be directly comparable from the plot."

Table 3.1. Percentages of Republican Votes in Residential Elections in Six Southern States for Selected Years

State	1902	1910	1918	1926	1934	1942
Missouri	38	38	48	66	66	43
Maryland	36	27	41	66	55	42
Kentucky	43	40	42	54	66	44
Louisiana	7	11	14	28	57	29
Mississippi	4	3	4	25	67	14
South Carolina	2	1	4	49	59	50

Example 3.4. The data in Table 3.1 are from Kleiner and Hartigan (1981). For these data, the preceding five graphical devices are illustrated in Figure 3.8. The relative magnitudes of the variables can be compared more readily using stars or profiles than faces. $\square$

3.2 MEAN VECTORS

It is a common practice in many texts to use an uppercase letter for a variable name and the corresponding lowercase letter for a particular value or observed value of the random variable, for example, $P(P = p)$. This notation is convenient in some univariate contexts, but it is often confusing in multivariate analysis, where we use uppercase letters for matrices. In the belief that it is easier to distinguish between a random vector and an observed value than between a vector and a matrix, throughout this text we follow the notation established in Chapter 2. Uppercase boldface letters are used for matrices of random variables or constants, lowercase boldface letters represent vectors of random variables or constants, and univariate random variables or constants are usually represented by lowercase nonbolded letters.

Let $\mathbf{y}$ represent a random vector of p variables measured on a sampling unit (subject or object). If there are n individuals in the sample, then a observation vectors are

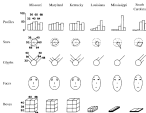


Figure 3A. Poses, size, shape, face, and bones of percentage of Republican votes in six presidential elections in six southern states. The values of the poses in the states in 1976. Asymptotic of rectangular-faced boxplots in 1972, shape of face, 1976, length of nose, 1980, variation of mouth, 1988, width of mouth, 1994, start of eyes, and 1998, length of eyebrows. (From the *Journal of the American Statistical Association*, 1998, p. 262.)

denoted by $y_1, y_2, \dots, y_p$, where

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix}.$$

The sample mean vector $\bar{y}$ can be found either as the average of the n observation vectors or by calculating the average of each of the p variables separately:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{pmatrix}. \quad (3.10)$$

where, for example, $\bar{Y}_1 = \sum_{i=1}^n y_{1i}/n$. Thus $\bar{Y}_1$ is the mean of the n observations on the first variable, $\bar{Y}_2$ is the mean of the second variable, and so on.

All n observation vectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ can be transposed to row vectors and listed in the data matrix $\mathbf{Y}$ as follows:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = (\text{matrix}) \begin{matrix} & \begin{matrix} \text{(variables)} \\ 1 & 2 & \dots & j & \dots & p \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ n \end{matrix} & \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1j} & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2j} & \dots & y_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{i1} & y_{i2} & \dots & y_{ij} & \dots & y_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nj} & \dots & y_{np} \end{pmatrix} \end{matrix} \quad (3.17)$$

Since n is usually greater than p , the data can be more conveniently tabulated by entering the observation vectors as rows rather than columns. Note that the first subscript i corresponds to units (subjects or objects) and the second subscript j refers to variables. This convention will be followed whenever possible.

In addition to the two ways of calculating $\bar{\mathbf{y}}$ given in (3.16), we can obtain $\bar{\mathbf{y}}$ from $\mathbf{Y}$. We sum the n entries in each column of $\mathbf{Y}$ and divide by n , which gives $\bar{\mathbf{y}}$. This can be indicated in matrix notation using (3.16),

$$\bar{\mathbf{y}} = \frac{1}{n} \mathbf{1}' \mathbf{Y}, \quad (3.18)$$

where $\mathbf{1}$ is a vector of 1 's, as defined in (2.11). For example, the second element of $\mathbf{1}' \mathbf{Y}$ is

$$(1, 1, \dots, 1) \begin{pmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n2} \end{pmatrix} = \sum_{i=1}^n y_{i2}.$$

We can transpose (3.18) to obtain

$$\bar{\mathbf{y}} = \frac{1}{n} \mathbf{Y}' \mathbf{1}. \quad (3.19)$$

We now turn to populations. The mean of y over all possible values in the population is called the *population mean vector* or *expected value* of $\mathbf{y}$. It is defined as a vector of expected values of each variable,

$$E(\mathbf{y}) = E\left(\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}\right) = \begin{pmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}, \quad (3.20)$$

where μ_j is the population mean of the j th variable.

It can be shown that the expected value of each $\bar{y}_j$ in $\bar{\mathbf{y}}$ is μ_j ; that is, $E(\bar{y}_j) = \mu_j$. Thus the expected value of $\bar{\mathbf{y}}$ (over all possible samples) is

$$E(\bar{\mathbf{y}}) = E\left(\begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{pmatrix}\right) = \begin{pmatrix} E(\bar{y}_1) \\ E(\bar{y}_2) \\ \vdots \\ E(\bar{y}_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}. \quad (3.21)$$

Therefore, $\bar{\mathbf{y}}$ is an unbiased estimator of $\boldsymbol{\mu}$. We emphasize again that $\bar{\mathbf{y}}$ is never equal to $\boldsymbol{\mu}$.

Example 3.5. Table 3.5 gives partial data from Kramer and Jensen (1989a). Three variables were measured (in milliequivalents per 100 g) at 10 different locations in the South. The variables are

y_1 = available calcium,

y_2 = exchangeable soil calcium,

y_3 = turnip greens calcium.

To find the mean vector $\bar{\mathbf{y}}$, we simply calculate the average of each column and obtain

$$\bar{\mathbf{y}} = (28.1, 7.18, 3.08)^T. \quad \square$$

Table 3.5. Calcium in Soil and Turnip Greens

Location Number	y_1	y_2	y_3
1	16	0.5	2.89
2	16	4.9	2.79
3	40	10.0	4.18
4	10	1.8	3.23
5	6	2.7	2.75
6	20	2.8	2.81
7	16	4.8	2.88
8	16	10.0	2.89
9	16	8.0	3.28
10	10	1.6	3.28

3.4 COVARIANCE MATRICES

The *sample covariance matrix* $\mathbf{S} = (s_{jk})$ is the matrix of sample variances and covariances of the p variables:

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}. \quad (3.22)$$

In $\mathbf{S}$ the sample variances of the p variables are on the diagonal, and all possible pairwise sample covariances appear off the diagonal. The j th row (column) contains the covariances of y_j with the other $p - 1$ variables.

We give three approaches to obtaining $\mathbf{S}$. The first of these is to simply calculate the individual elements s_{jk} . The sample variance of the j th variable, $s_{jj} = s_j^2$, is calculated as in (3.4) or (3.5), using the j th column of $\mathbf{Y}$:

$$s_{jj} = s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2 \quad (3.23)$$

$$= \frac{1}{n-1} \left(\sum_i y_{ij}^2 - n\bar{y}_j^2 \right), \quad (3.24)$$

where $\bar{y}_j$ is the mean of the j th variable, as in (3.16). The sample covariance of the j th and k th variables, s_{jk} , is calculated as in (3.6) or (3.18), using the j th and k th columns of $\mathbf{Y}$:

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k) \quad (3.25)$$

$$= \frac{1}{n-1} \left(\sum_i y_{ij}y_{ik} - n\bar{y}_j\bar{y}_k \right). \quad (3.26)$$

Note that in (3.23) the variance s_{jj} is expressed as s_j^2 , the square of the standard deviation s_j , and that $\mathbf{S}$ is symmetric because $s_{jk} = s_{kj}$ in (3.25). Other names used for the covariance matrix are variance matrix, variance-covariance matrix, and dispersion matrix.

By way of notational clarification, we note that in the univariate case, the sample variance is denoted by s^2 , but in the multivariate case, we denote the sample covariance matrix as $\mathbf{S}$, not as $\mathbf{S}^2$.

The sample covariance matrix $\mathbf{S}$ can also be expressed in terms of the observation vectors:

$$s = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \quad (3.27)$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right). \quad (3.28)$$

Since $(y_i - \bar{y})(x_i - \bar{x}) = (y_{i1} - \bar{y}_1)(x_{i1} - \bar{x}_1) \dots (y_{ip} - \bar{y}_p)(x_{ip} - \bar{x}_p)$, the element in the $(1, 1)$ position of $(y_i - \bar{y})(x_i - \bar{x}) = (y_{i1} - \bar{y}_1)^2$, and when this is summed over i as in (3.28), the result is the numerator of s_{11} in (3.23). Similarly, the $(1, 2)$ element of $(y_i - \bar{y})(x_i - \bar{x}) = (y_{i1} - \bar{y}_1)(x_{i2} - \bar{x}_2)$, which sums to the numerator of s_{12} in (3.23). Thus (3.27) is equivalent to (3.23) and (3.25), and Maxwell (3.28) produces (3.24) and (3.26).

We can also obtain S directly from the data matrix $\mathbf{Y}$ in (3.17), which provides a third approach. The first term in the right side of (3.26), $\sum_{i=1}^n y_{ij}x_{ij}$, is the product of the j th and i th columns of $\mathbf{Y}$, whereas the second term, $n\bar{x}_j\bar{y}_j$, is the (j, j) th element of $n\bar{\mathbf{y}}\bar{\mathbf{y}}'$. It is noted in (2.34) that $\mathbf{Y}'\mathbf{Y}$ is obtained as product of columns of $\mathbf{Y}$. By (3.18) and (3.19), $\bar{\mathbf{y}} = \mathbf{Y}'\mathbf{1}/n$ and $\bar{\mathbf{x}} = \mathbf{X}'\mathbf{1}/n$, and using (3.30), we have $n\bar{\mathbf{y}}\bar{\mathbf{y}}' = \mathbf{Y}'(\mathbf{1}\mathbf{1}')\mathbf{Y}/n$. Thus S can be written as

$$\begin{aligned} S &= \frac{1}{n-1} \left[\mathbf{Y}'\mathbf{Y} - \mathbf{Y}' \left(\frac{1}{n} \mathbf{1} \right) \mathbf{Y} \right] \\ &= \frac{1}{n-1} \mathbf{Y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y} \quad \text{[by (3.30)]}. \end{aligned} \quad (3.29)$$

Expression (3.29) is a convenient representation of S since it makes direct use of the data matrix $\mathbf{Y}$. However, the matrix $\mathbf{I} - \mathbf{J}/n$ is $n \times n$ and may be unwieldy in computation if n is large.

If $\mathbf{y}$ is a random vector taking on any possible value in a multivariate population, the population covariance matrix is defined as

$$\Sigma = \text{cov}(\mathbf{y}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}. \quad (3.30)$$

The diagonal elements $\sigma_{jj} = \sigma_j^2$ are the population variances of the y_j 's, and the off-diagonal elements σ_{jk} are the population covariances of all possible pairs of y_j 's.

The notation Σ for the covariance matrix is widely used and somewhat unfortunate because Σ is the uppercase version of σ . It should not be confused with the same symbol used for summation of a series. The difference should always be apparent from the context. To help further distinguish the two uses, the covariance matrix Σ will differ

in typeface and in size from the summation symbol Σ . Also, whenever they appear together, the summation symbol will have an index of summation, such as $\Sigma_{i=1}^n$.

The population covariance matrix in (3.33) can also be written as

$$\Sigma = E[(\mathbf{y} - \mu)(\mathbf{y} - \mu)'] \quad (3.34)$$

which is analogous to (3.27) for the sample covariance matrix. The $p \times p$ matrix $(\mathbf{y} - \mu)(\mathbf{y} - \mu)'$ is a random matrix. The expected value of a random matrix is defined as the matrix of expected values of the corresponding elements. To see that (3.34) produces population variances and covariances of the p variables as in (3.33), note that

$$\begin{aligned} \Sigma &= E[(\mathbf{y} - \mu)(\mathbf{y} - \mu)'] = E \left(\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_p - \mu_p \end{pmatrix} (y_1 - \mu_1, y_2 - \mu_2, \dots, y_p - \mu_p) \right) \\ &= E \left(\begin{pmatrix} (y_1 - \mu_1)^2 & (y_1 - \mu_1)(y_2 - \mu_2) & \cdots & (y_1 - \mu_1)(y_p - \mu_p) \\ (y_2 - \mu_2)(y_1 - \mu_1) & (y_2 - \mu_2)^2 & \cdots & (y_2 - \mu_2)(y_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (y_p - \mu_p)(y_1 - \mu_1) & (y_p - \mu_p)(y_2 - \mu_2) & \cdots & (y_p - \mu_p)^2 \end{pmatrix} \right) \\ &= \begin{pmatrix} E(y_1 - \mu_1)^2 & E(y_1 - \mu_1)(y_2 - \mu_2) & \cdots & E(y_1 - \mu_1)(y_p - \mu_p) \\ E(y_2 - \mu_2)(y_1 - \mu_1) & E(y_2 - \mu_2)^2 & \cdots & E(y_2 - \mu_2)(y_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(y_p - \mu_p)(y_1 - \mu_1) & E(y_p - \mu_p)(y_2 - \mu_2) & \cdots & E(y_p - \mu_p)^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}. \end{aligned}$$

It can be easily shown that Σ can be expressed in a form analogous to (3.34):

$$\Sigma = E(\mathbf{y}\mathbf{y}') - \mu\mu'. \quad (3.35)$$

Since $E(y_j y_k) = \sigma_{jk}$ for all j, k , the sample covariance matrix $\mathbf{S}$ is an unbiased estimator for Σ :

$$E(\mathbf{S}) = \Sigma. \quad (3.36)$$

As in the univariate case, we note that it is the average of all possible values of $\mathbf{S}$ that is equal to Σ . Generally, $\mathbf{S}$ will never be equal to Σ .

Example 3.8. To calculate the sample covariance matrix for the calcium data of Table 3.3 using the computational forms (3.24) and (3.25), we need the sum of squares of each column and the sum of products of each pair of columns. We illustrate the computation of s_{11} ,

$$\sum_{i=1}^{10} x_{1i}x_{1i} = (25)(2.80) + (25)(2.78) + \cdots + (30)(3.20) = 885.48.$$

From Example 3.5 we have $\bar{y}_1 = 28.1$ and $\bar{y}_2 = 3.895$. By (3.25), we obtain

$$r_{11} = \frac{1}{[10-1]}[885.48 - 10(28.1)(3.895)] = \frac{17.471}{9} = 1.9412.$$

Continuing in this fashion, we obtain

$$\mathbf{S} = \begin{pmatrix} 1.9412 & -0.658 & 1.56 \\ -0.658 & 0.225 & 1.68 \\ 1.56 & 1.68 & .25 \end{pmatrix}. \quad \square$$

3.7 CORRELATION MATRICES

The sample correlation between the j th and k th variable is defined in (3.13) as

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{s_{jk}}{r_{jj}r_{kk}}. \quad (3.30)$$

The sample correlation matrix is analogous to the covariance matrix with correlations in place of covariances:

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}. \quad (3.31)$$

The second row, for example, contains the correlations of y_2 with each of the y 's including the correlation of y_2 with itself, which is 1). Of course, the matrix $\mathbf{R}$ is symmetrical, since $r_{jk} = r_{kj}$.

The correlation matrix can be obtained from the covariance matrix, and vice versa. Define

$$\begin{aligned} \mathbf{D}_s &= \text{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \dots, \sqrt{s_{pp}}) \\ &= \text{diag}(s_{11}, s_{22}, \dots, s_{pp}) \end{aligned}$$

$$= \begin{pmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{pp} \end{pmatrix}. \quad (3.36)$$

Then by (3.35)

$$\mathbf{R} = \mathbf{D}_S^{-1/2} \mathbf{S} \mathbf{D}_S^{-1/2}, \quad (3.37)$$

$$\mathbf{S} = \mathbf{D}_S \mathbf{R} \mathbf{D}_S. \quad (3.38)$$

The population correlation matrix analogous to (3.35) is defined as

$$\mathbf{P}_p = (p_{jkl}) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}, \quad (3.39)$$

where

$$\rho_{jkl} = \frac{\sigma_{jkl}}{\sigma_j \sigma_k},$$

as in (3.12).

Example 3.5. In Example 3.4, we obtained the sample covariance matrix $\mathbf{S}$ for the calcium data in Table 3.1. To obtain the sample correlation matrix for the same data, we can calculate the individual elements using (3.34) or use the direct matrix operation in (3.37). The diagonal matrix $\mathbf{D}_S$ can be found by taking the square roots of the diagonal elements of $\mathbf{S}$,

$$\mathbf{D}_S = \begin{pmatrix} 11.8559 & 0 & 0 \\ 0 & 1.4959 & 0 \\ 0 & 0 & .5864 \end{pmatrix}.$$

Note that we have used the unrounded version of $\mathbf{S}$ for computation. Therefore (3.37),

$$\mathbf{R} = \mathbf{D}_S^{-1/2} \mathbf{S} \mathbf{D}_S^{-1/2} = \begin{pmatrix} 1.000 & .493 & .327 \\ .493 & 1.000 & .882 \\ .327 & .882 & 1.000 \end{pmatrix}.$$

Note that .882 > .493 > .327, which is a different order than that of the variances in $\mathbf{S}$ in Example 3.4. Thus we cannot compare variances, even within the same matrix $\mathbf{S}$. $\square$

3.8 MEAN VECTORS AND COVARIANCE MATRICES FOR SUBJECTS OF VARIABLES

3.8.1 Two Subjects

Sometimes a researcher is interested in two different kinds of variables, both measured on the same sampling unit. This corresponds to type 2 data in Section 1.4. For example, several classroom behaviors are observed for students, and during the same time-period (the basic experimental unit) several teacher behaviors are also observed. The researcher wishes to study the relationships between the pupil variables and the teacher variables.

We will denote the two submatrices by $\mathbf{p}$ and $\mathbf{q}$, with p variables in $\mathbf{p}$ and q variables in $\mathbf{q}$. Thus each observation vector in a sample is partitioned as

$$\begin{pmatrix} \mathbf{p}_i \\ \mathbf{q}_i \end{pmatrix} = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{pi} \\ x_{(p+1)i} \\ \vdots \\ x_{(p+q)i} \end{pmatrix}, \quad i = 1, 2, \dots, n. \quad (3.40)$$

Hence there are $p + q$ variables in each of n observation vectors. In Chapter 10 we will discuss regression of the y 's on the x 's, and in Chapter 11 we will define a measure of correlation between the y 's and the x 's.

For the sample of n observation vectors, the mean vector and covariance matrix have the form

$$\begin{pmatrix} \bar{\mathbf{p}} \\ \bar{\mathbf{q}} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \\ \bar{x}_{(p+1)} \\ \vdots \\ \bar{x}_{(p+q)} \end{pmatrix}, \quad (3.41)$$

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{pp} & \mathbf{S}_{pq} \\ \mathbf{S}_{qp} & \mathbf{S}_{qq} \end{pmatrix}, \quad (3.42)$$

where $\mathbf{S}_{pp}$ is $p \times p$, $\mathbf{S}_{qq}$ is $q \times q$, $\mathbf{S}_{qp}$ is $q \times p$, and $\mathbf{S}_{pq}$ is $p \times q$. Note that because of the symmetry of $\mathbf{S}$,

$$\mathbf{S}_{qp} = \mathbf{S}_{pq}'. \quad (3.43)$$

Thus (3.42) could be written

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_{xy} \\ \mathbf{z}_{xz} \end{pmatrix}. \quad (3.44)$$

To illustrate (3.41) and (3.42), let $p = 2$ and $q = 3$. Then

$$\begin{pmatrix} \mathbf{z} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \frac{z_1}{x_1} \\ \frac{z_2}{x_1} \end{pmatrix},$$

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_{xy} \\ \mathbf{z}_{xz} \end{pmatrix} = \left(\begin{array}{cc|ccc} \sigma_{y_1}^2 & \sigma_{y_1 y_2} & \sigma_{y_1 x_1} & \sigma_{y_1 x_2} & \sigma_{y_1 x_3} \\ \sigma_{y_2 y_1} & \sigma_{y_2}^2 & \sigma_{y_2 x_1} & \sigma_{y_2 x_2} & \sigma_{y_2 x_3} \\ \hline \sigma_{x_1 y_1} & \sigma_{x_1 y_2} & \sigma_{x_1}^2 & \sigma_{x_1 x_2} & \sigma_{x_1 x_3} \\ \sigma_{x_2 y_1} & \sigma_{x_2 y_2} & \sigma_{x_2 x_1} & \sigma_{x_2}^2 & \sigma_{x_2 x_3} \\ \sigma_{x_3 y_1} & \sigma_{x_3 y_2} & \sigma_{x_3 x_1} & \sigma_{x_3 x_2} & \sigma_{x_3}^2 \end{array} \right).$$

The pattern in each of $\mathbf{z}_{xy}$, $\mathbf{z}_{yz}$, $\mathbf{z}_{yx}$, and $\mathbf{z}_{xz}$ is clearly seen in this illustration. For example, the first row of $\mathbf{z}_{xy}$ has the covariance of y_1 with each of x_1, x_2, x_3 ; the second row exhibits covariances of y_2 with the three x 's. On the other hand, $\mathbf{z}_{yx}$ has in its first row the covariances of x_1 with y_1 and y_2 , and so on. Thus $\mathbf{z}_{xy} = \mathbf{z}_{yx}$, as noted in (3.43).

The analogous population results for a partitioned random vector are

$$E\left(\begin{pmatrix} \mathbf{z} \\ \mathbf{x} \end{pmatrix}\right) = \begin{pmatrix} E(\mathbf{z}) \\ E(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \quad (3.45)$$

$$\text{cov}\left(\begin{pmatrix} \mathbf{z} \\ \mathbf{x} \end{pmatrix}\right) = \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}, \quad (3.46)$$

where $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}$. The submatrix $\boldsymbol{\Sigma}_{yy}$ is a $p \times p$ covariance matrix containing the variances of $y_1, y_2, \dots, y_p$ on the diagonal and the covariance of each y_j with each y_k off the diagonal. Similarly, $\boldsymbol{\Sigma}_{xx}$ is the $q \times q$ covariance matrix of $x_1, x_2, \dots, x_q$. The matrix $\boldsymbol{\Sigma}_{yx}$ is $p \times q$ and contains the covariance of each y_j with each x_k . The covariance matrix $\boldsymbol{\Sigma}_{xy}$ is also denoted by $\text{cov}(\mathbf{y}, \mathbf{x})$, that is,

$$\text{cov}(\mathbf{y}, \mathbf{x}) = \boldsymbol{\Sigma}_{xy}. \quad (3.47)$$

Note the difference in meaning between $\text{cov}(\mathbf{z}) = \boldsymbol{\Sigma}$ in (3.46) and $\text{cov}(\mathbf{y}, \mathbf{x}) = \boldsymbol{\Sigma}_{xy}$ in (3.47). $\text{cov}(\mathbf{z})$ involves a single vector containing $p + q$ variables, and $\text{cov}(\mathbf{y}, \mathbf{x})$ involves two vectors.

If $\mathbf{y}$ and $\mathbf{x}$ are independent, then $\boldsymbol{\Sigma}_{xy} = \mathbf{0}$. This means that each y_j is uncorrelated with each x_k so that $\sigma_{x_k y_j} = 0$ for $j = 1, 2, \dots, p$; $k = 1, 2, \dots, q$.

Example 3.8.1. Ratner and Miller (1979; see also Andrews and Hothorn, 1983, pp. 215–216) measured three variables in a comparison of normal patients and diabetics. In Table 3.4 we give partial data for normal patients only. The three variables of major interest were

- x_1 = glucose intolerance,
- x_2 = insulin response to oral glucose,
- x_3 = insulin resistance.

The two additional variables of minor interest were

- x_4 = relative weight,
- x_5 = fasting plasma glucose.

The mean vector, partitioned as in (3.41), is

$$\begin{pmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{z}} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \\ \bar{x}_5 \end{pmatrix} = \begin{pmatrix} .518 \\ 86.44 \\ 540.85 \\ 171.37 \\ 85.78 \end{pmatrix}.$$

The covariance matrix, partitioned as in the illustration following (3.44), is

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xz} \\ \mathbf{S}_{zx} & \mathbf{S}_{zz} \end{pmatrix} = \begin{pmatrix} .0067 & .2667 & .7872 & -.2188 & 2.188 \\ .2667 & 79.95 & 26.23 & -25.98 & -20.84 \\ .7872 & 26.23 & 11.06 & 866.7 & 308.4 \\ -.2188 & -20.86 & 866.7 & 2382 & 1169 \\ 2.188 & -20.84 & 308.4 & 1167 & 2156 \end{pmatrix}.$$

Notice that $\mathbf{S}_{xx}$ and $\mathbf{S}_{zz}$ are symmetric and that $\mathbf{S}_{zx}$ is the transpose of $\mathbf{S}_{xz}$. □

3.8.2 Three or More Subjects

In some cases, three or more subjects of variables are of interest. If the observation vector $\mathbf{y}$ is partitioned as

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

Table 3.4. Relative Weight, Blood Glucose, and Insulin Levels

Point Number	F_1	F_2	F_3	G_1	G_2
1	88	88	186	128	88
2	85	87	200	117	76
3	84	103	119	148	105
4	1.04	88	135	199	188
5	1.00	88	115	180	140
6	78	88	341	157	185
7	81	100	150	221	113
8	1.11	89	351	180	189
9	88	87	179	142	88
10	78	87	266	111	88
11	88	81	114	221	50
12	75	87	305	178	88
13	86	78	290	135	142
14	84	88	171	200	95
15	74	88	111	208	88
16	88	88	314	202	182
17	1.11	88	364	182	76
18	85	88	189	185	87
19	82	87	286	116	88
20	80	88	148	125	58
21	85	88	178	135	47
22	74	88	354	154	58
23	85	95	147	164	81
24	87	88	117	190	124
25	75	82	386	209	74
26	1.11	74	368	228	205
27	1.28	88	308	145	158
28	1.15	100	112	172	148
29	1.00	88	115	175	145
30	78	88	111	222	88
31	1.00	78	390	154	88
32	1.00	88	136	148	189
33	71	75	112	169	32
34	78	88	111	268	188
35	88	85	171	174	78
36	88	88	175	154	88
37	1.17	100	305	182	54
38	85	78	111	241	175
39	87	100	195	128	88
40	1.00	88	177	202	188
41	1.00	102	178	168	117
42	88	88	380	282	188
43	88	84	211	54	71
44	78	88	208	121	29
45	74	95	188	75	42
46	81	88	128	195	58

where y_1 has p_1 variables, y_2 has p_2 , ..., y_k has p_k , with $p = p_1 + p_2 + \cdots + p_k$. Then the sample mean vector and covariance matrix are given by

$$\bar{\mathbf{y}} = \begin{pmatrix} \bar{y}_{11} \\ \bar{y}_{12} \\ \vdots \\ \bar{y}_{1p_1} \end{pmatrix}, \quad (3.48)$$

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kp} \end{pmatrix}. \quad (3.49)$$

The $p_1 \times p_k$ submatrix S_{1k} , for example, contains the covariances of the variables in y_1 with the variables in y_k .

The corresponding population results are

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \quad (3.50)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1p} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{k1} & \Sigma_{k2} & \cdots & \Sigma_{kp} \end{pmatrix}. \quad (3.51)$$

3.8 LINEAR COMBINATIONS OF VARIABLES

3.8.1 Sample Properties

We are frequently interested in linear combinations of the variables $y_1, y_2, \dots, y_p$. For example, two of the types of linear functions we use in later chapters are (1) linear combinations that maximize some function and (2) linear combinations that compare variables, for example, $y_1 - y_2$. In this section, we investigate the means, variances, and covariances of linear combinations.

Let $a_1, a_2, \dots, a_p$ be constants and consider the linear combination of the elements of the vector $\mathbf{y}$,

$$z = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p = \mathbf{a}'\mathbf{y}, \quad (3.52)$$

where $\mathbf{a}' = (a_1, a_2, \dots, a_p)$. If the same coefficient vector $\mathbf{a}$ is applied to each $\mathbf{y}_i$ in a sample, we have

$$\begin{aligned} z_i &= a_1 y_{i1} + a_2 y_{i2} + \cdots + a_p y_{ip} \\ &= \mathbf{a}' \mathbf{y}_i, \quad i = 1, 2, \dots, n. \end{aligned} \quad (3.53)$$

The sample mean of z can be found either by averaging the n values, $\bar{z}_1 = \mathbf{a}' \bar{\mathbf{y}}_1$, $\bar{z}_2 = \mathbf{a}' \bar{\mathbf{y}}_2, \dots$, $\bar{z}_n = \mathbf{a}' \bar{\mathbf{y}}_n$ or as a linear combination of $\bar{\mathbf{y}}$, the sample mean vector of $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$:

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \mathbf{a}' \bar{\mathbf{y}}. \quad (3.54)$$

The result in (3.54) is analogous to the univariate result (3.2), $\bar{z} = a\bar{y}$, where $z_i = ay_i, i = 1, 2, \dots, n$.

Similarly, the sample variance of $z_i = \mathbf{a}' \mathbf{y}_i, i = 1, 2, \dots, n$, can be found as the sample variance of $z_1, z_2, \dots, z_n$ or directly from $\mathbf{a}$ and $\mathbf{S}$, where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$:

$$s_z^2 = \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{n-1} = \mathbf{a}' \mathbf{S} \mathbf{a}. \quad (3.55)$$

Note that $s_z^2 = \mathbf{a}' \mathbf{S} \mathbf{a}$ is the multivariate analogue of the univariate result in (3.6), $s_z^2 = a^2 s_y^2$, where $z_i = ay_i, i = 1, 2, \dots, n$, and s_y^2 is the variance of $y_1, y_2, \dots, y_n$.

Since a variance is always nonnegative, we have $s_z^2 \geq 0$, and therefore $\mathbf{a}' \mathbf{S} \mathbf{a} \geq 0$ for every $\mathbf{a}$. Hence $\mathbf{S}$ is at least positive semidefinite (definition 2.7). If the variables are continuous and are not linearly related, and if $n-1 = p$ (so that $\mathbf{S}$ is full rank), then $\mathbf{S}$ is positive definite (with probability 1).

If we define another linear combination $w = \mathbf{b}' \mathbf{y} = b_1 y_1 + b_2 y_2 + \cdots + b_p y_p$, where $\mathbf{b}' = (b_1, b_2, \dots, b_p)$ is a vector of constants different from $\mathbf{a}'$, then the sample covariance of z and w is given by

$$s_{zw} = \frac{\sum_{i=1}^n (z_i - \bar{z})(w_i - \bar{w})}{n-1} = \mathbf{a}' \mathbf{S} \mathbf{b}. \quad (3.56)$$

The sample correlation between z and w is readily obtained as

$$r_{zw} = \frac{s_{zw}}{\sqrt{s_z^2 s_w^2}} = \frac{\mathbf{a}' \mathbf{S} \mathbf{b}}{\sqrt{(\mathbf{a}' \mathbf{S} \mathbf{a})(\mathbf{b}' \mathbf{S} \mathbf{b})}}. \quad (3.57)$$

We now denote the two constant vectors $\mathbf{a}$ and $\mathbf{b}$ as $\mathbf{a}_1$ and $\mathbf{a}_2$ to facilitate later expansion to more than two such vectors. Let

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix}$$

and define

$$\mathbf{x} = \begin{pmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

Then we can factor $\mathbf{y}$ from this expression by (2.46):

$$\mathbf{x} = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \end{pmatrix} \mathbf{y} = \mathbf{A}\mathbf{y}.$$

If we evaluate the bivariate $\mathbf{a}_i$ for each y -value $\mathbf{y}_i$ in the sample, we obtain $\mathbf{a}_i = \mathbf{A}\mathbf{y}_i, i = 1, 2, \dots, n$, and the average of $\mathbf{x}$ over the sample can be found from $\bar{\mathbf{y}}$:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^T \bar{\mathbf{y}} \\ \mathbf{a}_2^T \bar{\mathbf{y}} \end{pmatrix} \quad \text{by (2.54)} \quad (2.56)$$

$$= \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \end{pmatrix} \bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{y}} \quad \text{by (2.46)}. \quad (2.57)$$

We can use (2.56) and (2.57) to construct the sample covariance matrix for $\mathbf{x}$:

$$\begin{aligned} \mathbf{S}_x &= \begin{pmatrix} s_{z_1 z_1}^2 & s_{z_1 z_2}^2 \\ s_{z_2 z_1}^2 & s_{z_2 z_2}^2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a}_1^T \mathbf{S}_{\mathbf{y}} \mathbf{a}_1 & \mathbf{a}_1^T \mathbf{S}_{\mathbf{y}} \mathbf{a}_2 \\ \mathbf{a}_2^T \mathbf{S}_{\mathbf{y}} \mathbf{a}_1 & \mathbf{a}_2^T \mathbf{S}_{\mathbf{y}} \mathbf{a}_2 \end{pmatrix}. \end{aligned} \quad (2.58)$$

By (2.56), this factors into

$$\mathbf{S}_x = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \end{pmatrix} \mathbf{S}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{A}\mathbf{S}\mathbf{A}^T. \quad (2.59)$$

The bivariate results in (2.56) and (2.57) can be readily extended to more than two linear combinations. (See principal components in Chapter 12, for instance, where we attempt to transform the y 's to a few dimensions that capture most of the information in the y 's.) If we have k linear transformations, they can be expressed as

$$\begin{aligned} t_1 &= a_{11}y_1 + a_{12}y_2 + \cdots + a_{1p}y_p = \mathbf{a}_1^T \mathbf{y} \\ t_2 &= a_{21}y_1 + a_{22}y_2 + \cdots + a_{2p}y_p = \mathbf{a}_2^T \mathbf{y} \\ &\vdots \\ t_k &= a_{k1}y_1 + a_{k2}y_2 + \cdots + a_{kp}y_p = \mathbf{a}_k^T \mathbf{y} \end{aligned}$$

or in matrix notation,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} = \begin{pmatrix} a_1' \mathbf{y} \\ a_2' \mathbf{y} \\ \vdots \\ a_k' \mathbf{y} \end{pmatrix} = \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix} \mathbf{y} = \mathbf{A} \mathbf{y} \quad [\text{by (2.47)}],$$

where $k \leq k' = 1$, $\mathbf{A}$ is $k \times p$, and $\mathbf{y}$ is $p \times 1$ (we typically have $k \leq p$). If $\mathbf{x}_i = \mathbf{A} \mathbf{y}_i$ is calculated for all $\mathbf{y}_i$, $i = 1, 2, \dots, n$, then by (3.50) and (2.49), the sample mean vector of the $\mathbf{x}$'s is

$$\bar{\mathbf{x}} = \begin{pmatrix} a_1' \bar{\mathbf{y}} \\ a_2' \bar{\mathbf{y}} \\ \vdots \\ a_k' \bar{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix} \bar{\mathbf{y}} = \mathbf{A} \bar{\mathbf{y}}. \quad (3.52)$$

By an extension of (3.50), the sample covariance matrix of the $\mathbf{x}$'s becomes

$$\begin{aligned} \mathbf{S}_x &= \begin{pmatrix} a_1' S_{y_1} & a_1' S_{y_2} & \cdots & a_1' S_{y_p} \\ a_2' S_{y_1} & a_2' S_{y_2} & \cdots & a_2' S_{y_p} \\ \vdots & \vdots & \ddots & \vdots \\ a_k' S_{y_1} & a_k' S_{y_2} & \cdots & a_k' S_{y_p} \end{pmatrix} \\ &= \begin{pmatrix} a_1' (S_{y_1}, S_{y_2}, \dots, S_{y_p}) \\ a_2' (S_{y_1}, S_{y_2}, \dots, S_{y_p}) \\ \vdots \\ a_k' (S_{y_1}, S_{y_2}, \dots, S_{y_p}) \end{pmatrix} \\ &= \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix} (S_{y_1}, S_{y_2}, \dots, S_{y_p}) \quad [\text{by (2.47)}] \\ &= \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{pmatrix} \mathbf{S}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_p) \quad [\text{by (2.48)}] \\ &= \mathbf{A} \mathbf{S} \mathbf{A}'. \end{aligned} \quad (3.54)$$

Note that by (3.52) and (3.54), we have

$$V(\mathbf{A} \mathbf{S} \mathbf{A}') = \sum_{i=1}^k a_i^2 S_{y_i}. \quad (3.55)$$

A slightly more general linear transformation is

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}, \quad i = 1, 2, \dots, n. \quad (3.66)$$

The sample mean vector and covariance matrix of $\mathbf{y}$ are given by

$$\bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{x}} + \mathbf{b}, \quad (3.67)$$

$$\mathbf{S}_y = \mathbf{A}\mathbf{S}_x\mathbf{A}'. \quad (3.68)$$

Example 3.8.1. Tans (1975, p. 233, 1983, p. 47) reported the results of an experiment where subjects responded to “probe words” at five positions in a sentence. The variables are response times for the j th probe word, y_j , $j = 1, 2, \dots, 5$. The data are given in Table 3.5.

Table 3.5. Response Times for Five Probe Word Positions

Subject Number	y_1	y_2	y_3	y_4	y_5
1	30	36	38	33	42
2	27	29	26	17	23
3	27	22	43	27	36
4	42	36	32	34	27
5	27	18	33	34	29
6	43	32	43	27	40
7	40	22	36	27	36
8	36	23	33	26	36
9	36	23	27	24	36
10	26	33	33	32	36
11	29	26	27	26	23

These variables are continuous (same measurement units and similar means and variances), and the researcher may wish to examine some simple linear combinations. Consider the following linear combination for illustrative purposes:

$$\begin{aligned} z &= 3y_1 - 2y_2 + 4y_3 - y_4 + y_5 \\ &= (3, -2, 4, -1, 1)\mathbf{y} = \mathbf{a}'\mathbf{y}. \end{aligned}$$

If z is calculated for each of the 11 observations, we obtain $z_1 = 288$, $z_2 = 133$, ..., $z_{11} = 146$ with mean $\bar{z} = 197.0$ and variance $s_z^2 = 2894.5$. These same results can be obtained using (3.56) and (3.55). The sample mean vector and covariance matrix for the data are

$$\bar{\mathbf{y}} = \begin{pmatrix} 36.09 \\ 28.93 \\ 34.09 \\ 27.17 \\ 30.73 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 65.69 & 33.65 & 47.59 & 36.77 & 25.03 \\ 33.65 & 49.07 & 38.99 & 46.34 & 28.70 \\ 47.59 & 38.99 & 60.69 & 57.37 & 41.13 \\ 36.77 & 46.34 & 57.37 & 62.92 & 31.68 \\ 25.03 & 28.70 & 41.13 & 31.68 & 33.22 \end{pmatrix}.$$

Then, by (3.54),

$$\mathbf{t} = \mathbf{a}\bar{\mathbf{y}} = (3, -2, 4, -1, 1) \begin{pmatrix} 36.09 \\ 28.93 \\ 34.09 \\ 27.17 \\ 30.73 \end{pmatrix} = 197.6,$$

and by (3.55), $s_t^2 = \mathbf{a}'\mathbf{S}\mathbf{a} = 2084.0$.

We now define a second linear combination:

$$\begin{aligned} w &= y_1 + 3y_2 - y_3 + y_4 - 2y_5 \\ &= (1, 3, -1, 1, -2)\mathbf{y} = \mathbf{b}'\mathbf{y}. \end{aligned}$$

The sample mean and variance of w are $\bar{w} = \mathbf{b}'\bar{\mathbf{y}} = 44.48$ and $s_w^2 = \mathbf{b}'\mathbf{S}\mathbf{b} = 608.87$.

The sample covariance of t and w is, by (3.55), $s_{tw} = \mathbf{a}'\mathbf{S}\mathbf{b} = 40.2$.

Using (3.57), we find the sample correlation between t and w to be

$$r_{tw} = \frac{s_{tw}}{\sqrt{s_t^2 s_w^2}} = \frac{40.2}{\sqrt{(2084)(608.87)}} = .0026.$$

We now define the three linear functions

$$\begin{aligned} z_1 &= y_1 + y_2 + y_3 + y_4 + y_5 \\ z_2 &= 2y_1 - 3y_2 + y_3 - 2y_4 - y_5 \\ z_3 &= -y_1 - 2y_2 + y_3 - 2y_4 + 3y_5, \end{aligned}$$

which can be written in matrix form as

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & -3 & 1 & -2 & -1 \\ -1 & -2 & 1 & -2 & 3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix},$$

The sample mean vector for $\mathbf{x}$ is given by (3.62) as

$$\bar{\mathbf{x}} = A\bar{\mathbf{y}} = \begin{pmatrix} 133.73 \\ -35.64 \\ -15.43 \end{pmatrix},$$

and the sample covariance matrix of $\mathbf{x}$ is given by (3.64) as

$$S_{\mathbf{x}} = A S_{\mathbf{y}} A' = \begin{pmatrix} 995.43 & -502.09 & -211.04 \\ -502.09 & 811.45 & 268.08 \\ -211.04 & 268.08 & 702.87 \end{pmatrix}.$$

The covariance matrix $S_{\mathbf{x}}$ can be converted to a correlation matrix by use of (3.37):

$$R_{\mathbf{x}} = D_{\mathbf{x}}^{-1/2} S_{\mathbf{x}} D_{\mathbf{x}}^{-1/2} = \begin{pmatrix} 1.00 & -.50 & -.29 \\ -.50 & 1.00 & .33 \\ -.29 & .33 & 1.00 \end{pmatrix},$$

where

$$D_{\mathbf{x}} = \begin{pmatrix} 11.59 & 0 & 0 \\ 0 & 26.45 & 0 \\ 0 & 0 & 26.50 \end{pmatrix}$$

is obtained from the square roots of the diagonal elements of $S_{\mathbf{x}}$. □

3.5.2 Population Properties

The sample results in Section 3.5.1 for linear combinations have population counterparts. Let $\mathbf{z} = \mathbf{a}'\mathbf{y}$, where $\mathbf{a}$ is a vector of constants. Then the population mean of $\mathbf{z}$ is

$$E(\mathbf{z}) = E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'E(\mathbf{y}) = \mathbf{a}'\boldsymbol{\mu}, \quad (3.68)$$

and the population variance is

$$\sigma_z^2 = \text{var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\Sigma\mathbf{a}. \quad (3.69)$$

Let $\mathbf{w} = \mathbf{b}'\mathbf{y}$, where $\mathbf{b}$ is a vector of constants different from $\mathbf{a}$. The population covariance of $\mathbf{z} = \mathbf{a}'\mathbf{y}$ and $\mathbf{w} = \mathbf{b}'\mathbf{y}$ is

$$\text{cov}(\mathbf{z}, \mathbf{w}) = \sigma_{\mathbf{zw}} = \mathbf{a}'\Sigma\mathbf{b}. \quad (3.70)$$

By (3.12) the population correlation of $\mathbf{z}$ and $\mathbf{w}$ is

$$\begin{aligned} A_{yx} &= \text{cov}(\mathbf{a}'\mathbf{y}, \mathbf{b}'\mathbf{y}) = \frac{R_{yx}}{n_1 n_2} \\ &= \frac{\mathbf{a}'\mathbf{B}\mathbf{b}}{n_1'(\mathbf{a}'\mathbf{A}\mathbf{a})(\mathbf{b}'\mathbf{B}\mathbf{b})} \end{aligned} \quad (3.72)$$

If $A\mathbf{y}$ represents several linear combinations, the population mean vector and covariance matrix are given by

$$E(A\mathbf{y}) = A\bar{E}(\mathbf{y}) = A\boldsymbol{\mu}, \quad (3.73)$$

$$\text{cov}(A\mathbf{y}) = A\mathbf{B}A'. \quad (3.74)$$

The more general linear transformation $\mathbf{z} = A\mathbf{y} + \mathbf{b}$ has population mean vector and covariance matrix

$$E(A\mathbf{y} + \mathbf{b}) = A\bar{E}(\mathbf{y}) + \mathbf{b} = A\boldsymbol{\mu} + \mathbf{b}, \quad (3.75)$$

$$\text{cov}(A\mathbf{y} + \mathbf{b}) = A\mathbf{B}A'. \quad (3.76)$$

3.10 MEASURES OF OVERALL VARIABILITY

The covariance matrix contains the variances of the p variables and the covariances between all pairs of variables and is thus a multifaceted picture of the overall variation in the data. Sometimes it is desirable to have a single numerical value for the overall multivariate scatter. One such measure is the generalised sample variance, defined as the determinant of the covariance matrix:

$$\text{Generalised sample variance} = |\mathbf{B}|. \quad (3.77)$$

The generalised sample variance has a geometric interpretation. The extension of an ellipse to more than two dimensions is called a hyperellipsoid. A p -dimensional hyperellipsoid $(\mathbf{y} - \bar{\mathbf{y}})' \mathbf{B}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) = \alpha^2$, centred at $\bar{\mathbf{y}}$ and based on $\mathbf{B}^{-1}$ to standardise the distance to the centre, contains a proportion of the observations $y_1, y_2, \dots, y_n$ in the sample (if $\mathbf{B}$ were replaced by $\mathbf{\hat{B}}$, the value of α^2 could be determined by tables of the chi-square distribution; see property 3 in Section 4.7). The ellipsoid $(\mathbf{y} - \bar{\mathbf{y}})' \mathbf{B}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) = \alpha^2$ has area proportional to the square roots of the eigenvalues of $\mathbf{B}$. It can be shown that the volume of the ellipsoid is proportional to $|\mathbf{B}|^{1/2}$. If the smallest eigenvalue λ_p is zero, there is no axis in that direction, and the ellipsoid lies wholly in a $(p-1)$ -dimensional subspace of p -space. Consequently, the volume in p -space is zero. This can also be seen by (2.10), $|\mathbf{B}| = \lambda_1 \lambda_2 \cdots \lambda_p$. Hence, if $\lambda_p = 0$, $|\mathbf{B}| = 0$. A zero eigenvalue indicates a redundancy in the form of a linear relationship among the variables. (As will be seen in Section 12.7, the eigenvector corresponding to the zero eigenvalue reveals the form of the linear dependency.) One solution to the dilemma when $\lambda_p = 0$ is to remove one or more variables.

Another measure of overall variability, the total sample variance, is simply the trace of $\mathbf{S}$:

$$\text{Total sample variance} = s_{11} + s_{22} + \cdots + s_{pp} = \text{tr}(\mathbf{S}). \quad (3.78)$$

This measure of overall variability ignores correlation structure altogether but is found useful for comparison purposes in techniques such as principal components (Chapter 12).

In general, the both $\|\mathbf{S}\|$ and $\text{tr}(\mathbf{S})$, relatively large values reflect a broad scatter of $y_1, y_2, \dots, y_p$ about $\bar{y}$, whereas lower values indicate closer concentration about $\bar{y}$. In the case of $\|\mathbf{S}\|$, however, as noted previously, an extremely small value of $\|\mathbf{S}\|$ or $\|\mathbf{R}\|$ may indicate either small scatter or multicollinearity, a term indicating near linear relationships in a set of variables. Multicollinearity may be due to high pairwise correlations or to a high multiple correlation between one variable and several of the other variables. For other measures of intercorrelation, see Rencher (1995, Section 1.7).

3.11 ESTIMATION OF MISSING VALUES

It is not uncommon to find missing measurements in an observation vector, that is, missing values for one or more variables. A small number of rows with missing entries in the data matrix $\mathbf{Y}$ [see (3.17)] does not constitute a serious problem; we can simply discard each row that has a missing value. However, with this procedure, a small portion of missing data, if widely distributed, would lead to a substantial loss of data. For example, in a large data set with $n = 350$ and $p = 85$, only about 1.5% of the $350 \times 85 = 40,750$ measurements were missing. However, nearly half of the observation vectors (rows of $\mathbf{T}$) turned out to be incomplete.

The distribution of missing values in a data set is an important consideration. Randomly missing variable values scattered throughout a data matrix are less serious than a pattern of missing values that depends, in some extent, on the values of the missing variables.

We discuss two methods of estimating the missing values, or “filling the holes,” in the data matrix, also called *imputation*. Both procedures presume that the missing values occur at random. If the occurrence or nonoccurrence of missing values is related to the values of some of the variables, then the techniques may not estimate the missing responses very well.

The first method is very simple: substitute a mean for each missing value, specifically the average of the available data in the columns of the data matrix in which the unknown value lies. Replacing an observation by its mean reduces the variance and the absolute value of the covariance. Therefore, the sample covariance matrix $\mathbf{S}$ computed from the data matrix $\mathbf{Y}$ in (3.17) with means imputed for missing values is biased. However, it is positive definite.

The second technique is a regression approach. The data matrix $\mathbf{Y}$ is partitioned into two parts, one containing all rows with missing entries and the other comprising

all the complete rows. Suppose y_{ij} is the only missing entry in the i th row of $\mathbf{Y}$. Then using the data in the submatrix with complete rows, y_{ij} is regressed on the other variables to obtain a prediction equation: $y_{ij} = b_0 + b_1x_{i1} + \cdots + b_{j-1}x_{i,j-1} + b_{j+1}x_{i,j+1} + \cdots + b_px_{ip}$. Then the remaining entries in the i th row are entered as independent variables in the regression equation to obtain the predicted value, $\hat{y}_{ij}$. The regression method was first proposed by Berk (1953) and is a special case of the EM algorithm (Dempster, Laird, and Rubin 1977).

The regression method can be improved by iteration, carried out, for example, in the following way. Impute all missing entries in the data matrix using regression. After filling in the missing entries, use the full-data matrix to obtain new prediction equations. Use these prediction equations to calculate new predicted values $\hat{y}_{ij}$ for missing entries. Use the new data matrix to obtain revised prediction equations and new predicted values $\hat{y}_{ij}$. Continue this process until the predicted values stabilize.

A modification may be needed if the missing entries are so pervasive that it is difficult to find data to estimate the initial regression equations. In this case, the process could be started by using means as in the first method and then beginning the iteration.

The regression approach will ordinarily yield better results than the method of inserting means. However, if the other variables are not very highly correlated with the one to be predicted, the regression technique is essentially equivalent to imputing means. The regression method underestimates the variances and covariances, though, to a lesser extent than the method based on means.

Example 3.1.1. We illustrate the iterated regression method of estimating missing values. Consider the oil-rim data of Table 3.3 as reproduced here and suppose the entries in parentheses are missing:

Location Number	x_1	x_2	x_3
1	33	(1.5)	(180)
2	33	4.9	(370)
3	40	30.0	4.38
4	30	1.8	3.23
5	8	1.7	273
6	30	1.8	133
7	33	4.6	138
8	33	30.9	(50)
9	33	8.0	3.28
10	30	1.6	3.20

We first regress x_3 on x_1 and x_2 for observations 3–10 and obtain $\hat{x}_3 = b_0 + b_1x_1 + b_2x_2$. When this is evaluated for the two missing entries in the first row ($x_1 = 33$ and $x_2 = 1.5$), we obtain $\hat{x}_3 = 4.097$. Similarly, we regress x_2 on x_1 and x_3 for observations 3–10 to obtain $\hat{x}_2 = c_0 + c_1x_1 + c_2x_3$. Evaluating this for the two missing entries in the second row yields $\hat{x}_2 = 3.811$. We now insert these

estimates for the missing values and calculate the regression equations based on all 18 observations. Using the revised equation, $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 y_3$, we obtain a new predicted value, $\hat{y}_2 = 3.698$. Similarly, we obtain a revised regression equation for y_3 that gives a new predicted value, $\hat{y}_3 = 1.081$. With these values inserted, we calculate new equations and obtain new predicted values, $\hat{y}_2 = 3.672$ and $\hat{y}_3 = 1.076$. At the third iteration we obtain $\hat{y}_2 = 3.679$ and $\hat{y}_3 = 1.075$. There is very little change in subsequent iterations. These values are close to the actual values, $y_2 = 3.5$ and $y_3 = 2.5$, than the initial regression estimates, $\hat{y}_2 = 4.093$ and $\hat{y}_3 = 3.041$. They are also much better estimates than the means of the second and third columns, $\bar{y}_2 = 7.989$ and $\bar{y}_3 = 5.132$. $\square$

3.3.2 DISTANCE BETWEEN VECTORS

In a univariate setting, the distance between two points is simply the difference (or absolute difference) between their values. For statistical purposes, this difference may not be very informative. For example, we do not want to know how many observations apart two means are, but rather how many standard deviations apart they are. Thus we examine the standardised or statistical distance, such as

$$\frac{|x_1 - \mu_1|}{\sigma} \quad \text{or} \quad \frac{|\bar{y} - \mu|}{s_y}.$$

To obtain a useful distance measure in a multivariate setting, we must consider not only the variances of the variables but also their covariances or correlations. The simple (squared) Euclidean distance between two vectors, $(y_1 - y_2)^T(y_1 - y_2)$, is not useful in some situations because there is no adjustment for the variances of the covariates. For a statistical distance, we standardise by inserting the inverse of the covariance matrix:

$$d^2 = (y_1 - y_2)^T S^{-1} (y_1 - y_2). \quad (3.79)$$

Other examples are

$$D^2 = (\bar{y} - \mu)^T \Sigma^{-1} (\bar{y} - \mu) \quad (3.80)$$

$$\hat{d}^2 = (\bar{y} - \mu)^T \hat{\Sigma}^{-1} (\bar{y} - \mu) \quad (3.81)$$

$$\hat{d}^2 = (\mu_1 - \mu_2)^T \hat{\Sigma}^{-1} (\mu_1 - \mu_2). \quad (3.82)$$

These (squared) distances between two vectors were first proposed by Mahalanobis (1936) and are often referred to as Mahalanobis distances. If a random variable has a larger variance than another, it receives relatively less weight in a Mahalanobis distance. Similarly, two highly correlated variables do not contribute as much as two variables that are less correlated. In essence, then, the use of the inverse of the covariance matrix in a Mahalanobis distance has the effect of (1) standardising all variables to the same variance and (2) eliminating correlations. To illustrate this,

we use the square root matrix defined in (2.1.11) to rewrite (3.87) as

$$\begin{aligned}\Delta^2 &= (\mathbf{y} - \boldsymbol{\mu})(\Sigma^{-1/2})(\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})(\Sigma^{1/2}\Sigma^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu}) \\ &= [(\Sigma^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu})](\Sigma^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{z}'\mathbf{z},\end{aligned}$$

where $\mathbf{z} = (\Sigma^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu}) = (\Sigma^{1/2})^{-1}\mathbf{y} - (\Sigma^{1/2})^{-1}\boldsymbol{\mu}$. Next, by (3.36) it can be shown that

$$\text{cov}(\mathbf{z}) = \frac{1}{n}\mathbf{I}. \quad (3.88)$$

Hence the transformed variables $z_1, z_2, \dots, z_p$ are uncorrelated, and each has variance equal to $1/n$. If the appropriate covariance matrix for the random vector were used in a Mahalanobis distance, the variances would reduce to 1. For example, if $\text{cov}(\mathbf{y}) = \Sigma/n$ were used above in place of Σ , we would obtain $\text{cov}(\mathbf{z}) = \mathbf{I}$.

PROBLEMS

- 3.1 If $x_i = a_i y_i$ for $i = 1, 2, \dots, n$, show that $\bar{x} = a\bar{y}$ as in (3.34).
- 3.2 If $x_i = a_i y_i$ for $i = 1, 2, \dots, n$, show that $s_x^2 = a^2 s_y^2$ as in (3.35).
- 3.3 For the data in Figure 3.3, show that $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 0$.
- 3.4 Show that $(x - \bar{x})^2/(y - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, thus verifying (3.43).
- 3.5 For $p = 3$ show that

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y}) = \begin{pmatrix} s_{y1} & s_{y2} & s_{y3} \\ s_{y1} & s_{y2} & s_{y3} \\ s_{y1} & s_{y2} & s_{y3} \end{pmatrix},$$

which illustrates (3.27).

- 3.6 Show that $\mathbf{I} = n\mathbf{\bar{y}}$ as in (3.54), where $\mathbf{z}_i = \mathbf{a}'\mathbf{y}_i$, $i = 1, 2, \dots, n$.
- 3.7 Show that $\mathbf{z}_i^2 = n\mathbf{\bar{y}}_i$ as in (3.55), where $\mathbf{z}_i = \mathbf{a}'\mathbf{y}_i$, $i = 1, 2, \dots, n$.
- 3.8 Show that $\text{tr}(\mathbf{A}\mathbf{S}\mathbf{S}) = \sum_{i=1}^n n\mathbf{\bar{y}}_i$ as in (3.65).
- 3.9 Use (3.76) to verify (3.88), $\text{cov}(\mathbf{z}) = 1/n$, where $\mathbf{z} = (\Sigma^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu})$.
- 3.10 Use the calcium data in Table 3.3:
 - (a) Calculate $\mathbf{S}$ using the data matrix $\mathbf{Y}$ as in (3.28).
 - (b) Obtain $\mathbf{S}$ by calculating s_{11} , s_{12} , and s_{22} , as in (3.34) and (3.36).
 - (c) Find $\mathbf{S}$ using (3.37).
- 3.11 Use the calcium data in Table 3.3:
 - (a) Find the generalized sample variance (S) as in (3.77).
 - (b) Find the total sample variance $\text{tr}(\mathbf{S})$, as in (3.78).

3.12 Use the probe word data of Table 3.3:

- (a) Find the generalized sample variance ($\mathbf{S}$) as in (3.77).
- (b) Find the total sample variance ($n\mathbf{S}$) as in (3.78).

3.13 For the probe word data in Table 3.3, find $\mathbf{B}$ using (3.77).

3.14 For the variables in Table 3.3, define $z = 3x_1 - x_2 + 2x_3 = (3, -1, 2)_p$. Find $\bar{z}$ and s_z^2 in two ways:

- (a) Evaluate $\bar{z}$ for each row of Table 3.3 and find $\bar{z}$ and s_z^2 directly from $(z_1, z_2, \dots, z_n)$ using (3.1) and (3.5).
- (b) Use $\bar{z} = \mathbf{a}'\bar{\mathbf{x}}$ and $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$, as in (3.54) and (3.55).

3.15 For the variables in Table 3.3, define $w = -2x_1 + 3x_2 + x_3$ and define z as in Problem 3.14. Find r_{zw} in two ways:

- (a) Evaluate z and w for each row of Table 3.3 and find r_{zw} from the 10 pairs (z_i, w_i) , $i = 1, 2, \dots, 10$, using (3.16) and (3.13).
- (b) Find r_{zw} using (3.57).

3.16 For the variables in Table 3.3, find the correlation between x_1 and $\frac{1}{2}(x_2 + x_3)$ using (3.57).

Table 3.6. Russian Name Length of Four Signs for 20 Boys

Individual	Age			
	8 yr (x_1)	$10 \frac{1}{2} \text{ yr}$ (x_2)	12 yr (x_3)	$8 \frac{1}{2} \text{ yr}$ (x_4)
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.3
4	45.1	43.3	46.1	47.2
5	47.0	48.2	48.9	49.3
6	50.5	51.2	51.3	51.7
7	51.2	51.0	54.3	54.5
8	46.8	50.0	50.3	51.7
9	48.1	50.8	51.3	54.4
10	46.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.8
12	48.5	49.2	50.0	50.5
13	50.1	51.8	51.5	51.8
14	48.2	49.9	49.3	49.8
15	46.6	50.4	51.2	51.8
16	50.7	51.7	51.7	51.3
17	47.0	47.7	48.4	49.8
18	53.3	54.6	55.1	55.3
19	46.2	47.3	48.1	48.4
20	46.3	47.0	51.3	51.8

3.17 Define the following linear combinations for the variables in Table 3.2:

$$z_1 = y_1 + y_2 + y_3$$

$$z_2 = 2y_1 - 3y_2 + 2y_3$$

$$z_3 = -y_1 - 2y_2 - 3y_3$$

(a) Find Σ and $\mathbf{B}_z$ using (3.62) and (3.64).

(b) Find $\mathbf{B}_z$ from $\mathbf{B}_y$ using (3.17).

3.18 The data in Table 3.6 (Fletcher and Clatworthy 1962) consist of measurements y_1 , y_2 , y_3 , and y_4 of the crania bones at four different ages intervals of 20 days.

(a) Find $\bar{\mathbf{y}}$, $\mathbf{S}$, and $\mathbf{B}$.

(b) Find $\hat{\mathbf{S}}$ and $\text{tr}(\hat{\mathbf{S}})$.

Table 3.6. Measurements on the First and Second Adult Rat in a Sample of 28 Families

First lion		Second lion	
Head Length	Head Breadth	Head Length	Head Breadth
y_1	y_2	y_3	y_4
140	123	179	145
135	149	208	152
181	148	185	149
182	152	188	149
176	144	171	142
208	157	192	152
169	138	166	149
197	159	169	152
188	152	197	159
192	158	187	151
179	158	186	148
183	147	174	147
174	158	185	152
198	159	195	157
188	151	187	158
183	157	181	153
193	155	183	158
186	153	175	148
181	145	182	145
175	148	185	157
192	154	185	152
174	143	176	147
176	139	176	145
197	167	208	158
198	163	187	152

3.19 For the data in Table 3.5, define $z = y_1 + 2y_2 + y_3 - 2y_4$ and $w = -2y_1 + 3y_2 - y_3 + 2y_4$.

(a) Find E, W, s_z^2 , and s_w^2 using (3.94) and (3.95).

(b) Find s_{zw} and r_{zw} using (3.96) and (3.97).

3.20 For the data in Table 3.6 define

$$z_1 = 2y_1 + 3y_2 - y_3 + 4y_4$$

$$z_2 = -2y_1 - y_2 + 4y_3 - 2y_4$$

$$z_3 = 3y_1 - 2y_2 - y_3 + 3y_4$$

Find $\bar{z}$, S_z , and R_z using (3.62), (3.66), and (3.97), respectively.

3.21 The data in Table 3.7 consist of head measurements on first and second sons (Table B-2). Define y_1 and y_2 as the measurements on the first son and x_1 and x_2 for the second son.

(a) Find the mean vector for all four variables and partition it into $\begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix}$ as in (3.41).

(b) Find the covariance matrix for all four variables and partition it into

$$S = \begin{pmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{pmatrix},$$

as in (3.42).

3.22 Table 3.8 contains data from O'Brien and Mallon (1985), see also Anderson and Hothberg, 1985, p. 214) with measurements of blood glucose levels on three occasions for 50 women. The y 's represent fasting glucose measurements on the three occasions; the x 's are glucose measurements 1 hour after sugar intake. Find the mean vector and covariance matrix for all six variables and partition them into $\begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix}$, as in (3.41), and

$$S = \begin{pmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{pmatrix},$$

as in (3.42).

Table 3.8. Blood Glucose Measurements on Three Occasions

Fasting			One Hour after Sugar Intake		
y_1	y_2	y_3	x_1	x_2	x_3
60	69	62	107	98	98
56	63	64	103	78	107
60	69	76	88	99	108
55	68	60	89	87	114

(continued)

Table 3A. (Continued)

Fasting			One hour after sugar intake		
T_1	T_2	T_3	T_1	T_2	T_3
62	75	68	105	150	91
74	66	70	108	108	100
64	71	66	71	100	106
73	70	64	113	113	109
68	67	70	76	83	114
69	82	74	71	110	127
60	67	61	140	149	121
70	74	76	150	150	108
60	74	76	150	151	142
63	70	74	95	96	105
68	66	66	115	65	109
78	60	70	164	98	108
104	77	77	160	117	121
77	66	74	144	71	103
60	77	68	71	82	89
70	70	72	114	99	122
73	65	70	71	79	100
61	74	67	106	115	106
66	75	73	150	147	121
73	82	76	103	150	113
74	71	66	143	106	106
70	70	64	114	119	129
74	90	86	73	106	114
74	77	66	140	61	77
67	71	69	64	67	76
76	75	66	104	150	66
64	66	70	63	94	100
71	80	76	81	87	86
68	76	73	120	89	66
90	100	74	107	109	101
60	76	61	66	111	96
46	77	75	113	124	97
60	60	67	140	112	122
74	70	76	106	66	105
60	74	70	71	60	71
68	76	66	140	108	66
66	80	66	140	117	104
77	67	74	83	90	107
70	67	100	150	140	146
73	76	61	119	120	114
76	66	77	121	155	140
73	66	66	103	60	123
71	80	68	104	60	66
68	60	76	114	66	66
82	70	76	61	66	106

Data Measurements are in mg/dl x 100.

The Multivariate Normal Distribution

4.1 MULTIVARIATE NORMAL DENSITY FUNCTION

Many univariate tests and confidence intervals are based on the univariate normal distribution. Similarly, the majority of multivariate procedures have the multivariate normal distribution as their underpinning.

The following are some of the useful features of the multivariate normal distribution (see Section 4.2): (1) the distribution can be completely described using only means, variances, and covariances; (2) bivariate plots of multivariate data show linear trends; (3) if the variables are uncorrelated, they are independent; (4) linear functions of multivariate normal variables are also normal; (5) as in the univariate case, the convenient form of the density function lends itself to derivation of many properties and test statistics; and (6) even when the data are not multivariate normal, the multivariate normal may serve as a useful approximation, especially in inferences involving sample mean vectors, which are approximately multivariate normal by the central limit theorem (see Section 6.3.2).

Since the multivariate normal density is an extension of the univariate normal density and shares many of its features, we review the univariate normal density function in Section 4.1.1. We then describe the multivariate normal density in Sections 4.1.2–4.1.4.

4.1.1 Univariate Normal Density

If a random variable y , with mean μ and variance σ^2 , is normally distributed, its density is given by

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}, \quad -\infty < y < \infty. \quad (4.1)$$

When y has the density (4.1), we say that y is distributed as $N(\mu, \sigma^2)$, or simply y is $N(\mu, \sigma^2)$. This function is represented by the familiar bell-shaped curve illustrated in Figure 4.1 for $\mu = 10$ and $\sigma = 2.5$.



Figure 4.1. The normal density curve.

4.1.2 Multivariate Normal Density

If $\mathbf{y}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the density is given by

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})}, \quad (4.1)$$

where p is the number of variables. When $\mathbf{y}$ has the density (4.1), we say that $\mathbf{y}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or simply $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The term $(\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu})$ in the exponent of the univariate normal density (4.1) measures the squared distance from $\mathbf{y}$ to $\boldsymbol{\mu}$ in standard deviation units. Similarly, the term $(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ in the exponent of the multivariate normal density (4.1) is the squared generalised distance from $\mathbf{y}$ to $\boldsymbol{\mu}$, or the Mahalanobis distance,

$$\Delta^2 = (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}). \quad (4.2)$$

The characteristics of this distance between $\mathbf{y}$ and $\boldsymbol{\mu}$ were discussed in Section 3.12. Note that Δ , the square root of (4.2), is not a standard deviation unless $\mathbf{y} = \boldsymbol{\mu}$. The distance Δ increases with p , the number of variables (see Problem 4.4).

In the coefficient of the exponential function in (4.1), $|\boldsymbol{\Sigma}|^{1/2}$ appears as the analogue of $\sigma\sqrt{2\pi}$ in (4.1). In the next section, we discuss the effect of $|\boldsymbol{\Sigma}|$ on the density.

4.1.3 Generalized Population Variances

In Section 3.16, we referred to (S) as a generalized sample variance. Analogously, $(\boldsymbol{\Sigma})$ is a generalized population variance. If σ^2 is small in the univariate normal, the y values are concentrated near the mean. Similarly, a small value of $|\boldsymbol{\Sigma}|$ in the



Figure 4.2. Bivariate normal densities.

multivariate case indicates that the y 's are concentrated close to μ in p -space or that there is multicollinearity among the variables. The term, multicollinearity indicates that the variables are highly intercorrelated, in which case the effective dimensionality is less than p . (See Chapter 12 for a method of finding a reduced number of new dimensions that represent the data.) In the presence of multicollinearity, one or more eigenvalues of Σ will be near zero and $|\Sigma|$ will be small, since $|\Sigma|$ is the product of the eigenvalues [see (2.3.6)].

Figure 4.2 shows, for the bivariate case, a comparison of a distribution with small $|\Sigma|$ and a distribution with large $|\Sigma|$. An alternative way to portray the concentration of points in the bivariate normal distribution is with contour plots. Figure 4.3 shows contour plots for the two distributions in Figure 4.2. Each ellipse contains a different proportion of observation vectors y . The contours in Figure 4.3 can be found by setting the density function equal to a constant and solving for y , as illustrated in Figure 4.4. The bivariate normal density surface sliced at a constant height traces an ellipse, which contains a given proportion of the observations (Buncher 1998, Section 2.1.3).

In both Figures 4.2 and 4.3, small $|\Sigma|$ appears on the left and large $|\Sigma|$ appears on the right. In Figure 4.3a, there is a larger correlation between y_1 and y_2 . In Figure 4.3b, the variances are larger (in the natural directions). In general, for any mul-

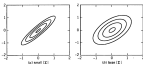


Figure 4.3. Contour plots for the distributions in Figure 4.2.

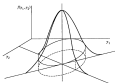


Figure 4.4. Constant density contours for bivariate normal.

for a variable j , a decrease in intercorrelations among the variables or an increase in the variances will lead to a larger $|\mathbf{S}|$.

4.1.4 Diversity of Applications of the Multivariate Normal

Nearly all the inferential procedures we discuss in this book are based on the multivariate normal distribution. We acknowledge that a major motivation for the widespread use of the multivariate normal is its mathematical tractability. From the multivariate normal assumption, a host of useful procedures can be derived, and many of these are available in software packages. Practical alternatives to the multivariate normal are fewer than in the univariate case. Because it is not as simple to order (or rank) multivariate observation vectors as it is for univariate observations, there are not as many nonparametric procedures available for multivariate data.

Although real data may not often be exactly multivariate normal, the multivariate normal will frequently serve as a useful approximation to the true distribution. Tests and graphical procedures are available for assessing normality (see Sections 4.4 and 4.5). Fortunately, many of the procedures based on multivariate normality are robust to departures from normality.

4.2 PROPERTIES OF MULTIVARIATE NORMAL RANDOM VARIABLES

We first review the properties of a random $p \times 1$ vector $\mathbf{y}$ from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

1. Normality of linear combinations of the variables in $\mathbf{y}$:

- (a) If $\mathbf{a}$ is a vector of constants, the linear function $\mathbf{a}'\mathbf{y} = a_1y_1 + a_2y_2 + \cdots + a_ry_r$ is univariate normal:

$$\text{If } \mathbf{y} \text{ is } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ then } \mathbf{a}'\mathbf{y} \text{ is } N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}).$$

The mean and variance of $\mathbf{a}'\mathbf{y}$ were given in (3.89) and (3.90) as $E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\mu}$ and $\text{var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ for any random vector $\mathbf{y}$. We now have the additional attribute that $\mathbf{a}'\mathbf{y}$ has a (univariate) normal distribution if $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- (b) If $\mathbf{A}$ is a constant $q \times p$ matrix of rank q , where $q \leq p$, the q linear combinations in $\mathbf{A}\mathbf{y}$ have a multivariate normal distribution:

$$\text{If } \mathbf{y} \text{ is } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ then } \mathbf{A}\mathbf{y} \text{ is } N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').$$

Here, again, $E(\mathbf{A}\mathbf{y}) = \mathbf{A}\boldsymbol{\mu}$ and $\text{cov}(\mathbf{A}\mathbf{y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, in general, as given in (3.13) and (3.14). But we now have the additional feature that the q variables in $\mathbf{A}\mathbf{y}$ have a multivariate normal distribution.

2. Standardized variables:

A 'standardized' vector $\mathbf{z}$ can be obtained in two ways:

$$\mathbf{z} = (\boldsymbol{\Sigma})^{-1/2}(\mathbf{y} - \boldsymbol{\mu}), \quad (4.4)$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}$ is factored using the Cholesky procedure in Section 2.7, or

$$\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu}), \quad (4.5)$$

where $\boldsymbol{\Sigma}^{1/2}$ is the symmetric square root matrix of $\boldsymbol{\Sigma}$ defined in (2.112) such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. In either (4.4) or (4.5), the standardized vector of random variables has all means equal to 0, all variances equal to 1, and all correlations equal to 0. In either case, it follows from property 1b that $\mathbf{z}$ is multivariate normal.

$$\text{If } \mathbf{y} \text{ is } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ then } \mathbf{z} \text{ is } N_p(\mathbf{0}, \mathbf{I}).$$

3. Chi-square distribution:

A chi-square random variable with p degrees of freedom is defined as the sum of squares of p independent standard normal random variables. Thus, if $\mathbf{z}$ is the standardized vector defined in (4.4) or (4.5), then $\sum_{i=1}^p z_i^2 = \mathbf{z}'\mathbf{z}$ has the χ^2 -distribution with p degrees of freedom, denoted as χ_p^2 or $\chi^2(p)$. From either (4.4) or (4.5) we obtain $\mathbf{z}'\mathbf{z} = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$. Hence,

$$\text{If } \mathbf{y} \text{ is } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ then } (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \text{ is } \chi_p^2. \quad (4.6)$$

4. Normality of marginal distributions:

- (iii) Any subset of the y 's in $\mathbf{y}$ has a multivariate normal distribution, with mean vector consisting of the corresponding subvector of $\boldsymbol{\mu}$ and covariance matrix composed of the corresponding submatrix of $\boldsymbol{\Sigma}$. To illustrate, let $\mathbf{y}_1 = (y_1, y_2, \dots, y_r)'$ denote the subvector containing the first r elements of $\mathbf{y}$ and $\mathbf{y}_2 = (y_{r+1}, \dots, y_p)'$ consist of the remaining $p - r$ elements. Then $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\mathbf{y}$ are partitioned as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{y}_1$ and $\boldsymbol{\mu}_1$ are $r \times 1$ and $\boldsymbol{\Sigma}_{11}$ is $r \times r$. Then $\mathbf{y}_1$ is multivariate normal:

$$\text{If } \mathbf{y} \text{ is } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ then } \mathbf{y}_1 \text{ is } N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}).$$

Here, again, $E(\mathbf{y}_1) = \boldsymbol{\mu}_1$ and $\text{cov}(\mathbf{y}_1) = \boldsymbol{\Sigma}_{11}$ hold for any random vector partitioned in this way. But if $\mathbf{y}$ is p -variate normal, then $\mathbf{y}_1$ is r -variate normal.

- (iv) As a special case of the preceding result, each y_j in $\mathbf{y}$ has the univariate normal distribution:

$$\text{If } \mathbf{y} \text{ is } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ then } y_j \text{ is } N(\mu_j, \sigma_{jj}^2), \quad j = 1, 2, \dots, p.$$

The converse of this is not true. If the density of each y_j in $\mathbf{y}$ is normal, it does not necessarily follow that $\mathbf{y}$ is multivariate normal.

In the next three properties, let the observation vector be partitioned into two subvectors denoted by $\mathbf{y}$ and $\mathbf{u}$, where $\mathbf{y}$ is $p \times 1$ and $\mathbf{u}$ is $q \times 1$. Or, alternatively, let $\mathbf{u}$ represent some additional variables to be considered along with those in $\mathbf{y}$. Then, as in (3.45) and (3.46),

$$E\left(\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix}\right) = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_u \end{pmatrix}, \quad \text{cov}\left(\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix}\right) = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yu} \\ \boldsymbol{\Sigma}_{uy} & \boldsymbol{\Sigma}_{uu} \end{pmatrix}.$$

In properties 5, 6, and 7, we assume that

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix} \text{ is } N_{p+q}\left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_u \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yu} \\ \boldsymbol{\Sigma}_{uy} & \boldsymbol{\Sigma}_{uu} \end{pmatrix}\right].$$

5. Independence:

- (i) The subvectors $\mathbf{y}$ and $\mathbf{u}$ are independent if $\boldsymbol{\Sigma}_{yu} = \mathbf{0}$.
- (ii) Two individual variables y_1 and y_2 are independent if $\sigma_{y_1 y_2} = 0$. Note that this is not true for many nonnormal random variables, as illustrated in Section 5.2.1.

4. Conditional distribution:

If $\mathbf{y}$ and $\mathbf{x}$ are not independent, then $\Sigma_{yy} \neq \mathbf{O}$, and the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, $f(\mathbf{y}|\mathbf{x})$, is multivariate normal with

$$E(\mathbf{y}|\mathbf{x}) = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x), \quad (4.7)$$

$$\text{cov}(\mathbf{y}|\mathbf{x}) = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \quad (4.8)$$

Note that $E(\mathbf{y}|\mathbf{x})$ is a vector of linear functions of $\mathbf{x}$, whereas $\text{cov}(\mathbf{y}|\mathbf{x})$ is a matrix that does not depend on $\mathbf{x}$. The linear trend in (4.7) holds for any pair of variables. Thus to use (4.7) as a check on normality, one can examine bivariate scatter plots of all pairs of variables and look for any nonlinear trends. In (4.7), we have the justification for using the covariance or correlation to measure the relationship between two bivariate normal random variables. As noted in Section 3.2.1, the covariance and correlation are good measures of relationship only for variables with linear trend and are generally unsuitable for nonnormal random variables with a curvilinear relationship. The matrix $\Sigma_{yx}\Sigma_{xx}^{-1}$ in (4.7) is called the matrix of regression coefficients because it relates $E(\mathbf{y}|\mathbf{x})$ to $\mathbf{x}$. The sample counterpart of this matrix appears in (10.52).

5. Distribution of the sum of two subvectors:

If $\mathbf{y}$ and $\mathbf{z}$ are the same size (both $p = 1$) and independent, then

$$\mathbf{y} + \mathbf{z} \sim N_p(\mu_y + \mu_z, \Sigma_{yy} + \Sigma_{zz}), \quad (4.9)$$

$$\mathbf{y} - \mathbf{z} \sim N_p(\mu_y - \mu_z, \Sigma_{yy} + \Sigma_{zz}). \quad (4.10)$$

In the remainder of this section, we illustrate property 5 for the special case of the bivariate normal. Let

$$\mathbf{u} = \begin{pmatrix} y \\ x \end{pmatrix}$$

have a bivariate normal distribution with

$$E(\mathbf{u}) = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \quad \text{cov}(\mathbf{u}) = \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix}.$$

By definition $f(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}, \mathbf{x})/h(\mathbf{x})$, where $h(\mathbf{x})$ is the density of x and $g(\mathbf{y}, \mathbf{x})$ is the joint density of y and x . Hence

$$g(\mathbf{y}, \mathbf{x}) = f(\mathbf{y}|\mathbf{x})h(\mathbf{x}),$$

and because the right side is a product, we seek a function of y and x that is independent of x and whose density can serve as $f(\mathbf{y}|\mathbf{x})$. Since linear functions of y and x are normal by property 1a, we consider $y - \rho x$ and seek the value of ρ so that $y - \rho x$ and x are independent.

Since $z = y - \beta x$ and x are normal and independent, $\text{cov}(x, z) = 0$. To find $\text{cov}(x, z)$, we express x and z as functions of $\mathbf{u}$,

$$x = \phi(1) \left(\frac{x}{\sigma} \right) = \phi(1) \mathbf{u} = a' \mathbf{u},$$

$$z = y - \beta x = (1 - \beta) \mathbf{u} = b' \mathbf{u}.$$

Now

$$\begin{aligned} \text{cov}(x, z) &= \text{cov}(a' \mathbf{u}, b' \mathbf{u}) \\ &= a' \Sigma b \quad \text{[eq. (3.71)]} \\ &= \phi(1) \left(\frac{\sigma^2}{\sigma_x} \frac{\sigma_y}{\sigma_x} \right) \left(-\frac{1}{\beta} \right) = (\sigma_y - \sigma_x^2) \left(-\frac{1}{\beta} \right) \\ &= \sigma_y - \beta \sigma_x^2. \end{aligned}$$

Since $\text{cov}(x, z) = 0$, we obtain $\beta = \sigma_y/\sigma_x^2$, and $z = y - \beta x$ becomes

$$z = y - \frac{\sigma_y}{\sigma_x^2} x.$$

By property 1a, the density of $y - (\sigma_y/\sigma_x^2)x$ is normal with

$$\begin{aligned} E \left(y - \frac{\sigma_y}{\sigma_x^2} x \right) &= \mu_y - \frac{\sigma_y}{\sigma_x^2} \mu_x, \\ \text{var} \left(y - \frac{\sigma_y}{\sigma_x^2} x \right) &= \text{var}(b' \mathbf{u}) = b' \Sigma b \\ &= \left(1 - \frac{\sigma_y}{\sigma_x^2} \right) \left(\frac{\sigma_x^2}{\sigma_x} \frac{\sigma_y}{\sigma_x} \right) \left(-\frac{1}{\sigma_x^2} \right) \\ &= \sigma_x^2 - \frac{\sigma_y^2}{\sigma_x^2}. \end{aligned}$$

For a given value of x , we can express y as $y = \beta x + (y - \beta x)$, where βx is a fixed quantity corresponding to the given value of x and $y - \beta x$ is a random deviation. Thus, $f(x)$ is normal, with

$$\begin{aligned} E[y(x)] &= \beta x + E(y - \beta x) = \beta x + \mu_y - \beta \mu_x \\ &= \mu_y + \beta(x - \mu_x) = \mu_y + \frac{\sigma_y}{\sigma_x^2} (x - \mu_x), \\ \text{var}[y(x)] &= \sigma_x^2 - \frac{\sigma_y^2}{\sigma_x^2}. \end{aligned}$$

4.3 ESTIMATION IN THE MULTIVARIATE NORMAL

4.3.1 Maximum Likelihood Estimation

When a distribution such as the multivariate normal is assumed to hold for a population, estimates of the parameters are often found by the method of maximum likelihood. This technique is conceptually simple. The observation vectors $y_1, y_2, \dots, y_n$ are considered to be known and values of μ and Σ are sought that maximize the joint density of the y 's, called the likelihood function. For the multivariate normal, the maximum likelihood estimates of μ and Σ are

$$\hat{\mu} = \bar{y}, \quad (4.11)$$

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})' \\ &= \frac{1}{n} \mathbf{M} \\ &= \frac{n-1}{n} S, \end{aligned} \quad (4.12)$$

where $\mathbf{M} = \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})'$ and S is the sample covariance matrix defined in (3.21) and (3.27). Since $\hat{\Sigma}$ has divisor n instead of $n-1$, it is biased [see (3.13)], and we usually use S in place of $\hat{\Sigma}$.

We now give a justification of $\bar{y}$ as the maximum likelihood estimator of μ . Because the y_i 's constitute a random sample, they are independent, and the joint density is the product of the densities of the y_i 's. The likelihood function is, therefore,

$$\begin{aligned} L(y_1, y_2, \dots, y_n; \mu, \Sigma) &= \prod_{i=1}^n f(y_i; \mu, \Sigma) \\ &= \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(y_i - \mu)' \Sigma^{-1} (y_i - \mu)} \\ &= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)' \Sigma^{-1} (y_i - \mu)}. \end{aligned} \quad (4.13)$$

To see that $\hat{\mu} = \bar{y}$ maximizes the likelihood function, we begin by adding and subtracting $\bar{y}$ in the exponent in (4.13),

$$-\frac{1}{2} \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \mu)' \Sigma^{-1} (y_i - \bar{y} + \bar{y} - \mu).$$

When this is expanded in terms of $y_i - \bar{y}$ and $\bar{y} - \mu$, two of the four resulting terms vanish because $\sum_{i=1}^n (y_i - \bar{y}) = \mathbf{0}$, and (4.13) becomes

$$L = \frac{1}{(2\pi)^{p/2} |\Sigma|^{n/2}} e^{-\frac{1}{2}(\bar{y} - \mu)' \Sigma^{-1}(\bar{y} - \mu) - \frac{1}{2}(\bar{S} - \mu' \Sigma^{-1} \bar{y} - \mu' \Sigma^{-1} \bar{y} - \mu' \Sigma^{-1} \bar{y} - \mu' \Sigma^{-1} \bar{y})}, \quad (4.14)$$

Since Σ^{-1} is positive definite, we have $-\mu'(\bar{y} - \mu)' \Sigma^{-1}(\bar{y} - \mu)/2 \leq 0$ and $0 \leq e^{-\mu'(\bar{y} - \mu)' \Sigma^{-1}(\bar{y} - \mu)/2} \leq 1$, with the maximum occurring when the exponent is 0. Therefore, L is maximized when $\bar{\mu} = \bar{y}$.

The maximum likelihood estimator of the population covariance matrix Σ , Den (3.26) is the sample covariance matrix, that is,

$$\hat{\Sigma}_p = \mathbf{S}.$$

Relationships among multivariate variables are linear, as can be seen in (4.3). Thus the estimators $\hat{\Sigma}$ and $\hat{\Sigma}_p$ serve well for the multivariate normal because they measure only linear relationships (see Sections 3.2.1 and 4.2). These estimators are not as useful for some nonnormal distributions.

4.2.2 Distribution of $\bar{y}$ and $\mathbf{S}$

For the distribution of $\bar{y} = \sum_{i=1}^n \mathbf{y}_i/n$, we can distinguish two cases:

1. When $\bar{y}$ is based on a random sample $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ from a multivariate normal distribution $N_p(\mu, \Sigma)$, then $\bar{y}$ is $N_p(\mu, \Sigma/n)$.
2. When $\bar{y}$ is based on a random sample $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ from a nonnormal multivariate population with mean vector μ and covariance matrix Σ , then for large n , $\bar{y}$ is approximately $N_p(\mu, \Sigma/n)$. More formally, this result is known as the *multivariate central limit theorem*. If $\bar{y}$ is the mean vector of a random sample $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ from a population with mean vector μ and covariance matrix Σ , then as $n \rightarrow \infty$, the distribution of $\sqrt{n}(\bar{y} - \mu)$ approaches $N_p(\mathbf{0}, \Sigma)$.

There are p variances in $\mathbf{S}$ and $\binom{p}{2}$ covariances, for a total of

$$p + \binom{p}{2} = p + \frac{p(p-1)}{2} = \frac{p(p+1)}{2}$$

distinct entries. The joint distribution of these $p(p+1)/2$ distinct variables in $\mathbf{W} = (n-1)\mathbf{S} = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ is the *Wishart distribution*, denoted by $W_p(n-1, \Sigma)$, where $n-1$ is the degrees of freedom.

The Wishart distribution is the multivariate analogue of the χ^2 -distribution, and it has similar uses. As noted in property 3 of Section 4.2, a χ^2 random variable is defined formally as the sum of squares of independent standard normal (univariate) random variables:

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \frac{(Z_i - \mu)^2}{\sigma^2} \sim \chi^2(n).$$

If $\bar{y}$ is substituted for μ , then $\sum_{i=1}^n (y_i - \bar{y})^2/n^2 = (n-1)(s^2/n^2) = s^2/n^2(n-1)$. Similarly, the formal definition of a Wishart random variable is

$$\sum_{i=1}^n (y_i - \mu)(y_i - \mu)^T \text{ is } W_p(n, \Sigma), \quad (4.15)$$

where $y_1, y_2, \dots, y_n$ are independently distributed as $N_p(\mu, \Sigma)$. When $\bar{y}$ is substituted for μ , the distribution remains Wishart with one less degree of freedom:

$$(n-1)S = \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T \text{ is } W_p(n-1, \Sigma). \quad (4.16)$$

Finally, we note that when sampling from a multivariate normal distribution, $\bar{y}$ and S are independent.

4.4 ASSESSING MULTIVARIATE NORMALITY

Many tests and graphical procedures have been suggested for evaluating whether a data set likely originated from a multivariate normal population. One possibility is to check each variable separately for univariate normality. Goodness reviews for both the univariate and multivariate cases have been given by Unmuth (1991, pp. 174–220) and Fisher (1994, pp. 341–353). We give a representative sample of univariate and multivariate methods in Sections 4.4.1 and 4.4.2, respectively.

4.4.1 Investigating Univariate Normality

When we have several variables, checking each for univariate normality should not be the sole approach, because (1) the variables are correlated and (2) normality of the individual variables does not guarantee joint normality. On the other hand, multivariate normality implies individual normality. Hence, if even one of the separate variables is not normal, the vector is not multivariate normal. An initial check on the individual variables may therefore be useful.

A basic graphical approach for checking normality is the $\bar{Q}$ - $\bar{Q}$ plot comparing quantiles of a sample against the population quantiles of the univariate normal. If the points are close to a straight line, there is no indication of departure from normality. Deviation from a straight line indicates nonnormality (at least for a large sample). In fact, the type of nonlinear pattern may reveal the type of departure from normality. Some possibilities are illustrated in Figure 4.8.

Quantiles are similar to the more familiar percentiles, which are expressed in terms of percent; a test score at the 90th percentile, for example, is above 90% of the test scores and below 10% of them. Quantiles are expressed in terms of fractions or proportions. Thus the 90th percentile score becomes the .9 quantile score.

The sample quantiles for the $\bar{Q}$ - $\bar{Q}$ plot are obtained as follows. First we rank the observations $y_1, y_2, \dots, y_n$ and denote the ordered values by $y_{(1)}, y_{(2)}, \dots, y_{(n)}$. Then

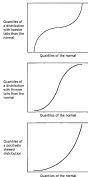


Figure 4.5. Typical Q-Q-plots for unimodal data.

first $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$. Then the point $y_{(i)}$ is the i 's sample quantile. For example, if $n = 20$, $y_{(5)}$ is the $\frac{5}{20} = .25$ quantile, because .25 of the sample is less than or equal to $y_{(5)}$. The fraction i/n is often changed to $(i - \frac{1}{2})/n$ as a continuity correction. If $n = 20$, $(i - \frac{1}{2})/n$ ranges from .125 to .875 and more evenly covers the interval from 0 to 1. With this convention, $y_{(5)}$ is designated as the $(i - \frac{1}{2})/n$ sample quantile.

The population quantiles for the Q-Q plot are similarly defined corresponding to $(i - \frac{1}{2})/n$. If we denote these by $q_1, q_2, \dots, q_n$, then q_i is the value below which a

proportion $i - \frac{1}{2}/n$ of the observations in the population lie, that is, $Q = \frac{i - \frac{1}{2}}{n}$ is the probability of getting an observation less than or equal to q_i . Formally, q_i can be found for the standard normal random variable y with distribution $N(0, 1)$ by solving

$$\Phi(q_i) = P(Y = q_i) = \frac{i - \frac{1}{2}}{n}, \quad (4.17)$$

which would require numerical integration or tables of the cumulative standard normal distribution, $\Phi(x)$. Another benefit of using $Q = \frac{i - \frac{1}{2}}{n}$ instead of i/n is that $n/n = 1$ would make $g_n = \infty$.

The population need not have the same mean and variance as the sample, since changes in mean and variance merely change the slope and intercept of the plotted line in the $Q-Q$ plot. Therefore, we use the standard normal distribution, and the g_i values can easily be found from a table of cumulative standard normal probabilities. We then plot the pairs $(q_i, x_{(i)})$ and examine the resulting $Q-Q$ plot for linearity.

Special graph paper, called normal probability paper, is available that eliminates the need to look up the g_i values. We need only plot $y = \frac{i - \frac{1}{2}}{n}$ in place of g_i , that is, plot the pairs $(y = \frac{i - \frac{1}{2}}{n}, x_{(i)})$ and look for linearity as before. As an even more alternative, most general-purpose statistical software programs provide normal probability plots of the pairs $(q_i, x_{(i)})$.

The $Q-Q$ plots provide a good visual check on normality and are considered to be adequate for this purpose by many researchers. For those who desire a more objective procedure, several hypothesis tests are available. We give three of those that have good properties and are computationally tractable.

We discuss first a classical approach based on the following measures of skewness and kurtosis:

$$\sqrt{g_1} = \frac{\sqrt{6} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}, \quad (4.18)$$

$$g_2 = \frac{6 \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}. \quad (4.19)$$

These are sample estimates of the population skewness and kurtosis parameters $\sqrt{g_1}$ and g_2 , respectively. When the population is normal, $\sqrt{g_1} = 0$ and $g_2 = 3$. If $\sqrt{g_1} < 0$, we have negative skewness; if $\sqrt{g_1} > 0$, the skewness is positive. Positive skewness is illustrated in Figure 4.6. If $g_2 < 3$, we have negative kurtosis, and if $g_2 > 3$, there is positive kurtosis. A distribution with negative kurtosis is characterized by being flatter than the normal distribution, that is, less peaked, with heavier flanks and thinner tails. A distribution with positive kurtosis has a higher peak than the normal, with an excess of values near the mean and in the tails but with thinner flanks. Positive and negative kurtosis are illustrated in Figure 4.7.

The test of normality can be carried out using the exact percentage points for $\sqrt{g_1}$ in Table A.1 for $4 \leq n \leq 25$, as given by Muthukrishnan (1977). Alternatively, for $n \geq 8$

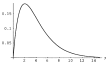


Figure 4.6. A distribution with positive skewness.

the function g , is defined by

$$g(\sqrt{d_1}) = d \sinh^{-1} \left(\frac{\sqrt{d_1}}{2} \right), \quad (4.28)$$

is approximately $2\sqrt{d_1}/3$, where

$$\sinh^{-1}(x) = \ln(x + \sqrt{x^2 + 1}). \quad (4.29)$$

Table A.1, from D'Agostino and Pearson (1973), gives values for d and $1/3$. To use b_2 as a test of normality, we can use Table A.3, from D'Agostino and Tibshirani (1990), which gives simulated percentiles of b_2 for selected values of n in the range $7 \leq n \leq 50$. Charts of percentiles of b_2 for $20 \leq n \leq 200$ can be found in D'Agostino and Pearson (1973).

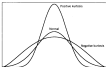


Figure 4.7. Distributions with positive and negative skewness compared to the normal.

Our second test for normality was given by D'Agostino (1971). The observations $y_1, y_2, \dots, y_n$ are ordered as $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$, and we calculate

$$D = \frac{\sum_{i=1}^n \left[i - \frac{1}{2}(n+1) \right] y_{(i)}}{\sqrt{n^3 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (4.22)$$

$$r = \frac{\sqrt{6} [D - (2/\sqrt{6})]}{0.004398}. \quad (4.23)$$

A table of percentiles for r , given by D'Agostino (1972) for $1.0 \leq n \leq 250$, is provided in Table A.4.

The final test we report is by Lin and Michalek (1980). The test statistic is

$$t = \tanh^{-1}(r) = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right), \quad (4.24)$$

where r is the sample correlation of the n pairs (y_i, x_i) , $i = 1, 2, \dots, n$, with x_i defined as

$$x_i = \frac{1}{n} \left[\sum_{j=1}^n y_j^2 - \frac{\left(\sum_{j=1}^n y_j \right)^2}{n-1} \right]^{1/2}. \quad (4.25)$$

If the y 's are normal, t is approximately $N(0, 1/n)$. A more accurate upper 5.0% percentile is given by

$$z_{.95} = z_{.95}^* \left[n_{.95} + \frac{1}{24} (n_{.95}^3 - 3n_{.95}) n_{.95} \right], \quad (4.26)$$

with

$$z_{.95}^* = \frac{1}{n} - \frac{7.124}{n^2} + \frac{10.806}{n^3}, \quad n_{.95} = \Phi^{-1}(z_{.95}), \quad z_{.95} = -\frac{11.79}{n} + \frac{10.66}{n^2},$$

where Φ is the distribution function of the $N(0, 1)$ distribution, that is, $\Phi(x)$ is the probability of an observation less than or equal to x , from (4.17). The inverse function Φ^{-1} is essentially a quantile. For example, $z_{.95} = -1.645$ and $z_{.05} = 1.645$.

4.4.3 Investigating Multivariate Normality

Checking for multivariate normality is conceptually not as straightforward as assessing univariate normality, and consequently the state of the art is not as well developed. The complexity of this issue can be illustrated in the context of a goodness-of-fit test for normality. For a goodness-of-fit test in the univariate case,

the range covered by a sample $y_1, y_2, \dots, y_n$ is divided into several intervals, and we count how many y 's fall into each interval. These observed frequencies (counts) are compared to the expected frequencies under the assumption that the sample came from a normal distribution with the same mean and variance as the sample. If the n observations $y_1, y_2, \dots, y_n$ are multivariate, however, the procedure is not so simple. We now have a p -dimensional region that would have to be divided into many more subregions than in the univariate case, and the expected frequencies for these subregions would be less easily obtained. With so many subregions, relatively few would contain observations.

Thus because of the inherent “sparseness” of multivariate data, a goodness-of-fit test would be impractical. The points $y_1, y_2, \dots, y_n$ are more distant from each other in p -space than in any one of the p individual dimensions. Unless n is very large, a multivariate sample may not provide a very complete picture of the distribution from which it was taken.

As a consequence of the sparseness of the data in p -space, the tests for multivariate normality may not be very powerful. However, some check on the distribution is often desirable. Numerous procedures have been proposed for assessing multivariate normality. We now discuss three of them.

The first procedure is based on the standardized distance from each y_i to $\bar{y}$,

$$D_i^2 = (y_i - \bar{y})'W^{-1}(y_i - \bar{y}), \quad i = 1, 2, \dots, n. \quad (4.27)$$

Omernikowski and Kottlering (1972) showed that if the y_i 's are multivariate normal, then

$$u_i = \frac{nD_i^2}{(n-1)p} \quad (4.28)$$

has a beta distribution, which is related to the F distribution. To obtain a Q-Q plot, the values $u_1, u_2, \dots, u_n$ are ranked to give $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$, and we plot $(u_{(i)}, i/n)$, where the quantiles v_i of the beta are given by the solution to

$$\int_0^v \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} dx = \frac{i-a}{a-a-b+1}, \quad (4.29)$$

where $a = \frac{1}{2}p$, $b = \frac{1}{2}(n-p-1)$,

$$\alpha = \frac{p-2}{2p}, \quad (4.30)$$

$$\beta = \frac{n-p-3}{2(n-p-1)}. \quad (4.31)$$

A nonlinear pattern in the plot would indicate a departure from normality. The quantiles of the beta as defined in (4.29) are easily obtained in many software packages.

A formal significance test is also available for $D_{\text{obs}}^2 = \max_i D_i^2$. Table A.6 gives the upper 5% and 1% critical values for $p = 1, 3, 4, 5$ from Barnett and Lewis (1978).

Some writers have suggested that the distribution of D_i^2 in (4.27) can be adequately approximated by a χ^2_1 since $(\mathbf{y} - \boldsymbol{\mu})(\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}))$ is χ^2_p [see (4.10)]. However, in Sections 5.3-2, it is shown that this approximation is very poor for even moderate values of p . Small (1978) showed that plots of D_i^2 vs. χ^2_p quantiles are misleading.

The second procedure involves scatter plots in two dimensions. If p is not too high, the bivariate plots of each pair of variables are often reduced in size and shown on one page, arranged to correspond to the entries in a correlation matrix. In this visual matrix, the eye readily picks out those pairs of variables that show a curved trend, outliers, or other nonnormal appearance. This plot is illustrated in Example 4.5.1 in Section 4.5.2. The procedure is based on properties 4 and 5 of Section 4.1, from which we infer that (1) each pair of variables has a bivariate normal distribution and (2) bivariate normal variables follow a straight-line trend.

A popular option in many graphical programs is the ability to dynamically rotate a plot of three variables. While the points are rotating on the screen, a three-dimensional effect is created. The shape of the three-dimensional cloud of points is easily perceived, and we can detect various features of the data. The only drawbacks to this technique are that (1) it is a dynamic display and cannot be printed and (2) if p is very large, the number of subsets of three variables becomes unmanageable, although the number of pairs may still be tractable for plotting. These numbers are compared in Table 4.1, where $\binom{p}{2}$ and $\binom{p}{3}$ represent the number of subsets of sizes 2 and 3, respectively. Thus, in many cases, the scatter plots for pairs of variables will continue to be used, even though three-dimensional plotting techniques are available.

The third procedure for assessing multivariate normality is a generalization of the univariate test based on the skewness and kurtosis measures $\sqrt{B_1}$ and B_2 as given by (4.18) and (4.20). The test is due to Mardia (1970). Let $\mathbf{y}$ and $\mathbf{x}$ be independent and identically distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then skewness and kurtosis for multivariate populations are defined by Mardia as

$$B_{1,p} = E[(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]^3, \quad (4.32)$$

$$B_{2,p} = E[(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})]^4. \quad (4.33)$$

Table 4.1. Comparison of Number of Subsets of Sizes 2 and 3

p	$\binom{p}{2}$	$\binom{p}{3}$
4	15	28
6	28	96
10	45	120
12	66	228
15	105	455

Since third-order central moments for the multivariate normal distribution are zero, $\beta_{3,p} = 0$ when $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It can also be shown that if $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\beta_{2,p} = p(p+2). \quad (4.34)$$

To estimate $\beta_{1,p}$ and $\beta_{2,p}$ using a sample $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$, we first define

$$\mathbf{d}_i = (\mathbf{y}_i - \bar{\mathbf{y}})(\bar{\mathbf{S}}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}})), \quad (4.35)$$

where $\bar{\mathbf{S}} = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'/n$ is the maximum likelihood estimator (4.12). Then estimates of $\beta_{1,p}$ and $\beta_{2,p}$ are given by

$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}', \quad (4.36)$$

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^n d_{ii}', \quad (4.37)$$

Table A.5 (Mardia, 1970, 1974) gives percentage points of $\beta_{1,p}$ and $\beta_{2,p}$ for $p = 2, 3, 4$, which can be used in testing the multivariate normality. For other values of p or when $n \geq 50$, the following approximate tests are available. For $b_{1,p}$ the statistic

$$z_1 = \frac{(p+1)(n+1)(n+2)}{n(p+1)(p+1)-6} b_{1,p}^2 \quad (4.38)$$

is approximately χ^2 with $\frac{1}{2}p(p+1)(p+2)$ degrees of freedom. We reject the hypothesis of multivariate normality if $z_1 \geq \chi_{\alpha}^2$. With $b_{2,p}$, on the other hand, we wish to reject for large values (distribution too peaked) or small values (distribution too flat). For the upper 2.5% points of $b_{2,p}$ use

$$z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{6(p^2+2)/n}}, \quad (4.39)$$

which is approximately $N(0, 1)$. For the lower 2.5% points we have two cases: (1) when $50 \leq n \leq 400$, use

$$z_3 = \frac{b_{2,p} - p(p+2)(n+p+1)/n}{\sqrt{6(p^2+2)(n-1)}}, \quad (4.40)$$

which is approximately $N(0, 1)$; (2) when $n \geq 400$ use z_1 as given by (4.39).

4.5 OUTLIERS

The detection of outliers has been of interest to statisticians and other scientists for over a century. Some authors have claimed that the researcher can typically expect

up to 10% of the observations to have errors in measurement or recording. Occasional stray observations from a different population than the target population are also fairly common. We review some major concepts and suggested procedures for univariate outliers in Section 4.5.1 before moving to the multivariate case in Section 4.5.2. An alternative to detection of outliers is to use robust estimators of μ and Σ (see Rousseeuw 1979, Section 1.4.8) that are less sensitive to extreme observations than are the standard estimators $\bar{y}$ and S .

4.5.1 Outliers in Univariate Samples

Excellent treatments of outliers have been given by Boxerman and Cook (1980), Hawkins (1980), and Barnett and Lewis (1978). We abstract a few highlights from Boxerman and Cook. Many techniques have been proposed for detecting outliers in the residuals from regression or designed experiments, but we will focus our attention only with simple random samples from the normal distribution.

There are two principal approaches for dealing with outliers. The first is identification, which usually involves detection of the outlier(s) that may also provide important information about the model or the data. The second method is accommodation, in which the method of analysis or the model is modified. Robust methods, in which the influence of outliers is reduced, provide an example of modification of the analysis. An example of a correction to the model is a mixture model that combines two normals with different variances. For example, Marks and Rao (1978) accommodated a particular type of outlier by a mixture of two normal distributions.

In small or moderate-sized univariate samples, visual methods of identifying outliers are the most frequently used. Tests are also available if a less subjective approach is desired.

Two types of slippage models have been proposed to account for outliers. Under the mean slippage model, all observations have the same variance, but one or more of the observations arise from a distribution with a different (population) mean. In the variance slippage model, one or more of the observations arise from a model with larger (population) variance but the same mean. Thus in the mean slippage model, the bulk of the observations arise from $N(\mu, \sigma^2)$, whereas the outliers originate from $N(\mu, \sigma^2)$. For the variance slippage model, the main distribution would again be $N(\mu, \sigma^2)$, with the outliers coming from $N(\mu, a\sigma^2)$ where $a > 1$. These models have led to the development of tests for rejection of outliers. We now briefly discuss some of these tests.

For a single outlier in a sample $y_1, y_2, \dots, y_n$, most tests are based on the maximum standardized residual,

$$\max_j t_j = \max_j \left| \frac{y_j - \bar{y}}{s} \right|, \quad (4.40)$$

If the largest or smallest observation is rejected, one could then examine the $n - 1$ remaining observations for another possible outlier, and so on. This procedure is

robustness, conservative test. However, if there are two or more outliers, the less-extreme ones will often make it difficult to detect the most extreme one, due to inflation of both mean and variance. This effect is called *masking*.

Papadatos (1961) showed that the maximum studentized residual (H -test) is more powerful than most other techniques for detecting intermediate or large shifts in the mean and gave the following guidelines for small shifts:

1. For outliers with small positive shifts in the mean, tests based on sample skewness are best.
2. For outliers with small shifts in the mean in either direction, tests based on the sample kurtosis are best.
3. For outliers with small positive shifts in the variance, tests based on the sample kurtosis are best.

Because of the masking problem in conservative tests, block tests have been proposed for simultaneous rejection of $k > 1$ outliers. These tests work well if k is known, but in practice, k is usually not known. If the value we compute for k is too small, we incur the risk of failing to detect any outliers because of masking. If we set k too large, there is a high risk of rejecting more outliers than there really are, an effect known as *oversteering*.

4.3.2 Outliers in Multivariate Samples

In the case of multivariate data, the problems in detecting outliers are intensified for several reasons:

1. For $p > 2$ the data cannot be readily plotted to pinpoint the outliers.
2. Multivariate data cannot be ordered as can a univariate sample, where extremes show up readily on either end.
3. An observation vector may have a large recording error in one of its components or smaller errors in several components.
4. A multivariate outlier may reflect slippage in mean, variance, or correlation. This is illustrated in Figure 4.8. Observation 1 causes a small shift in means and variances of both y_1 and y_2 but has little effect on the correlation. Observation 2 has little effect on means and variances, but it reduces the correlation somewhat. Observation 3 has a major effect on means, variances, and correlation.

One approach to multivariate outlier identification or accommodation is to use robust methods of estimation. Such methods minimize the influence of outliers in estimation or model fitting. However, an outlier sometimes furnishes valuable information, and the quest for outliers can be very worthwhile.

We present two methods of multivariate outlier identification, both of which are related to methods of assessing multivariate normality. (A third approach based



Figure 4.6. Simulated sample showing three types of outliers.

on principal components is given in Section 5.1.4.) The first method, due to Wilks (1953), is designed for detection of a single outlier; Wilks' statistic is

$$w = \max_i \frac{|y_i - \bar{y}|^2 \mathbf{S}_{-i}}{(n-1)\mathbf{S}}, \quad (4.42)$$

where $\mathbf{S}$ is the usual sample covariance matrix, and $\mathbf{S}_{-i}$ is obtained from the same sample with the i th observation deleted. The statistic w can also be expressed in terms of $D_{\max}^2 = \max_i (y_i - \bar{y})' \mathbf{S}^{-1} (y_i - \bar{y})$ as

$$w = 1 - \frac{n D_{\max}^2}{(n-1)^2}, \quad (4.43)$$

thus having a hint for an outlier in the distance $D_{\max}^2$ used in Section 4.4.2 in a graphical procedure for checking multivariate normality. Table A.6 gives the upper 5% and 1% critical values for $D_{\max}^2$ from Barnett and Lewis (1978).

Yang and Lee (1977) provide an F -test of w as given by (4.43). Define

$$F_i = \frac{n-p-1}{p} \left[\frac{1}{1 - n D_i^2 / (n-1)^2} - 1 \right], \quad i = 1, 2, \dots, n. \quad (4.44)$$

Then the F_i 's are independently and identically distributed as $F_{p, n-p-1}$, and a test can be constructed in terms of $\max_i F_i$:

$$P(\max_i F_i > f) = 1 - P(\text{all } F_i \leq f) = 1 - [P(F \leq f)]^n.$$

Therefore, the test can be carried out using an F -table. Note that

$$\max_i F_i = F_{\max} = \frac{n-p-1}{p} \left(\frac{1}{w} - 1 \right), \quad (4.45)$$

where w is given in (4.43).

The second test we discuss is designed for detection of several outliers. Schnberger and Margolin (1962) showed that the locally best invariant test for mean slippage is based on Mascha's (1930) sample invariant $b_{2,p}$ as defined by (4.35) and (4.37). To be more specific, among all tests invariant to a class of transformations of the type $\mathbf{x} = A\mathbf{y} + \mathbf{b}$, where A is nonsingular (see Problem 4.3), the test using $b_{2,p}$ is most powerful for small shifts in the mean vector. This result holds if the proportion of outliers is no more than 21.1%. With some restrictions on the pattern of the outliers, the permissible fraction of outliers ranges as high as 75%. The hypothesis is H_0 : no outliers are present. This hypothesis is rejected for large values of $b_{2,p}$.

A table of critical values of $b_{2,p}$ and some approximate tests were described in Section 4.4.7 following (4.37). Thus the test doubles as a check for multivariate normality and for the presence of outliers. One advantage of this test for outliers is that we do not have to specify the number of outliers and run the attendant risk of masking or swamping. Schnberger and Margolin (1962) pointed out that this feature "increases the importance of performing an overall test that is sensitive to a broad range of outlier configurations. There is also empirical evidence that the invariant test performs well in situations of practical interest when compared with other individual outlier procedures."

Sidak (1964) extended the result of Schnberger and Margolin to cover the general case of elliptically symmetric distributions. An *elliptically symmetric* distribution is one in which $f(\mathbf{x}) = |\Sigma|^{-1/2} g(\mathbf{x} - \mu)\Sigma^{-1}(\mathbf{x} - \mu)$. By varying the function g , distributions with shorter or longer tails than the normal can be obtained. The critical values of $b_{2,p}$ must be adjusted to correspond to the distribution, but rejection for large values would be a locally best invariant test.

Example 4.3.2. We use the rumex home data set of Table 1.6 to illustrate a search for multivariate outliers, while at the same time checking for multivariate normality. An examination of each column of Table 1.6 does not reveal any apparent univariate outliers. To check for multivariate outliers, we first calculate S_{ij}^2 in (4.27) for each observation vector. The results are given in Table 4.2. We see that S_{11}^2 , S_{12}^2 , and S_{22}^2 seem to stand out as possible outliers. In Table A.6, the upper 5% critical value for the maximum value, $D_{\max}^2$, is given as 11.65. In our case, the largest S_{ij}^2 is $S_{11}^2 = 11.45$, which does not exceed the critical value. This does not surprise us, since the test was designed to detect a single outlier, and we may have as many as three.

We compute u_1 and v_1 in (4.28) and (4.29) and plot them in Figure 4.3. The figure shows a departure from linearity due to these values and possibly a fourth.

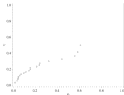
Table 4.1. Values of D_i^2 for the Random Stone Data in Table 3.6

Observation Number	D_i^2	Observation Number	D_i^2
1	.7568	11	1.8501
2	1.2080	12	10.4718
3	1.7891	13	1.3941
4	10.819	14	.6994
5	.8700	15	.3340
6	28.600	16	.8121
7	4.5515	17	1.1083
8	7.6807	18	4.3553
9	11.0504	19	1.1088
10	8.1819	20	10.0511

We next calculate $\hat{\sigma}_{1,p}$ and $\hat{\sigma}_{2,p}$, as given by (4.36) and (4.37):

$$\hat{\sigma}_{1,p} = 11.518, \quad \hat{\sigma}_{2,p} = 28.894.$$

In Table A.5, the upper .01 critical value for $\hat{\sigma}_{1,p}$ is 9.9; the upper .005-critical value for $\hat{\sigma}_{2,p}$ is 27.3. Thus both $\hat{\sigma}_{1,p}$ and $\hat{\sigma}_{2,p}$ exceed their critical values, and we have significant skewness and kurtosis, apparently caused by the three observations with large values of D_i^2 .

Figure 4.9. Q-Q plot of x_i and y_i for the random stone data of Table 3.6.

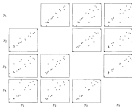


Figure 4.14: Scatter plots for the same data as in Table 3.5.

The bivariate scatter plots are given in Figure 4.14. Three values are clearly separate from the other observations in the plot of y_3 versus y_4 . In Table 3.5, the 9th, 12th, and 20th values of y_3 are not unusual, nor are the 9th, 12th, and 20th values of y_4 . However, the increase from y_3 to y_4 is exceptional in each case. If these values are not due to errors in recording the data and if this sample is representative, then we appear to have a mixture of two populations. This should be taken into account in making inferences. $\square$

PROBLEMS

4.1 Consider the two covariance matrices

$$\Sigma_1 = \begin{pmatrix} 14 & 8 & 3 \\ 8 & 9 & 2 \\ 3 & 2 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 6 & 5 & 1 \\ 5 & 8 & 2 \\ 1 & 2 & 1 \end{pmatrix}.$$

Show that $|\Sigma_1| > |\Sigma_2|$ and that $\text{tr}(\Sigma_1) < \text{tr}(\Sigma_2)$. Thus the generalized variance of population 1 is greater than the generalized variance of population 2, even though the total variance is less. Comment on why this is true in terms of the variances and correlations.

- 4.2 For $u = (T')^{-1}(y - \mu)$ in (4.4), show that $E(u) = 0$ and $\text{cov}(u) = I$.
- 4.3 Show that the form of the likelihood function in (4.13) follows from the previous expression.
- 4.4 For $(y - \mu)' \Sigma^{-1} (y - \mu)$ in (4.5) and (4.6), show that $E(y - \mu)' \Sigma^{-1} (y - \mu) = p$. Assume $E(y) = \mu$ and $\text{cov}(y) = \Sigma$. Normality is not required.
- 4.5 Show that by adding and subtracting $\bar{y}$, the exponent of (4.31) has the form given in (4.14), that is,

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \mu)' \Sigma^{-1} (y_i - \bar{y} + \bar{y} - \mu) &= \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y})' \Sigma^{-1} (y_i - \bar{y}) \\ &\quad + \frac{n}{2} (\bar{y} - \mu)' \Sigma^{-1} (\bar{y} - \mu). \end{aligned}$$

- 4.6 Show that $\sqrt{n} \bar{y}$ and S_n , as given in (4.18) and (4.19), are invariant to the transformation $z_i = ay_i + b$.
- 4.7 Show that if y is $N_p(\mu, \Sigma)$, then $\beta_{1,p} = p(p+2)$ as in (4.34).
- 4.8 Show that $\beta_{1,p}$ and $\beta_{2,p}$, as given by (4.36) and (4.37), are invariant under the transformation $y_i = Ay_i + b$, where b is nonsingular. Thus $\beta_{1,p}$ and $\beta_{2,p}$ do not depend on the units of measurement.
- 4.9 Show that $F_{(n,p)} = [n - p - 1]r/(1-r)$ as in (4.45).
- 4.10 Suppose y is $N_3(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 1 \\ 1 \\ 4 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 6 & 1 & -2 \\ 1 & 13 & 4 \\ -2 & 4 & 4 \end{pmatrix}.$$

- (a) Find the distribution of $z = 2y_1 - y_2 + 3y_3$.
- (b) Find the joint distributions of $z_1 = y_1 + y_2 + y_3$ and $z_2 = y_1 - y_2 + 2y_3$.
- (c) Find the distribution of y_2 .
- (d) Find the joint distributions of y_1 and y_3 .
- (e) Find the joint distributions of y_1 , y_3 , and $\frac{1}{2}(y_1 + y_2)$.
- 4.11 Suppose y is $N_3(\mu, \Sigma)$, with μ and Σ given in the previous problem.
- (a) Find a vector u such that $u = (T')^{-1}(y - \mu)$ is $N_3(0, I)$ as in (4.4).
- (b) Find a vector u such that $u = (\Sigma^{1/2})^{-1}(y - \mu)$ is $N_3(0, I)$ as in (4.5).
- (c) What is the distribution of $(y - \mu)' \Sigma^{-1} (y - \mu)$?
- 4.12 Suppose y is $N_4(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} -2 \\ 3 \\ -1 \\ 8 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 11 & -8 & 3 & 8 \\ -8 & 9 & -3 & 8 \\ 3 & -3 & 1 & 3 \\ 8 & 8 & 3 & 8 \end{pmatrix}.$$

- (iii) Find the distribution of $z = 4y_1 - 2y_2 + y_3 - 3y_4$.
- (iv) Find the joint distributions of $z_1 = y_1 + y_2 + y_3 + y_4$ and $z_2 = -2y_1 + 3y_2 + y_3 - 2y_4$.
- (v) Find the joint distributions of $z_1 = 3y_1 + y_2 - 4y_3 - y_4$, $z_2 = -y_1 - 2y_2 + y_3 - 2y_4$, and $z_3 = 2y_1 + 2y_2 + 4y_3 - 3y_4$.
- (vi) What is the distribution of y_3 ?
- (vii) What is the joint distribution of y_3 and y_4 ?
- (viii) Find the joint distributions of y_1 , $\frac{1}{2}(y_1 + y_2)$, $\frac{1}{2}(y_1 + y_2 + y_3)$, and $\frac{1}{2}(y_1 + y_2 + y_3 + y_4)$.

4.13 Suppose $\mathbf{y}$ is $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given in the previous problem.

- (i) Find a vector $\mathbf{a}$ such that $\mathbf{a} = (\mathbf{T}^T)^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is $N_2(\mathbf{0}, \mathbf{I})$ as in (3.4).
- (ii) Find a vector $\mathbf{a}$ such that $\mathbf{a} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is $N_2(\mathbf{0}, \mathbf{I})$ as in (4.5).
- (iii) What is the distribution of $(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$?

4.14 Suppose $\mathbf{y}$ is $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ -2 \\ 4 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & -3 & 0 \\ -3 & 6 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Which of the following random variables are independent?

- (a) y_1 and y_2
 (b) y_1 and y_3
 (c) y_2 and y_3
 (d) (y_1, y_2) and y_3
 (e) (y_1, y_2) and y_3

4.15 Suppose $\mathbf{y}$ is $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{pmatrix} -6 \\ 2 \\ 5 \\ -1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 8 & 0 & -1 & 0 \\ 0 & 5 & 0 & 1 \\ -1 & 0 & 5 & 0 \\ 0 & 1 & 0 & 7 \end{pmatrix}.$$

Which of the following random variables are independent?

- | | | |
|---------------------|----------------------------|-----------------------------------|
| (a) y_1 and y_2 | (d) y_3 and y_4 | (k) y_1 and y_2 and y_3 |
| (b) y_1 and y_3 | (e) (y_1, y_2) and y_3 | (l) y_1 and y_2 and y_4 |
| (c) y_1 and y_4 | (f) (y_1, y_2) and y_4 | (m) (y_3, y_2) and (y_3, y_4) |
| (d) y_2 and y_3 | (g) (y_3, y_4) and y_4 | (n) (y_1, y_2) and (y_3, y_4) |
| (e) y_2 and y_4 | (h) y_1 and (y_3, y_4) | |

4.16 Assume $\mathbf{y}$ and $\mathbf{u}$ are subvectors, each 2×1 , where $\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix}$ is $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ -1 \\ 3 \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 3 & 5 & -3 & 2 \\ 3 & 6 & 8 & 4 \\ \hline -3 & 0 & 5 & -2 \\ 2 & 4 & -2 & 4 \end{pmatrix}.$$

- (a) Find $E(\mathbf{y})$ by (4.7).
 (b) Find $\text{cov}(\mathbf{y})$ by (4.8).

4.17 Suppose $\mathbf{y}$ and $\mathbf{u}$ are subvectors, each that $\mathbf{y}$ is 2×1 and $\mathbf{u}$ is 3×1 , with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ partitioned accordingly:

$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ -2 \\ 4 \\ -5 \\ 6 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 14 & -8 & 15 & 0 & 3 \\ -8 & 18 & 8 & 6 & -2 \\ \hline 15 & 8 & 50 & 8 & 5 \\ 0 & 6 & 8 & 4 & 0 \\ 3 & -2 & 8 & 0 & 1 \end{pmatrix}.$$

Assume that $\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix}$ is distributed as $N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- (a) Find $E(\mathbf{y})$ by (4.7).
 (b) Find $\text{cov}(\mathbf{y})$ by (4.8).

4.18 Suppose that $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ is a random sample from a normal multivariate population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. If n is large, what is the approximate distribution of each of the following?

- (a) $\sqrt{n}(\bar{\mathbf{y}} - \boldsymbol{\mu})$
 (b) $\bar{\mathbf{y}}$

4.19 For the rumen time data treated in Example 4.3.2, check each of the five variables for univariate normality using the following techniques:

- (a) Q-Q plots;
 (b) $\sqrt{n} \hat{b}_1$ and b_2 as given by (4.18) and (4.19);
 (c) D'Agostino's test using D and F given in (4.22) and (4.23);
 (d) The test by Lin and Moulkour using z defined in (4.24).

4.20 For the calcium data in Table 5.3, check for multivariate normality and outliers using the following tests:

- (a) Calculate D_i^2 as in (4.27) for each observation.
 (b) Compare the largest value of D_i^2 with the critical value in Table A.6.
 (c) Compute v_1 and v_2 in (4.28) and (4.29) and plot them. Is there an indication of nonlinearity or outliers?
 (d) Calculate $b_{1,p}$ and $b_{2,p}$ in (4.36) and (4.37) and compare them with critical values in Table A.5.

4.21 For the probe word data in Table 3.5, check each of the five variables for univariate normality and outliers using the following tests:

- $\bar{Q}-\bar{Q}$ plot;
- $\chi^2/\bar{Y}$ and b_2 as given by (4.10) and (4.13);
- D'Agostino's test using D and F given in (4.22) and (4.24);
- The test by Lin and Blutholzer using z defined in (4.24).

4.22 For the probe word data in Table 3.5, check for multivariate normality and outliers using the following tests:

- Calculate D_i^2 as in (4.27) for each observation.
- Compare the largest value of D_i^2 with the critical value in Table A.6.
- Compute s_1 and s_2 in (4.28) and (4.29) and plot them. Is there an indication of nonnormality or outliers?
- Calculate $b_{1,p}$ and $b_{2,p}$ in (4.36) and (4.37) and compare them with critical values in Table A.5.

4.23 Six hematology variables were measured on 21 workers (Roxton 1987):

- | | |
|----------------------------------|----------------------------------|
| y_1 = hemoglobin concentration | y_4 = lymphocyte count |
| y_2 = packed cell volume | y_5 = eosinophil count |
| y_3 = white blood-cell count | y_6 = serum lead concentration |

The data are given in Table 4.3. Check each of the six variables for univariate normality using the following tests:

- $\bar{Q}-\bar{Q}$ plot;
- $\chi^2/\bar{Y}$ and b_2 as given by (4.10) and (4.13);
- D'Agostino's test using D and F given in (4.22) and (4.24);
- The test by Lin and Blutholzer using z defined in (4.24).

Table 4.3. Hematology Data

Observation Number	y_1	y_2	y_3	y_4	y_5	y_6
1	13.4	38	4300	14	20	17
2	14.6	48	5000	13	30	20
3	13.5	42	4500	19	21	18
4	15.0	46	4000	15	16	18
5	14.0	44	5100	17	24	19
6	14.0	44	4800	10	26	19
7	16.4	49	4300	11	17	18
8	14.8	44	4400	16	20	20
9	13.2	40	4100	27	13	21
10	13.5	48	5400	24	42	20

(continued)

Table 4.3. (Continued)

Observation Number	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$	$\bar{y}_6$
11	15.1	47	7000	26	27	13
12	16.9	50	8100	28	17	13
13	14.8	44	4700	24	30	13
14	16.2	45	7000	26	25	19
15	14.7	40	4000	23	13	17
16	14.7	42	3400	9	22	13
17	16.5	45	5400	19	32	17
18	15.4	45	6700	26	36	14
19	15.1	45	4000	17	20	17
20	14.5	46	4200	14	28	26
21	13.9	46	5200	8	34	16
22	16.0	47	4700	23	14	16
23	15.4	50	8000	27	19	17
24	14.5	40	5000	20	34	19
25	14.6	44	4200	15	24	19
26	14.9	45	4000	9	32	17
27	15.8	45	5200	16	30	20
28	14.8	43	3700	18	14	23
29	14.4	43	6000	17	17	23
30	14.6	46	4700	23	23	27
31	15.3	45	7100	43	20	23
32	14.9	45	3400	17	15	24
33	15.6	47	6000	23	32	21
34	14.4	44	7700	31	30	19
35	14.7	46	3700	11	26	19
36	14.8	43	6200	23	19	22
37	15.4	45	6000	30	25	16
38	16.2	50	8100	32	16	16
39	15.0	45	4000	17	26	24
40	15.1	47	6000	22	19	16
41	16.0	46	4000	20	22	22
42	15.5	46	5500	20	26	19
43	14.8	41	6200	26	36	21
44	14.5	41	4900	26	20	26
45	15.0	45	7200	40	28	23
46	14.2	46	5800	22	34	22
47	14.9	45	8400	41	17	17
48	16.1	48	3100	12	15	16
49	14.5	45	4000	20	16	26
50	16.4	49	6700	33	22	14
51	14.7	44	7600	18	34	16

- 4.34 For the homanology data in Table 4.1, check for multivariate normality using the following techniques:
- Calculate D_i^2 as in (4.27) for each observation.
 - Compare the largest value of D_i^2 with the critical value in Table A.6 (extrapolate).
 - Compute α_1 and α_2 in (4.28) and (4.29) and plot them. Is there an indication of nonlinearity or outliers?
 - Calculate $b_{1,p}$ and $b_{2,p}$ in (4.36) and (4.37) and compare them with critical values in Table A.8.

Tests on One or Two Mean Vectors

5.1 MULTIVARIATE VERSUS UNIVARIATE TESTS

Hypothesis testing in a multivariate context is more complex than in a univariate setting. The number of parameters may be staggering. The p -variate normal distribution, for example, has p means, p variances, and $\frac{p(p-1)}{2}$ covariances, where $\frac{p(p-1)}{2}$ represents the number of pairs among the p variables. The total number of parameters is

$$p + p + \binom{p}{2} = \frac{1}{2}p(p+3).$$

For $p = 10$, for example, the number of parameters is 65, for each of which, a hypothesis could be formulated. Additionally, we might be interested in testing hypotheses about subsets of these parameters or about functions of them. In some cases, we have the added dilemma of choosing among competing test statistics (see Chapter 6).

We first discuss the motivation for testing p variables multivariately rather than (or in addition to) univariately, as, for example, in hypotheses about $\mu_1, \mu_2, \dots, \mu_p$ in μ . There are at least four arguments for a multivariate approach to hypothesis testing:

- i. The use of p univariate tests inflates the Type I error rate, α , whereas the multivariate test preserves the exact α level. For example, if we do $p = 10$ separate univariate tests at the .05 level, the probability of at least one false rejection is greater than .05. If the variables were independent (they rarely are), we would have (under H_0)

$$\begin{aligned} P(\text{at least one rejection}) &= 1 - P(\text{all 10 tests accept } H_0) \\ &= 1 - (.95)^{10} = .40. \end{aligned}$$

The resulting overall α of $\mathcal{A}$ is not an acceptable error rate. Typically, the 10 variables are correlated, and the overall α would be somewhere between .05 and .40.

3. The univariate tests completely ignore the correlations among the variables, whereas the multivariate tests make direct use of the correlations.
3. The multivariate test is more powerful in many cases. The power of a test is the probability of rejecting M_0 when it is false. In some cases, all p of the univariate tests fail to reach significance, but the multivariate test is significant because small effects on some of the variables combine to jointly indicate significance. However, for a given sample size, there is a limit to the number of variables a multivariate test can handle without losing power. This is discussed further in Section 5.3.2.
4. Many multivariate tests involving means have as a byproduct the construction of a linear combination of variables that reveals more about how the variables combine to reject the hypothesis.

5.2 TESTS ON μ WITH Σ KNOWN

The test on a mean vector assuming a known Σ is introduced to illustrate the issues involved in multivariate testing and to serve as a foundation for the unknown Σ case. We first review the univariate case, in which we work with a single variable y that is distributed as $N(\mu, \sigma^2)$.

5.2.1 Review of Univariate Test for $H_0: \mu = \mu_0$ with σ Known

The hypothesis of interest is that the mean of y is equal to a given value, μ_0 , versus the alternative that it is not equal to μ_0 :

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu \neq \mu_0.$$

We do not consider one-sided alternative hypotheses because they do not readily generalize to multivariate tests. We assume a random sample of n observations $y_1, y_2, \dots, y_n$ from $N(\mu, \sigma^2)$ with σ^2 known. We calculate $\bar{y} = \sum_{i=1}^n y_i/n$ and compare it to μ_0 using the test statistic

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}, \quad (5.1)$$

which is distributed as $N(0, 1)$ if H_0 is true. For $\alpha = .05$, we reject H_0 if $|z| \geq 1.96$. Equivalently, we can use z^2 , which is distributed as χ^2 with one degree of freedom, and reject H_0 if $z^2 \geq (1.96)^2 = 3.84$. If n is large, we are assured by the central limit theorem that $\bar{y}$ is approximately normal, even if the observations are not from a normal distribution.

5.2.1 Multivariate Test for $H_0: \mu = \mu_0$ with Σ Known

In the multivariate case we have several variables measured on each sampling unit, and we wish to hypothesize a value for the means of each variable, $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$. More explicitly, we have

$$H_0: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}, \quad H_1: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \neq \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix},$$

where each μ_{0j} is specified from previous experience or is a target value. The vector equality in H_0 implies $\mu_j = \mu_{0j}$ for all $j = 1, 2, \dots, p$. The vector inequality in H_1 implies at least one $\mu_j \neq \mu_{0j}$. Thus, for example, if $\mu_j = \mu_{0j}$ for all j except 2, for which $\mu_2 \neq \mu_{02}$, then we wish to reject H_0 .

To test H_0 , we use a random sample of n observation vectors $y_1, y_2, \dots, y_n$ from $N_p(\mu, \Sigma)$, with Σ known, and calculate $\bar{y} = \sum_{i=1}^n y_i/n$. The test statistic is

$$Z^2 = n(\bar{y} - \mu_0)' \Sigma^{-1} (\bar{y} - \mu_0). \quad (5.2)$$

If H_0 is true, Z^2 is distributed as χ^2_p by (4.6), and we therefore reject H_0 if $Z^2 > \chi^2_{p, \alpha}$. Note that for one variable, Z^2 (the square of (3.1)) has a chi-square distribution with 1 degree of freedom, whereas, for p variables, Z^2 in (5.2) is distributed as a chi-square with p degrees of freedom.

If Σ is unknown, we could use $\hat{\Sigma}$ in its place in (5.2), and Z^2 would have an approximate χ^2 distribution. But n would have to be larger than in the analogous univariate situation, in which $t = (\bar{y} - \mu_0)/(s/\sqrt{n})$ is approximately $N(0, 1)$ for $n > 30$. The value of n needed for $n(\bar{y} - \mu_0)' \hat{\Sigma}^{-1} (\bar{y} - \mu_0)$ to have an approximate χ^2 distribution depends on p . This is clarified further in Section 5.3.2.

Example 5.2.1. In Table 3.1, height and weight were given for a sample of 30 college-age males. Let us assume that this sample originated from the bivariate normal $N_2(\mu, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}.$$

Suppose we wish to test $H_0: \mu = (70, 170)'$. From Example 3.2.1, $\bar{y}_1 = 71.45$ and $\bar{y}_2 = 164.7$. We thus have

$$\begin{aligned} Z^2 &= n(\bar{y} - \mu_0)' \Sigma^{-1} (\bar{y} - \mu_0) \\ &= (20) \begin{pmatrix} 71.45 - 70 \\ 164.7 - 170 \end{pmatrix}' \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}^{-1} \begin{pmatrix} 71.45 - 70 \\ 164.7 - 170 \end{pmatrix} \\ &= (20)(1.45, -5.3) \begin{pmatrix} .1 & -.01 \\ -.01 & .002 \end{pmatrix} \begin{pmatrix} 1.45 \\ -5.3 \end{pmatrix} = 8.4028. \end{aligned}$$

Using $\alpha = .05$, $z_{.025}^2 = 5.99$, and we therefore reject $H_0: \mu = (70, 170)'$ because $Z^2 = 8.4026 > 5.99$.

The rejection region for $\bar{Y} = (Y_1, Y_2)'$ is on or outside the ellipse in Figure 5.1, that is, the test statistic Z^2 is greater than 5.99 if and only if $\bar{Y}$ is outside the ellipse. If $\bar{Y}$ falls inside the ellipse, H_0 is accepted. Thus, distance from μ_0 as well as direction must be taken into account. When the distance is standardized by $\Sigma^{-1/2}$, all points on the curve are "statistically equidistant" from the center.

Note that the test is sensitive to the covariance structure. If $\text{cov}(Y_1, Y_2)$ were negative, Y_2 would tend to decrease as Y_1 increases, and the ellipse would be tilted in the other direction. In this case, $\bar{Y}$ would be in the acceptance region.

Let us now investigate the consequences of testing each variable separately. Using $z_{.025} = 1.96$ for $\alpha = .05$, we have

$$z_1 = \frac{Y_1 - \mu_{10}}{\sigma_{Y_1}/\sqrt{n}} = 1.450 < 1.96,$$

$$z_2 = \frac{Y_2 - \mu_{20}}{\sigma_{Y_2}/\sqrt{n}} = -.7489 > -1.96.$$

Thus both tests accept the hypothesis. In this case neither of the Y 's is far enough from the hypothesized value to cause rejection. But when the positive correlation between Y_1 and Y_2 is taken into account in the multivariate test, the two evidences

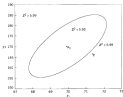


Figure 5.1. Elliptical acceptance region.

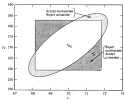


Figure 5.2. Acceptance and rejection regions for univariate and multivariate tests.

against μ_0 continue to cause rejection. This illustrates the third advantage of multivariate tests given in Section 5.1.

Figure 5.2 shows the rectangular acceptance region for the univariate tests superimposed on the elliptical multivariate acceptance region. The rectangle was obtained by calculating the two acceptance regions

$$\mu_{11} - 1.96 \frac{\sigma_1}{\sqrt{n}} < \bar{X}_1 < \mu_{11} + 1.96 \frac{\sigma_1}{\sqrt{n}},$$

$$\mu_{22} - 1.96 \frac{\sigma_2}{\sqrt{n}} < \bar{X}_2 < \mu_{22} + 1.96 \frac{\sigma_2}{\sqrt{n}}.$$

Points inside the ellipse but outside the rectangle will be rejected in at least one univariate dimension but will be accepted multivariately. This illustrates the inflation of α resulting from univariate tests, as discussed in the first motive for multivariate testing in Section 5.1. This phenomenon has been referred to as *Rao's paradox*. For further discussion see Rao (1960), Hooley (1987), and Morrison (1990, p. 174). Points outside the ellipse but inside the rectangle will be rejected univariately but accepted multivariately in both dimensions. This illustrates the third motive for multivariate testing given in Section 5.1, namely, that the multivariate test is more powerful in some situations.

Thus in either case represented by the shaded areas, we should use the multivariate test result R , not the univariate results. In the one case, the multivariate test is more powerful than the univariate tests; in the other case, the multivariate test preserves α

whereas the variariate tests inflate α . Consequently, when the multivariate and variariate results disagree, our tendency is to trust the multivariate result. In Section 5.5, we discuss various procedures for ascertaining the contribution of the individual variables after the multivariate test has rejected the hypothesis. $\square$

5.2 TESTS ON μ WHEN Σ IS UNKNOWN

In Section 5.1, we said little about properties of the tests, because the tests discussed were of slight practical consequence due to the assumption that Σ is known. We will be more concerned with test properties in Sections 5.3 and 5.4, first in the one-sample case and then in the two-sample case. The reader may wonder why we include one-sample tests, since we seldom, if ever, have need of a test for $H_0: \mu = \mu_0$. However, we will return this case for two reasons:

1. Many general principles are more easily illustrated in the one-sample framework than in the two-sample case.
2. Some very useful tests can be cast in the one-sample framework. Two examples are (1) $H_0: \mu_{\mathcal{A}} = 0$ used in the paired-comparison test covered in Section 5.7 and (2) $H_0: C_{\mathcal{A}} = 0$ used in profile analysis in Section 5.8, in analysis of repeated measures in Section 5.9, and in growth curves in Section 5.10.

5.2.1 Review of Univariate t -Test for $H_0: \mu = \mu_0$ with Σ Unknown

We first review the familiar one-sample t -test in the univariate case, with only one variable measured on each sampling unit. We assume that a random sample $y_1, y_2, \dots, y_n$ is available from $N(\mu, \sigma^2)$. We estimate μ by $\bar{y}$ and σ^2 by s^2 , where $\bar{y}$ and s^2 are given by (3.1) and (3.6). To test $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$, we use

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{y} - \mu_0)}{s}, \quad (5.2)$$

If H_0 is true, t is distributed as t_{n-1} , where $n-1$ is the degree of freedom. We reject H_0 if $|t| > \sqrt{n}(\bar{y} - \mu_0)/s \geq t_{\alpha/2, n-1}$, where $t_{\alpha/2, n-1}$ is a critical value from the t -table.

The first expression in (5.2), $t = (\bar{y} - \mu_0)/s/\sqrt{n}$, is the characteristic form of the t -statistic, which represents a sample standardized distance between $\bar{y}$ and μ_0 . In this form, the hypothesized mean is subtracted from $\bar{y}$ and the difference is divided by $s/\sqrt{n}$. Since $y_1, y_2, \dots, y_n$ is a random sample from $N(\mu, \sigma^2)$, the random variables $\bar{y}$ and s are independent. We will use an analogous characteristic form for the F -statistic in the multivariate case in Section 5.3.2.

5.3.1 Hotelling's T^2 -Test for $H_0: \mu = \mu_0$ with Σ Unknown

We now move to the multivariate case in which p variables are measured on each sampling unit. We assume that a random sample $y_1, y_2, \dots, y_n$ is available from $N_p(\mu, \Sigma)$, where y_i contain the p measurements on the i th sampling unit (subject

or object). We estimate μ by $\bar{y}$ and Σ by S , where $\bar{y}$ and S are given by (3.18), (3.19), (3.22), (3.23), and (3.24). In order to test $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$, we use an extension of the univariate t -statistic in (3.2). In expanded form, the univariate t can be rewritten as

$$t^2 = \frac{n(\bar{y} - \mu_0)^2}{s^2} = n(\bar{y} - \mu_0)(s^2)^{-1}(\bar{y} - \mu_0). \quad (3.4)$$

When $\bar{y} - \mu_0$ and s^2 are replaced by $\bar{y} - \mu_0$ and S , we obtain the test statistic

$$T^2 = n(\bar{y} - \mu_0)'S^{-1}(\bar{y} - \mu_0). \quad (3.5)$$

Alternatively, T^2 can be obtained from Z^2 in (3.2) by replacing Σ with S .

The distribution of T^2 was obtained by Hotelling (1931), assuming M_0 is true and sampling is from $N_p(\mu, \Sigma)$. The distribution is indexed by two parameters, the dimension p and the degrees of freedom $v = n - 1$. We reject M_0 if $T^2 > T_{\alpha, p, v}^2$ and accept M_0 otherwise. Critical values of the T^2 -distribution are found in Table A.7, taken from Raman and Iversen (1989a).

Note that the terminology "accept M_0 " is used for expediency convenience to describe our decision when we do not reject the hypothesis. Strictly speaking, we do not accept M_0 in the sense of actually believing it is true. If the sample data were extremely large and we accepted M_0 , we could be reasonably certain that the true μ is close to the hypothesized value μ_0 . Otherwise, accepting M_0 means only that we have failed to reject M_0 .

The T^2 -statistic can be viewed as the sample standardized distance between the observed sample mean vector and the hypothesized mean vector. If the sample mean vector is notably distant from the hypothesized mean vector, we become suspicious of the hypothesized mean vector and wish to reject M_0 .

The test statistic is a scalar quantity, since $T^2 = n(\bar{y} - \mu_0)'S^{-1}(\bar{y} - \mu_0)$ is a quadratic form. As with the χ^2 -distribution of Z^2 , the density of T^2 is skewed because the lower limit is zero and there is no upper limit.

The characteristic form of the T^2 -statistic (3.5) is

$$T^2 = (S - \mu_0)' \left(\frac{S}{n} \right)^{-1} (S - \mu_0). \quad (3.6)$$

The characteristic form has two features:

1. S is the sample covariance matrix of $\bar{y}$ and serves as a standardizing matrix in the distance function.
2. Since $y_1, y_2, \dots, y_n$ are distributed as $N_p(\mu, \Sigma)$, it follows that $\bar{y}$ is $N_p(\mu, \frac{1}{n}\Sigma)$, $(n-1)S$ is $W(n-1, \Sigma)$, and $\bar{y}$ and S are independent (see Section 4.3.2).

In (5.3), the univariate t -statistic represents the number of standard deviations $\bar{y}$ is separated from μ_0 . In appearance, the F^2 -statistic (5.6) is similar, but no such simple interpretation is possible. If we add a variable, the distance in (5.6) increases. (By analogy, the hypotenuse of a right triangle is longer than either of the legs.) Thus we need a test statistic that indicates the significance of the distance from $\bar{y}$ to μ_0 , while allowing for the number of dimensions (see comment 1 at the end of this section about the F^2 -table). Since the resulting F^2 -statistic cannot be readily interpreted in terms of the number of standard deviations $\bar{y}$ is from μ_0 , we do not have an intuitive feel for its significance as we do with the univariate t . We must compare the calculated value of F^2 with the table value. In addition, the F^2 -table provides some insights into the behavior of the F^2 -distribution. Four of these insights are noted at the end of this section.

If a test leads to rejection of $H_0: \mu = \mu_0$, the question arises as to which variable or variables contributed most to the rejection. This issue is discussed in Section 5.5 for the two-sample F^2 -test of $H_0: \mu_1 = \mu_2$, and the results there can be easily adapted to the one-sample test of $H_0: \mu = \mu_0$. For confidence intervals on the individual μ_j 's in μ , see Raftery (1978, Section 3.4).

The following are some key properties of the F^2 -test:

1. We must have $n - 1 = p$. Otherwise, S is singular and F^2 cannot be computed.
2. In both the one-sample and two-sample cases, the degrees of freedom for the F^2 -statistic will be the same as for the analogous univariate t -test, that is, $v = n - 1$ for one sample and $v = n_1 + n_2 - 2$ for two samples (see Section 3.4.2).
3. The alternative hypothesis is two-sided. Because the space is multidimensional, we do not consider one-sided alternative hypotheses, such as $\mu > \mu_0$. However, even though the alternative hypothesis $H_1: \mu \neq \mu_0$ is essentially two-sided, the critical region is one-tailed (we reject H_0 for large values of F^2). This is typical of many multivariate tests.
4. In the univariate case, $t_{n-1}^2 = F_{1,n-1}$. The statistic F^2 can also be converted to an F -statistic as follows:

$$\frac{v - p + 1}{vp} T_{p,v}^2 = F_{p,v-p+1}. \quad (5.7)$$

Note that the dimension p (number of variables) of the F^2 -statistic becomes the first of the two degrees-of-freedom parameters of the F . The number of degrees of freedom for F^2 is denoted by v , and the F transformation is given in terms of a general v , since other applications of F^2 will have v different from $n - 1$ (see, for example, Sections 5.4.1 and 5.1.2).

Equation (5.7) gives an easy way to find critical values for the F^2 -test. However, we have provided critical values of F^2 in Table A.7 because of the insights they provide into the behavior of the F^2 -distribution in particular and multivariate tests in general. The following are some insights that can readily be gleaned from the F^2 -table:

1. The first column of Table A.7 contains squares of t -table values; that is, $T_{\alpha/2, n-1}^2 = t_{\alpha/2, n-1}^2$. (We use $t_{\alpha/2}^2$ because the univariate test of $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$ is two-tailed.) Thus for $p = 1$, T^2 reduces to t^2 . This can easily be seen by comparing (5.2) with (5.4).
2. The last row of each page of Table A.7 contains χ^2 critical values, that is, $T_{p, \infty}^2 = \chi_p^2$. Thus as n increases, S approaches $\mathbf{I}$, and

$$T^2 = n(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{Y}} - \boldsymbol{\mu}_0)$$

approaches $\chi^2 = n(\bar{\mathbf{Y}} - \boldsymbol{\mu}_0)' \mathbf{I}^{-1} (\bar{\mathbf{Y}} - \boldsymbol{\mu}_0)$ in (5.2), which is distributed as χ_p^2 .

3. The values increase along each row of Table A.7; that is, for a fixed n , the critical value $T_{\alpha/2, n}^2$ increases with p . It was noted above that in any given sample, the calculated value of T^2 increases if a variable is added. However, since the critical value also increases, a variable should not be added unless it adds a significant amount to T^2 .
4. As p increases, large values of n are required for the distribution of T^2 to approach χ^2 . In the univariate case, t in (5.3) is considered a good approximation to the standard normal z in (5.1) when $n - 1$ is at least 30. In the first column ($p = 1$) of Table A.7, we see $T_{.05, 1, 30}^2 = 4.171$ and $T_{.05, 1, \infty}^2 = 3.841$, with a ratio of $4.171/3.841 = 1.086$. For $p = 5$, n must be 100 to obtain the same ratio: $T_{.05, 5, 100}^2/T_{.05, 5, \infty}^2 = 1.086$. For $p = 10$, we need $n = 200$ to obtain a similar value of the ratio: $T_{.05, 10, 200}^2/T_{.05, 10, \infty}^2 = 1.076$. Thus one must be very cautious in stating that T^2 has an approximate χ^2 distribution for large n . The α level (Type I error rate) could be substantially inflated. For example, suppose $p = 10$ and we assume that $n = 30$ is sufficiently large for a χ^2 approximation to hold. Then we would reject H_0 for $T^2 \geq 39.387$ with a target α -level of .05. However, the correct critical value is 50.644, and the chance of 18,300 would yield an actual α of $P(T_{18, 30}^2 \geq 39.387) = .314$.

Example 5.3.1. In Table 3.3 we have $n = 18$ observations on $p = 5$ variables. Desirable levels for η_1 and η_2 are 13.0 and 6.8, respectively, and the expected level of η_3 is 2.85. We can, therefore, test the hypothesis

$$H_0: \boldsymbol{\mu} = \begin{pmatrix} 15.0 \\ 6.8 \\ 2.85 \end{pmatrix}.$$

In Example 3.5 and 3.6, $\bar{\mathbf{y}}$ and $\mathbf{S}$ were obtained as

$$\bar{\mathbf{y}} = \begin{pmatrix} 28.1 \\ 7.18 \\ 3.09 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 145.34 & 49.68 & 1.84 \\ 49.68 & 31.25 & 3.68 \\ 1.84 & 3.68 & .25 \end{pmatrix}.$$

To test H_0 , we use (3.5):

$$\begin{aligned} T^2 &= n\bar{y} - \mu_0'S^{-1}(\bar{y} - \mu_0) \\ &= n \begin{pmatrix} 28.1 & - & 15.0 \\ 7.19 & - & 6.0 \\ 3.09 & - & 1.85 \end{pmatrix}' \begin{pmatrix} 140.94 & -49.68 & 1.94 \\ 49.68 & 31.25 & 1.68 \\ 1.94 & 1.68 & .25 \end{pmatrix}^{-1} \begin{pmatrix} 28.1 & - & 15.0 \\ 7.19 & - & 6.0 \\ 3.09 & - & 1.85 \end{pmatrix} \\ &= 34.551. \end{aligned}$$

From Table A.1, we obtain the critical value $F_{10,4,95}^* = 16.766$. Since the observed value of T^2 exceeds the critical value, we reject the hypothesis. $\square$

3.4 COMPARING TWO MEAN VECTORS

We first review the univariate two-sample t -test and then proceed with the analogous multivariate test.

3.4.1 Review of Univariate Two-Sample t -Test

In the univariate case we obtain a random sample $y_{11}, y_{12}, \dots, y_{1n_1}$ from $N(\mu_1, \sigma^2)$ and a second random sample $y_{21}, y_{22}, \dots, y_{2n_2}$ from $N(\mu_2, \sigma^2)$. We assume that the two samples are independent and that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, σ_1, σ_2 , with σ^2 unknown. [The assumptions of independence and equal variances are necessary in order for the t -statistic in (3.8) to have a t -distribution.] From the two samples we calculate $\bar{y}_1, \bar{y}_2$, $SS_1 = \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 = (n_1 - 1)s_1^2$, $SS_2 = \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2 = (n_2 - 1)s_2^2$, and the pooled variance

$$s_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

where $n_1 + n_2 - 2$ is the sum of the weights $n_1 - 1$ and $n_2 - 1$ in the numerator. With this observation, s_p^2 is an unbiased estimator for the common variance, σ^2 . That is, $E(s_p^2) = \sigma^2$.
To test

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2,$$

we use

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad (3.8)$$

which has a t -distribution with $n_1 + n_2 - 2$ degrees of freedom when Δ_0 is true. We therefore reject H_0 if $|T| \geq t_{\alpha/2, n_1+n_2-2}$.

Note that (3.6) exhibits the characteristic form of a t -statistic. In this form, the denominator is the sample standard deviation of the numerator, that is,

$$s_{\text{pool}} \sqrt{1/n_1 + 1/n_2}$$

is an estimate of

$$\begin{aligned} \sigma_{Y_1 - Y_2} &= \sigma \sqrt{\text{Var}(Y_1 - Y_2)} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \\ &= \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_1}} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_1}}. \end{aligned}$$

3.4.2 Multivariate Two-Sample T^2 -Test

We now consider the case where p variables are measured on each sampling unit in two samples. We wish to test

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2.$$

We obtain a random sample $y_{11}, y_{12}, \dots, y_{1n_1}$ from $N_p(\mu_1, \Sigma_1)$ and a second random sample $y_{21}, y_{22}, \dots, y_{2n_2}$ from $N_p(\mu_2, \Sigma_2)$. We assume that the two samples are independent and that $\Sigma_1 = \Sigma_2 = \Sigma$, say, with Σ unknown. These assumptions are necessary in order for the T^2 -statistic in (3.6) to have a T^2 -distribution. A test of $H_0: \Sigma_1 = \Sigma_2$ is given in Section 3.5.2. For an approximate test of $H_0: \mu_1 = \mu_2$ that can be used when $\Sigma_1 \neq \Sigma_2$, see Boncher (1998, Section 3.5).

The sample mean vectors are $\bar{Y}_1 = \sum_{i=1}^{n_1} y_{1i}/n_1$ and $\bar{Y}_2 = \sum_{i=1}^{n_2} y_{2i}/n_2$. Define $\mathbf{W}_1$ and $\mathbf{W}_2$ to be the matrices of sums of squares and cross products for the two samples:

$$\mathbf{W}_1 = \sum_{i=1}^{n_1} (y_{1i} - \bar{Y}_1)(y_{1i} - \bar{Y}_1)' = (n_1 - 1)\mathbf{S}_1,$$

$$\mathbf{W}_2 = \sum_{i=1}^{n_2} (y_{2i} - \bar{Y}_2)(y_{2i} - \bar{Y}_2)' = (n_2 - 1)\mathbf{S}_2.$$

Since $(n_1 - 1)\mathbf{S}_1$ is an unbiased estimator of $(n_1 - 1)\Sigma$ and $(n_2 - 1)\mathbf{S}_2$ is an unbiased estimator of $(n_2 - 1)\Sigma$, we can pool them to obtain an unbiased estimator of the common population covariance matrix, Σ :

$$\hat{\mathbf{S}}_{\text{pool}} = \frac{1}{n_1 + n_2 - 2} (\mathbf{W}_1 + \mathbf{W}_2)$$

$$= \frac{1}{n_1 + n_2 - 2} (n_1 - 1)S_1 + (n_2 - 1)S_2.$$

Thus $E(S_0) = \Sigma$.

The square of the univariate t -statistic (5.8) can be expressed as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{Y}_1 - \bar{Y}_2) (\hat{\sigma}_0^2)^{-1} (\bar{Y}_1 - \bar{Y}_2).$$

This can be generalized to p variables by substituting $\bar{Y}_1 - \bar{Y}_2$ for $\bar{Y}_1 - \bar{Y}_2$ and $\hat{\sigma}_0^2$ for $\hat{\sigma}_0^2$ to obtain

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{Y}_1 - \bar{Y}_2) (S_0^{-1}) (\bar{Y}_1 - \bar{Y}_2), \quad (5.9)$$

which is distributed as $F_{p, n_1+n_2-2}^2$ when $H_0: \mu_1 = \mu_2$ is true. To carry out the test, we collect the two samples, calculate T^2 by (5.9), and reject H_0 if $T^2 \geq F_{p, n_1+n_2-2}^2$. Critical values of T^2 are listed in Table A.7. For tables of the power of the T^2 -test (probability of rejecting H_0 when it is false) and illustrations of their use, see Raudenbush (1998, Section 3.10).

The T^2 -statistic (5.9) can be expressed in characteristic form as the standardized distance between $\bar{Y}_1$ and $\bar{Y}_2$:

$$T^2 = (\bar{Y}_1 - \bar{Y}_2)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_0 \right]^{-1} (\bar{Y}_1 - \bar{Y}_2), \quad (5.10)$$

where $(1/n_1 + 1/n_2)S_0$ is the sample covariance matrix for $\bar{Y}_1 - \bar{Y}_2$ and S_0 is independent of $\bar{Y}_1 - \bar{Y}_2$ because of sampling from the multivariate normal. For a discussion of robustness of T^2 to departure from the assumptions of multivariate normality and homogeneity of covariance matrices ($\Sigma_1 = \Sigma_2$), see Raudenbush (1998, Section 3.7).

Some key properties of the two-sample T^2 -test are given in the following list:

1. It is necessary that $n_1 + n_2 - 2 > p$ for S_0 to be nonsingular.
2. The statistic T^2 is, of course, a scalar. The $(p + p(p - 1)/2)$ quantities in $\bar{Y}_1$, $\bar{Y}_2$, and S_0 have been reduced to a single scalar on which T^2 is large if the sample evidence favors $H_1: \mu_1 \neq \mu_2$ and small if the evidence supports $H_0: \mu_1 = \mu_2$; we reject H_0 if the standardized distance between $\bar{Y}_1$ and $\bar{Y}_2$ is large.
3. Since the lower limit of T^2 is zero and there is no upper limit, the density is skewed. In fact, as noted in (5.11), T^2 is directly related to F , which is a well-known skewed distribution.
4. For degrees of freedom of T^2 we have $n_1 + n_2 - 2$, which is the same as for the corresponding univariate t -statistic (5.8).

5. The alternative hypothesis $H_1: \mu_1 \neq \mu_2$ is two-sided. The critical region $F^T \leq F_\alpha^L$ is one-tailed, however, as is typical of many multivariate tests.
6. The F^L -statistic can be readily transformed to an F -statistic using (3.11):

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} F^L = F_{p, n_1 + n_2 - p - 1}, \quad (3.11)$$

where again the dimension p of the F^L -statistic becomes the first degree-of-freedom parameter for the F -statistic.

Example 3.4.1. Four psychological tests were given to 32 men and 32 women. The data are recorded in Table 5.1 (Bent 1962). The variables are

$$\begin{array}{ll} y_1 = \text{picture inconsistency} & y_2 = \text{word recognition} \\ y_3 = \text{paper form board} & y_4 = \text{vocabulary} \end{array}$$

The mean vectors and covariance matrices of the two samples are

$$\begin{aligned} \bar{\mathbf{y}}_1 &= \begin{pmatrix} 15.97 \\ 15.94 \\ 21.19 \\ 21.75 \end{pmatrix}, & \bar{\mathbf{y}}_2 &= \begin{pmatrix} 12.34 \\ 15.81 \\ 16.66 \\ 21.34 \end{pmatrix}, \\ \mathbf{S}_1 &= \begin{pmatrix} 5.192 & 4.545 & 6.522 & 5.250 \\ 4.848 & 15.35 & 8.798 & 6.266 \\ 6.522 & 6.765 & 28.67 & 14.47 \\ 5.250 & 6.266 & 14.47 & 16.65 \end{pmatrix}, \\ \mathbf{S}_2 &= \begin{pmatrix} 5.158 & 7.549 & 4.884 & 4.191 \\ 7.549 & 18.60 & 10.22 & 5.446 \\ 4.884 & 10.22 & 26.04 & 13.49 \\ 4.191 & 5.446 & 13.49 & 28.00 \end{pmatrix}. \end{aligned}$$

The sample covariance matrices do not appear to indicate a disparity in the population covariance matrices. (A significance test to check this assumption is carried out in Example 3.4.2, and the hypothesis $H_0: \Sigma_1 = \Sigma_2$ is not rejected.) The pooled covariance matrix is

$$\begin{aligned} \mathbf{S}_{12} &= \frac{1}{32 + 32 - 2} [(32 - 1)\mathbf{S}_1 + (32 - 1)\mathbf{S}_2] \\ &= \begin{pmatrix} 5.164 & 6.047 & 5.693 & 4.761 \\ 6.047 & 15.89 & 8.492 & 5.856 \\ 5.693 & 6.492 & 26.36 & 13.88 \\ 4.761 & 5.856 & 13.98 & 28.11 \end{pmatrix}. \end{aligned}$$

Table A.1. Four Psychological Test Scores on 32 Males and 32 Females

Males				Females			
x_{1i}	x_{2i}	y_{1i}	y_{2i}	x_{1i}	x_{2i}	y_{1i}	y_{2i}
12	17	24	14	13	14	12	20
17	15	31	26	14	12	14	26
15	14	29	23	15	19	21	20
13	12	30	18	13	17	10	18
20	17	35	28	11	20	16	18
18	20	26	21	12	8	14	18
12	13	26	22	10	12	18	24
12	9	22	22	10	8	12	22
14	7	30	17	12	20	19	22
17	15	30	27	11	18	11	27
17	17	26	28	12	18	25	27
17	20	26	24	14	18	15	26
18	18	26	24	14	18	28	28
18	18	22	28	13	18	8	14
18	18	21	27	14	8	13	23
13	14	26	21	13	16	23	28
18	17	23	26	10	21	26	26
10	14	19	17	14	17	14	14
18	20	30	29	10	16	15	23
18	20	34	26	13	18	25	24
13	17	30	24	3	8	16	20
16	16	36	16	14	16	22	26
11	12	23	23	17	17	22	26
10	12	26	16	16	12	16	14
10	12	23	21	15	14	20	26
18	18	34	24	12	18	12	9
10	12	26	27	14	17	24	23
13	16	29	24	13	17	16	26
18	18	22	23	11	16	18	26
18	18	23	23	7	7	19	18
17	20	21	21	12	12	7	26
19	18	30	28	6	5	6	12

By (3.8), we obtain

$$T^2 = \frac{\hat{\sigma}_1 \hat{\sigma}_2}{\hat{\sigma}_1 + \hat{\sigma}_2} (\hat{\theta}_1 - \hat{T}_2 / \hat{\sigma}_p^{-1} \hat{\theta}_1 - \hat{T}_2)^2 = 67.6045.$$

From interpolations in Table A.1, we obtain $F_{20,4,0.05}^2 = 13.379$, and we therefore reject $H_0: \mu_1 = \mu_2$. (See Example 5.3 for a discussion of which variables contribute most to separation of the two groups.) $\square$

5.4.3 Likelihood Ratio Tests

The maximum likelihood approach to estimation was introduced in Section 4.3.1. As noted there, the likelihood function is the joint density of $y_1, y_2, \dots, y_p$. The values of the parameters that maximize the likelihood function are the maximum likelihood estimators.

The likelihood ratio method of test construction uses the ratio of the maximum value of the likelihood function assuming H_0 is true to the maximum under H_1 , which is essentially unrestricted. Likelihood ratio tests usually have good power and sometimes have optimum power over a wide class of alternatives.

When applied to multivariate normal samples and $H_0: \mu_1 = \mu_2$, the likelihood ratio approach leads directly to Hotelling's T^2 -test in (5.9). Similarly, in the one-sample case, the T^2 -statistic in (5.5) is the likelihood ratio test. Thus the F^2 -test, which we introduced rather informally, is the best test according to certain criteria.

5.5 TESTS ON INDIVIDUAL VARIABLES CONDITIONAL ON REJECTION OF H_0 BY THE F^2 -TEST

If the hypothesis $H_0: \mu_1 = \mu_2$ is rejected, the implication is that $\mu_{1j} \neq \mu_{2j}$ for at least one $j = 1, 2, \dots, p$. But there is no guarantee that $H_0: \mu_{1j} = \mu_{2j}$ will be rejected for some j by a suitable test. However, if we consider a linear combination of the variables, $z = a'y$, then there is at least one coefficient vector a for which

$$t(a) = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{[(1/a_1 + 1/a_2)a']}} \quad (5.12)$$

will reject the corresponding hypothesis $H_0: \mu_{1j} = \mu_{2j}$ or $H_0: a'\mu_1 = a'\mu_2$. By (5.54), $\bar{y}_1 = a'\bar{y}_1$ and $\bar{y}_2 = a'\bar{y}_2$, and from (5.55) the variance estimator s_a^2 is the pooled estimator of Σ_{pooled} . Thus (5.12) can be written as

$$t(a) = \frac{a'\bar{y}_1 - a'\bar{y}_2}{\sqrt{[(1/a_1 + 1/a_2)a'\Sigma_{\text{pooled}}a]}} \quad (5.13)$$

Since $t(a)$ can be negative, we work with $t^2(a)$. The linear function $z = a'y$ is a projection of y onto a line through the origin. We seek the line (direction) on which the difference $\bar{y}_1 - \bar{y}_2$ is maximized when projected. The projected-difference $a'(\bar{y}_1 - \bar{y}_2)$ (standardized by $a'\Sigma_{\text{pooled}}a$ as in (5.13)) will be less in any other direction than that parallel to the line joining $\bar{y}_1$ and $\bar{y}_2$. The value of a that projects onto this line, or, equivalently, maximizes $t^2(a)$ in (5.12), is (any multiple of)

$$a = \Sigma_{\text{pooled}}^{-1}(\bar{y}_1 - \bar{y}_2). \quad (5.14)$$

Since a in (5.14) projects $\bar{y}_1 - \bar{y}_2$ onto a line parallel to the line joining $\bar{y}_1$ and $\bar{y}_2$, we would expect that $t^2(a) = F^2$, and this is indeed the case (see Problem 5.3).

When $\mathbf{a} = \mathbf{S}_d^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$, then $z = \mathbf{a}'\mathbf{y}$ is called the *discriminant function*. Sometimes the vector $\mathbf{a}$ itself in (5.14) is loosely referred to as the discriminant function.

If $H_0: \mu_1 = \mu_2$ is rejected by T^2 in (5.9), the discriminant function $\mathbf{a}'\mathbf{y}$ will lead to rejection of $H_0: \mu_1 = \mu_2$ using (5.15), with $\mathbf{a} = \mathbf{S}_d^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$. We can then examine each x_j in $\mathbf{a}$ for an indication of the contribution of the corresponding y_j in rejection of H_0 . This follow-up examination of each x_j should be done only if $H_0: \mu_1 = \mu_2$ is rejected by T^2 . The discriminant function will appear again in Section 5.6.2 and in Chapters 8 and 9.

We list those and other procedures that could be used to check each variable following rejection of H_0 by a two-sample T^2 -test:

1. Univariate t -tests, one for each variable,

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{[s_{11} + s_{22}]/(n_1 + n_2)}\sqrt{v_{jj}}}, \quad j = 1, 2, \dots, p, \quad (5.17)$$

where v_{jj} is the j th diagonal element of $\mathbf{S}_d$. Reject $H_0: \mu_{1j} = \mu_{2j}$ if $|t_j| = t_{\alpha/2, n_1+n_2-2}$. For confidence intervals on $\mu_{1j} - \mu_{2j}$, see Rencher (1995, Section 5.6).

2. To adjust the α -level resulting from performing the p tests in (5.15), we could use a Bonferroni critical value $t_{\alpha/(2p), n_1+n_2-2}$ for (5.15) (Bonferroni 1936). A critical value $t_{\alpha/(2p)}$ is much greater than the corresponding $t_{\alpha/2}$, and the resulting overall α -level is conservative. Bonferroni critical values $t_{\alpha/(2p), n}$ are given in Table A.8, from Bailey (1977).
3. Another critical value that could be used with (5.15) is $\bar{F}_{\alpha, p, n_1+n_2-2}$, where $\bar{F}_{\alpha}$ is the square root of F_{α}^2 from Table A.7; that is, $\bar{F}_{\alpha, p, n_1+n_2-2} = \sqrt{F_{\alpha, p, n_1+n_2-2}^2}$. This allows for all p variables to be tested as well as all possible linear combinations, as in (5.17), even linear combinations chosen after seeing the data. Consequently, the use of $\bar{F}_{\alpha}$ is even more conservative than using $t_{\alpha/(2p)}$, that is, $\bar{F}_{\alpha, p, n_1+n_2-2} > t_{\alpha/(2p), n_1+n_2-2}$.
4. Partial F - or t -tests (just of each variable adjusted for the other variables; see (5.32) in Section 5.4)
5. Standardized discriminant function coefficients (see Section 5.5)
6. Correlations between the variables and the discriminant function (see Section 5.5.4)
7. Stepwise discriminant analysis (see Section 5.5)

The first three methods are univariate approaches that do not use correlations or covariances among the variables in the computation of the test statistic. The last four methods are multivariate in the sense that the correlation structure is explicitly taken into account in the computation.

Method 6, involving the correlation between each variable and the discriminant function, is recommended in many texts and software packages. However, Rencher

(1988) has shown that these correlations are proportional to individual t - or F -tests (see Section 8.1.4). Thus, this method is equivalent to method 1 and is a univariate rather than a multivariate approach. Method 7 is often used to identify a subset of important variables or even to rank the variables according to order of entry. But Bender and Larssen (1988) have shown that stepwise methods have a high risk of selecting spurious variables, unless the sample size is very large.

We now consider the univariate procedures 1, 2, and 3. The probability of rejecting one or more of the p univariate tests when H_0 is true is called the overall α or experimentwise error rate. If we do univariate tests only, with no F^2 -test, then the tests based on $t_{n-1,p}$ and T_p in procedures 2 and 3 are conservative (overall α too low), and tests based on $t_{n-1,2}$ in procedure 1 are liberal (overall α too high). However, when these tests are carried out only after rejection by the F^2 -test (such tests are sometimes called protected tests), the experimentwise error rates change. Obviously the tests will reject less often under H_0 if they are carried out only if F^2 rejects. Thus the tests using $t_{n-1,2}$ and T_p become even more conservative, and the test using $t_{n-1,p}$ becomes more acceptable.

Hummel and Sligo (1971) studied the experimentwise error rate for univariate t -tests following rejection of H_0 by the F^2 -test (protected tests). Using $\alpha = .05$, they found that using $t_{n-1,2}$ for a critical value yields an overall α acceptably close to the nominal .05. In fact, it is slightly conservative, making this the preferred univariate test within the limits of their study. They also compared this procedure with that of performing univariate tests without a prior F^2 -test (unprotected tests). For this case, the overall α is too high, as expected. Table 5.1 gives an excerpt of Hummel and Sligo's results. The sample size is for each of the two samples; the r^2 in column 6 is for every pair of variables.

Hummel and Sligo therefore recommended performing the multivariate F^2 -test followed by univariate t -tests. This procedure appears to have the desired overall α level and will clearly have better power than tests using T_p or $t_{n-1,p}$ as a critical value. Table 5.2 also highlights the importance of using univariate t -tests only if the multivariate F^2 -test is significant. The inflated α 's resulting if t -tests are used without regard to the outcome of the F^2 -test are clearly evident. Thus among the three univariate procedures (procedures 1, 2, and 3), the first appears to be preferred.

Among the multivariate approaches (procedures 4, 5, and 7), we prefer the fifth procedure, which compares the (absolute value of) coefficients in the discriminant function to find the effect of each variable in separating the two groups of observations. These coefficients will often tell a different story from the univariate tests, because the univariate tests do not take into account the correlations among the variables or the effect of each variable on F^2 in the presence of the other variables. A variable will typically have a different effect in the presence of other variables than it has by itself. In the discriminant function, $1 = a_0\bar{Y} = a_1\bar{Y}_1 + a_2\bar{Y}_2 + \cdots + a_p\bar{Y}_p$, where $a = S_p^{-1}(\bar{Y}_1 - \bar{Y}_2)$, the coefficients $a_1, a_2, \dots, a_p$ indicate the relative importance of the variables in a multivariate context, something the univariate t -tests cannot do. If the variables are not commensurate (similar in scale and variance), the coefficients should be standardized, as in Section 8.5; this allows for more valid comparisons among the variables. Bender and Lord (1988) provided a decomposition

Table 3.1. Comparison of Experimentwise Error Rates (Nominal $\alpha = .05$)

Sample Size	Number of Variables	Classifiers ^a			
		.10	.30	.50	.70
Discriminate Tests Only ^b					
10	3	.145	.102	.114	.087
10	4	.267	.190	.170	.111
10	5	.600	.347	.309	.170
20	3	.115	.109	.117	.080
20	4	.320	.200	.170	.110
20	5	.590	.303	.223	.140
50	3	.130	.124	.102	.080
50	4	.330	.190	.160	.115
50	5	.570	.270	.200	.140
Multivariate Test Followed by Univariate Tests ^c					
10	3	.060	.039	.044	.020
10	4	.060	.039	.040	.017
10	5	.060	.030	.028	.018
20	3	.037	.044	.020	.020
20	4	.037	.037	.042	.021
20	5	.040	.042	.040	.021
50	3	.030	.040	.034	.020
50	4	.037	.039	.028	.027
50	5	.030	.030	.020	.020

^aIgnoring multivariate tests.^bControl rate only if multivariate test rejects.

tion of the information is the standardized discriminant function coefficients. For a detailed analysis of the effect of each variable in the presence of the other variables, see Rencher (1995; 1998, Sections 3.3.3 and 3.3.3).

Example 3.4. For the psychological data in Table 3.1, we obtained $\bar{\mathbf{y}}_1$, $\bar{\mathbf{y}}_2$, and $\mathbf{S}_p$ in Example 3.4.2. The discriminant function coefficient vector is obtained from (3.14) as

$$\mathbf{a} = \mathbf{S}_p^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = \begin{pmatrix} .3104 \\ -.2050 \\ .4668 \\ -.3007 \end{pmatrix}.$$

Thus the linear combination that best separates the two groups is

$$a_1\mathbf{y} = .3104y_1 - .2050y_2 + .4668y_3 - .3007y_4.$$

in which y_1 and y_2 appear to contribute most to separation of the two groups. After standardization, the relative contribution of the variables changes somewhat, see the answer to Problem 8.7 in Appendix B.) $\square$

5.6 COMPUTATION OF T^2

If one has a program available with matrix manipulations capability, it is simple matter to compute T^2 using (5.9). However, this approach is somewhat cumbersome for those not accustomed to the use of such a programming language, and many would prefer a more automated procedure. But very few general-purpose statistical programs provide for direct calculation of the two-sample T^2 statistic, perhaps because it is so easy to obtain from other procedures. We will discuss two types of widely available procedures that can be used to compute T^2 .

5.6.1 Obtaining T^2 from a MANOVA Program

Multivariate analysis of variance (MANOVA) is discussed in Chapter 6, and the reader may wish to return to the present section after becoming familiar with that material. One-way MANOVA involves a comparison of mean vectors from several samples. Typically, the number of samples is three or more, but the procedure will also accommodate two samples. The two-sample T^2 test is thus a special case of MANOVA.

Four common test statistics are defined in Section 6.1: Wilks' Λ , the Lawley-Hotelling $U^{(H)}$, Pillai's $T^{(H)}$, and Roy's largest root R . Without concerning ourselves here with how these are defined or calculated, we show how to use each to obtain the two-sample T^2 :

$$T^2 = (n_1 + n_2 - 2) \frac{1 - \Lambda}{\Lambda}, \quad (5.16)$$

$$T^2 = (n_1 + n_2 - 2) U^{(H)}, \quad (5.17)$$

$$T^2 = (n_1 + n_2 - 2) \frac{T^{(H)}}{1 - T^{(H)}}, \quad (5.18)$$

$$T^2 = (n_1 + n_2 - 2) \frac{R}{1 - R}. \quad (5.19)$$

For the special case of two groups, $T^{(H)} = F$. These relationships are demonstrated in Section 6.1.3. If the MANOVA program gives eigenvalues of $E^{-1}H$ (E and H are defined in Section 6.1.2), the eigenvalue corresponding to the largest eigenvalue will be equal to (a constant multiple of) the discriminant function $K_p^{-1}(\bar{y}_1 - \bar{y}_2)$.

5.6.2 Obtaining T^2 from Multiple Regression

In this section, the y 's become independent variables in a regression model. For each observation vector y_{1i} and y_{2i} in a two-sample T^2 , define a "dummy" group variable

is

$$\begin{aligned}w_i &= \frac{a_1^2}{a_1 + a_2} \quad \text{for each of } y_{11}, y_{12}, \dots, y_{1n_1} \text{ in sample 1} \\ &= -\frac{a_2}{a_1 + a_2} \quad \text{for each of } y_{21}, y_{22}, \dots, y_{2n_2} \text{ in sample 2.}\end{aligned}$$

Then $W = 0$ for all $a_1 + a_2$ observations. The prediction equation for the regression of w on the y 's can be written as

$$\hat{w}_i = b_0 + b_1 y_{i1} + b_2 y_{i2} + \dots + b_p y_{ip},$$

where i ranges over all $a_1 + a_2$ observations and the least squares estimate b_0 is (see (3.15))

$$b_0 = W - b_1 T_1 - b_2 T_2 - \dots - b_p T_p.$$

Substituting this into the regression equation, we obtain

$$\begin{aligned}\hat{w}_i &= W + b_1(y_{i1} - T_1) + b_2(y_{i2} - T_2) + \dots + b_p(y_{ip} - T_p) \\ &= b_1(y_{i1} - T_1) + b_2(y_{i2} - T_2) + \dots + b_p(y_{ip} - T_p) \quad \text{since } W = 0.\end{aligned}$$

Let $\mathbf{b}' = (b_1, b_2, \dots, b_p)$ be the vector of regression coefficients and $\mathbf{R}^2$ be the squared multiple correlation. Then we have the following relationships:

$$F^2 = (n_1 + n_2 - 2) \frac{\mathbf{R}^2}{1 - \mathbf{R}^2}, \quad (5.20)$$

$$\mathbf{a} = \mathbf{S}_p^{-1} (\mathbf{T}_1 - T_2) = \frac{n_1 + n_2}{n_1 n_2} (n_1 + n_2 - 2) F^2 \mathbf{b}. \quad (5.21)$$

Thus with ordinary multiple regression, one can easily obtain F^2 and the discriminant function $\mathbf{S}_p^{-1} (\mathbf{T}_1 - T_2)$. We simply define w_i as above for each of the $a_1 + a_2$ observations, regress the w 's on the y 's, and use the resulting $\mathbf{R}^2$ in (5.20). For $\mathbf{b}$, define the intercept from the regression coefficients for use in (5.21). Actually, since only the relative values of the elements of $\mathbf{a} = \mathbf{S}_p^{-1} (\mathbf{T}_1 - T_2)$ are of interest, it is not necessary to correct from $\mathbf{b}$ to $\mathbf{a}$ in (5.21). We can use $\mathbf{b}$ directly or standardize the values $b_1, b_2, \dots, b_p$ as in Section 3.5.

Example 5.4.2. We illustrate the regression approach to computation of F^2 using the psychological data in Table 5.1. We set $w = a_1/(a_1 + a_2) = \frac{25}{55} = \frac{5}{11}$ for each observation in group 1 (males) and equal to $-a_2/(a_1 + a_2) = -\frac{1}{11}$ in the second group (females). When w is regressed on the 64 y 's, we obtain

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} -.791 \\ .091 \\ -.020 \\ .047 \\ -.031 \end{pmatrix}, \quad R^2 = .8413,$$

by (5.20),

$$T^2 = (n_1 + n_2 - 2) \frac{R^2}{1 - R^2} = \frac{(55)(.8413)}{1 - .8413} = 97.604,$$

as was obtained before in Example 5.4.1. Note that $\mathbf{W} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (.086, -.020, .047, -.031)$, with the intercept deleted, is proportional to the discriminant function coefficient vector $\mathbf{a}$ from Example 5.3, as we would expect from (5.23). $\square$

5.7 PAIRED-OBSERVATIONS TEST

As usual, we begin with the univariate case to set the stage for the multivariate presentation.

5.7.1 Univariate Case

Suppose two samples are not independent because there exists a natural pairing between the i th observation y_i in the first sample and the i th observation x_i in the second sample for all i , as, for example, when a treatment is applied twice to the same individual or when subjects are matched according to some criterion, such as IQ or family background. With such pairing, the samples are often referred to as *paired observations* or *matched pairs*. The two samples thus obtained are correlated, and the two-sample test statistic in (5.9) is not appropriate because the samples need be independent in order for (5.9) to have a t -distribution. [The two-sample test in (5.9) is somewhat robust to heterogeneity of variances and to lack of normality but not to dependence.] We reduce the two samples to one by working with the differences between the paired observations, as in the following layout for two treatments applied to the same subject:

Pair Number	Treatment		Difference
	1	2	$d_i = y_i - x_i$
1	y_1	x_1	d_1
2	y_2	x_2	d_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n	d_n

To obtain a t test, it is not sufficient to assume individual normality for each of y and x . To allow for the covariance between y and x , we need the additional assump-

tion that y and x have a bivariate normal distribution with

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}.$$

It then follows by property 1a in Section 4.1 that $d_i = y_i - x_i$ is $N(\mu_d = \mu_y - \mu_x, \sigma_d^2)$, where $\sigma_d^2 = \sigma_y^2 - 2\sigma_{yx} + \sigma_x^2$. From $d_1, d_2, \dots, d_n$ we calculate

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i \quad \text{and} \quad s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2.$$

To test $H_0: \mu_y = \mu_x$, that is, $H_0: \mu_d = 0$, we use the one-sample statistic

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}, \quad (5.22)$$

which is distributed as t_{n-1} if H_0 is true. We reject H_0 in favor of $H_1: \mu_d \neq 0$ if $|t| > t_{\alpha/2, n-1}$. It is not necessary to assume $\sigma_y^2 = \sigma_x^2$ because there are no restrictions on Σ .

This test has only $n-1$ degrees of freedom compared with $2(n-1)$ for the two-independent-sample t test (5.8). In general, the pairing reduces the within-sample variation s_d and thereby increases the power.

If we mistakenly treated the two-samples as independent and used (5.8) with $n_1 = n_2 = n$, we would have

$$t = \frac{\bar{y} - \bar{x}}{s_{xy} \sqrt{2/n}} = \frac{\bar{y} - \bar{x}}{\sqrt{2s_d^2/n}}.$$

However,

$$E\left(\frac{2s_d^2}{n}\right) = 2E\left[\frac{(n-1)s_d^2 + (n-1)s_d^2}{(n+n-2n)}\right] = \frac{\sigma_y^2 + \sigma_x^2}{n},$$

whereas $\text{var}(\bar{y} - \bar{x}) = (\sigma_y^2 + \sigma_x^2 - 2\sigma_{yx})/n$. Thus if the test statistic for independent samples (5.8) is used for paired data, it does not have a t -distribution and, in fact, underestimates the true average t -value (assuming H_0 is false), since $\sigma_y^2 + \sigma_x^2 = \sigma_d^2 + \sigma_{yx}$ if $\sigma_{yx} > 0$, which would be typical in this situation. One could therefore use

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{(\sigma_y^2 + \sigma_x^2 - 2\sigma_{yx})/n}}, \quad (5.23)$$

but $t = \sqrt{n/2}t_d$ in (5.22) is equal to it, and somewhat simpler to use.

5.3.1 Multivariate Case

Here we assume the same natural pairing of sampling units as in the univariate case, but we measure p variables on each sampling unit. Thus $\mathbf{y}_i$ from the first sample is paired with $\mathbf{x}_i$ from the second sample, $i = 1, 2, \dots, n$. In terms of two treatments applied to each sampling unit, this situation is as follows:

Pair Number	Treatment 1	Treatment 2	Difference
1	$\mathbf{y}_1$	$\mathbf{x}_1$	$\mathbf{d}_1$
2	$\mathbf{y}_2$	$\mathbf{x}_2$	$\mathbf{d}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\mathbf{y}_n$	$\mathbf{x}_n$	$\mathbf{d}_n$

In Section 5.1.1, we made the assumption that y and x have a bivariate normal distribution, in which y and x are correlated. Here we assume $\mathbf{y}$ and $\mathbf{x}$ are correlated and have a multivariate normal distribution:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim N_{2p} \left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \right].$$

To test $H_0: \boldsymbol{\mu}_y = \mathbf{0}$, which is equivalent to $H_0: \boldsymbol{\mu}_y = \boldsymbol{\mu}_x$ since $\boldsymbol{\mu}_d = E(\mathbf{y} - \mathbf{x}) = \boldsymbol{\mu}_y - \boldsymbol{\mu}_x$, we calculate

$$\bar{\mathbf{d}} = \frac{1}{n} \sum_{i=1}^n \mathbf{d}_i \quad \text{and} \quad \mathbf{S}_d = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})'$$

We then have

$$T^2 = \bar{\mathbf{d}}' \left(\frac{\mathbf{S}_d}{n} \right)^{-1} \bar{\mathbf{d}} = n \bar{\mathbf{d}}' \mathbf{S}_d^{-1} \bar{\mathbf{d}}. \quad (5.24)$$

Under H_0 , this paired comparison T^2 -statistic is distributed as $F_{p, n-p}^2$. We reject H_0 if $T^2 > F_{p, n-p}^2$. Note that $\mathbf{S}_d$ estimates $\text{cov}(\mathbf{y} - \mathbf{x}) = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx} - \boldsymbol{\Sigma}_{xy} + \boldsymbol{\Sigma}_{xx}$, for which an equivalent estimator would be $\mathbf{S}_{yy} - \mathbf{S}_{yx} - \mathbf{S}_{xy} + \mathbf{S}_{xx}$ [see (3.42)].

The random process in Section 5.3.1 for univariate paired observation data also apply here. If the two samples of multivariate observations are correlated because of a natural pairing of sampling units, the test in (5.24) should be used rather than the two-sample T^2 -test in (5.9), which assumes two independent samples. Misuse of (5.9) in place of (5.24) will lead to loss of power.

Since the assumption $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}$ is not needed for (5.24) to have a T^2 -distribution, this test can be used for independent samples when $\boldsymbol{\Sigma}_2 \neq \boldsymbol{\Sigma}_1$ (as long as $n_1 = n_2$). The observations in the two samples would be paired in the order they were obtained or in an arbitrary order. However, in the case of independent

samples, the pairing achieves no gain in power to offset the loss of $n - 1$ degrees of freedom.

By analogy with (5.14), the discriminant function coefficient vector for paired observation data becomes

$$\mathbf{a} = \mathbf{S}_D^{-1} \bar{\mathbf{D}}. \quad (5.22)$$

For tests on individual variables, we have

$$t_j = \frac{\bar{D}_j}{\sqrt{S_{D,jj}/n}}, \quad j = 1, 2, \dots, p. \quad (5.26)$$

The critical value for t_j is $t_{\alpha/2, n-1}$ or $t_{\alpha, n-1}$ depending on whether a T^2 -test is carried out first (see Section 5.5).

Example 5.5.1. To compare two types of coating for resistance to corrosion, 15 pieces of pipe were coated with each type of coating (Kramer and Jensen 1969b). Two pipes, one with each type of coating, were buried together and left for the same length of time at 15 different locations, providing a natural pairing of the observations. Corrosion for the first type of coating was measured by two variables,

y_1 = maximum depth of pit in thousandths of an inch,

y_2 = number of pits,

Table 5.1. Depth of Maximum Pit and Number of Pits of Coated Pipes

Location	Coating 1		Coating 2		Difference	
	Depth	Number	Depth	Number	Depth	Number
	x_1	x_2	x_1	x_2	d_1	d_2
1	73	31	58	33	15	-4
2	43	19	48	14	-5	5
3	47	11	43	19	4	8
4	53	16	48	19	15	-3
5	68	16	47	14	11	2
6	47	10	52	16	15	4
7	52	19	29	19	23	0
8	48	10	43	17	-5	-1
9	61	14	53	14	8	0
10	56	13	52	17	4	4
11	56	19	57	14	-1	5
12	64	19	48	19	16	0
13	69	16	57	16	-2	-4
14	65	13	65	7	0	6
15	72	19	66	13	6	6

with x_1 and x_2 defined analogously for the second reading. The data and differences are given in Table 5.3. Thus we have, for example, $y_1' = (73, 34)'$, $y_2' = (51, 33)'$, and $d_1' = y_1' - y_2' = (22, -1)'$. For the 15 difference vectors, we obtain

$$\bar{d} = \begin{pmatrix} 8.000 \\ 1.067 \end{pmatrix}, \quad S_d = \begin{pmatrix} 121.571 & 17.071 \\ 17.071 & 21.761 \end{pmatrix}.$$

By (5.24),

$$T^2 = (15)(8.000, 1.067) \begin{pmatrix} 121.571 & 17.071 \\ 17.071 & 21.761 \end{pmatrix}^{-1} \begin{pmatrix} 8.000 \\ 1.067 \end{pmatrix} = 10.319.$$

Since $T^2 = 10.319 > T_{0.05, 1, 14}^2 = 9.197$, we reject $H_0: \mu_d = 0$ and conclude that the two readings differ in their effect on corrosion. $\square$

5.8 TEST FOR ADDITIONAL INFORMATION

In this section, we are again considering two independent samples, as in Section 5.4.2. We start with a basic $p \times 1$ vector y of measurements on each sampling unit, and ask whether a $q \times 1$ subvector x measured in addition to y (on the same unit) will significantly increase the separation of the two samples as shown by T^2 . It is not necessary that we add new variables. We may be interested in determining whether some of the variables we already have are redundant in the presence of other variables in terms of separating the groups. We have designated the subset of interest by x for notational convenience.

It is assumed that the two samples are from multivariate normal populations with a common covariance matrix; that is,

$$\begin{aligned} \begin{pmatrix} y_{11} \\ x_{11} \end{pmatrix}, \begin{pmatrix} y_{12} \\ x_{12} \end{pmatrix}, \dots, \begin{pmatrix} y_{1n_1} \\ x_{1n_1} \end{pmatrix} & \quad \text{are from } P_{p+q}(p_1, \Sigma), \\ \begin{pmatrix} y_{21} \\ x_{21} \end{pmatrix}, \begin{pmatrix} y_{22} \\ x_{22} \end{pmatrix}, \dots, \begin{pmatrix} y_{2n_2} \\ x_{2n_2} \end{pmatrix} & \quad \text{are from } P_{p+q}(p_2, \Sigma), \end{aligned}$$

where

$$\begin{aligned} p_1 &= E \begin{pmatrix} y_{1i} \\ x_{1i} \end{pmatrix} = \begin{pmatrix} \mu_{y1} \\ \mu_{x1} \end{pmatrix}, \quad p_2 = E \begin{pmatrix} y_{2i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} \mu_{y2} \\ \mu_{x2} \end{pmatrix}, \\ \Sigma &= \text{cov} \begin{pmatrix} y_{1i} \\ x_{1i} \end{pmatrix} = \text{cov} \begin{pmatrix} y_{2i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}. \end{aligned}$$

We partition the sample mean vectors and covariance matrix accordingly:

$$\begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix}, \quad \begin{pmatrix} S_y \\ S_x \end{pmatrix}, \quad S_d = \begin{pmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{pmatrix},$$

where S_d is the pooled sample covariance matrix from the two samples.

We wish to test the hypothesis that x_1 and x_2 are redundant for separating the two groups, that is, that the extra q variables do not contribute anything significant beyond the information already available in y_1 and y_2 for separating the groups. This is in the spirit of a *full and reduced model* test in regression (see (3.31) and Section 10.2.5b). However, here we are working with a subset of dependent variables as contrasted to the subset of independent variables in the regression setting. Thus both y and x are subvectors of dependent variables. In this setting, the independent variables would be grouping variables 1 and 2 corresponding to μ_1 and μ_2 .

We are not asking if the x 's can significantly separate the two groups by themselves, but whether they provide additional separation beyond the separation already achieved by the y 's. If the x 's were independent of the y 's, we would have $F_{p+q}^2 = F_y^2 + F_x^2$, but this does not hold, because they are correlated. We must compare F_{p+q}^2 for the full set of variables $(y_1, \dots, y_p, x_1, \dots, x_q)$ with F_y^2 based on the reduced set $(y_1, \dots, y_p)$. We are inquiring if the increase from F_y^2 to F_{p+q}^2 is significant.

By definition, the F^2 -statistic based on the full set of $p+q$ variables is given by

$$F_{p+q}^2 = \frac{n_1 n_2}{n_1 + n_2} \left[\left(\begin{pmatrix} \bar{y}_1 \\ \bar{x}_1 \end{pmatrix} \right) - \left(\begin{pmatrix} \bar{y}_2 \\ \bar{x}_2 \end{pmatrix} \right) \right]' S_y^{-1} \left[\left(\begin{pmatrix} \bar{y}_1 \\ \bar{x}_1 \end{pmatrix} \right) - \left(\begin{pmatrix} \bar{y}_2 \\ \bar{x}_2 \end{pmatrix} \right) \right], \quad (3.27)$$

whereas F^2 for the reduced set of p variables is

$$F_p^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{y}_1 - \bar{y}_2)' S_y^{-1} (\bar{y}_1 - \bar{y}_2). \quad (3.28)$$

Then the test statistic for the significance of the increase from F_y^2 to F_{p+q}^2 is given by

$$F^2(x|y) = (v-p) \frac{F_{p+q}^2 - F_y^2}{v + F_y^2}, \quad (3.29)$$

which is distributed as $F_{p, v-p}^2$. We reject the hypothesis of redundancy of x if $F^2(x|y) \geq F_{p, v-p}^2$.

By (3.23), $F^2(x|y)$ can be converted to an F -statistic:

$$F = \frac{v-p-q+1}{q} \frac{F_{p+q}^2 - F_y^2}{v + F_y^2}, \quad (3.30)$$

which is distributed as $F_{p, v-p-q+1}$, and we reject the hypothesis if $F \geq F_{p, v-p-q+1}$.

In both cases $v = n_1 + n_2 - 2$. (Note that the first degrees-of-freedom parameter in both (3.29) and (3.30) is q , the number of x 's. The second parameter in (3.29) is $v-p$ because the statistic is adjusted for the p variables in y .)

To prove directly that the statistic defined in (3.29) has an F -distribution, we can use a basic relationship from multiple regression (see (1.3.15)):

$$F_{q, r-p-q+1}^2 = \frac{(R_{p+q}^2 - R_p^2)(r - p - q + 1)}{(1 - R_{p+q}^2)q}, \quad (5.30)$$

where R_{p+q}^2 is the squared multiple correlation from the full model with $p + q$ independent variables, and R_p^2 is from the reduced model with p independent variables. If we solve for R^2 in terms of F^2 from (5.30) and substitute this into (5.31), we readily obtain the test statistic in (5.29).

If we are interested in the effect of adding a single x , then $q = 1$, and both (5.29) and (5.30) reduce to

$$F^2(x|y) = (r - p) \frac{F_{p+1}^2 - F_p^2}{r + F_p^2}, \quad (5.32)$$

and we reject the hypothesis of redundancy of x if $F^2(x|y) > F_{\alpha, 1, r-p}^2$.

Example 5.8. We use the psychological data of Table 5.1 to illustrate tests on subvectors. We begin by testing the significance of y_3 and y_4 above and beyond y_1 and y_2 . On the notation of the present section, y_1 and y_2 become x_1 and x_2 . For these subvectors, $p = 2$ and $q = 2$. The value of F_{p+q}^2 for all four variables is given by (5.27) was obtained in Example 5.4.2 as 93.6075. For y_3 and y_4 , we obtain, by (5.28),

$$\begin{aligned} F_F^2 &= \frac{y_3 y_4}{y_3 + y_4} (y_3 - y_4)' R_{FF}^{-1} (y_3 - y_4) \\ &= \frac{(32)^2}{32 + 32} \begin{pmatrix} 15.97 & -12.94 \\ 15.81 & -13.91 \end{pmatrix}' \begin{pmatrix} 7.18 & 8.09 \\ 6.05 & 15.69 \end{pmatrix}^{-1} \begin{pmatrix} 15.97 & -12.94 \\ 15.81 & -13.91 \end{pmatrix} \\ &= 31.8126. \end{aligned}$$

By (5.29), the test statistic is

$$\begin{aligned} F^2(y_3, y_4|y_1, y_2) &= (r - p) \frac{F_{p+q}^2 - F_p^2}{r + F_p^2} \\ &= (62 - 2) \frac{93.6075 - 31.8126}{62 + 31.8126} = 41.953. \end{aligned}$$

We reject the hypothesis that $\mathbf{u} = (y_3, y_4)'$ is redundant, since $41.953 > F_{\alpha, 2, 59}^2 = 18.137$. We conclude that $\mathbf{u} = (y_3, y_4)'$ adds a significant amount of separation to $\mathbf{f} = (y_1, y_2)'$.

To test the effect of each variable, adjusted for the other three, we use (5.32). In this case, $p = 3$, $r = 62$, and $r - p = 59$. The results are given below, where $F_{p+1}^2 = 93.6075$ and F_p^2 in each case is based on the three variables, excluding the variable in question. For example, $F_p^2 = 70.8548$ for y_3 is based on y_1, y_2 , and y_4 .

and $r^2(y_1, y_2, y_3, y_4) = 1.812$.

Variable	T_j^2	$\lambda = p \frac{T_{\text{max}}^2 - T_j^2}{n + T_{\text{max}}^2}$
y_1	78.8733	7.844
y_2	88.8348	2.612
y_3	53.6273	46.913
y_4	74.9536	9.938

When we compare these four test statistic values with the critical value $r_{.05, 95}^2 = 4.002$, we see that each variable except y_2 makes a significant contribution to T^2 . Note that y_3 contributes most, followed by y_1 and then y_4 . This order differs from that given by the raw discriminant function in Example 5.5 but agrees with the order for the standardized discriminant function given in the answer to Problem 8.7 in Appendix B. $\square$

5.4 PROFILE ANALYSIS

If $\mathbf{y}$ is $(Y_p/\mu, \Sigma)$ and the variables in $\mathbf{y}$ are commensurate (measured in the same units and with approximately equal variances as, for example, in the probe word data in Table 3.5), we may wish to compare the means $\mu_1, \mu_2, \dots, \mu_p$ in μ . This might be of interest when a measurement is taken on the same research unit at p successive times. Such situations are often referred to as repeated measures designs or growth curves, which are discussed in some generality in Sections 6.9 and 6.10. In the present section, we discuss one- and two-sample profile analysis. Profile analysis for several samples is covered in Section 5.5.

The pattern obtained by plotting $\mu_1, \mu_2, \dots, \mu_p$ as ordinates and connecting the points is called a *profile*; we usually draw straight lines connecting the points $(1, \mu_1)$, $(2, \mu_2), \dots, (p, \mu_p)$. Profile analysis is an analysis of the profile; it is a comparison of two or more profiles. Profile analysis is often discussed in the context of administering a battery of p psychological or other tests.

In growth curves analysis, where the variables are measured at time intervals, the responses have a natural order. In profile analysis where the variables arise from test scores, there is ordinarily no natural order. A distinction is not always made between repeated measures of the same variable through time and profile analysis of several different commensurate variables on the same individual.

5.5.1 One-Sample Profile Analysis

We begin with a discussion of the profile of the mean vector μ from a single sample. A plot of μ might appear as in Figure 5.3, where we plot $(1, \mu_1)$, $(2, \mu_2), \dots, (p, \mu_p)$ and connect the points.

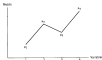


Figure 8.3. Profile of a mean vector.

In order to compare the means $\mu_1, \mu_2, \dots, \mu_p$ in μ , the basic hypothesis is that the profile is in level or flat:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_p = \mu_0, \quad H_1: \mu_j \neq \mu_0 \quad \text{for some } j \neq 0.$$

The data matrix $\mathbf{Y}$ is given in (8.17). We cannot use univariate analysis of variance to test H_0 because the columns in $\mathbf{Y}$ are not independent. For a multivariate approach that allows for correlated variables, we first express H_0 as $p-1$ comparisons,

$$H_{0j}: \begin{pmatrix} \mu_1 - \mu_1 \\ \mu_2 - \mu_1 \\ \vdots \\ \mu_{p-1} - \mu_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

or as

$$H_{0j}: \begin{pmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \\ \vdots \\ \mu_p - \mu_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

These two expressions can be written in the form $H_{0j}: \mathbf{C}_1 \mu = \mathbf{0}$ and $H_{0j}: \mathbf{C}_2 \mu = \mathbf{0}$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are the $(p-1) \times p$ matrices:

$$\mathbf{C}_1 = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 \end{pmatrix}, \quad \mathbf{C}_2 = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{pmatrix}.$$

In fact, any $(p-1) \times p$ matrix C of rank $p-1$ such that $CC^T = 0$ can be used in $H_0: C\mu = 0$ to produce $H_0: \mu_1 = \mu_2 = \cdots = \mu_p$. If $C\mu = 0$, each row c_j^T of C sums to zero by (2.38). A linear combination $c_j^T\mu = c_{j1}\mu_1 + c_{j2}\mu_2 + \cdots + c_{jp}\mu_p$ is called a contrast in the μ 's if the coefficients sum to zero, that is, if $\sum_{j=1}^p c_{jj} = 0$. The $p-1$ contrasts in $C\mu$ must be linearly independent in order to require $H_0: \mu_1 = \mu_2 = \cdots = \mu_p$ in $H_0: C\mu = 0$. Thus rank(C) = $p-1$.

From a sample $y_1, y_2, \dots, y_n$, we obtain estimates $\bar{y}$ and S of population parameters μ and Σ . To test $H_0: C\mu = 0$, we transform each $y_j, j = 1, 2, \dots, n$, to $z_j = Cy_j$, which is $(p-1) \times 1$. By (2.42) and (2.44), the sample mean vector and covariance matrix of $z_j = Cy_j, j = 1, 2, \dots, n$, are $\bar{z} = C\bar{y}$ and $S_z = CSC^T$, respectively. If y is $N_p(\mu, \Sigma)$, then by property 1b in Section 4.2, $z = Cy$ is $N_{p-1}(C\mu, CSC^T)$. Thus when $H_0: C\mu = 0$ is true, $\bar{z}$ is $N_{p-1}(0, CSC^T/n)$, and

$$T^2 = n\bar{z}Y \left(\frac{CSC^T}{n} \right)^{-1} n\bar{z}Y = n\bar{z}Y(CSC^T)^{-1}C\bar{y} \quad (5.33)$$

is distributed as $T^2_{p-1, n-p-1}$. We reject $H_0: C\mu = 0$ if $T^2 \geq T^2_{\alpha, p-1, n-p-1}$. The dimension $p-1$ corresponds to the number of rows of C . Thus $\bar{z} = C\bar{y}$ is $(p-1) \times 1$ and $S_z = CSC^T$ is $(p-1) \times (p-1)$. Note that the C 's in (5.33) don't "cancel" because C is $(p-1) \times p$ and does not have an inverse. In fact, T^2 in (5.33) is less than $T^2 = n\bar{y}B^{-1}\bar{y}$ [see Rautava (1978, p. 84)].

If the variables have a natural ordering, as, for example, in the raters' bone data in Table 5.4, we could test for a linear trend or polynomial curve in the means by suitably choosing the rows of C . This is discussed in connection with growth curves in Section 6.1.3. Other comparisons of interest can be made as long as they are linearly independent.

5.5.2 Two-Sample Profile Analysis

Suppose two independent groups or samples receive the same set of p tests or measurements. If these tests are comparable, for example, all on a scale of 0 to 100, the variables will often be commensurate.

Rather than testing the hypothesis that $\mu_0 = \mu_1$, we wish to be more specific in comparing the profiles obtained by connecting the points $(j, \mu_{1j}), j = 1, 2, \dots, p$, and $(j, \mu_{0j}), j = 1, 2, \dots, p$. There are three hypotheses of interest in comparing the profiles of two samples. The first of these hypotheses addresses the question, Are the two profiles similar in appearance, or more precisely, are they parallel? We illustrate this hypothesis in Figure 5.4. If the two profiles are parallel, then one group scored uniformly better than the other group on all p tests.

The parallelism hypothesis can be defined in terms of the slopes. The two profiles are parallel if the two slopes for each segment are the same. If the two profiles are parallel, the two increments for each segment are the same, and it is not necessary to use the actual slopes to express the hypothesis. We can simply compare the increments from one point to the next. The hypothesis can then be expressed as $H_{02}: \mu_{1j} -$

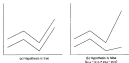


Figure 3.4. Comparison of two profiles under the hypothesis of parallelism.

$\mu_{1,j+1} = \mu_{2,j+1}$ for $j = 2, 3, \dots, p-1$

$$R_{01} = \begin{pmatrix} \mu_{11} - \mu_{11} \\ \mu_{11} - \mu_{11} \\ \vdots \\ \mu_{1p} - \mu_{1p-1} \end{pmatrix} = \begin{pmatrix} \mu_{21} - \mu_{21} \\ \mu_{21} - \mu_{21} \\ \vdots \\ \mu_{2p} - \mu_{2p-1} \end{pmatrix},$$

which can be written as $R_{01} = C\mu_{01} = C\mu_{02}$ using the contrast matrix

$$C = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

From two samples, $y_{11}, y_{12}, \dots, y_{1n_1}$ and $y_{21}, y_{22}, \dots, y_{2n_2}$, we obtain $\bar{y}_1, \bar{y}_2$, and S_{01} as estimates of μ_{01}, μ_{02} , and Σ . As in the two-sample F -test, we assume that each y_{1j} in the first sample is $N_p(\mu_1, \Sigma)$ and each y_{2j} in the second sample is $N_p(\mu_2, \Sigma)$. If C is a $(p-1) \times p$ contrast matrix, as before, then Cy_{1j} and Cy_{2j} are distributed as $N_{p-1}(C\mu_{01}, C\Sigma C')$ and $N_{p-1}(C\mu_{02}, C\Sigma C')$, respectively. Under $R_{01} = C\mu_{01} = C\mu_{02} = 0$, the random vector $Cy_{1j} - Cy_{2j}$ is $N_{p-1}(0, C\Sigma C')(1/n_1 + 1/n_2)$, and

$$\begin{aligned} F &= (Cy_1 - Cy_2)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) C S_{01} C' \right]^{-1} (Cy_1 - Cy_2) \\ &= \frac{S_0 S_1}{n_1 + n_2} (\bar{y}_1 - \bar{y}_2)' C' (C S_{01} C')^{-1} C (\bar{y}_1 - \bar{y}_2) \end{aligned} \quad (3.24)$$

is distributed as $F_{p-1, n_1+n_2-p}^2$. Note that the denominator $p-1$ is the number of rows of C .

By analogy with the discussion in Section 3.5, if H_{01} is rejected, we can follow up with univariate tests on the individual components of $C(\bar{Y}_1 - \bar{Y}_2)$. Alternatively, we can calculate the discriminant function

$$\mathbf{a} = (C\mathbf{N}_p C^T)^{-1} C(\bar{Y}_1 - \bar{Y}_2) \quad (5.28)$$

as an indication of which slope differences contributed most to rejection of H_{01} in the presence of the other components of $C(\bar{Y}_1 - \bar{Y}_2)$. There should be less need in this case to standardize the components of $\mathbf{a}$, as suggested in Section 3.5, because the variables are assumed to be commensurate. The vector $\mathbf{a}$ is $(p-1) \times 1$, corresponding to the $p-1$ segments of the profile. Thus if the second component of $\mathbf{a}$, for example, is largest in absolute value, the divergence in slopes between the first profile and the second segment contributes most to rejection of H_{01} .

If the data are arranged as in Table 5.4, we see an analogy to a two-way ANOVA model. A plot of the means is often made in a two-way ANOVA, a lack of parallelism corresponding to interaction between the two factors. Thus, the hypothesis H_{01} is analogous to the group-by-test (variable)-interaction hypothesis.

However, the usual ANOVA assumption of independence of observations does not hold here because the variables (tests) are correlated. The ANOVA assumption of independence and homogeneity of variances would require $\text{cov}(y) = \Sigma = \sigma^2 \mathbf{I}$. Hence the test of H_{01} cannot be carried out using a univariate ANOVA approach, since $\Sigma \neq \sigma^2 \mathbf{I}$. We therefore proceed with the multivariate approach using F^2 .

The second hypothesis of interest in comparing two profiles is, Are the two populations or groups at the same level? This hypothesis corresponds to a group (population)-main effect in the ANOVA analogy. We can express this hypothesis in terms of the average level of group 1, compared to the average level of group 2:

$$H_{02}: \frac{\mu_{11} + \mu_{12} + \cdots + \mu_{1p}}{p} = \frac{\mu_{21} + \mu_{22} + \cdots + \mu_{2p}}{p}$$

Table 5.4. Data Layout for Two-Sample Profile Analysis

		Tests (variables)			
		1	2	...	p
Group 1					
$\bar{y}_{11}$	=	$\bar{y}_{111}$	$\bar{y}_{112}$	...	$\bar{y}_{11p}$
$\bar{y}_{12}$	=	$\bar{y}_{121}$	$\bar{y}_{122}$	...	$\bar{y}_{12p}$
$\vdots$		$\vdots$	$\vdots$		$\vdots$
$\bar{y}_{1p}$	=	$\bar{y}_{1p1}$	$\bar{y}_{1p2}$	...	$\bar{y}_{1pp}$
Group 2					
$\bar{y}_{21}$	=	$\bar{y}_{211}$	$\bar{y}_{212}$	...	$\bar{y}_{21p}$
$\bar{y}_{22}$	=	$\bar{y}_{221}$	$\bar{y}_{222}$	...	$\bar{y}_{22p}$
$\vdots$		$\vdots$	$\vdots$		$\vdots$
$\bar{y}_{2p}$	=	$\bar{y}_{2p1}$	$\bar{y}_{2p2}$	...	$\bar{y}_{2pp}$

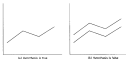


Figure 3.8. Hypothesis H_{02} of equal group effect, assuming parallelism.

By (2.37), this can be expressed as

$$H_{02}: \bar{\mu}_1 = \bar{\mu}_2$$

If H_{02} is true, H_{01} can be pictured as in Figure 3.5a. If H_{02} is false, then the two profiles differ by a constant (given that H_{01} is true), as in Figure 3.5b.

The hypothesis H_{02} can be true when H_{01} does not hold. Thus the average level of population 1 can equal the average level of population 2 without the two profiles being parallel, as illustrated in Figure 3.6. In this case, the "group main effect" is somewhat harder to interpret, with the case in the analogous two-way ANOVA, where main effects are more difficult to describe in the presence of significant interaction. However, the test may still furnish useful information if a useful description of the results is provided.



Figure 3.6. Hypothesis H_{02} of equal group effect without parallelism.

To test $H_{02} : f(\mu_1 - \mu_2) = 0$, we estimate $f(\mu_1 - \mu_2)$ by $f(\bar{Y}_1 - \bar{Y}_2)$, which is $n(y_1 f(2g(1/n_1 + 1/n_2)))$ when H_{02} is true. We can therefore use

$$t = \frac{f(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{f(2g(1/n_1 + 1/n_2))}} \quad (5.36)$$

and reject H_{02} if $|t| \geq t_{\alpha/2, n_1+n_2-2}$.

The third hypothesis of interest, now depending in the test (or variable) matrix B , is: Are the profiles flat? Assuming parallelism (assuming H_{01} is true), the “flatness” hypothesis can be pictured as in Figure 5.7. If H_{01} is not true, the test could be carried out separately for each group using the test in Section 5.4.1. If H_{02} is true, the two profiles in Figure 5.7a and Figure 5.7b will be coincident.

To express the third hypothesis in a form suitable for testing, we note from Figure 5.7a that the average of the two-group means is the same for each test:

$$H_{03} : \frac{1}{2}(x_{11} + x_{21}) = \frac{1}{2}(x_{12} + x_{22}) = \cdots = \frac{1}{2}(x_{1p} + x_{2p}) \quad (5.37)$$

or

$$H_{03} : \frac{1}{2}C(\mu_1 + \mu_2) = 0, \quad (5.38)$$

where C is a $(p-1) \times p$ matrix such that $Cj = 0$. From Figure 5.7a, we see that H_{03} could also be expressed as $x_{11} = x_{21} = \cdots = x_{1p}$ and $x_{21} = x_{22} = \cdots = x_{2p}$, or

$$H_{03} : C\mu_1 = 0 \quad \text{and} \quad C\mu_2 = 0.$$

To estimate $\frac{1}{2}(\mu_1 + \mu_2)$, we use the sample grand mean vector based on a weighted average:

$$\bar{y} = \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{n_1 + n_2}.$$

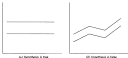


Figure 5.7. Hypothesis H_{03} of equal tests (parallelism) assuming parallelism.

It can easily be shown that under H_{01} (and H_{02}), $E(\mathbf{C}\bar{\mathbf{Y}}) = \mathbf{0}$ and $\text{cov}(\bar{\mathbf{Y}}) = \mathbf{\Sigma}_y/(\alpha_1 + \alpha_2)$. Therefore, $\mathbf{C}\bar{\mathbf{Y}}$ is $N_{p-1}(\mathbf{0}, \mathbf{C}\mathbf{\Sigma}_y\mathbf{C}^t)/(\alpha_1 + \alpha_2)$, and

$$\begin{aligned} T^2 &= (\mathbf{C}\bar{\mathbf{Y}})^t \left(\frac{\mathbf{C}\mathbf{\Sigma}_y\mathbf{C}^t}{\alpha_1 + \alpha_2} \right)^{-1} (\mathbf{C}\bar{\mathbf{Y}}) \\ &= (\alpha_1 + \alpha_2)(\mathbf{C}\bar{\mathbf{Y}})^t(\mathbf{C}\mathbf{\Sigma}_y\mathbf{C}^t)^{-1}\mathbf{C}\bar{\mathbf{Y}} \end{aligned} \quad (5.36)$$

is distributed as $T^2_{p-1, \alpha_1 + \alpha_2, 0}$ when both H_{01} and H_{02} are true. It can be readily shown that M_{02} is unaffected by a difference in the profile levels (unaffected by the status of H_{02}).

Example 5.3.1. We use the psychological data in Table 5.1 to illustrate two-sample profile analysis. The values of $\bar{\mathbf{y}}_1$, $\bar{\mathbf{y}}_2$, and M_{02} are given in Example 5.4.2. The profiles of the two mean vectors $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are plotted in Figure 5.8. There appears to be a lack of parallelism.

To test for parallelism, $M_{01} = \mathbf{C}\bar{\mathbf{y}}_{p1} = \mathbf{C}\bar{\mathbf{y}}_{p2}$, we use the matrix

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

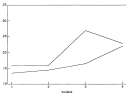


Figure 5.8. Profiles for the psychological data in Table 5.1.

and obtain

$$C(\mathbf{y}_2 - \mathbf{f}_2) = \begin{pmatrix} -1.62 \\ 8.59 \\ -9.72 \end{pmatrix}, \quad CS_{\mathbf{y}_2}C' = \begin{pmatrix} 18.85 & -7.65 & -1.64 \\ -7.65 & 27.26 & -12.74 \\ -1.64 & -12.74 & 25.71 \end{pmatrix}.$$

Then, by (5.34),

$$F^2 = \frac{(12)(11)}{12+12} (\mathbf{y}_1 - \mathbf{f}_2)' C (CS_{\mathbf{y}_2}C')^{-1} C(\mathbf{y}_1 - \mathbf{f}_2) = 14.798.$$

Upon comparison of this value with $F_{10,1,20}^2 = 12.798$ obtained by interpolation in Table A.7, we reject the hypothesis of parallelism.

In Figure 5.8 the lack of parallelism is most notable in the second and third segments. This can also be seen in the relatively large values of the second and third components of

$$C(\mathbf{y}_1 - \mathbf{f}_2) = \begin{pmatrix} -1.62 \\ 8.59 \\ -9.72 \end{pmatrix}.$$

To see which of these made the greatest statistical contribution, we can examine the discriminant function coefficient vector given in (5.35) as

$$\mathbf{a} = (CS_{\mathbf{y}_2}C')^{-1}C(\mathbf{y}_1 - \mathbf{f}_2) = \begin{pmatrix} -1.16 \\ .104 \\ -.593 \end{pmatrix}.$$

Thus the third segment contributed most to rejection in the presence of the other two segments.

To test for equal levels, $H_{02}: \mathbf{f}(\boldsymbol{\mu}_1) = \mathbf{f}(\boldsymbol{\mu}_2)$, we use (5.36),

$$\begin{aligned} F &= \frac{f(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{\sqrt{\frac{1}{f} S_{\mathbf{y}_2} (1/n_1 + 1/n_2)}} \\ &= \frac{10.269}{\sqrt{(14.278)(1/12 + 1/12)}} = 5.2657. \end{aligned}$$

Comparing this with $F_{10,2,21} = 2.658$, we reject the hypothesis of equal levels.

To test the flatness hypothesis, $H_{03}: \frac{1}{2}C(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{f}_0$ we first calculate

$$\bar{\mathbf{y}} = \frac{\frac{12}{12}\bar{\mathbf{y}}_1 + \frac{12}{12}\bar{\mathbf{y}}_2}{12+12} = \frac{\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2}{2} = \begin{pmatrix} 14.16 \\ 14.91 \\ 11.81 \\ 12.34 \end{pmatrix}.$$

Using

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

we obtain, by (3.10),

$$T^2 = (52 + 32)(CY)'(CY_0C')^{-1}CY = 154.006,$$

which exceeds $F_{0.05,1,60}^2 = 12.785$, so we reject the hypothesis of balance. However, since the parallelism hypothesis was rejected, a more appropriate approach would be to test each of the two groups separately for balance using the test of Section 5.5.1. By (5.33), we obtain

$$T^2 = n_1(CY_1)'(CY_0C')^{-1}(CY_1) = 123.128,$$

$$T^2 = n_2(CY_2)'(CY_0C')^{-1}(CY_2) = 103.483.$$

Both of these exceed $F_{0.05,1,30}^2 = 14.626$, and we have significant lack of balance. $\square$

PROBLEMS

- 5.1 Show that the characteristic form of T^2 in (5.8) is the same as the original form in (5.3).
- 5.2 Show that the T^2 -statistic in (5.9) can be expressed in the characteristic form given in (5.10).
- 5.3 Show that $F^2(\mathbf{a}) = T^2$, where $\mathbf{a}$ is given by (5.13), T^2 is given by (5.9), and $\mathbf{a} = \mathbf{S}_0^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$ as in (5.14).
- 5.4 Show that the paired observation t -test in (5.21), $t = \bar{\mathbf{d}}'(\mathbf{S}_D/c)\bar{\mathbf{d}}$, has the t_{n-1} distribution.
- 5.5 Show that $\mathbf{d}_0' = \sum_{i=1}^n (\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})' = \mathbf{d}_1' + \mathbf{d}_2' - 2\mathbf{d}_{12}$, as in a comparison of (5.22) and (5.28).
- 5.6 Show that $T^2 = n\bar{\mathbf{d}}'\mathbf{S}_D^{-1}\bar{\mathbf{d}}$ in (5.24) has the characteristic form $T^2 = \bar{\mathbf{d}}'(\mathbf{S}_D/c)^{-1}\bar{\mathbf{d}}$.
- 5.7 Use (5.7) to show that $T^2(CY)$ in (5.29) can be converted to T^2 as in (5.10).
- 5.8 Show that the first statistic in (5.50) for additional information is a slope and beyond $\mathbf{y}$ has an F -distribution by solving for R^2 in terms of T^2 from (5.28) and substituting this into (5.31).
- 5.9 In Section 5.3.2, show that under H_{00} and H_{01} , $E(CY) = \mathbf{0}$ and $\text{cov}(\bar{\mathbf{y}}) = \bar{\mathbf{\Sigma}}_0/(n_1 + n_2)$, where $\bar{\mathbf{y}} = (n_1\bar{\mathbf{y}}_1 + n_2\bar{\mathbf{y}}_2)/(n_1 + n_2)$ and $\bar{\mathbf{\Sigma}}$ is the common covariance matrix of the two populations from which $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are sampled.

5.10 Verify that $T^2 = (n_1 + n_2)(XY)'(X\Omega_p X')^{-1}XY$ in (5.35) has the T^2_{p-1, n_1+n_2-1} distribution.

5.11 Test $H_0: \mu^T = (5, 11)$ using the data

$$\mathbf{Y} = \begin{pmatrix} 3 & 10 \\ 6 & 11 \\ 8 & 14 \\ 10 & 9 \end{pmatrix}.$$

5.12 Use the probe/word data in Table 3.5:

- Test $H_0: \mu = (30, 25, 40, 25, 30)'$.
- If H_0 is rejected, test each variable separately, using (5.3).

5.13 For the probe/word data in Table 3.5, test $H_0: \mu_1 = \mu_2 = \cdots = \mu_5$ using T^2 in (5.33).

5.14 Use the rumex leaf data in Table 3.6:

- Test $H_0: \mu = (44, 49, 59, 31)'$.
- If H_0 is rejected, test each variable separately, using (5.3).

5.15 For the rumex leaf data in Table 3.6, test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ using T^2 in (5.33).

5.16 Four measurements were made on two species of the beetle *Labidion* (1952). The variables were

- y_1 = distance of antennae grooves from posterior border of pronotum (μm),
- y_2 = length of elytra (in 10 mm),
- y_3 = length of second antennal joint (μm),
- y_4 = length of third antennal joint (μm).

The data are given in Table 3.5.

- Test $H_0: \mu_1 = \mu_2$ using T^2 .
- If the T^2 -test in part (a) rejects H_0 , carry out a t -test on each variable, as in (5.12).
- Calculate the discriminant function coefficient vector $\mathbf{a} = \Sigma_p^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$.
- Show that if the vector $\mathbf{a}$ found in part (c) is substituted into r^2 (a) from (5.15), the result is the same as the value of T^2 found in part (a).
- Obtain T^2 using the regression approach in Section 5.6.2.
- Test the significance of each variable adjusted for the other three.
- Test the significance of y_3 and y_4 adjusted for y_1 and y_2 .

5.17 Carry out a profile analysis on the beetle data in Table 3.5.

Table 5.5. Four Measurements on Two Species of Fishes

Mollus. elongata					Mollus. undulatus				
Experiment Number	x_1	x_2	x_3	x_4	Experiment Number	x_1	x_2	x_3	x_4
1	199	345	137	156	1	393	305	194	206
2	192	360	152	147	2	156	137	133	190
3	117	256	141	150	3	364	300	156	254
4	111	309	145	119	4	171	171	162	211
5	171	189	128	136	5	393	297	169	224
6	192	262	147	173	6	393	308	160	223
7	213	278	136	203	7	177	301	166	223
8	192	235	128	195	8	198	308	141	197
9	170	344	126	150	9	390	286	146	214
10	201	256	146	195	10	177	269	171	150
11	159	345	138	150	11	176	177	156	211
12	168	363	147	150	12	393	311	156	206
13	186	332	151	167	13	176	283	141	200
14	192	283	138	190	14	368	297	162	214
15	200	294	138	198	15	364	263	147	192
16	192	277	150	177	16	393	308	157	204
17	200	297	136	175	17	392	276	154	206
18	391	255	146	195	18	393	176	149	216
19	151	267	141	156	19	171	171	140	150
					20	197	303	170	208

5.16 Twenty engineers' aptitudes and 20 pilots were given six tests (Gross 1959). The variables were

- x_1 = intelligence,
- x_2 = form relations,
- x_3 = dynamic motion,
- x_4 = dexterity,
- x_5 = sensory motor coordination,
- x_6 = generalization.

The data are given in Table 5.6.

(a) Test $H_0: \mu_1 = \mu_2$.

(b) If the F^L -test in part (a) rejects H_0 , carry out a t -test for each variable, as in (5.12).

(c) Test each variable adjusted for the other five.

(d) Test the significance of $\beta_4, \beta_5, \beta_6$ adjusted for $\beta_1, \beta_2, \beta_3$.

Table 5.6. Comparison of this Data on Engineer Apprentices and Pilots

Engineer Apprentices						Pilots					
X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6
125	22	74	225	54	234	132	17	77	202	50	289
108	30	80	175	40	300	123	32	79	192	44	345
122	40	87	306	41	313	129	30	86	258	55	349
77	57	85	178	80	304	131	23	87	204	48	340
168	58	71	173	38	303	118	34	88	209	42	288
168	37	87	241	39	349	47	23	87	218	60	317
124	39	82	194	71	342	129	32	87	207	39	328
138	36	89	300	83	342	129	29	102	204	58	300
149	50	94	198	30	277	130	26	104	256	58	290
129	34	72	162	47	308	147	47	82	240	50	322
154	57	87	190	60	344	124	37	88	227	58	347
145	50	88	308	51	338	133	40	83	214	50	308
112	40	80	252	29	279	100	35	83	183	57	242
128	39	75	180	39	333	144	37	84	207	39	280
118	31	80	182	88	333	149	38	78	208	62	271
141	42	80	195	36	343	123	27	89	203	46	294
135	49	75	152	42	349	136	30	83	257	34	341
151	57	90	225	54	308	87	36	100	252	30	325
97	48	85	104	31	343	143	37	105	258	37	245
107	42	82	190	27	307	104	32	78	187	50	284

5.19 Data were collected in an attempt to find a screening procedure to detect carriers of Duchenne muscular dystrophy, a disease transmitted from female carriers to some of their male offspring (Archons and Horberg, 1988, pp. 273–278). The following variables were measured on a sample of noncarriers and a sample of carriers:

X_1 = age,

X_2 = month in which measurements are taken,

X_3 = creatine kinase,

X_4 = hemoglobin,

X_5 = lactate dehydrogenase,

X_6 = pyruvate kinase.

The data are given in Table 5.7.

(a) Test $H_0: \mu_1 = \mu_2$ using X_3 , X_4 , X_5 , and X_6 .

(b) The variables X_3 and X_4 are relatively inexpensive to measure compared to X_5 and X_6 . Do X_3 and X_4 contribute an important amount to T^2 above and beyond X_3 and X_4 ?

Table 5.5. Comparison of Curries and Recoveries of Molecular Hydrogen

[Table not available in this electronic edition.]

- (a) The levels of y_1 , y_2 , y_3 , and y_4 may depend on age and income, x_1 and x_2 . Do y_1 and y_2 contribute a significant amount to R^2 when adjusted for y_3 , y_4 , x_1 and x_2 ?

8.20 Various aspects of economic cycles were measured for consumers' goods and producers' goods by Tinber (1945). The variables are

- y_1 = length of cycle,
 y_2 = percentage of rising prices,
 y_3 = cyclical amplitude,
 y_4 = rate of change.

The data for several items are given in Table 5.8.

Table 5.8. Cyclical Measurements of Consumer Goods and Producer Goods

Item	y_1	y_2	y_3	y_4	Item	y_1	y_2	y_3	y_4
Consumer Goods					Producer Goods				
1	72	58	8	.8	1	31	37	12.8	.8
2	69.5	48	13	1.0	2	100	56	17	.5
3	74	57	14	1.0	3	100	52	16.5	.7
4	67	68	15	.9	4	100.5	65	20.5	.9
5	44	57	14	.9	5	74	51	18	.6
6	41	52	18	1.0	6	78.5	59	18	1.2
7	34.5	58	4	.8	7	48	60	21	1.6
8	34.5	48	8.5	1.0	8	118	46	25.5	1.4
9	28	54	3	1.2	9	84	66	13	.8
					10	108	55	17	1.8

- (a) Test $H_0: \rho_1 = \rho_2$ using F^2 .
 (b) Calculate the discriminant function coefficient matrix.
 (c) Test for significance of each variable adjusted for the other three.

8.21 Each of 15 students wrote an informal and a formal essay (Bromer 1971, p. 189). The variables recorded were the number of words and the number of verbs.

- y_1 = number of words in the informal essay,
 y_2 = number of verbs in the informal essay,
 x_1 = number of words in the formal essay,
 x_2 = number of verbs in the formal essay.

Table 2.8. Number of Words and Number of Vectors

Student	Inflected		Formal		$\alpha_1 = \beta_1 = \alpha_2$	$\alpha_2 = \beta_2 = \alpha_1$
	Words	Vectors	Words	Vectors		
	β_1	β_2	α_1	α_2		
1	148	20	137	15	+11	+5
2	139	24	164	25	-3	-1
3	144	19	154	27	-40	-4
4	104	18	208	16	-104	-13
5	121	17	178	24	-57	-7
6	89	11	128	20	-39	-9
7	119	17	154	18	-35	-1
8	123	13	158	16	-35	-3
9	76	16	102	21	-26	-5
10	117	29	164	25	+3	+4
11	148	10	209	24	-61	-2
12	171	21	151	16	5	+5
13	83	7	123	13	-40	-6
14	113	20	161	22	-48	-2
15	178	13	173	23	+3	-10

Table 2.9. Survival Times for Breast-Cancer Patients and Matched Controls

Breast-Cancer Patients		Matched Controls	
β_1	β_2	α_1	α_2
81	79	72	33
401	420	134	18
28	15	94	28
436	450	58	58
340	87	48	10
106	113	143	49
83	50	113	38
64	50	90	24
105	113	30	18
121	36	260	24
166	126	136	28
37	27	47	27
113	213	66	32
128	138	100	27
72	39	113	38
240	218	198	60

The data are given in Table 3.5. Since each student wrote both types of essays, the observation vectors are paired, and we use the paired comparison test.

- (a) Test $H_0: \mu_d = 0$.
- (b) Find the discriminant function coefficient vector.
- (c) Do a univariate t -test on each d_j .

5.21 A number of patients with bronchus cancer were treated with mesothane and compared with matched patients who received no mesothane (Cameron and Pauling 1958). The data are given in Table 3.10. The variables measured were

$x_1 - x_1 =$ survival time (days) from date of first hospital admission,

$x_2 - x_2 =$ survival time from date of metastability.

Compare μ_1 and μ_2 with n_1 and n_2 using a paired comparison T^2 test.

5.22 Use the glucose data in Table 3.8:

- (a) Test $H_0: \mu_Y = \mu_X$ using a paired comparison test.
- (b) Test the significance of each variable adjusted for the other two.

CHAPTER 6

Multivariate Analysis of Variance

In this chapter we extend univariate analysis of variance to multivariate analysis of variance, in which we measure more than one variable on each experimental unit. For multivariate analysis of covariance, see Rencher (1995, Section 4.10).

6.1 ONE-WAY MODELS

We begin with a review of univariate analysis of variance (ANOVA) before covering multivariate analysis of variance (MANOVA) with several dependent variables.

6.1.1 Univariate One-Way Analysis of Variance (ANOVA)

In the balanced one-way ANOVA, we have a random sample of n observations from each of k normal populations with equal variances, as in the following layout:

	Sample 1 from $N(\mu_1, \sigma^2)$	Sample 2 from $N(\mu_2, \sigma^2)$	...	Sample k from $N(\mu_k, \sigma^2)$
	x_{11}	x_{12}	...	x_{1k}
	x_{21}	x_{22}	...	x_{2k}
	$\vdots$	$\vdots$		$\vdots$
	x_{n1}	x_{n2}	...	x_{nk}
Total	T_1	T_2	...	T_k
Mean	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
Variance	s_1^2	s_2^2	...	s_k^2

The k samples or the populations from which they arise are sometimes referred to as *groups*. The groups may correspond to treatments applied by the researcher in an experiment. We have used the “tilde” notation for totals and means for each group:

$$T_i = \sum_{j=1}^n x_{ji}, \quad \bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ji}, \quad (6.1)$$

The k samples are assumed to be independent. The assumptions of independence and constant variance are necessary to obtain an F -test.

The model for each observation is

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\ &= \mu_i + \epsilon_{ij}, \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n, \end{aligned} \quad (3.2)$$

where $\mu_i = \mu + \alpha_i$ is the mean of the i th population. We wish to compare the sample means $\bar{Y}_i$, $i = 1, 2, \dots, k$, to see if they are sufficiently different to lead one to believe the populations means differ. The hypothesis can be expressed as $H_0: \mu_1 = \mu_2 = \dots = \mu_k$. Note that the notation for subscripts differs from that of previous chapters, in which the subscript i represented the observation. In this chapter, we use the first subscript in a model such as (3.2) to represent the observation.

If the hypothesis is true, all Y_{ij} are from the same population, $N(\mu, \sigma^2)$, and we can obtain two estimates of σ^2 , one based on the sample variances $s_1^2, s_2^2, \dots, s_k^2$ [see (3.4) and (3.5)] and the other based on the sample means $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_k$. The pooled "within-sample" estimator of σ^2 is

$$s_p^2 = \frac{1}{k} \sum_{i=1}^k s_i^2 = \frac{\sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2}{k(n-1)} = \frac{\sum_{i=1}^k Y_{ij}^2 - \sum_{i=1}^k n\bar{Y}_i^2}{k(n-1)} \quad (3.3)$$

Our second estimate of σ^2 (under H_0) is based on the variance of the sample means,

$$s_b^2 = \frac{\sum_{i=1}^k n\bar{Y}_i^2 - \bar{Y}^2}{k-1}, \quad (3.4)$$

where $\bar{Y} = \sum_{i=1}^k \bar{Y}_i/k$ is the overall mean. If H_0 is true, s_b^2 estimates σ^2/n [see remarks following (3.7) in Section 3.1], and therefore $E(s_b^2) = \sigma^2/n^2(n) = \sigma^2$, from which the estimate of σ^2 is

$$ns_b^2 = \frac{n \sum_{i=1}^k n\bar{Y}_i^2 - \bar{Y}^2}{k-1} = \frac{\sum_{i=1}^k Y_i^2/n - \bar{Y}^2/n}{k-1}, \quad (3.5)$$

where $Y_i = \sum_{j=1}^n Y_{ij} = \sum_{j=1}^n Y_{ij}$ is the overall total. If H_0 is false, $E(ns_b^2) = \sigma^2 + n \sum_{i=1}^k \alpha_i^2/(k-1)$, and ns_b^2 will tend to reflect a larger spread in $Y_1, Y_2, \dots, Y_k$. Since s_p^2 is based on variability within each sample, it estimates σ^2 whether or not H_0 is true; thus $E(s_p^2) = \sigma^2$ in either case.

When sampling from normal distributions, s_p^2 , a pooled estimator based on the k values of s_i^2 , is independent of s_b^2 , which is based on the $\bar{Y}_i$'s. We can justify this assertion by noting that Y_{ij} and s_i^2 are independent in each sample (when sampling from the normal distribution) and that the k samples are independent of each other.

Since s_x^2 and s_y^2 are independent and both estimate σ^2 , their ratio forms an F -statistic (see Section 3.3.1)

$$\begin{aligned} F &= \frac{s_x^2}{s_y^2} = \frac{(\sum_{i=1}^n x_i^2/n - \bar{x}^2/n) / (n-1)}{(\sum_{i=1}^n y_i^2/n - \bar{y}^2/n) / (n-1)} \\ &= \frac{SSB/(k-1)}{SSW/(n-k)} \end{aligned} \quad (5.6)$$

$$= \frac{SSB}{MSB} \quad (5.7)$$

where $SSB = \sum_{i=1}^n x_i^2/n - \bar{x}^2/n$ and $SSW = \sum_{i=1}^n y_i^2/n - \bar{y}^2/n$ are the "between"-sample sum of squares (due to the means) and "within"-sample sum of squares, respectively, and MSB and MSW are the corresponding sample meansquares. The F -statistic (5.6) is distributed as $F_{(k-1), (n-k)}$ when H_0 is true. We reject H_0 if $F > F_\alpha$. The F -statistic (5.6) can be shown to be a simple function of the likelihood ratio.

5.1.2 Multivariate One-Way Analysis of Variance Model (MANOVA)

We often measure several dependent variables on each experimental unit instead of just one variable. In the multivariate case, we assume that k independent random samples of size n are obtained from p -variate normal populations with equal covariance matrices, as in the following layout for balanced one-way multivariate analysis of variance. (In practice, the observations vectors $\mathbf{y}_{ij}$ would ordinarily be listed in row form, and sample 2 would appear below sample 1, and so on. See, for example, Table 6.2.)

	Sample 1 from $N_p(\mu_1, \Sigma)$	Sample 2 from $N_p(\mu_2, \Sigma)$	...	Sample k from $N_p(\mu_k, \Sigma)$
	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	...	$\mathbf{y}_{1k}$
	$\mathbf{y}_{21}$	$\mathbf{y}_{22}$	...	$\mathbf{y}_{2k}$
	$\vdots$	$\vdots$		$\vdots$
	$\mathbf{y}_{n1}$	$\mathbf{y}_{n2}$	...	$\mathbf{y}_{nk}$
Total	$\mathbf{T}_1$	$\mathbf{T}_2$	...	$\mathbf{T}_k$
Mean	$\bar{\mathbf{y}}_1$	$\bar{\mathbf{y}}_2$	...	$\bar{\mathbf{y}}_k$

Totals and means are defined as follows:

$$\text{Total of the } i\text{th sample: } \mathbf{T}_i = \sum_{j=1}^n \mathbf{y}_{ij}$$

$$\text{Overall total: } \mathbf{T}_\cdot = \sum_{i=1}^k \sum_{j=1}^n \mathbf{y}_{ij}$$

$$\text{Mean of the } i\text{th sample: } \bar{\mathbf{y}}_i = \mathbf{T}_i/n$$

$$\text{Overall mean: } \bar{\mathbf{y}}_\cdot = \mathbf{T}_\cdot/kn$$

The model for each observation vector is

$$\begin{aligned} \mathbf{y}_{ij} &= \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\alpha}_{ij} \\ &= \boldsymbol{\mu} + \boldsymbol{\alpha}_{ij}, \quad i = 1, 2, \dots, k, \quad j = 1, 2, \dots, n. \end{aligned} \quad (5.8)$$

In terms of the p variables in $\mathbf{y}_{ij}$, (5.8) becomes

$$\begin{pmatrix} y_{ij1} \\ y_{ij2} \\ \vdots \\ y_{ijp} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{ip} \end{pmatrix} = \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{ip} \end{pmatrix} + \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \\ \vdots \\ \mu_{jp} \end{pmatrix},$$

so that the model for the r th variable ($r = 1, 2, \dots, p$) in each vector $\mathbf{y}_{ij}$ is

$$y_{ijr} = \mu_r + \alpha_{ir} + \alpha_{jpr} = \mu_r + \alpha_{jpr}.$$

We wish to compare the mean vectors of the k samples for significant differences. The hypothesis is, therefore,

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_k \quad \text{vs.} \quad H_1: \text{at least two } \boldsymbol{\mu}\text{'s are unequal.}$$

Equality of the mean vectors implies that the k means are equal for each variable, that is, $\mu_{1r} = \mu_{2r} = \dots = \mu_{kr}$ for $r = 1, 2, \dots, p$. If two means differ for just one variable, for example, $\mu_{1p} \neq \mu_{2p}$, then H_0 is false and we wish to reject it. We can see this by examining the elements of the population mean vectors:

$$H_0: \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix} = \dots = \begin{pmatrix} \mu_{k1} \\ \mu_{k2} \\ \vdots \\ \mu_{kp} \end{pmatrix}.$$

Thus H_0 implies p sets of equalities:

$$\begin{aligned} \mu_{11} &= \mu_{21} = \dots = \mu_{k1}, \\ \mu_{12} &= \mu_{22} = \dots = \mu_{k2}, \\ &\vdots \\ \mu_{1p} &= \mu_{2p} = \dots = \mu_{kp}. \end{aligned}$$

All $p(k-1)$ equalities must hold for H_0 to be true; failure of only one equality will falsify the hypothesis.

In the next section, we have "between" and "within" sums of squares SS_B and SS_E. If $\mu_1, \mu_2, \dots, \mu_k$, these are given by

$$\begin{aligned} \text{SSB} &= n \sum_{i=1}^k (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^k \frac{\bar{X}_{i.}^2}{n} - \frac{\bar{X}_{..}^2}{n}, \\ \text{SSB} &= \sum_{i=1}^k \sum_{j=1}^p (\bar{X}_{ij.} - \bar{X}_{..})^2 = \sum_{j=1}^p \bar{X}_{.j.}^2 - \sum_{j=1}^p \frac{\bar{X}_{.j.}^2}{n}. \end{aligned}$$

By analogy, in the multivariate case, we have "between" and "within" matrices **B** and **W**, defined as

$$\begin{aligned} \mathbf{B} &= n \sum_{i=1}^k (\bar{\mathbf{X}}_{i.} - \bar{\mathbf{X}}_{..})(\bar{\mathbf{X}}_{i.} - \bar{\mathbf{X}}_{..})' \\ &= \sum_{i=1}^k \frac{1}{n} \mathbf{P}_i \mathbf{R}_i' - \frac{1}{n^2} \mathbf{P} \mathbf{R}', \end{aligned} \quad (5.8)$$

$$\begin{aligned} \mathbf{W} &= \sum_{i=1}^k \sum_{j=1}^{r_i} (\mathbf{X}_{ij.} - \bar{\mathbf{X}}_{i.})(\mathbf{X}_{ij.} - \bar{\mathbf{X}}_{i.})' \\ &= \sum_{j=1}^p \mathbf{X}_{.j.} \mathbf{X}_{.j.}' - \sum_{j=1}^p \frac{1}{n} \mathbf{P}_j \mathbf{R}_j'. \end{aligned} \quad (5.9)$$

The $p \times p$ "between" matrix **B** has a between sum of squares on the diagonal for each of the p variables. Off-diagonal elements are analogous sums of products for each pair of variables. Assuming there are no linear dependencies in the variables, the rank of **B** is the smaller of p and v_B , where v_B , $v_W \leq k$, where v_B represents the degrees of freedom for hypothesis, in the one-way case $v_B = k - 1$. Thus **B** can be singular. The $p \times p$ "error" matrix **W** has a within sum of squares for each variable on the diagonal, with analogous sums of products off-diagonal. The rank of **W** is p , unless v_W is less than p .

Thus **B** has the form

$$\mathbf{B} = \begin{pmatrix} \text{SSB}_{11} & \text{SPB}_{12} & \cdots & \text{SPB}_{1p} \\ \text{SPB}_{21} & \text{SSB}_{22} & \cdots & \text{SPB}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \text{SPB}_{p1} & \text{SPB}_{p2} & \cdots & \text{SSB}_{pp} \end{pmatrix}, \quad (5.10)$$

where, for example,

$$\begin{aligned} \text{SSB}_{22} &= n \sum_{i=1}^k (\bar{X}_{i2} - \bar{X}_{..})^2 = \sum_{i=1}^k \frac{\bar{X}_{i2}^2}{n} - \frac{\bar{X}_{..}^2}{n}, \\ \text{SPB}_{22} &= n \sum_{i=1}^k (\bar{X}_{i1} - \bar{X}_{..})(\bar{X}_{i2} - \bar{X}_{..}) = \sum_{i=1}^k \frac{\bar{X}_{i1} \bar{X}_{i2}}{n} - \frac{\bar{X}_{..} \bar{X}_{..}}{n}. \end{aligned}$$

In these expressions, the subscript 1 or 2 indicates the first or second variable. Thus, for example, $\bar{Y}_{11}$ is the second element in $\bar{\mathbf{Y}}_1$:

$$\bar{\mathbf{Y}}_1 = \begin{pmatrix} \bar{Y}_{11} \\ \bar{Y}_{12} \\ \vdots \\ \bar{Y}_{1p} \end{pmatrix}.$$

The matrix $\mathbf{E}$ can be expressed in a form similar to (8.11):

$$\mathbf{E} = \begin{pmatrix} SSP_{11} & SSP_{12} & \cdots & SSP_{1p} \\ SSP_{21} & SSP_{22} & \cdots & SSP_{2p} \\ \vdots & \vdots & & \vdots \\ SSP_{p1} & SSP_{p2} & \cdots & SSP_{pp} \end{pmatrix}, \quad (8.12)$$

where, for example,

$$SSP_{11} = \sum_{i=1}^k \sum_{j=1}^p (y_{ij1} - \bar{Y}_{11})^2 = \sum_{ij} r_{ij1}^2 - \sum_i \frac{r_{i1}^2}{n},$$

$$SSP_{12} = \sum_{i=1}^k \sum_{j=1}^p (y_{ij1} - \bar{Y}_{11})(y_{ij2} - \bar{Y}_{12}) = \sum_{ij} r_{ij1} r_{ij2} - \sum_i \frac{r_{i1} r_{i2}}{n}.$$

Note that the elements of $\mathbf{E}$ are sums-of-squares and products, not variances and covariances. To estimate Σ , we use $\hat{\Sigma}_{12} = \mathbf{E}/(nI - 1)$, so that

$$E\left(\frac{\mathbf{E}}{nI - 1}\right) = \Sigma.$$

8.1.3 Wilks' Test Statistic

The likelihood ratio test of $H_0: \mu_1 = \mu_2 = \cdots = \mu_p$ is given by

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{B}|}, \quad (8.13)$$

which is known as Wilks' Λ . (It has also been called Wilks' U .) We reject H_0 if $\Lambda \leq \Lambda_{\alpha, p, r, n-p-r}$. Note that rejection is for small values of Λ . Exact critical values $\Lambda_{\alpha, p, r, n-p-r}$ for Wilks' Λ are found in Table A.8, taken from Wall (1967). The parameters in Wilks' Λ distribution are

p = number of variables (dimension),

r_0 = degrees-of-freedom for hypothesis,

r_2 = degrees-of-freedom for error.

Wilks' Λ compares the within sum of squares and products matrix $\mathbf{W}$ to the total sum of squares and products matrix $\mathbf{E} + \mathbf{W}$. This is similar to the univariate F -statistic in (3.6) that compares the between sum of squares to the within sum of squares. By using determinants, the test statistic Λ is reduced to a scalar. Thus the multivariate information in $\mathbf{E}$ and $\mathbf{W}$ about separation of mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_p$ is distilled into a single scalar, through which we can determine if the separation of mean vectors is significant. This is typical of multivariate tests in general.

The mean vectors occupy a space of dimension $r = \min(p, v_y)$, and within this space various configurations of these mean vectors are possible. This suggests the possibility that another test statistic may be more powerful than Wilks' Λ . Competing test statistics are discussed in Sections 5.1.4 and 5.1.5.

Some of the properties and characteristics of Wilks' Λ are as follows:

1. In order for the determinants in (5.13) to be positive, it is necessary that $v_y \geq p$.
2. For any MANOVA model, the degrees of freedom v_B and v_W are always the same as in the analogous univariate case. In the balanced one-way model, for example, $v_B = k - 1$ and $v_W = d(p - 1)$.
3. The parameters p and v_W can be interchanged; the distribution of Λ_{p, v_W, v_B} is the same as that of $\Lambda_{v_W, p, v_B + v_W - p}$.
4. Wilks' Λ in (5.13) can be expressed in terms of the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_r$ of $\mathbf{E}^{-1}\mathbf{W}$ as follows:

$$\Lambda = \prod_{i=1}^r \frac{1}{1 + \lambda_i}. \quad (5.14)$$

The number of nonzero eigenvalues of $\mathbf{E}^{-1}\mathbf{W}$ is $r = \min(p, v_y)$, which is the rank of $\mathbf{W}$. The matrix $\mathbf{W}\mathbf{E}^{-1}$ has the same eigenvalues as $\mathbf{E}^{-1}\mathbf{W}$ (see Section 2.11.5) and could be used in its place to obtain Λ . However, we prefer $\mathbf{E}^{-1}\mathbf{W}$ because we will use its eigenvalues later.

5. The range of Λ is $0 \leq \Lambda \leq 1$, and the test based on Wilks' Λ is an inverse test: in the sense that we reject H_0 for small values of Λ . If the sample mean vectors were equal, we would have $\mathbf{W} = \mathbf{O}$ and $\Lambda = |\mathbf{W}|/|\mathbf{E} + \mathbf{O}| = 0$. On the other hand, as the sample mean vectors become more widely spread apart compared to the within-sample variation, $\mathbf{W}$ becomes much "larger" than $\mathbf{E}$, and Λ approaches zero.
6. In Table A.5, the critical values decrease for increasing p . Thus the addition of variables will reduce the power unless the variables contribute to rejection of the hypothesis by producing a significant reduction in Λ .
7. When $v_y = 1$ or 2 or when $p = 1$ or 2, Wilks' Λ transforms to an exact F -statistic. The transformations from Λ to F for these special cases are given in Table 5.1. The hypothesis is rejected when the transformed value of Λ exceeds

Table 6.1. Transformations of Wilks' Λ to Exact Upper Tail F -Tests

Parameter	Statistic Having F -Distribution	Degrees of Freedom
Any $p, v_E = 1$	$\frac{1 - \beta \sqrt{v_E - p + 1}}{\beta}$	$p, v_E - p + 1$
Any $p, v_E = 2$	$\frac{1 - \sqrt{\beta} \sqrt{v_E - p + 1}}{\sqrt{\beta}}$	$2p, 2(v_E - p + 1)$
$p = 1, \text{any } v_E$	$\frac{1 - \beta \sqrt{v_E}}{\beta}$	v_E, v_E
$p = 2, \text{any } v_E$	$\frac{1 - \sqrt{\beta} \sqrt{v_E - 1}}{\sqrt{\beta}}$	$2v_E, 2(v_E - 1)$

the upper α -level percentage point of the F -distribution, with degrees of freedom as shown.

8. For values of p and v_E other than those in Table 6.1, an approximate F -statistic is given by

$$F = \frac{1 - \beta_1^{1/2} \frac{d_1}{d_2}}{\beta_1^{1/2} \frac{d_1}{d_2}}, \quad (6.17)$$

with d_1 and d_2 degrees of freedom, where

$$\begin{aligned} d_1 &= p v_E, & d_2 &= m - \frac{1}{2}(p v_E - 2), \\ m &= v_E + v_E - \frac{1}{2}(p + v_E + 1), & t &= \sqrt{\frac{p^2 v_E^2 - 4}{p^2 + v_E^2 - 5}}. \end{aligned}$$

When $p v_E = 2$, t is set equal to 1. The approximate F in (6.17) reduces to the exact F -values given in Table 6.1, when either v_E or p is 1 or 2.

A (less accurate) approximate test is given by

$$\chi^2 = -1 v_E - \frac{1}{2}(p + v_E + 1) \ln \Lambda, \quad (6.18)$$

which has an approximate χ^2 -distribution with $p v_E$ degrees of freedom. We reject H_0 if $\chi^2 > \chi_{\alpha}^2$. This approximation is accurate to three decimal places when $p^2 + v_E^2 \leq \frac{1}{2} F$, where $F = v_E - \frac{1}{2}(p + v_E + 1)$.

9. If the multivariate test based on Λ rejects H_0 , it could be followed by an F -test, as in (6.5) instead of the p individual t 's. We can formulate a hypothesis comparing the means across the k groups for each variable, namely, $\mu_{11}, \mu_{12}, \dots, \mu_{1p}, \mu_{21}, \mu_{22}, \dots, \mu_{2p}, \dots, \mu_{k1}, \mu_{k2}, \dots, \mu_{kp}$. It does not necessarily follow that any

of the F -tests on the p individual variables will reject the corresponding H_0 . Conversely, it is possible that one or more of the F 's will reject H_0 when the Λ -test accepts H_0 . In either case, where the multivariate test and the univariate tests disagree, we use the multivariate test result rather than the univariate results. This is similar to the relationship between χ^2 -tests and t -tests shown in Figure 5.2.

In the three bivariate samples plotted in Figure 5.1, we illustrate the case where Λ rejects H_0 : $\mu_1 = \mu_2 = \mu_3$, but the F 's accept both of H_{01} : $\mu_{11} = \mu_{21} = \mu_{31}$, $r = 1, 2$, that is, for y_1 and y_2 . There is no significant separation of the three samples in either the y_1 or y_2 direction alone. Other follow-up procedures are given in Sections 5.1.6 and 5.4.

18. The Wilks' Λ -test is the likelihood ratio test. Other approaches to test construction lead to different tests. These such tests are given in Sections 5.1.4 and 5.1.5.

5.1.4 Roy's Test

In the scalar-invariant approach, we seek the linear combination $z_0 = a'y_0$ that maximizes the spread of the transformed means $\bar{z}_0 = a'\bar{y}_0$ relative to the within-sample spread of points. Thus we seek the vector a that maximizes

$$J = \frac{a' \sum_{i=1}^g \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 (n-1)}{\sum_{i=1}^g \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 (n-1)}. \quad (5.13)$$

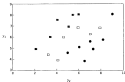


Figure 5.1. Three samples with significant Wilks' Λ test and significant F 's.

which, by analogy to $\sigma_1^2 = \sigma^2 \ln(1.55)$, can be written as

$$F = \frac{\mathbf{a}'_1 \mathbf{H} \mathbf{a}_1 / (k-1)}{\mathbf{a}'_1 \mathbf{E} \mathbf{a}_1 / (kn-k)}. \quad (5.18)$$

This is maximized by $\mathbf{a}_1$, the eigenvector corresponding to λ_1 , the largest eigenvalue of $\mathbf{K}^{-1} \mathbf{H}$ (see Section 5.3.2), and we have

$$\max_{\mathbf{a}} F = \frac{\mathbf{a}'_1 \mathbf{H} \mathbf{a}_1 / (k-1)}{\mathbf{a}'_1 \mathbf{E} \mathbf{a}_1 / (kn-k)} = \frac{d(k-1)}{d-1} \lambda_1. \quad (5.19)$$

Since $\max_{\mathbf{a}} F$ in (5.19) is maximized over all possible linear functions, it no longer has an F -distribution. To test $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ based on λ_1 , we use Roy's union-intersection test, also called Roy's largest-root test. The test statistic is given by

$$\theta = \frac{\lambda_1}{1 + \lambda_1}. \quad (5.20)$$

Critical values for θ are given in Table A.10 (Fleissner and Hartley, 1972; Pillai, 1964, 1965). We reject $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ if $\theta \geq \theta_{\alpha, d, k, p}$. The parameters α , m , and N are defined as

$$d = \min(r_H, F), \quad m = \frac{1}{2}(r_H - p + 1), \quad N = \frac{1}{2}(r_E - p + 1).$$

For $p = 1$, use (5.34) and (5.37) in Section 5.1.7 to obtain an F -test.

The eigenvector $\mathbf{a}_1$ corresponding to λ_1 is used in the discriminant function, $z = \mathbf{a}'_1 \mathbf{y}$. Since this is the function that best separates the transformed means $\mathbf{y}_i = \mathbf{a}'_1 \mathbf{y}_i$, $i = 1, 2, \dots, k$ (relative to the within-sample spread, see (5.17)), the coefficients $a_{11}, a_{12}, \dots, a_{1p}$ in the linear combination $z = \mathbf{a}'_1 \mathbf{y}$ can be examined for an indication of which variable contributes most to separating the means. The discriminant function is discussed further in Sections 5.3.3 and 6.4 and in Chapter 6.

We do not have a satisfactory F -approximation for θ or λ_1 , but an "upper bound" on F that is provided in some software programs is given by

$$F = \frac{(m-d-1)\lambda_1}{d}, \quad (5.21)$$

with degrees of freedom d and $r_E - d - 1$, where $d = \min(p, r_H)$. The term upper bound indicates that the F in (5.21) is greater than the "true F ", that is, $F > F_{E, m-d-1}$. Therefore, we test with H_0 accepted by (5.21), but if rejection of H_0 is indicated, the information is virtually worthless.

Some computer programs do not provide eigenvalues of nonsymmetric matrices, such as $\mathbf{K}^{-1} \mathbf{H}$. However, the eigenvalues of $\mathbf{K}^{-1} \mathbf{H}$ are the same as the eigenvalues of the symmetric matrices $(\mathbf{H}^{1/2})^{-1} \mathbf{H} \mathbf{H}^{1/2})^{-1}$ and $(\mathbf{U})^{-1} \mathbf{H} \mathbf{U}^{-1}$, where $\mathbf{H}^{1/2}$ is the square root matrix of $\mathbf{H}$ given in (2.112) and $\mathbf{U} \mathbf{U}' = \mathbf{E}$ is the Cholesky factorization

of $\mathbf{H}$ (Section 3.7). We demonstrate this for the Cholesky approach. We first multiply the defining relationship $(\mathbf{E}^{-1}\mathbf{H} - \lambda)\mathbf{u} = \mathbf{0}$ by $\mathbf{E}$ to obtain

$$\mathbf{H} - \lambda\mathbf{E}\mathbf{u} = \mathbf{0}. \quad (6.22)$$

Then substituting $\mathbf{E} = \mathbf{U}\mathbf{U}'$ into (6.22), multiplying by $(\mathbf{U}')^{-1}$, and inserting $\mathbf{U}^{-1}\mathbf{U} = \mathbf{I}$, we have

$$\begin{aligned} (\mathbf{H} - \lambda\mathbf{U}\mathbf{U}')\mathbf{u} &= \mathbf{0}, \\ (\mathbf{U}')^{-1}(\mathbf{H} - \lambda\mathbf{U}\mathbf{U}')\mathbf{u} &= (\mathbf{U}')^{-1}\mathbf{0} = \mathbf{0}, \\ [(\mathbf{U}')^{-1}\mathbf{H} - \lambda\mathbf{I}](\mathbf{U}')^{-1}\mathbf{u} &= \mathbf{0}, \\ [(\mathbf{U}')^{-1}\mathbf{H}\mathbf{U}^{-1} - \lambda\mathbf{I}]\mathbf{u} &= \mathbf{0}. \end{aligned} \quad (6.23)$$

Thus $(\mathbf{U}')^{-1}\mathbf{H}\mathbf{U}^{-1}$ has the same eigenvalues as $\mathbf{E}^{-1}\mathbf{H}$ and has eigenvectors of the form $\mathbf{U}\mathbf{u}$, where $\mathbf{u}$ is an eigenvector of $\mathbf{E}^{-1}\mathbf{H}$. Note that $(\mathbf{U}')^{-1}\mathbf{H}\mathbf{U}^{-1}$ is positive semidefinite, and thus $\lambda_i \geq 0$ for all eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$.

6.1.5 Pillai and Lawley-Hotelling Tests

There are two additional test statistics for $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ based on the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_r$ of $\mathbf{E}^{-1}\mathbf{H}$. The Pillai statistic is given by

$$V^{(P)} = n\{(\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}\} = \sum_{i=1}^r \frac{\lambda_i}{1 + \lambda_i}. \quad (6.24)$$

We reject H_0 for $V^{(P)} \geq V_{\alpha}^{(P)}$. The upper percentage points, $V_{\alpha}^{(P)}$, are given in Table A.11 (Schuurmans, Kriekemink, and Chaturvedi, 1975), indexed by n , m , and k , which are defined as in Section 6.1.1 for Roy's test. For $r = 1$, use (6.24) and (6.22) in Section 6.1.7 to obtain an F -test.

Pillai's test statistic in (6.24) is an extension of Roy's statistic $\theta = \lambda_1/(1 + \lambda_1)$. If the mean vectors do not lie in one dimension, the information in the additional terms $\lambda_i/(1 + \lambda_i)$, $i = 2, 3, \dots, r$, may be helpful in rejecting H_0 .

For parameter values not included in Table A.11, we can use an approximate F -statistic:

$$F_1 = \frac{(2N + r + 1)V^{(P)}}{(2m + r + 1)(r - 1)V_1}, \quad (6.25)$$

which is approximately distributed as $F_{(r-1), (2m+r+1)(r-1), 1}$. Two alternative F -approximations are given by

$$F_2 = \frac{n'V_1 - r_1 + nV^{(P)}}{p_1m(r - 1)V_1}, \quad (6.26)$$

with prq and $r(q - r + 1)$ degrees of freedom, and

$$F_3 = \frac{(rq - p + r)F^{(3)}}{d(q - F^{(3)})}, \quad (5.27)$$

with rd and $r(q - p + r)$ degrees of freedom, where $d = \min(p, rq)$. It can be shown that F_3 in (5.27) is the same as F_3 in (5.25).

The Lawley-Hotelling statistic (Lawley 1938, Hotelling 1951) is defined as

$$L^{(3)} = \text{tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^d \lambda_i \quad (5.28)$$

and is also known as Hotelling's generalized F^2 -statistic (see a comment at the end of Section 6.1.7). Table A.12 (Davis 1970a, b, 1980) gives upper percentage points of the test statistic

$$\frac{\text{SE}_L^{(3)}}{v_d}.$$

We reject H_0 for large values of the test statistic. Note that in Table A.12, $p \leq rq$ and $p \leq rq$. If $p = rq$, use $(rq, p, rq + rq - p)$ instead of (p, rq, rq) . (Cyclicality pattern in the parameters is found in 'Wiles'.) If you properly 3 in Section 5.1.3.4. If $rq = 1$ and $p > 1$, use the relationship $L^{(3)} = F^{(2)}/v_d$ (see (5.36) in Section 5.1.3). For other values of the parameters not included in Table A.12, we can use an approximate F -statistic

$$F_1 = \frac{L^{(3)}}{c}, \quad (5.29)$$

which is approximately distributed as $F_{a,b}$, where

$$\begin{aligned} a &= prq, & b &= 4 + \frac{a+2}{b-1}, & c &= \frac{a(b-2)}{b(rq-p-1)}, \\ b &= \frac{(rq + rq - p - 1)(rq - 1)}{(rq - p - 2)(rq - p)}. \end{aligned}$$

Alternative F -approximations are given by

$$F_2 = \frac{2c(N+1)F^{(3)}}{F^2(2n+n+1)}, \quad (5.30)$$

with $c(2n+n+1)$ and $2c(N+1)$ degrees of freedom, and

$$F_3 = \frac{(r(rq - rq - 1) + 2)F^{(3)}}{4prq}, \quad (5.31)$$

with prq and $r(r_1 + r_2 + \cdots + r_k - 1)$ degrees of freedom. If $p = r_1$, then F_2 in (6.34) is the same as F_2 in (6.30).

6.1.6 Unbalanced One-Way MANOVA

The balanced one-way model can easily be extended to the unbalanced case, in which there are n_{ij} observation vectors in the i th group. The model in (5.8) becomes

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_{ij}, \quad i = 1, 2, \dots, k, \quad j = 1, 2, \dots, n_{ij}.$$

The mean vectors become $\bar{\mathbf{Y}}_i = \sum_{j=1}^{n_{ij}} \mathbf{X}_{ij}/n_i$ and $\bar{\mathbf{Y}}_{..} = \sum_{i=1}^k \sum_{j=1}^{n_{ij}} \mathbf{X}_{ij}/N$, where $N = \sum_{i=1}^k n_i$. Similarly, the total vectors are defined as $\mathbf{y}_i = \sum_{j=1}^{n_{ij}} \mathbf{x}_{ij}$ and $\mathbf{y}_{..} = \sum_{i,j} \mathbf{x}_{ij}$. The $\mathbf{B}$ and $\mathbf{E}$ matrices are calculated as

$$\mathbf{B} = \sum_{i=1}^k n_i (\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}}_{..})' = \sum_{i=1}^k \frac{1}{n_i} \mathbf{y}_i \mathbf{y}_i' - \frac{1}{N} \mathbf{y}_{..} \mathbf{y}_{..}', \quad (6.32)$$

$$\mathbf{E} = \sum_{i=1}^k \sum_{j=1}^{n_{ij}} (\mathbf{X}_{ij} - \bar{\mathbf{Y}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{Y}}_i)' = \sum_{i=1}^k \sum_{j=1}^{n_{ij}} \mathbf{X}_{ij} \mathbf{X}_{ij}' - \sum_{i=1}^k \frac{1}{n_i} \mathbf{y}_i \mathbf{y}_i'. \quad (6.33)$$

Wilks' Λ and the other tests have the same form as in Sections 6.1.3, 6.1.4, 6.1.5, using $\mathbf{B}$ and $\mathbf{E}$ from (6.32) and (6.33). In each test we have

$$r_B = k - 1, \quad r_E = N - k = \sum_{i=1}^k n_i - k.$$

Note that $N = \sum_{i=1}^k n_i$ differs from R used as a parameter in Roy's and Pillai's tests in Sections 6.1.4 and 6.1.5.

6.1.7 Summary of the Four Tests and Relationship to Γ^2

We compare the four test statistics in terms of the eigenvalues $\lambda_1 = \lambda_2 = \cdots = \lambda_r$ of $\mathbf{E}^{-1}\mathbf{B}$, where $r = \min(r_B, p)$:

$$\text{Pillai: } \Gamma^{(p)} = \sum_{i=1}^r \frac{\lambda_i}{1 + \lambda_i},$$

$$\text{Lawley-Hotelling: } U^{(p)} = \sum_{i=1}^r \lambda_i,$$

$$\text{Wilks' lambda: } \Lambda = \prod_{i=1}^r \frac{1}{1 + \lambda_i},$$

$$\text{Roy's largest root: } \theta = \frac{\lambda_1}{1 + \lambda_1}.$$

Note that for all four tests we must have $v_0 = p$. As noted in Section 5.1.3 and elsewhere, p is the number of variables, v_0 is the degrees of freedom for the hypothesis, and v_2 is the degrees of freedom for error.

Why do we use four different tests? All four are exact tests; that is, when H_0 is true, each test has probability α of rejecting H_0 . However, the tests are not equivalent, and in a given sample they may lead to different conclusions even when H_0 is true; some may reject H_0 while others accept H_0 . This is due to the multidimensional nature of the space in which the mean vectors $\mu_1, \mu_2, \dots, \mu_k$ lie. A comparison of power and other properties of the tests is given in Section 6.2.

When $v_2 = 1$, then r is also equal to 1, and there is only one nonzero eigenvalue λ_1 . In this case, all four test statistics are functions of each other and give equivalent results. In terms of F , for example, the other three become

$$L^{(H)} = L_1 = \frac{t}{1 - t}, \quad (5.34)$$

$$L^{(S)} = t, \quad (5.35)$$

$$L = 1 - t. \quad (5.36)$$

In the case of $v_2 = 1$, all four statistics can be transformed to an exact F using

$$F = \frac{v_0 - p + 1}{p} L^{(H)}, \quad (5.37)$$

which is distributed as $F_{p, v_0 - p + 1}$.

The equivalence of all four test statistics to Hotelling's T^2 when $v_2 = 1$ was noted in Section 5.5.1. We now demonstrate the relationship $T^2 = (n_1 + n_2 - 2)L^{(H)}$ in (5.17). For $\mathbf{H}$ and $\mathbf{B}$, we use (5.32) and (5.33), which after canceling n_1 (since we do not require $n_1 = n_2$ in T^2). In this case, with only two groups, $\mathbf{H} = \sum_{i=1}^2 n_i (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_*)'(\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_*)'$ can be expressed as

$$\mathbf{H} = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' = c(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)', \quad (5.38)$$

where $c = n_1 n_2 / (n_1 + n_2)$. Then by (5.34) and (5.38), $L^{(H)}$ becomes

$$\begin{aligned} L^{(H)} &= \lambda_1 = \lambda(\mathbf{K}^{-1}\mathbf{H}) \\ &= n_1 c \mathbf{K}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \quad \text{by (5.38)} \\ &= c n_1 (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \mathbf{K}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \quad \text{by (2.97)} \\ &= \frac{c}{n_1 + n_2 - 2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \left(\frac{\mathbf{B}}{n_1 + n_2 - 2} \right)^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \\ &= \frac{T^2}{n_1 + n_2 - 2}. \end{aligned} \quad (5.39)$$

since $E(\alpha_1 + \alpha_2 + \alpha_3) = S_{00}$ (see Section 5.4.2). Equations (5.16), (5.18), and (5.19) follow immediately from this result using (5.14)–(5.16).

Because of the direct relationship in (5.20) between $\hat{\sigma}^{(2)}$ and T^2 for the case of two groups, the Lawley–Hotelling statistic $\hat{\sigma}^{(2)}$ is often called the generalized T^2 -statistic.

Example 6.2.9. In a classical experiment carried out from 1918 to 1934, apple trees of different varieties were compared (Anderson and Hurlbert, 1965, pp. 367–369). The data for eight trees from each of six varieties are given in Table 6.2. The variables are

y_1 = trunk girth at 4 years (mm $\times 100$),

y_2 = extension growth at 4 years (in),

y_3 = trunk girth at 15 years (mm $\times 100$),

y_4 = weight of tree above ground at 15 years (lb $\times 100$).

The matrices $\mathbf{H}$, $\mathbf{E}$, and $\mathbf{E} + \mathbf{H}$ are given by

$$\mathbf{H} = \begin{pmatrix} .076 & .071 & .312 & .208 \\ .057 & 4.280 & 2.335 & 1.637 \\ .312 & 2.335 & 6.114 & 3.781 \\ .208 & 1.637 & 3.781 & 2.693 \end{pmatrix},$$

$$\mathbf{E} = \begin{pmatrix} .328 & 1.697 & .594 & .217 \\ 1.697 & 12.346 & 4.364 & 2.110 \\ .594 & 4.364 & 4.281 & 2.692 \\ .217 & 2.110 & 2.482 & 1.725 \end{pmatrix},$$

$$\mathbf{E} + \mathbf{H} = \begin{pmatrix} .404 & 2.234 & .886 & .426 \\ 2.234 & 16.342 & 6.719 & 3.747 \\ .886 & 6.719 & 10.405 & 6.250 \\ .426 & 3.747 & 6.263 & 4.215 \end{pmatrix}.$$

In this case, the mean vectors represent six points in four-dimensional space. We can compare the mean vectors for significant differences using ‘Wilks’ Λ as given by (6.13):

$$\Lambda = \frac{|\mathbf{H}|}{|\mathbf{E} + \mathbf{H}|} = \frac{.0576}{0.2567} = .224.$$

In this case, the parameters of the ‘Wilks’ Λ distribution are $p = 4$, $v_0 = 6 - 1 = 5$, and $v_2 = 6(5 - 1) = 24$. We reject $H_0: \mu_1 = \mu_2 = \dots = \mu_6$ because

$$\Lambda = .224 < \lambda_{.05,4,5,24} = .455.$$

Table S.2. Receptor Data

[Table not available for this electronic edition.]

Note the use of $v_2 = 40$ in place of $v_2 = 42$. This is a conservative approach that allows a table value to be used without interpolation.

To obtain an approximate F , we first calculate

$$\begin{aligned}t &= \sqrt{\frac{\rho^2 v_2^2 - 4}{\rho^2 + v_2^2 - 3}} = \sqrt{\frac{4^2 50^2 - 4}{4^2 + 50^2 - 3}} = 3.3166, \\w &= v_2 + v_3 = \frac{1}{2}(p + v_2 + 1) = 42 + 3 = \frac{1}{2}(4 + 3 + 1) = 42, \\d_{11} &= p v_2 = 4(5) = 20, \quad d_{12} = w - \frac{1}{2}(p v_2 - 1) = 39.5.\end{aligned}$$

Then the approximate F is given by (6.15),

$$F = \frac{1 - F^{(1)} \frac{d_{11}}{d_{12}}}{F^{(2)} \frac{d_{11}}{d_{12}}} = \frac{1 - (.344)^{(3.3166)} \frac{150.3}{39}}{(.154)^{(3.3166)} \frac{150.3}{39}} = 4.937,$$

which exceeds $F_{.05, 15.1, 39} = 2.53$, and we reject H_0 .

The four eigenvalues of $\mathbf{E}^{-1}\mathbf{B}$ are 1.876, .791, .225, and .005. With these we can calculate the other three test statistics. For Pillai's statistic we have, by (6.24),

$$V^{(4)} = \sum_{i=1}^4 \frac{\lambda_i}{1 + \lambda_i} = 1.505.$$

To find a critical value for $V^{(4)}$ in Table A.11, we need

$$\begin{aligned}r &= \min(v_2, p) = 4, \quad m = \frac{1}{2}(v_2 - p + 1) = 0, \\B &= \frac{1}{2}(v_2 - p + 1) = 39.5.\end{aligned}$$

Then $V_{.05}^{(4)} = .645$ (by interpolation). Since $1.505 > .645$, we reject H_0 .

For the Lawley-Hotelling statistic we obtain, by (6.28),

$$U^{(4)} = \sum_{i=1}^4 \lambda_i = 2.821.$$

To make the test, we calculate the test statistic

$$\frac{v_2}{v_3} U^{(4)} = \frac{40}{3} (2.821) = 34.508.$$

The .05 critical value for $v_2 U^{(4)}/v_3$ is given in Table A.12 as 7.8188 (using $v_2 = 40$), and we therefore reject H_0 .

Fey's test statistic is given by (5.28) as

$$F = \frac{s_1}{1 + s_1} = \frac{1.876}{1 + 1.876} = .681,$$

which exceeds the .05 critical value .377 obtained (by interpolation) from Table A.10, and we reject H_0 . $\square$

6.1.5 Measures of Multivariate Association

In multiple regression, a measure of association between the dependent variable y and the independent variables $x_1, x_2, \dots, x_k$ is given by the squared multiple correlation

$$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}}. \quad (5.40)$$

Similarly, in one-way univariate ANOVA, Fisher's correlation ratio η^2 is defined as

$$\eta^2 = \frac{\text{between sum of squares}}{\text{total sum of squares}}.$$

This is a measure of model fit similar to R^2 and gives the proportion of variation in the dependent variable y attributable to differences among the means of the groups. It answers the question, How well can we predict y by knowing what group it comes from? Thus η^2 can be considered to be a measure of association between the dependent variable y and the grouping variable i associated with μ_i or α_i in the model (5.2). In fact, if the grouping variable is represented by $k - 1$ dummy variables (also called indicators, or categorical variables), then we have a dependent variable related to several independent variables as in multiple regression.

A dummy variable takes on the value 1 for sampling units in a group (sample) and 0 for all other sampling units. (Values other than 0 and 1 could be used.) Thus for k samples (groups), the $k - 1$ dummy variables are

$$x_i = \begin{cases} 1 & \text{if sampling unit is in } i\text{th group,} \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, 2, \dots, k - 1.$$

Only $k - 1$ dummy variables are needed because if $x_1 = x_2 = \dots = x_{k-1} = 0$, the sampling unit must be from the k th group (see Section 11.4.2 for an illustration). The dependent variable y can be regressed on the $k - 1$ dummy variables $x_1, x_2, \dots, x_{k-1}$ to produce results equivalent to the usual ANOVA calculations.

In (one-way) MANOVA, we need to measure the strength of the associations between several dependent variables and several independent (grouping) variables. Various measurements of multivariate association have been proposed. Wilks (1952) suggested a "generalized η^2 ":

$$\text{MANOVA } \eta^2 = \eta_{1k}^2 = 1 - \Lambda, \quad (5.41)$$

based on the use of $\mathbf{B}$ and $\mathbf{B} + \mathbf{B}$ as generalizations of sums of squares. We use $1 - \lambda$ because λ is small if the spread in the means is large.

We now consider an η^2 based on Roy's statistic, R . We noted in Section 5.1.4 that the discriminant function is the linear function $\mathbf{z} = \mathbf{a}_1' \mathbf{p}$ that maximizes the spread of the means $\bar{z}_i = \mathbf{a}_1' \bar{\mathbf{x}}_i$, $i = 1, 2, \dots, J$, where $\mathbf{a}_1$ is the eigenvector of $\mathbf{B}^{-1}\mathbf{B}$ corresponding to the largest eigenvalue λ_1 . We measure the spread among the means by $\text{SSB} = n \sum_{i=1}^J \bar{z}_i^2 - J \bar{z}^2$, divided by the within-sample spread $\text{SSW} = \sum_{i,j} z_{ij}^2 - J \bar{z}^2$. The maximum value of this ratio is given by λ_1 (see (8.19)). Thus

$$\lambda_1 = \frac{\text{SSB}(\lambda_1)}{\text{SSW}(\lambda_1)},$$

and by (8.20),

$$\theta^2 = \frac{\lambda_1}{1 + \lambda_1} = \frac{\text{SSB}(\lambda_1)}{\text{SSB}(\lambda_1) + \text{SSW}(\lambda_1)}. \quad (8.42)$$

Hence θ serves directly as a measure of multivariate association.

$$\eta_1^2 = \theta = \frac{\lambda_1}{1 + \lambda_1}. \quad (8.43)$$

It can be shown that the square root of this quantity,

$$\eta_1 = \sqrt{\frac{\lambda_1}{1 + \lambda_1}}, \quad (8.44)$$

is the maximum correlation between a linear combination of the p -dependent variables and a linear combination of the $J - 1$ dummy group variables (see Section 11.3.1). This type of correlation is often called a *canonical correlation* (see Chapter 11) and is defined for each eigenvalue $\lambda_1, \lambda_2, \dots, \lambda_p$ as $\eta_i (\lambda_i) = \sqrt{\lambda_i / (1 + \lambda_i)}$.

We now consider some measures of multivariate association suggested by Cramer and Nicolson (1979) and Muller and Peterson (1984). It is easily shown (Section 11.6.2) that λ can be expressed as

$$\lambda = \prod_{i=1}^J \frac{1}{1 + \lambda_i} = \prod_{i=1}^J (1 - \eta_i^2), \quad (8.45)$$

where $\eta_i^2 = \lambda_i / (1 + \lambda_i)$ is the i th squared canonical correlation described earlier. The geometric mean of a set of positive numbers $a_1, a_2, \dots, a_k$ is defined as $(a_1 a_2 \cdots a_k)^{1/k}$. Thus $\lambda^{1/J}$ is the geometric mean of the $(1 - \eta_i^2)$'s, and another measure of multivariate association based on λ , in addition to that in (8.42), is

$$\lambda_2 = 1 - \lambda^{1/J}. \quad (8.46)$$

In fact, as noted by Muller and Peterson, the F -approximation given in (5.15),

$$F = \frac{1 - R^{1/2}}{R^{1/2}} \frac{B_1}{B_1'},$$

is very similar to the univariate F -statistic (3.4) (for testing significance in multiple regression),

$$F = \frac{R^2/(k \text{ df model})}{(1 - R^2)/(n \text{ df error})}, \quad (5.47)$$

based on R^2 in (3.43).

Pillai's statistic is easily expressible as the sum of the squared canonical correlations:

$$V^{(k)} = \sum_{i=1}^k \frac{\lambda_i}{1 + \lambda_i} = \sum_{i=1}^k r_i^2, \quad (5.48)$$

and the average of the r_i^2 can be used as a measure of multivariate association:

$$A_F = \frac{\sum_{i=1}^k r_i^2}{p} = \frac{V^{(k)}}{p}. \quad (5.49)$$

In terms of A_F the F -approximation given in (5.26) becomes

$$F_2 = \frac{A_F(p \alpha)}{(1 - A_F(p \alpha) r_F - r_F + 1)}, \quad (5.50)$$

which has an obvious parallel to (5.47).

For the Lawley-Hotelling statistic $V^{(k)}$, a multivariate measure of association can be defined as

$$A_{LH} = \frac{V^{(k)} \alpha}{[1 + V^{(k)} \alpha]}. \quad (5.51)$$

If $p = 1$, (5.51) reduces to (5.45). In fact, (5.45) is a special case of (5.51) because $V^{(k)}(x) = \sum_{i=1}^k \lambda_i(x)$ is the arithmetic average of the λ_i 's. It is easily seen that the F -approximation F_2 for the Lawley-Hotelling statistic given in (5.31) can be expressed in terms of A_{LH} as

$$F_2 = \frac{A_{LH} p \alpha}{(1 - A_{LH} (1/r_F) r_F - r_F + 1/r_F)}, \quad (5.52)$$

which resembles (5.47).

Example 6.1.8. We illustrate some measures of association for the root-stock data in Table 6.1:

$$\begin{aligned}\kappa_{12}^2 &= 1 - \beta = .945, \\ \eta_{12}^2 &= \theta = .892, \\ A_{12} &= 1 - \beta^{1/4} = 1 - (.156)^{1/4} = .374, \\ A_{\theta} &= 1^{1/4}/4 = 1.262/4 = .316, \\ A_{\eta\theta} &= \frac{0^{1/4}/4}{1 + 0^{1/4}/4} = \frac{2.821/4}{1 + 2.821/4} = .422.\end{aligned}$$

There is a wide range of values among these measures of association. $\square$

6.2 COMPARISON OF THE FOUR MANOVA TEST STATISTICS

When $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ is true, all the main vectors are at the same point. Therefore, all four MANOVA test statistics have the same Type I error rate, α , as noted in Section 6.1.7; that is, all have the same probability of rejection when H_0 is true. However, when H_0 is false, the four tests have different probabilities of rejection. We noted in Section 6.1.7 that in a given sample the four tests need not agree, even if H_0 is true. One test could reject H_0 and the others accept H_0 , for example.

Historically, Wilks' Λ has played the dominant role in significance tests in MANOVA because it was the first to be derived and has well-known χ^2 and F approximations. It can also be partitioned in certain ways we will find useful later. However, it is not always the most powerful among the four tests. The probability of rejecting H_0 when it is false is known as the power of the test.

In univariate ANOVA with $p = 1$, the means $\mu_1, \mu_2, \dots, \mu_k$ can be uniquely ordered along a line in one dimension, and the usual F -test is uniformly most powerful. In the multivariate case, on the other hand, with $p > 1$, the mean vectors are points in $x = \min(p, n-1)$ dimensions. We have four tests, not one of which is uniformly most powerful. The relative power of the four test statistics depends on the configuration of the mean vectors $\mu_1, \mu_2, \dots, \mu_k$ in the x -dimensional space. A given test will be more powerful for one configuration of mean vectors than another.

If $n_k < p$, then $x = n_k$ and the mean vectors lie in an x -dimensional subspace of the p -dimensional space of the observations. The points may, in fact, occupy a subspace of the x dimensions. For example, they may be confined to a line (one dimension) or a plane (two-dimensional). This is illustrated in Figure 6.2.

An indication of the pattern of the mean vectors is given by the eigenvalues of $K^{-1}B$. If there is one large eigenvalue and the others are small, the mean vectors lie close to a line. If there are two large eigenvalues, the mean vectors lie mostly in two dimensions, and so on.

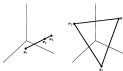


Figure 4.1. Two possible configurations for three mean vectors in 3-space.

Because Roy's test uses only the largest eigenvalue of $K^{-1}H$, it is more powerful than the others if the mean vectors are collinear. The other three tests have greater power than Roy's when the mean vectors are diffuse (spread out in several dimensions).

In terms of power, the tests are ordered $P \geq U^{(1)} \geq A \geq U^{(2)}$ for the collinear case. In the diffuse case and for intermediate structure between collinear and diffuse, the ordering of power is reversed, $U^{(2)} \geq A \geq U^{(1)} \geq P$. The latter ordering also holds for accuracy of the Type I error rate when the population covariance matrices $\Sigma_1, \Sigma_2, \dots, \Sigma_k$ are not equal. These orderings are comparisons of power. For actual computation of power in a given experimental setting or to find the sample size needed to yield a desired level of power, see Kowshik (1999, Section 3.4).

Generally, if group sizes are equal, the tests are sufficiently robust with respect to heterogeneity of covariance matrices so that we need not worry. If the n_i 's are unequal and we have heterogeneity, then the ordering of the MANCOV test may be affected as follows. If the larger variances and covariances are associated with the larger samples, the true α -level is reduced and the tests become conservative. On the other hand, if the larger variances and covariances come from the smaller samples, α is inflated, and the tests become liberal. Box's M -test in Section 3.3.2 can be used to test for heterogeneity of covariance matrices.

In conclusion, the use of Roy's P is not recommended in any situation, except the collinear case under standard assumptions. In the diffuse case its performance is inferior to that of the other three, both when the assumptions hold and when they do not. If the data come from nonnormal populations exhibiting skewness or positive kurtosis, any of the other three tests perform acceptably well. Among these three, $U^{(2)}$ is superior to the other two when there is heterogeneity of covariance matrices. Indeed $U^{(2)}$ is first in all settings except those for the collinear case. However, A is not far behind, except when there is severe heterogeneity of covariance matrices. It seems likely that Wilks' A will continue its dominant role because of its flexi-

bility and historical precedence. (For references for this section, see Rencher (1995, Section 4.2).)

In practice, most MANOVA software programs routinely calculate all four test statistics, and they usually reach the same conclusion. In those cases when they differ as to acceptance or rejection of the hypothesis, one can examine the eigenvalues and covariance matrices and evaluate the conflicting conclusions in light of the test properties discussed previously.

Example 6.2. We inspect the eigenvalues of $K^{-1}B$ for the roostack data of Table 6.1 for an indication of the configuration of the six mean vectors in a four-dimensional space. The eigenvalues are 1.876, .791, .229, .028. The first eigenvalue, 1.876, constitutes a proportion

$$\frac{1.876}{1.876 + .791 + .229 + .028} = .641$$

of the sum of the eigenvalues. Therefore, the first eigenvalue does not dominate the others, and the mean vectors are not collinear. The first two eigenvalues account for a proportion

$$\frac{1.876 + .791}{1.876 + \cdots + .028} = .913$$

of the sum of the eigenvalues, and thus the six mean vectors lie largely in two dimensions. Since the mean vectors are not collinear, the test statistics Λ , $L^{(1)}$, and $L^{(2)}$ will be more appropriate than P in this case. $\square$

6.2 CONSTRAINTS

As in Sections 6.1.1–6.1.5, we consider only the balanced model where $n_1 = n_2 = \cdots = n_g = n$. We begin with a review of constraints in the univariate setting before moving to the multivariate case.

6.2.1 Univariate Constraints

A *contrast* in the population means is defined as a linear combination

$$\bar{\mu} = c_1\mu_1 + c_2\mu_2 + \cdots + c_g\mu_g, \quad (6.53)$$

where the coefficients satisfy

$$\sum_{i=1}^g c_i = 0. \quad (6.54)$$

An unbiased estimator of $\bar{\mu}$ is given by

$$\hat{\bar{\mu}} = c_1\bar{y}_1 + c_2\bar{y}_2 + \cdots + c_g\bar{y}_g. \quad (6.55)$$

The sample means $\bar{Y}_k$ were defined in (5.1). Since the Y_k 's are independent with variance σ^2/n , the variance of $\bar{Y}$ is

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{n} \sum_{k=1}^d c_k^2,$$

which can be estimated by

$$s_{\bar{Y}}^2 = \frac{\text{MSB}}{n} \sum_{k=1}^d c_k^2, \quad (5.56)$$

where MSB was defined in (5.6) and (5.7) as $\text{SSB}/(n-1)$.

The usual hypothesis to be tested by a contrast is

$$H_0: \bar{Y} = c_1\mu_1 + c_2\mu_2 + \cdots + c_d\mu_d = 0.$$

For example, suppose $d = 4$ and that a contrast of interest to the researcher is $3\mu_1 - \mu_2 - \mu_3 - \mu_4$. If this contrast is set equal to zero, we have

$$3\mu_1 = \mu_2 + \mu_3 + \mu_4 \quad \text{or} \quad \mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4),$$

and the experiment is comparing the first mean with the average of the other three. A contrast is often called a *comparison among the treatment means*.

Assuming normality, $H_0: \bar{Y} = 0$ can be tested by

$$t = \frac{\bar{Y} - 0}{s_{\bar{Y}}}, \quad (5.57)$$

which is distributed as t_{n-d} . Alternatively, since $t^2 = F$, we can use

$$\begin{aligned} F &= \frac{t^2}{1} = \frac{\left(\sum_{k=1}^d c_k \bar{Y}_k\right)^2}{\text{MSB} \sum_{k=1}^d c_k^2/n} \\ &= \frac{n \sum_{k=1}^d c_k^2 \bar{Y}_k^2 / \sum_{k=1}^d c_k^2}{\text{MSB}}, \end{aligned} \quad (5.58)$$

which is distributed as $F_{1,n-d}$. The numerator of (5.58) is often referred to as the *sum of squares for the contrast*.

If two contrasts $A = \sum_{k=1}^d a_k \mu_k$ and $B = \sum_{k=1}^d b_k \mu_k$ are such that $\sum_{k=1}^d a_k b_k = 0$, the contrasts are said to be *orthogonal*. The two estimated contrasts can be written in the form $\sum_{k=1}^d a_k \bar{Y}_k = a'\bar{Y}$ and $\sum_{k=1}^d b_k \bar{Y}_k = b'\bar{Y}$, where $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_d)'$. Then $\sum_{k=1}^d a_k b_k = a'b = 0$, and by the discussion following (3.14), the coefficient vectors a and b are perpendicular.

When two contrasts are orthogonal, the two corresponding sums of squares are independent. In fact, for k treatments, we can find $k - 1$ orthogonal contrasts that partition the treatment sum of squares $SS(T)$ into $k - 1$ independent sums of squares, each with one degree of freedom. In the unbalanced case (Section 5.1.6), orthogonal contrasts such that $\sum_i a_i b_i = 0$ do not partition $SS(T)$ into $k - 1$ independent sums of squares. For a discussion of contrasts in the unbalanced case, see Hothorn (1998, Sections 4.8.2 and 4.8.3; 2000, Section 14.2.2).

5.3.1 Multivariate Contrasts

There are two types of contrasts in a multivariate setting. We have previously encountered contrasts in Section 5.3.1, where we considered the hypothesis $H_0: C\mu = 0$ with $CJ = 0$. Each row of C sums to zero, and $C\mu$ is therefore a set of contrasts comparing the elements $\mu_1, \mu_2, \dots, \mu_p$ of μ with each other. In this section, on the other hand, we consider contrasts comparing several mean vectors, not the elements within a vector.

A contrast among the population mean vectors is defined as

$$\tilde{\mu} = c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k, \quad (5.58)$$

where $\sum_{i=1}^k c_i = 0$. An unbiased estimator of $\tilde{\mu}$ is given by the corresponding contrast in the sample mean vectors:

$$\tilde{\bar{x}} = c_1\bar{x}_1 + c_2\bar{x}_2 + \dots + c_k\bar{x}_k. \quad (5.59)$$

The sample mean vectors $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ as defined in Section 5.1.2 were assumed to be independent and to have common-covariance matrix, $\text{cov}(\bar{x}_i) = \Sigma/n$. Thus the covariance matrix for $\tilde{\bar{x}}$ is given by

$$\text{cov}(\tilde{\bar{x}}) = c_1^2 \frac{\Sigma}{n} + c_2^2 \frac{\Sigma}{n} + \dots + c_k^2 \frac{\Sigma}{n} = \frac{\Sigma}{n} \sum_{i=1}^k c_i^2, \quad (5.60)$$

which can be estimated by

$$\frac{\bar{S}_p}{n} \sum_{i=1}^k c_i^2 = \left(\frac{\mathbf{K}}{n} \right) \left(\frac{\sum_{i=1}^k c_i^2}{n} \right),$$

where $\bar{S}_p = \mathbf{K}/n_p$ is an unbiased estimator of Σ .

The hypothesis $H_0: \tilde{\mu} = 0$ or $H_0: c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k = 0$ makes comparisons among the population mean vectors. For example, $\mu_1 - 2\mu_2 + \mu_3 = 0$ is equivalent to

$$\mu_2 = \frac{1}{2}(\mu_1 + \mu_3).$$

and we are computing μ_2 as the average of μ_{21} and μ_{22} . Of course this implies that every element of μ_2 must equal the corresponding element of $\frac{1}{2}(\mu_{21} + \mu_{22})$:

$$\begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(\mu_{21} + \mu_{22}) \\ \frac{1}{2}(\mu_{21} + \mu_{22}) \\ \vdots \\ \frac{1}{2}(\mu_{21} + \mu_{22}) \end{pmatrix}.$$

Under appropriate multivariate normality assumptions, $H_0: c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k = \mathbf{0}$ or $H_0: \mathbf{B}\boldsymbol{\mu} = \mathbf{0}$ can be tested with the one-sample F^T statistic

$$\begin{aligned} F^T &= \mathbf{B} \left(\frac{\mathbf{S}_{\mathbf{B}}}{n} \sum_{i=1}^k \mathbf{c}_i^2 \right)^{-1} \mathbf{B} \\ &= \frac{n}{\sum_{i=1}^k \mathbf{c}_i^2} \left(\sum_{i=1}^k \mathbf{c}_i \mathbf{B}_i \right)' \left(\frac{\mathbf{K}}{n} \right)^{-1} \left(\sum_{i=1}^k \mathbf{c}_i \mathbf{B}_i \right), \end{aligned} \quad (5.52)$$

which is distributed as $F_{k, n-k}^T$. In the one-way model under discussion here, $n_{\mathbf{B}} = k(n-1)$.

An equivalent test of H_0 can be made with Wilks' Λ . By analogy with the numerator of (5.52), the hypothesis matrix due to the contrast is given by

$$\mathbf{H}_1 = \frac{n}{\sum_{i=1}^k \mathbf{c}_i^2} \left(\sum_{i=1}^k \mathbf{c}_i \mathbf{B}_i \right) \left(\sum_{i=1}^k \mathbf{c}_i \mathbf{B}_i \right)', \quad (5.53)$$

The rank of $\mathbf{H}_1$ is 1, and the test statistic is

$$\Lambda = \frac{|\mathbf{B}|}{|\mathbf{K} + \mathbf{H}_1|}, \quad (5.54)$$

which is distributed as $\Lambda_{k, n-k, 1}$. The other three MANOVA test statistics can also be applied here using the single nonzero eigenvalue of $\mathbf{K}^{-1}\mathbf{H}_1$. Because $r_{\mathbf{B}} = 1$ in this case, all four MANOVA statistics and F^T give the same results; that is, all five transformations to the same F value using the formulations in Section 6.1.3. If 1 = 1 orthogonal contrasts are used, they partition the $\mathbf{B}$ matrix into 1 = 1 independent matrices $\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_{k-1}$. Each $\mathbf{B}_i$ matrix has one degree of freedom because $\text{rank}(\mathbf{B}_i) = 1$.

Example 6.3.2. We consider the following two orthogonal contrasts for the stock stock data in Table 6.2:

$$\begin{array}{cccccc} 2 & -1 & -1 & -1 & -1 & 2 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{array}$$

The first compares μ_1 and μ_2 with the other four mean vectors. The second compares μ_1 vs. μ_2 . Thus $H_{01}: 2\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5 + 2\mu_6 = 0$ can be written as

$$M_{01}: 2\mu_1 + 2\mu_6 = \mu_2 + \mu_3 + \mu_4 + \mu_5.$$

Dividing both sides by 4 to express this in terms of averages, we obtain

$$M_{01}: \frac{1}{2}(\mu_1 + \mu_6) = \frac{1}{4}(\mu_2 + \mu_3 + \mu_4 + \mu_5).$$

Similarly, the hypothesis for the second contrast can be expressed as

$$M_{02}: \mu_1 = \mu_2.$$

The mean vectors are given by

$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$	$\bar{X}_6$
1.14	1.05	1.11	1.10	1.08	1.04
2.98	3.03	3.05	2.88	2.96	3.11
3.24	4.93	4.46	5.01	4.31	3.60
.67	1.28	1.35	1.04	1.18	.74

For the first contrast, we obtain H_1 from (6.65) as

$$\begin{aligned} H_1 &= \frac{8}{\sum_{i=1}^6 c_i^2} \left(\sum_{i=1}^6 c_i \bar{X}_i \right) \left(\sum_{i=1}^6 c_i \bar{X}_i \right)' \\ &= \frac{8}{12} (2\bar{X}_1 - \bar{X}_2 - \cdots + 2\bar{X}_6) (2\bar{X}_1 - \bar{X}_2 - \cdots + 2\bar{X}_6)' \\ &= \frac{8}{12} \begin{pmatrix} -.005 \\ -.078 \\ -2.939 \\ -1.680 \end{pmatrix} \begin{pmatrix} -.005 & -.078 & -2.939 & -1.680 \end{pmatrix} \\ &= \begin{pmatrix} .006 & .062 & .080 & .106 \\ .062 & .078 & 1.642 & 1.079 \\ .080 & 1.642 & 4.239 & 2.829 \\ .106 & 1.079 & 2.829 & 1.881 \end{pmatrix}. \end{aligned}$$

Then

$$F = \frac{|B|}{|E + B|} = \frac{.6371}{1.4824} = .433,$$

which is less than $F_{.05,4,1,48} = .779$ from Table A.5. We therefore reject H_{01} .

To test the significance of the second contrast, we have

$$\begin{aligned} B_2 &= \frac{8}{3}(\bar{y}_{1.} - \bar{y}_{2.}, \bar{y}_{3.} - \bar{y}_{4.})' \\ &= \frac{8}{3} \begin{pmatrix} .304 \\ .360 \\ .342 \\ .336 \end{pmatrix} \quad (, .304, .360, .342, .336) \\ &= \begin{pmatrix} .641 & .308 & .636 & .655 \\ .309 & .326 & .435 & .415 \\ .656 & .405 & .681 & .678 \\ .666 & .413 & .678 & .674 \end{pmatrix}. \end{aligned}$$

Then

$$A = \frac{|B_2|}{|K + B_2|} = \frac{.6571}{.3153} = .779,$$

which is less than $\beta_{.05, 4, 1, 31} = .779$, and we reject H_{02} . (2)

6.1 TESTS ON INDIVIDUAL VARIABLES FOLLOWING REJECTION OF H_0 BY THE OVERALL MANOVA TEST

In Section 6.1, we considered tests of equality of mean vectors, $H_0: \mu_1 = \mu_2 = \dots = \mu_k$, which implies equality of means for each of the p variables:

$$H_{0r}: \mu_{1r} = \mu_{2r} = \dots = \mu_{kr}, \quad r = 1, 2, \dots, p.$$

This hypothesis could be tested for each variable by itself with an ordinary univariate ANOVA F -test, as noted in property 5 in Section 6.1.3. For example, if there are three mean vectors,

$$\mu_0 = \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{pmatrix},$$

we have $H_{01}: \mu_{01} = \mu_{11} = \mu_{21}$, $H_{02}: \mu_{02} = \mu_{12} = \mu_{22}$, $\dots$, $H_{0p}: \mu_{0p} = \mu_{1p} = \mu_{2p}$. Each of these p hypotheses can be tested with a simple ANOVA F -test.

If an F -test is made on each of the p variables regardless of whether the overall MANOVA test of $H_0: \mu_0 = \mu_1 = \mu_2$ rejects H_0 , then the overall α -level will increase beyond the nominal value because we are making p tests. As in Section 5.3, we define the overall α or experimental error rate as the probability of rejecting one or more of the p univariate tests when $H_0: \mu_0 = \mu_1 = \mu_2$ is true. We could

"protect" against inflation of the experimentwise error rate by performing tests on individual variables only if the overall MANOVA test of $H_0: \mu_1 = \mu_2 = \mu_3$ rejects H_0 . In this procedure, the probability of rejection for the tests on individual variables is reduced, and these tests become more conservative.

Randall and Snee (1966) compared these two procedures for testing the individual variables in a one-way MANOVA model. Since the focus was on α -levels, only the case where H_0 is true was considered. Specifically, the two procedures were as follows:

1. A univariate F -test is made on each variable, testing $H_{0r}: \mu_{1r} = \mu_{2r} = \cdots = \mu_{kr}$, $r = 1, 2, \dots, p$. In this context, the p univariate tests constitute an experiment, and one or more rejections are counted as one experimentwise error. No multivariate test is made.
2. The overall MANOVA hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ is tested with Wilks' Λ , and if H_0 is rejected, p univariate F -tests on $H_{01}, H_{02}, \dots, H_{0p}$ are carried out. Again, one or more rejections among the F -tests are counted as one experimentwise error.

The amount of intercorrelation among the multivariate normal variables was indicated by $\sum_{i=1}^p (1/\lambda_i)/p$, where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigenvalues of the population correlation matrix $\mathbf{P}_p$. Note that $\sum_{i=1}^p (1/\lambda_i)/p = 1$ for the uncorrelated case ($\mathbf{P}_p = \mathbf{I}$) and $\sum_{i=1}^p (1/\lambda_i)/p > 1$ for the correlated case ($\mathbf{P}_p \neq \mathbf{I}$). When the variables are highly intercorrelated, one or more of the eigenvalues will be near zero (see Section 4.1.3), and $\sum_{i=1}^p (1/\lambda_i)/p$ will be large.

The error rates of these two procedures were investigated for several values of p , n , k , and $\sum_{i=1}^p (1/\lambda_i)/p$, where p is the number of variables, n is the number of observation vectors in each group, k is the number of groups, and $\sum_{i=1}^p (1/\lambda_i)/p$ is the measure of intercorrelation defined above. In procedure 1, the probability of rejecting one or more univariate tests when H_0 is true varied from .09 to .31 (or was .05 in each test). Such experimentwise error rates are clearly unacceptable when the nominal value of α is .05. However, this approach is commonly used when the researcher is not familiar with the MANOVA approach or does not have access to appropriate software.

Table 6.3 contains the error rates for procedure 2, univariate F -tests following a rejection by Wilks' Λ . The values range from .012 to .057, consistently close to the target value of .05. No apparent trends or patterns are seen; the values do not seem to depend on p , k , n , or the amount of intercorrelation as measured by $\sum_{i=1}^p (1/\lambda_i)/p$. Thus when univariate tests are made only following a rejection of the overall test, the experimentwise error rate is about right.

Based on these results, we recommend making an overall MANOVA test followed by F -tests on the individual variables at the same α -level as the MANOVA test (only if the MANOVA test leads to rejection of H_0).

Another procedure that can be used following rejection of the MANOVA test is an examination of the discriminant function coefficients. The discriminant function was defined in Section 5.1.4 as $z = \mathbf{a}'_1 \mathbf{x}$, where $\mathbf{a}_1$ is the eigenvector asso-

Table 4.3. Empirical Power Error Rates for Procedure 2: Univariate F -Test Following Rejection by Wilks' Λ

n	p	$\sum_{i=1}^p \lambda_i / \lambda_{p+1}$							
		1		10		100		1000	
		$\alpha = 5$	$\alpha = 5$	$\alpha = 5$	$\alpha = 5$	$\alpha = 5$	$\alpha = 5$	$\alpha = 5$	$\alpha = 5$
5	5	.080	.081	.082	.083	.086	.087	.088	.089
5	5	.081	.081	.083	.083	.086	.086	.087	.088
5	7	.080	.082	.083	.083	.086	.087	.088	.089
10	5	.080	.081	.080	.083	.083	.085	.086	.087
10	5	.080	.080	.080	.083	.083	.085	.087	.088
10	7	.084	.084	.081	.083	.087	.088	.089	.090
20	5	.080	.082	.082	.084	.088	.089	.090	.091
20	5	.080	.083	.082	.083	.087	.088	.090	.091
20	7	.080	.081	.084	.086	.089	.090	.090	.091

vised with the largest eigenvalue λ_1 of $\mathbf{E}^{-1}\mathbf{H}$. Additionally, there are other discriminant functions using eigenvalues corresponding to the other eigenvalues. Since the first discriminant function maximally separates the groups, we can examine its coefficients for the contribution of each variable to group separation. Thus in $z = a_1y_1 + a_2y_2 + \cdots + a_p y_p$, if a_1 is larger than the other a_i 's, we believe y_1 contributes more than any of the other variables to separation of groups. A method of standardization of the a_i 's to adjust for differences in the scale among the variables is given in Section 8.5.

The information in a_1 , then $z = a_1' y$ about the contribution of y_1 to separation of the groups is fundamentally different from the information provided in a univariate F -test that considers y_1 alone (see property P in Section 6.1.3). The relative size of a_1 shows the contribution of y_1 in the presence of the other variables and takes into account (1) the correlation of y_1 with the others y 's and (2) the contributions of y_1 to Wilks' Λ alone and beyond the contribution of the other y 's. In contrast, the individual F -test on y_1 ignores the presence of the other variables. Because we are primarily interested in the collective behavior of the variables, the discriminant function coefficients provide more pertinent information than the tests on individual variables. For a detailed analysis of the effect of each variable in the presence of other variables, see Rencher (1995, 1998, Sections 8.1-6).

Holbert (1979) compared the (standardized) coefficients to some correlations that can be shown to be related to individual variable tests (see Section 8.7.3). In a limited simulation, the discriminant coefficients were found to be more useful than the univariate tests in identifying those variables that contribute least to separation of groups. Considerable variation was found from sample to sample in ranking the relative potency of the variables.

Example 6.4. In Example 6.1.2, the hypothesis $\mathbf{H}_0: \mu_1 = \mu_2 = \cdots = \mu_4$ was rejected for the woodcock data of Table 6.2. We can then use test the four individual

variables using the .05 level of significance. For the first variable, $y_1 = 4$ -year trunk girth, we obtain the following ANOVA table:

Source	Sum of Squares	df	Mean Square	F
Between	.87160	3	.29053	1.91
Error	.10988	42	.002616	
Total	.98148	45		

For $F = 1.91$ the p -value is .1094, and we do not reject H_0 . For the other three variables we have

Variable	F	p-Value
$y_1 = 4$ -year extension growth	1.91	.064
$y_2 = 10$ -year trunk girth	11.97	< .0001
$y_3 = 10$ -year weight	12.16	< .0001

Thus for three of the four variables, the six means differ significantly. We examine the standardized discriminant function coefficients for this set of data in Chapter 8 (Problem 8.12). $\square$

6.2 TWO-WAY CLASSIFICATION

We consider only balanced models, where each cell in the model has the same number of observations, n . For the unbalanced case with unequal cell sizes, see Raudenbush (1978, Sections 4.6).

6.2.1 Review of Univariate Two-Way ANOVAs

In the univariate two-way model, we measure one dependent variable y on each experimental unit. The balanced two-way fixed-effects model with factors A and B is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad (6.65)$$

$$= \mu_{ij} + \epsilon_{ijk} \quad (6.66)$$

$$i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J, \quad k = 1, 2, \dots, n,$$

where α_i is the effect (or μ_{ijk}) of the i th level of A , β_j is the effect of the j th level of B , γ_{ij} is the corresponding interaction effect, and μ_{ij} is the population mean for the i th level of A and the j th level of B . In order to obtain F -tests, we further assume that the ϵ_{ijk} 's are independently distributed as $N(0, \sigma^2)$.

Let $\bar{y}_{i\cdot} = \sum_{j,k} y_{ijk}/n$ be the mean at the i th level of A and define $\bar{y}_{\cdot j}$ and $\bar{y}_{\cdot\cdot}$ similarly. Then if we use side conditions $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$,

the effect of the i th level of A can be defined as $\alpha_i = \bar{y}_{i.} - \bar{y}_{..}$, with similar definitions of β_j and γ_{ij} . We can show that $\sum_i \alpha_i = 0$ if $\alpha_i = \bar{y}_{i.} - \bar{y}_{..}$ as follows:

$$\begin{aligned}\sum_{i=1}^I \alpha_i &= \sum_{i=1}^I (\bar{y}_{i.} - \bar{y}_{..}) = \sum_{i=1}^I \bar{y}_{i.} - I\bar{y}_{..} \\ &= n\bar{y}_{..} - n\bar{y}_{..} = 0.\end{aligned}$$

Many texts recommend that the interaction AB be tested first, and that if it is found to be significant, then the main effects should not be tested. However, with the side conditions imposed earlier (side conditions are not necessary in order to obtain tests), the effect of A is defined as the average effect over the levels of B , and the effect of B is defined similarly. With this definition of main effects, the tests for A and B make sense even if AB is significant. Admittedly, interpretation requires more care, and the effect of a factor may vary if the number of levels of the other factor is altered. But in many cases useful information can be gained about the main effects in the presence of interaction.

We illustrate the preceding statement that $\alpha_i = \bar{y}_{i.} - \bar{y}_{..}$ represents the effect of the i th level of A averaged over the levels of B . Suppose A has two levels and B has three. We represent the means of the six cells as follows:

		B			Mean
		1	2	3	
A	1	μ_{11}	μ_{12}	μ_{13}	$\bar{y}_{1.}$
	2	μ_{21}	μ_{22}	μ_{23}	$\bar{y}_{2.}$
Mean		$\bar{y}_{.1}$	$\bar{y}_{.2}$	$\bar{y}_{.3}$	$\bar{y}_{..}$

The means of the rows (corresponding to levels of A) and columns (levels of B) are also given. Then $\alpha_1 = \bar{y}_{1.} - \bar{y}_{..}$ can be expressed as the average of the effect of the 1st level of A at the three levels of B . For example,

$$\begin{aligned}\alpha_1 &= \frac{1}{3}(\mu_{11} - \bar{y}_{.1}) + \frac{1}{3}(\mu_{12} - \bar{y}_{.2}) + \frac{1}{3}(\mu_{13} - \bar{y}_{.3}) \\ &= \frac{1}{3}(\mu_{11} + \mu_{12} + \mu_{13}) - \frac{1}{3}(\bar{y}_{.1} + \bar{y}_{.2} + \bar{y}_{.3}) = \bar{y}_{1.} - \bar{y}_{..}.\end{aligned}$$

To estimate α_i , we can use $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$, with similar estimates for β_j and γ_{ij} . The notation $\bar{y}_{i.}$ indicates that $y_{i.j}$ is averaged over the levels of j and k to obtain the mean of all nk observations at the i th level of A , namely, $\bar{y}_{i.} = \sum_{j,k} y_{i.jk}/nk$. The means $\bar{y}_{.j}$, $\bar{y}_{.k}$, and $\bar{y}_{..}$ have analogous definitions.

To construct tests for the significance of factors A and B and the interaction AB , we use the usual sums of squares and degrees of freedom as shown in Table 8.4. Computational forms of the sums of squares can be found in many standard (university) methods texts.

Table 5.4. Univariate Two-Way Analysis of Variance

Source	Sum of Squares	df
A	$SSA = ab \sum_{i=1}^I (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$a - 1$
B	$SSB = ab \sum_{j=1}^J (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$b - 1$
$A:B$	$SSAB = a \sum_{i=1}^I (\bar{Y}_{i.} - \bar{Y}_{..})^2 + b \sum_{j=1}^J (\bar{Y}_{.j} - \bar{Y}_{..})^2$	$(a - 1)(b - 1)$
Error	$SSS = \sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{i.})^2$	$ab(b - 1)$
Total	$SST = \sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{..})^2$	$abn - 1$

The sums of squares in Table 5.4 (for the balanced model) have the relationship

$$SST = SSA + SSB + SSAB + SSS,$$

and the four sums of squares on the right are independent. The sums of squares are divided by their corresponding degrees of freedom to obtain mean squares MSA , MSB , $MSAB$, and MSS . For the fixed effects model, each of MSA , MSB , and $MSAB$ is divided by MSS to obtain an F -test. In the case of factor A , for example, the hypothesis can be expressed as

$$H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0,$$

and the test statistic is $F = MSA/MSS$, which is distributed as $F_{a-1, ab(b-1)}$.

In order to define contrasts among the levels of each main effect, we can conveniently use the model in the form given in (5.66),

$$Y_{ij} = \mu_i + \beta_j + \epsilon_{ij}.$$

A contrast among the levels of A is defined as $\sum_{i=1}^a c_i \beta_i$, where $\sum_{i=1}^a c_i = 0$. An estimate of the contrast is given by $\sum_{i=1}^a c_i \bar{Y}_{i.}$, with variance $\sigma^2 \sum_{i=1}^a c_i^2 / ab$, since each $\bar{Y}_{i.}$ is based on ab observations and the β_i 's are independent. To test $H_0: \sum_{i=1}^a c_i \beta_i = 0$, we can use an F -statistic corresponding to (5.59),

$$F = \frac{ab \sum_{i=1}^a c_i \bar{Y}_{i.}^2 / \sum_{i=1}^a c_i^2}{SSS} \quad (5.67)$$

with 1 and $ab(b-1)$ degrees of freedom. To preserve the experimentwise error rate, significant tests for more than one contrast could be conducted in the spirit of Section 5.4, that is, contrasts should be chosen prior to seeing the data, and tests should be made only if the overall F -test for factor A rejects H_{0A} .

Contrasts $\sum_{j=1}^b c_j \beta_j$ among the levels of B are tested in an entirely analogous manner.

5.5.2 Multivariate Two-Way MANOVA

A balanced two-way fixed-effects MANOVA model for p dependent variables can be expressed in vector form analogous to (5.65) and (5.66):

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} = \mu_0 + \epsilon_{ijk}, \\ i &= 1, 2, \dots, I, \quad j = 1, 2, \dots, J, \quad k = 1, 2, \dots, n, \end{aligned} \quad (5.58)$$

where α_i is the effect of the i th level of A on each of the p variables in $\mathbf{y}_{ijk}$, β_j is the effect of the j th level of B , and γ_{ij} is the AB interaction effect. We set side conditions $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ and assume the ϵ_{ijk} 's are independently distributed as $N_p(\mathbf{0}, \Sigma)$. Under the side condition $\sum_i \alpha_i = 0$, the effect of A is averaged over the levels of B , that is, $\alpha_i = \bar{\alpha}_i = \bar{\mathbf{y}}_{i.}$, where $\bar{\mathbf{y}}_{i.} = \sum_j \mathbf{y}_{ij}/b$ and $\bar{\alpha}_i = \sum_{ij} \mathbf{y}_{ij}/ab$. There are similar definitions for β_j and γ_{ij} .

As in the univariate case, the mean vector $\bar{\mathbf{y}}_{i.}$ indicates an average over the subscripts replaced by dots, that is, $\bar{\mathbf{y}}_{i.} = \sum_{jk} \mathbf{y}_{ijk}/ab$. The means $\bar{\mathbf{y}}_{.j}$, $\bar{\mathbf{y}}_{ij}$, and $\bar{\mathbf{y}}_{..}$ have analogous definitions: $\bar{\mathbf{y}}_{.j} = \sum_{ik} \mathbf{y}_{ijk}/ia$, $\bar{\mathbf{y}}_{ij} = \sum_k \mathbf{y}_{ijk}/n$, $\bar{\mathbf{y}}_{..} = \sum_{ijk} \mathbf{y}_{ijk}/iab$. The sums of squares and products matrices are given in Table 5.5. Note that the degrees of freedom in Table 5.5 are the same as in the univariate case in Table 5.4. For the two-way model with balanced data, the total sum of squares and products matrix is partitioned as

$$\mathbf{T} = \mathbf{B}_A + \mathbf{B}_B + \mathbf{B}_{AB} + \mathbf{R}. \quad (5.59)$$

The structure of any of the hypothesis matrices is similar to that of $\mathbf{B}$ in (5.11). For example, $\mathbf{B}_A$ has on the diagonal the sum of squares for factor A for each of the p variables. The off-diagonal elements of $\mathbf{B}_A$ are corresponding sums of products for all pairs of variables. Thus the rs th diagonal element of $\mathbf{B}_A$, corresponding to the r th variable, $r = 1, 2, \dots, p$, is given by

$$b_{A,r} = ab \sum_{i=1}^I (\bar{y}_{i,r} - \bar{y}_{..r})^2 = \sum_{i=1}^I \frac{b_{i,r}^2}{ab} - \frac{t_{i,r}^2}{ab}, \quad (5.70)$$

where $\bar{y}_{i,r}$ and $\bar{y}_{..r}$ represent the r th components of $\bar{\mathbf{y}}_{i.}$ and $\bar{\mathbf{y}}_{..}$, respectively, and $b_{i,r}$ and $t_{i,r}$ are totals corresponding to $\bar{y}_{i,r}$ and $\bar{y}_{..r}$. The rs th off-diagonal element of $\mathbf{B}_A$ is

$$b_{A,r} = ab \sum_{i=1}^I (\bar{y}_{i,r} - \bar{y}_{..r})(\bar{y}_{i,s} - \bar{y}_{..s}) = \sum_{i=1}^I \frac{b_{i,r}b_{i,s}}{ab} - \frac{t_{i,r}t_{i,s}}{ab}, \quad (5.71)$$

Table 5.5. Multivariate Two-Way Analysis of Variance

Source	Sum of Squares and Products Effects	df
A	$\mathbf{B}_A = ab \sum_{i=1}^I (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})'$	$a-1$
B	$\mathbf{B}_B = ab \sum_{j=1}^J (\bar{\mathbf{y}}_{.j} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{.j} - \bar{\mathbf{y}}_{..})'$	$b-1$
AB	$\mathbf{B}_{AB} = a \sum_{i=1}^I (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})' \\ \quad + (\bar{\mathbf{y}}_{.j} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{.j} - \bar{\mathbf{y}}_{..})'$	$(a-1)(b-1)$
Error	$\mathbf{R} = \sum_{ijk} (\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{i.})(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{i.})'$	$abn(a-1)$
Total	$\mathbf{T} = \sum_{ijk} (\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{..})(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{..})'$	$abn-1$

From (5.65) and Table 5.3, we obtain

$$\begin{aligned}h_{AB\tau} &= \sum_{ij} \frac{Y_{ij\tau}^2}{n} - \frac{T_{.j\tau}^2}{na^2b} - h_{A\tau} - h_{B\tau}, \\h_{AB\tau\tau} &= \sum_{ij} \frac{Y_{ij\tau}Y_{ij\tau'}}{n} - \frac{T_{.j\tau}T_{.j\tau'}}{na^2b} - h_{A\tau\tau} - h_{B\tau\tau}.\end{aligned}\quad (5.72)$$

For the $\mathbf{E}$ matrix, computational formulas are based on (5.68):

$$\mathbf{E} = \mathbf{T} - \mathbf{H}_A - \mathbf{H}_B - \mathbf{H}_{AB}.$$

Thus the elements of $\mathbf{E}$ have the form

$$\begin{aligned}e_{\tau\tau} &= \sum_{ij} T_{ij\tau}^2 - \frac{T_{.j\tau}^2}{na^2b} - h_{A\tau\tau} - h_{B\tau\tau} - h_{AB\tau\tau}, \\e_{\tau\tau'} &= \sum_{ij} T_{ij\tau}T_{ij\tau'} - \frac{T_{.j\tau}T_{.j\tau'}}{na^2b} - h_{A\tau\tau'} - h_{B\tau\tau'} - h_{AB\tau\tau'}.\end{aligned}\quad (5.73)$$

The hypothesis matrices for interaction and main effects in this fixed-effects model can be compared to $\mathbf{E}$ to make a test. Thus for 'Wilk's' Λ , we use $\mathbf{E}$ to test each of Λ , B , and AB :

$$\begin{aligned}\Lambda_A &= \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_A|} \sim \beta_{p, a-1, ab(n-1)}, \\ \Lambda_B &= \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_B|} \sim \beta_{p, b-1, ab(n-1)}, \\ \Lambda_{AB} &= \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_{AB}|} \sim \beta_{p, (a-1)(b-1), ab(n-1)}.\end{aligned}$$

In each case, the indicated distribution holds when H_0 is true. To calculate the other three MANOVA test statistics for A , B , and AB , we use the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}_A$, $\mathbf{E}^{-1}\mathbf{H}_B$, and $\mathbf{E}^{-1}\mathbf{H}_{AB}$.

If the interaction is not significant, interpretation of the main effects is simpler. However, the comments in Section 5.5.1 about testing main effects in the presence of interactions apply in the multivariate model as well. If we define each main effect as the average effect over the levels of the other factors, then main effects can be tested even if the interaction is significant. One must be more careful with the interpretation in case of a significant interaction, but there is information to be gained.

By analogy with the univariate two-way ANOVA in Section 5.5.1, a contrast among the levels of factor A can be defined in terms of the mean vectors as follows: $\sum_{i=1}^a c_i \bar{\mathbf{y}}_{i.}$, where $\sum_i c_i = 1$ and $\bar{\mathbf{y}}_{i.} = \sum_{j=1}^b \mathbf{y}_{ij}/b$. Similarly, $\sum_{j=1}^b c_j \bar{\mathbf{y}}_{.j}$ represents a contrast among the levels of B . The hypothesis that these contrasts are

$\mathbf{B}$ can be tested by F^1 or any of the four MANOVA test statistics, as in (5.52), (5.53), and (5.64). To test $M_0: \sum_{j=1}^p \alpha_j \mu_{.j} = \mathbf{0}$, for example, we construct

$$F^1 = \frac{nb}{\sum_{i=1}^n r_i^2} \left(\sum_{j=1}^p \alpha_j \bar{\mathbf{r}}_{.j} \right)' \left(\frac{\mathbf{B}}{r_0} \right)^{-1} \left(\sum_{j=1}^p \alpha_j \bar{\mathbf{r}}_{.j} \right), \quad (5.74)$$

which is distributed as $F_{p, n-p}^1$ when M_0 is true. Alternatively, the hypothesis matrix

$$\mathbf{B}_0 = \frac{nb}{\sum_{i=1}^n r_i^2} \left(\sum_{j=1}^p \alpha_j \bar{\mathbf{r}}_{.j} \right) \left(\sum_{j=1}^p \alpha_j \bar{\mathbf{r}}_{.j} \right)' \quad (5.75)$$

can be used in

$$\Lambda = \frac{|\mathbf{B}|}{|\mathbf{B} + \mathbf{B}_0|},$$

which, under M_0 , is $\Lambda_{p, 1, n-p}$ with $r_0 = nb(n-1)$ in the two-way model. The other three MANOVA test statistics can also be constructed from $\mathbf{B}^{-1}\mathbf{B}_0$. All five test statistics will give equivalent results because $r_0 = 1$.

If follow-up tests on individual variables are desired, we can infer from Renscher and Smith (1985), as reported in Section 5.4, that if the MANOVA test on factor A or B leads to rejection of M_0 , then we can proceed with the univariate F -tests on the individual variables with assurance that the experimentwise error rate will be close to α .

To determine the contribution of each variable in the presence of the others, we can examine the first discriminant function obtained from eigenvectors of $\mathbf{B}^{-1}\mathbf{B}_A$ or $\mathbf{B}^{-1}\mathbf{B}_B$, as in Section 5.4 for one-way MANOVAs. The first discriminant function for $\mathbf{B}^{-1}\mathbf{B}_A$, for example, is $z = w'y$, where w is the eigenvector associated with the largest eigenvalue of $\mathbf{B}^{-1}\mathbf{B}_A$. Let $z = \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_p y_p$. If α_i is larger than the other α 's, then y_i contributes more than the other variables to the significance of $\Lambda_{A, 1}$. (Of course, here, the α_i 's should be standardized as in Section 5.3.) Note that the first discriminant function obtained from $\mathbf{B}^{-1}\mathbf{B}_A$ will not have the same pattern as the first discriminant function from $\mathbf{B}^{-1}\mathbf{B}_B$. This is not surprising since we expect that the relative contribution of the variables to separating the levels of factor A will be different from the relative contribution to separating the levels of B .

A randomized block design or a two-way MANOVA without replication can easily be analyzed in a manner similar to that for the two-way model with replications given here; therefore, no specific details will be given.

Example 4.5.1. Table 5.1 contains data reported by Poston (1962) and analyzed by Kramer and Brown (1970). The experiment involved a 2×4 design with 4 replications, for a total of 32 observation vectors. The factors were rotational velocity [A_1 (fast) and A_2 (slow)] and lubricants (four types). The experimental units were

Table 8.6. Two-Way Classification of Measurements on Bar Steel

Lab/Steel	A_1		A_2	
	F_1	F_2	F_1	F_2
B_1	7.80	99.4	7.12	85.1
	7.15	88.8	7.86	89.8
	7.89	85.8	7.65	75.8
	7.85	88.8	7.65	77.9
B_2	8.80	82.5	8.18	86.8
	8.80	92.4	8.28	76.8
	7.65	82.4	7.65	83.1
	7.79	87.4	7.65	86.4
B_3	7.28	79.8	7.19	81.2
	8.86	89.1	7.18	72.8
	7.78	88.2	7.78	78.8
	7.89	88.8	7.65	71.8
B_4	7.89	94.1	7.88	81.2
	7.89	88.8	7.84	78.8
	7.85	89.8	7.85	86.4
	7.89	88.8	7.78	76.4

32 homogeneous pieces of bar steel. Two variables were measured on each piece of bar steel:

y_1 = ultimate tensile,

y_2 = ultimate strain.

We display the totals for each variable for use in computations. The numbers inside the box are cell totals (over the four replications), and the marginal totals are for each level of A and B :

	Totals for y_1			Totals for y_2		
	A_1	A_2		A_1	A_2	
B_1	30.60	28.08	58.68	334.8	327.8	662.6
B_2	32.80	31.34	64.14	344.7	330.0	674.7
B_3	31.78	28.45	60.23	352.8	305.0	657.8
B_4	30.22	28.32	58.54	355.4	323.8	679.2
	125.40	115.19	240.59	1487.8	1286.6	2674.4

Using computational forms for B_{12} , in (8.79), the (1, 2) element of $\mathbf{H}_B$ (corresponding to F_2) is given by

$$A_{440} = \frac{(125.40)^2 + (119.39)^2}{(4 \times 4)} = \frac{(244.59)^2}{32} = 1.208.$$

For the (2, 2) element of $\mathbb{B}_4$ (corresponding to x_4^2), we have

$$A_{442} = \frac{(1407.6)^2 + (1206.8)^2}{16} = \frac{(2673.8)^2}{32} = 64.23.$$

For the (1, 2) element of $\mathbb{B}_4$ (corresponding to x_1x_2), we use (5.71) for b_{412} to obtain

$$\begin{aligned} A_{441} &= \frac{(125.40)(1407.6) + (119.39)(1206.8)}{16} = \frac{(244.59)(2673.8)}{32} \\ &= 27.208. \end{aligned}$$

Thus

$$\mathbb{B}_4 = \begin{pmatrix} 1.208 & 27.208 \\ 27.208 & 64.23 \end{pmatrix}.$$

We obtain $\mathbb{B}_5$ similarly:

$$B_{551} = \frac{(79.09)^2 + \cdots + (79.34)^2}{(4 \times 2)} = \frac{(244.59)^2}{32} = 1.694,$$

$$B_{552} = \frac{(881.9)^2 + \cdots + (879.3)^2}{8} = \frac{(2873.8)^2}{32} = 74.874,$$

$$B_{5512} = \frac{(59.69)(881.8) + \cdots + (59.34)(879.3)}{8} = \frac{(244.59)(2673.8)}{32} = -9.862,$$

$$\mathbb{B}_5 = \begin{pmatrix} 1.694 & -9.862 \\ -9.862 & 74.874 \end{pmatrix}.$$

For $\mathbb{B}_{52}$ we have, by (5.72),

$$B_{5211} = \frac{(30.81)^2 + \cdots + (29.32)^2}{4} = \frac{(244.59)^2}{32} = 1.205 + 1.694 = .131,$$

$$B_{5222} = \frac{(356.8)^2 + \cdots + (328.9)^2}{4} = \frac{(2673.8)^2}{32} = 64.23 + 74.874 = 32.204,$$

$$\begin{aligned} B_{5212} &= \frac{(30.81)(354.9) + \cdots + (29.32)(325.9)}{4} = \frac{(244.59)(2673.8)}{32} \\ &= 27.208 + (-9.862) = 1.345, \end{aligned}$$

$$\mathbb{B}_{52} = \begin{pmatrix} .131 & 1.345 \\ 1.345 & 32.204 \end{pmatrix}.$$

The error matrix $\mathbf{E}$ is obtained using the computational forms given for s_{ii} and s_{ij} in (5.11). For example, s_{11} and s_{12} are computed as

$$\begin{aligned}s_{11} &= (7.80)^2 + (7.70)^2 + \cdots + (7.70)^2 = \frac{(344.89)^2}{12} = 1,209 \\ &\quad - 1,594 = .615, \\ s_{12} &= (7.80)(76.4) + \cdots + (7.70)(76.4) = \frac{(344.89)(2673.8)}{12} = 27,308 \\ &\quad - (-1,882) = 1,585 = -1,890.\end{aligned}$$

Proceeding in this fashion, we obtain

$$\mathbf{E} = \begin{pmatrix} .615 & -1,890 \\ -1,890 & 116,390 \end{pmatrix},$$

with $v_2 = ab(n-1) = (2)(3)(4-1) = 24$.

To test the main effect of A with Wilks' Λ , we compute

$$\Lambda_A = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_A|} = \frac{5681.2}{5268.2} = .474 < \Lambda_{.05,2,1,24} = .771,$$

and we conclude that velocity has a significant effect on y_1 or y_2 or both.

For the B main effect, we have

$$\Lambda_B = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_B|} = \frac{5681.2}{5268.8} = .906 > \Lambda_{.05,1,1,24} = .591.$$

We conclude that the effect of lubricants is not significant.

For the AB interaction, we obtain

$$\Lambda_{AB} = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_{AB}|} = \frac{3602.1}{3665.3} = .982 > \Lambda_{.05,1,1,24} = .591.$$

Hence we conclude that the interaction effect is not significant.

We now obtain the other three MANOVA test statistics for each test. For A , the only nonzero eigenvalue of $\mathbf{E}^{-1}\mathbf{H}_A$ is 1.110. Thus

$$\begin{aligned}\mathbf{P}^{(A)} &= \frac{\lambda_1}{1 + \lambda_1} = .526, & \mathbf{L}^{(A)} &= \lambda_1 = 1.110, \\ \theta &= \frac{\lambda_1}{1 + \lambda_1} = .526.\end{aligned}$$

In this case, all three tests give results equivalent to that of Λ_A because $v_A = r = 1$.

For β , $r_F = 3$ and $p = r = 2$. The eigenvalues of $\mathbf{K}^{-1}\mathbf{B}_\beta$ are .418 and .028, and we obtain

$$\begin{aligned} \mathbf{F}^{(B)} &= \sum_{j=1}^{r_F} \frac{\lambda_j}{1 + \lambda_j} = .364, \\ \sigma^2^{(B)} &= \sum_{j=1}^{r_F} \lambda_j = .446, & \frac{\partial \mathbf{E}(\mathbf{F}^{(B)})}{\partial \beta} &= 3.902, \\ \hat{\beta} &= \frac{\lambda_1}{1 + \lambda_1} = .295. \end{aligned}$$

With $r = \frac{1}{2}$, $m = 0$, and $\mathbf{B} = \mathbf{I}(2)$, we have $\mathbf{F}^{(B)} = .459$ and $\hat{\beta}_{LS} = .364$. The .05 critical value of $r_F \mathbf{E}^{(B)} / r_F$ is 3.1759. Thus $\mathbf{F}^{(B)}$, $\hat{\sigma}^{(B)}$, and $\hat{\beta}$ lead to acceptance of H_0 , as does $\hat{\beta}$. On the first test, $\hat{\beta}$ appears to be closer to rejection. This is because $\lambda_1 / (\lambda_1 + \lambda_2) = .446 / (.418 + .028) = .334$, indicating that the mean weights for factor $\mathbf{B}$ are essentially uniform, in which case Roy's test is most powerful. If the mean-weights $\bar{\mathbf{y}}_{11}$, $\bar{\mathbf{y}}_{12}$, $\bar{\mathbf{y}}_{21}$, and $\bar{\mathbf{y}}_{22}$ for the four levels of $\mathbf{B}$ move a little farther apart, we would have a situation in which the four MANOVA tests do not lead to the same conclusion.

For $\mathbf{AB}$, the eigenvalues of $\mathbf{K}^{-1}\mathbf{B}_{\mathbf{AB}}$ are .0034 and .0073, from which

$$\begin{aligned} \mathbf{F}^{(AB)} &= \sum_{j=1}^{r_F} \frac{\lambda_j}{1 + \lambda_j} = .0033, & \mathbf{F}^{(B)} &= .4710, & \frac{\partial \mathbf{E}(\mathbf{F}^{(B)})}{\partial \beta} &= .000, \\ \hat{\beta} &= \frac{\lambda_1}{1 + \lambda_1} = .0033. \end{aligned}$$

The critical values remain the same as for factor β , and all three tests accept H_0 , as does "Wilks" Λ . With a nonsignificant interaction, interpretation of the main effects is simplified. $\square$

6.6 OTHER MODELS

6.6.1 Higher-Order Fixed Effects

A higher-order (balanced) fixed-effects model in factorial experiment presents no new difficulties. As an illustration, consider a three-way classification with three factors A , B , and C and all interactions AB , AC , BC , and ABC . The observation vector $\mathbf{y}$ has p variables as usual. The MANCOV model allowing for main effects and interactions can be written as

$$\mathbf{y}_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \delta_{ik} + \gamma_{jk} + \delta_{ijk} \quad (6.76)$$

where, for example, α_i is the effect of the i th level of factor A on each of the p variables in $\mathbf{y}_{ijk}$ and δ_{ij} is the AB interaction effect on each of the p variables.

Similarly, η_{jA} , η_{jB} , and η_{jAB} represent the AC , BC , and ABC interactions on each of the p variables.

The matrices of sums of squares and products for main effects, interactions, and error are defined in a manner similar to that for the matrices detailed for the two-way model in Section 6.5.2. The sum of squares (not the diagonal) for each variable is calculated only for main α in a univariate ANOVA for a three-way model. The sums of products (off-diagonal) are obtained analogously. That construction parallels that for the two-way model, using the matrix for error to test all factors and interactions.

Degrees of freedom for each factor are the same as in the corresponding three-way univariate model. All four MANOVA test statistics can be computed for each test. Contrasts can be defined and tested in a manner similar to that in Section 6.5.1. Follow-up procedures on the individual variables (F -tests and discriminant functions) can be used as discussed for the one-way or two-way models in Sections 6.4 and 6.5.2.

6.6.2 Mixed Models

There is a MANOVA counterpart for every univariate ANOVA design. This applies to fixed, random, and mixed models and to experimental structures that are crossed, nested, or a combination. Keiffers (1982) has provided a formal proof that univariate mixed models can be generalized to multivariate mixed models. Schott and Saw (1984) have shown that for the one-way multivariate random effects model, the likelihood ratio approach leads to the same test statistics (involving the eigenvalues of $E^{-1}B$) as in the fixed-effects model.

In the (balanced) MANOVA mixed model, the expected mean square matrices have exactly the same pattern as expected mean squares for the corresponding univariate ANOVA model. Thus a table of expected mean squares for the terms in the corresponding univariate model provides direction for choosing the appropriate error matrix to test each term in the MANOVA model. However, if the matrix indicated for "error" has fewer degrees of freedom than p , it will not have an inverse and the test cannot be made.

To illustrate, suppose we have a (balanced) two-way MANOVA model with A fixed and B random. Then the (univariate) expected mean squares (EMS) and Wilks' Λ -tests are as follows:

Source	EMS	Λ
A	$\sigma^2 + n\sigma_{\alpha}^2 + nb\sigma_{\beta}^2$	$(B_{A,0})/(B_{A,0} + B_{A,1})$
B	$\sigma^2 + na\sigma_{\beta}^2$	$(B_{B,0})/(B_{B,0} + B_{B,1})$
AB	$\sigma^2 + na\sigma_{\alpha\beta}^2$	$(B_{AB,0})/(B_{AB,0} + B_{AB,1})$
Error	σ^2	

In the expected mean squares for factor A , we have used the notation σ_{α}^2 in place of $\sum_{i=1}^a \alpha_i^2/(a-1)$. The test for A using $B_{A,0}$ for error matrix will be indeterminate (at the limit $B_{A,0}$) if $\nu_{A,B} \leq p$, where $\nu_{A,B} = (a-1)(b-1)$. In this case, $\nu_{A,B}$ will often fail to exceed p . For example, suppose A has two levels and B has three.

Then $r_{adj} = 1$, which will certainly be less than p . In such a case, we would have little recourse except to compute univariate tests on the p individual variables. However, we would not have the multivariate test to protect against carrying out too many univariate tests and thereby inflating the experimentwise α (see Section 5.4). To protect against inflation of α when making p tests, we could use a Bonferroni correction, as in procedure 3 in Section 5.5. In the case of F -tests, we do not have a table of Bonferroni critical values, as we do for t -tests (Table A.6), but we can achieve an equivalent result by comparing the p -values for the F -tests against α/p instead of against α .

As another illustration, consider the analysis for a (balanced) multivariate split-plot design. For simplicity, we show the associated univariate model in place of the multivariate model. We use the factor names A , B , AC , $\dots$ to indicate parameters in the model:

$$y_{ijk} = \mu + A_i + B_{(j)} + C_k + AC_{ik} + BC_{(jk)} + \epsilon_{ijk},$$

where A and C are fixed and B is random. Noting is indicated by bracketed subscripts; for example, B and BC are named in A. Table 6.7 shows the expected mean squares and corresponding Wilks tests.

Table 6.7. Wilks' Λ -Tests for a Typical Split-Plot Design

Source	df	Expected Mean Squares	Wilks' Λ
A	$a - 1$	$\sigma^2 + a\sigma_B^2 + ab\sigma_C^2$	$(B_1)/(\bar{B}_1 + B_2)$
B	$a(b - 1)$	$\sigma^2 + ab\sigma_C^2$	$(B_2)/(\bar{B}_1 + B_2)$
C	$c - 1$	$\sigma^2 + ac\sigma_{(C)}^2 + ab\sigma_{(C)}^2$	$(B_{3+})/(\bar{B}_1 + \bar{B}_{3+})$
AC	$(a - 1)(c - 1)$	$\sigma^2 + ac\sigma_{(C)}^2 + bc\sigma_{(C)}^2$	$(B_{3+})/(\bar{B}_{3+} + \bar{B}_{4+})$
BC	$a(b - 1)(c - 1)$	$\sigma^2 + ac\sigma_{(C)}^2$	$(B_4)/(\bar{B}_{3+} + \bar{B}_4)$
Error	$abc(r - 1)$	σ^2	

Since we use $\bar{B}_1$ and $\bar{B}_{3+}$, as well as $\bar{B}_4$, to make tests, the following must hold:

$$a(b - 1) \geq p, \quad a(b - 1)(c - 1) \geq p, \quad abc(r - 1) \geq p.$$

To construct the other three MANOVA tests, we use eigenvalues of the following matrices:

Source	Matrix
A	$B_1^{-1}B_2$
B	$B^{-1}B_2$
C	$B_{3+}^{-1}B_4$
AC	$B_{3+}^{-1}B_{4+}$
BC	$B^{-1}B_{4+}$

With a table of expected mean squares, such as those in Table 6.7, it is a simple matter to determine the error matrix in each case. For a given factor or interac-

tion, such as A , B , or AC , the appropriate error matrix is ordinarily the one whose expected mean square matches that of the given factor (except for the last term). For example, factor C , with expected mean square $\sigma^2 + \sigma\alpha_{BC}^2 + \sigma\alpha\epsilon_C^2$, is tested by BC , whose expected mean square is $\sigma^2 + \sigma\alpha_{BC}^2$. If $H_0: \alpha\epsilon_C^2 = 0$ is true, then C and BC have the same expected mean square.

In some mixed and random models, certain terms have no available error term. When this happens in the univariate case, we can construct an approximate test using Satterthwaite's (1946) or other synthetic mean square approach. For a similar approach in the multivariate case, see Khuri, Mathew, and Nal (1994).

6.7 CHECKING ON THE ASSUMPTIONS

In Section 6.2 we discussed the robustness of the four MANOVA test statistics to nonnormality and interdependency of covariance matrices. The MANOVA tests (except for Roy's) are rather robust to these departures from the assumptions, although, in general, as dimensionality increases, robustness decreases.

Even though MANOVA procedures are fairly robust to departures from multivariate normality, we may want to check for gross violations of this assumption. Any of the tests or plots from Section 4.4 could be used. For a two-way design, for example, the tests could be applied separately to the n observations in each individual cell (if n is sufficiently large) or to all the residuals. The residual vector after fitting the model $F_{ijk} = \mu_{ij} + \alpha_{ijk}$ would be

$$\hat{\alpha}_{ijk} = Y_{ijk} - F_{ijk}, \quad i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J, \quad k = 1, 2, \dots, n.$$

It is also advisable to check for outliers, which can lead to either a Type I or Type II error. The tests of Section 4.5 can be run separately for each cell (for sufficiently large n) or for all of the other residuals, $\hat{\alpha}_{i,j} = Y_{ijk} - \hat{\mu}_{ij}$.

A test of the equality of covariance matrices can be made using Box's M test given in Section 7.3.2. (Note the cautions expressed there about the sensitivity of this test to nonnormality and unequal sample sizes.)

The assumption of independence of the observation vector $\mathbf{y}_{ijk}$ is even more important than the assumptions of normality and equality of covariance matrices. We are referring, of course, to independence from one observation vector to another. The residuals within a vector are assumed to be correlated, as usual. In the univariate case, Box-Coxen's (1961) showed that a moderate amount of dependence among the observations produces an actual α much greater than the nominal α . This effect is to be expected, since the dependence leads to an underestimate of the variance, so that MSE is reduced and the F statistic is inflated. We can assume that this effect on error rates carries over to MANOVA.

In univariate ANOVA, a simple measure of dependence among the n observations in a one-way model is the intraclass correlation:

$$r_c = \frac{MSE - MSE'}{MSE - (n - 1)MSE'} \quad (6.77)$$

where MSB and MSW are the between and within mean squares for the variable and n is the number of observations per group. This could be calculated for each variable in a MANOVA to check for independence.

In many experimental settings, we do not anticipate a lack of independence. But in certain cases the observations are dependent. For example, if the sampling units are people, they may influence each other as they interact together. In some educational studies, researchers must use entire classrooms as sampling units rather than use individual students. Another example of dependence is furnished by observations that are serially correlated, as in a time series, for example. Each observation depends to a certain extent on the preceding one, and its random movement is somewhat dampened as a result.

6.2 PROFILE ANALYSIS

The two-sample profile analysis of Section 5.3.2 can be extended to k groups. Again we assume that the variables are commensurate, as, for example, when each subject is given a battery of tests. Other assumptions, cautions, and comments expressed in Section 5.3.2 apply here as well.

The basic model is the balanced one-way MANOVA

$$X_{ij} = \mu_j + \epsilon_{ij}, \quad i = 1, 2, \dots, d, \quad j = 1, 2, \dots, p.$$

To test $H_0: \mu_1 = \mu_2 = \dots = \mu_k$, we use the usual H and E matrices given in (5.17) and (5.18). If the variables are commensurate, we can be more specific and extend H_0 to an examination of the k profiles obtained by plotting the p values $x_{i1}, x_{i2}, \dots, x_{ip}$ in each μ_i , as was done with two μ 's in Section 5.3.2 (see, for example, Figure 5.8). We are interested in the same three hypotheses as before:

H_{01} : The k profiles are parallel.

H_{02} : The k profiles are all at the same level.

H_{03} : The k profiles are flat.

The hypothesis of parallelism for two groups was expressed in Section 5.3.2 as $H_{01}: C\mu_1 = C\mu_2$, where C is any $(p-1) \times p$ matrix of rank $p-1$ such that $C\mathbf{1} = \mathbf{0}$. For example,

$$C = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

For k groups, the analogous hypothesis of parallelism is

$$H_{01}: C\mu_1 = C\mu_2 = \cdots = C\mu_k. \quad (5.78)$$

The hypothesis (5.76) is equivalent to the hypothesis $H_0: \mu_{1j} = \mu_{2j} = \cdots = \mu_{kj}$ in a one-way MANOVA on the transformed variable $\mathbf{y}_{ij} = \mathbf{C}\mathbf{y}_{ij}$. Since $\mathbf{C}$ has $p-1$ rows, $\mathbf{C}\mathbf{y}_{ij}$ is $(p-1) \times 1$, $\mathbf{C}\mathbf{y}_{1j}$ is $(p-1) \times 1$, and $\mathbf{C}\mathbf{B}\mathbf{C}'$ is $(p-1) \times (p-1)$. By property 1b in Section 4.2, $\mathbf{y}_{ij}$ is distributed as $N_{p-1}(\mathbf{C}\mu_{ij}, \mathbf{C}\Sigma\mathbf{C}')$.

By analogy with (5.64), the hypothesis and error matrices for testing H_0 in (5.76) are

$$\mathbf{B}_0 = \mathbf{C}\mathbf{B}\mathbf{C}', \quad \mathbf{B}_e = \mathbf{C}\mathbf{B}\mathbf{C}'.$$

We then have

$$\Lambda = \frac{|\mathbf{C}\mathbf{B}\mathbf{C}'|}{|\mathbf{C}\mathbf{B}\mathbf{C}' + \mathbf{C}\mathbf{B}\mathbf{C}'|} = \frac{|\mathbf{C}\mathbf{B}\mathbf{C}'|}{|\mathbf{C}(\mathbf{B} + \mathbf{B})\mathbf{C}'|}, \quad (5.79)$$

which is distributed as $\Lambda_{p-1, p-1, p-1}$ (where $v_1 = 1-1$ and $v_2 = J(p-1)$). The other three MANOVA test statistics can be obtained from the eigenvalues of $(\mathbf{C}\mathbf{B}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{B}\mathbf{C}')$. The test for H_{01} can easily be adjusted for unbalanced data, as in Section 5.1.6. We would calculate $\mathbf{H}$ and $\mathbf{K}$ by (5.32) and (5.33) and use $v_2 = \sum_{i=1}^J n_i - k$.

The hypothesis that two profiles are at the same level in $H_{01}: \bar{\mathbf{f}}\mu_1 = \bar{\mathbf{f}}\mu_2$ (see Section 5.1.2), which generalizes immediately to J profiles at the same level,

$$H_0: \bar{\mathbf{f}}\mu_1 = \bar{\mathbf{f}}\mu_2 = \cdots = \bar{\mathbf{f}}\mu_J. \quad (5.80)$$

For two groups we need a univariate t , as defined in (5.35), to test H_{01} . Similarly, for J groups, we can employ an F -test for one-way ANOVA, comparing J groups with observations $\{\bar{\mathbf{f}}\mathbf{y}_{ij}\}$. Minimal levels, we can utilize (5.79) with $\mathbf{C} = \bar{\mathbf{f}}$,

$$\Lambda = \frac{|\bar{\mathbf{f}}\mathbf{B}|}{|\bar{\mathbf{f}}\mathbf{B} + \bar{\mathbf{f}}\mathbf{B}|}, \quad (5.81)$$

which is distributed as $\Lambda_{1, Jp-1, Jp-1}$ ($Jp-1$ is a scalar). This is, of course, equivalent to the F -test on $\{\bar{\mathbf{f}}\mathbf{y}_{ij}\}$, since by Table 6.1 in Section 6.1.3,

$$F = \frac{1 - \Lambda}{\Lambda} \frac{v_2}{v_1} \quad (5.82)$$

is distributed as F_{v_1, v_2} .

The third hypothesis, that of "fairness," essentially states that the average of the J group means is the same for each variable (see (5.17)):

$$\begin{aligned} \mu_{0j} &= \frac{\mu_{1j} + \mu_{2j} + \cdots + \mu_{Jj}}{J} = \frac{\mu_{11} + \mu_{21} + \cdots + \mu_{J1}}{J} \\ &= \cdots = \frac{\mu_{1J} + \mu_{2J} + \cdots + \mu_{Jj}}{J}. \end{aligned}$$

or by analogy with (5.31),

$$R_{Bj} = \frac{C(\mu_1 + \mu_2 + \cdots + \mu_k)}{k} = 0, \quad (5.83)$$

where C is a $(p-1) \times p$ matrix of rank $p-1$ such that $Cj = 0$ [see (5.78)]. The fairness hypothesis can also be stated as, the means of all p variables in each group are the same, or $\mu_{11} = \mu_{12} = \cdots = \mu_{1p}$, $j = 1, 2, \dots, k$. This can be expressed as $R_{Bj} = C\mu_1 = C\mu_2 = \cdots = C\mu_k = 0$.

To test H_{01} as given by (5.82), we can extend the F^2 -test in (5.59). The grand mean vector $(\mu_1 + \mu_2 + \cdots + \mu_k)/k$ in (5.83) can be estimated as in Section 5.1.2 by

$$\bar{y}_\cdot = \frac{\sum_{j=1}^k \bar{y}_{j\cdot}}{k}.$$

Under H_{01} (and H_{02}), $C\bar{y}_\cdot$ is $N_{p-1}(0, C\Sigma C')$ [6], and R_{Bj} can be tested by

$$T^2 = k(C\bar{y}_\cdot)'(C\Sigma C')^{-1}C\bar{y}_\cdot, \quad (5.84)$$

where Σ/ν_2 is an estimate of Σ . As in the two-sample case, H_{01} is unaffected by the status of H_{02} . When both H_{01} and H_{02} are true, T^2 in (5.84) is distributed as F_{p-1, ν_2}^2 .

Example 5.8. Three vitamin E-diet supplements with levels zero, low, and high were compared for their effect on growth of guinea pigs (Crowder and Hand 1980, pp. 21–27). Five guinea pigs received each supplement level and their weights were recorded at the end of weeks 1, 3, 4, 8, 9, and 7. These weights are given in Table 5.8.

Table 5.8. Weights of Guinea Pigs under Three Levels of Vitamin E Supplements

Group	Animal	Week 1	Week 3	Week 4	Week 8	Week 9	Week 7
1	1	455	465	510	594	495	495
1	2	467	508	495	596	562	587
1	3	44.5	550	580	587	582	610
1	4	485	545	554	583	611	612
1	5	480	500	610	628	582	576
2	6	514	560	565	624	592	597
2	7	440	480	536	604	587	588
2	8	495	570	568	585	576	607
2	9	505	540	610	607	671	702
2	10	508	555	591	602	640	675
3	11	466	560	622	622	652	650
3	12	498	640	664	692	588	600
3	13	478	610	568	699	676	658
3	14	545	568	580	660	630	649
3	15	472	498	540	524	502	580

The three mean vectors are

$$\bar{\mathbf{y}}_1 = (986.8, 519.4, 568.8, 561.6, 566.8, 572.5),$$

$$\bar{\mathbf{y}}_2 = (994.8, 521.8, 574.2, 562.0, 563.8, 564.8),$$

$$\bar{\mathbf{y}}_3 = (971.8, 514.8, 579.8, 571.8, 568.2, 573.2),$$

and the overall mean vector is

$$\bar{\mathbf{y}}_{..} = (985.2, 515.8, 574.3, 566.8, 570.3, 573.1).$$

A profile plot of the means $\bar{\mathbf{y}}_1$, $\bar{\mathbf{y}}_2$, and $\bar{\mathbf{y}}_3$ is given in Figure 6.5. There is a high degree of parallelism in the three profiles, with the possible exception of week 8 for group 1.

The E and H matrices are as follows:

$$\mathbf{E} = \begin{pmatrix} 8481.2 & 8536.8 & 4879.8 & 8511.8 & 8716.8 & 8468.2 \\ 8536.8 & 17150.4 & 18293.8 & 16436.8 & 17634.2 & 16025.6 \\ 4879.8 & 18293.8 & 12992.8 & 17677.8 & 17381.8 & 17697.2 \\ 8511.8 & 16436.8 & 17677.8 & 26966.0 & 26236.8 & 26823.2 \\ 8716.8 & 17634.2 & 17381.8 & 26236.8 & 36888.8 & 31565.8 \\ 8468.2 & 16025.6 & 17697.2 & 26823.2 & 31565.8 & 33508.8 \end{pmatrix}.$$

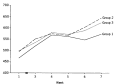


Figure 6.5. Profile of the three groups for the pig data of Table 6.8.

$$\mathbf{B} = \begin{pmatrix} 2669.2 & 2677.2 & 856.4 & 913.0 & 4725.2 & 5621.6 \\ 2177.2 & 2497.8 & 418.0 & 411.6 & 4428.8 & 5637.8 \\ 859.4 & 470.8 & 562.5 & 280.4 & 1152.1 & 1352.9 \\ 913.8 & 411.6 & 280.4 & 253.4 & 1056.4 & 1352.8 \\ 4725.2 & 4428.8 & 1132.1 & 1098.4 & 8350.8 & 10830.8 \\ 5621.6 & 5637.8 & 1561.5 & 1352.0 & 10850.8 & 12730.1 \end{pmatrix}.$$

Using

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

in the test statistic (5.79), we have, as a test for parallelism,

$$\begin{aligned} \Lambda &= \frac{|\mathbf{CBCT}'|}{|\mathbf{C}(\mathbf{B} + \mathbf{B})\mathbf{C}'|} = \frac{3.8238 \times 10^{26}}{2.1583 \times 10^{26}} \\ &= 1.771 > \Lambda_{0.05,2,11} = 1.53. \end{aligned}$$

Thus we do not reject the parallelism hypothesis.

To test the hypothesis that the three profiles are at the same level, we use (5.84),

$$\begin{aligned} \Lambda &= \frac{|\mathbf{F}_0|}{|\mathbf{F}_0 + \mathbf{F}_1|} = \frac{632,683.2}{801,699.7 + 111,288.1} \\ &= .8664 = \Lambda_{0.05,2,11} = .687. \end{aligned}$$

Hence we do not reject the levels hypothesis. This can also be seen by using (5.82) to transform Λ to F ,

$$F = \frac{(1 - \Lambda)/q_0}{\Lambda/q_1} = \frac{(1 - .8664)/2}{(.8664)/2} = 1.2553,$$

which is clearly nonsignificant ($p = .378$).

To test the "flatness" hypothesis, we use (5.84)

$$\begin{aligned} F^2 &= 4\alpha(\mathbf{C}\mathbf{B}_0^{-1}(\mathbf{C}\mathbf{B}\mathbf{C}')^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{B}_0^{-1}\mathbf{C}' \\ &= 15 \begin{pmatrix} -48.88 \\ -78.27 \\ 7.47 \\ -12.47 \\ -33.88 \end{pmatrix}' \begin{pmatrix} 334.5 & -13.2 & 267.5 & -208.9 & 276.2 \\ -13.2 & 288.1 & -174.9 & 221.0 & -208.9 \\ 267.5 & -174.9 & 645.5 & -240.8 & 183.8 \\ -208.9 & 221.0 & -240.8 & 1121.6 & -648.5 \\ 276.2 & -208.9 & 183.8 & -648.5 & 618.8 \end{pmatrix}^{-1} \end{aligned}$$

$$= \begin{pmatrix} -49.08 \\ -39.27 \\ 7.47 \\ -11.47 \\ -31.08 \end{pmatrix} \\ = 281.13 = F_{3,1,1,1,1}^2 = 49.708.$$

Thus only the fitness hypothesis is rejected in this case. $\square$

6.5 REPEATED MEASURES DESIGNS

6.5.1 Multivariate vs. Univariate Approach

In repeated measures designs, the research unit is typically a human or animal subject. Each subject is measured under several treatments or at different points of time. The treatments might be tests, drug levels, various kinds of stimuli, and so on. If the treatments are such that the order of presentation to the various subjects can be varied, then the order should be randomized to avoid an ordering bias. If subjects are measured at successive time points, it may be of interest to determine the degree of polynomial required to fit the curves. This is treated in Section 6.10 as part of an analysis of growth curves.

When comparing means of the treatments applied to each subject, we are examining the within-subjects factor. There will also be a between-subjects factor if there are several groups of subjects that we wish to compare. In Sections 6.9.2–6.9.8, we consider designs up to a complexity level of two within-subjects factors and two between-subjects factors.

We now discuss univariate and multivariate approaches to hypothesis testing in repeated measures designs. As a framework for this discussion, consider the layout in Table 6.9 for a repeated measures design with one repeated measures (within-subjects) factor, A , and one grouping (between-subjects) factor, B .

This design has often been analyzed as a univariate mixed-model nested design, also called a split-plot design, with subjects nested in factor B (whole-plot), which is crossed with factor A (repeated measures, or split-plot). The univariate model for each y_{ijr} would be

$$y_{ijr} = \mu + B_i + B_{ijr} + A_r + B_i A_r + \epsilon_{ijr}, \quad (6.85)$$

where the factor level designations (B_i , B_{ijr} , A_r , and $B_i A_r$) from (6.85) and Table 6.9 are used as parameter values and the subscript (i, j, r) on ϵ indicates that subjects are nested in factor B . In Table 6.9, the observations y_{ijr} for $r = 1, 2, \dots, p$ are enclosed in parentheses and denoted by y_{ij}^r to emphasize that these p variables are measured on one subject and thus constitute a vector of correlated variates. The ranges of the subscripts can be seen in Table 6.9: $i = 1, 2, \dots, b$; $j = 1, 2, \dots, n_i$; and $r =$

Table 4.9. Data Layout for k -Groups Repeated Measures Experiment

Factor B (Group)	Subjects	Factor A (Repeated Measure)				
		A_1	A_2	...	A_k	
B_1	S_{11}	X_{111}	X_{112}	...	X_{11k}	$= Y'_{11}$
	S_{12}	X_{121}	X_{122}	...	X_{12k}	$= Y'_{12}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
	S_{1n}	X_{1n1}	X_{1n2}	...	X_{1nk}	$= Y'_{1n}$
B_2	S_{21}	X_{211}	X_{212}	...	X_{21k}	$= Y'_{21}$
	S_{22}	X_{221}	X_{222}	...	X_{22k}	$= Y'_{22}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
	S_{2n}	X_{2n1}	X_{2n2}	...	X_{2nk}	$= Y'_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
B_k	S_{k1}	X_{k11}	X_{k12}	...	X_{k1k}	$= Y'_{k1}$
	S_{k2}	X_{k21}	X_{k22}	...	X_{k2k}	$= Y'_{k2}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
	S_{kn}	X_{kn1}	X_{kn2}	...	X_{knk}	$= Y'_{kn}$

1, 2, ..., p . With factors A and B fixed and subjects random, the univariate ANOVA is given in Table 4.10.

However, our initial reaction would be to rule out the univariate ANOVA because the y 's in each row are correlated and the assumption of independence is critical, as noted in Section 4.3. We will discuss below some assumptions under which the univariate analysis would be appropriate.

In the multivariate approach, the p responses $y_{11}, y_{12}, \dots, y_{1p}$ (repeated measures) for subject S_{11} constitute a vector $\mathbf{y}_{11}$, as shown in Table 4.9. The multivariate model for $\mathbf{y}_{11}$ is a simple one-way MANOVA model,

$$\mathbf{y}_{11} = \boldsymbol{\mu} + \mathbf{A}_1 + \boldsymbol{\epsilon}_{11}, \quad (4.38)$$

Table 4.10. Univariate ANOVA for Data Layout in Table 4.9

Source	df	MS	F
B (between)	$k - 1$	MSB	(MSB/MSW)
B (subjects)	$k(p - 1)$	MSB	
A (within or repeated)	$p - 1$	MSA	(MSA/MSW)
ABA	$(k - 1)(p - 1)$	MSBA	(MSBA/MSW)
Error (SA interaction)	$k(p - 1)(p - 1)$	MSW	

where β_k is a vector of p main-effects corresponding to the p variables in y_{kj} for factor k , and e_{kj} is an error vector for subject y_{kj} . This model seems to include only factor B , but we show in Section 6.9.3 how to use an approach similar to provide analysis in Section 6.8 to obtain tests on factor A and the $A \times B$ interaction. The MANOVA assumption that $\text{cov}(y_{ij}) = \mathbf{I}$ for all i and j allows the p repeated measures to be correlated in any pattern, since $\mathbf{I}$ is completely general. On the other hand, the ANOVA assumptions of independence and homogeneity of variances can be expressed as $\text{cov}(y_{ij}) = \sigma^2 \mathbf{I}$. We would be very surprised if repeated measurements on the same subject were independent.

The univariate ANOVA approach has been found to be appropriate under less stringent conditions than $\mathbf{I} = \sigma^2 \mathbf{I}$. Wilks (1946) showed that the ordinary F -tests of ANOVA remain valid for a covariance structure of the form

$$\begin{aligned}\text{cov}(\mathbf{y}_{ij}) &= \mathbf{I} + \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ 1 & \rho & 1 & \cdots & 1 \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix} \\ &= \sigma^2 (\mathbf{I} + \rho \mathbf{J} + \rho \mathbf{J}),\end{aligned}\quad (6.87)$$

where $\mathbf{J}$ is a square matrix of 1's, as defined in (2.12) [see Rencher (2000, pp. 126–131)]. The covariance pattern (6.87) is variously known as *uniformly correlated symmetry*, or the *intraclass correlation model*. It allows for the variables to be correlated but restricts every variable to have the same variance and every pair of variables to have the same covariance. In a carefully designed experiment with appropriate randomization, this assumption may hold under the hypothesis of no A effect. Alternatively, we could use a test of the hypothesis that $\mathbf{I}$ has the pattern (6.87) (see Section 7.2.3). If this hypothesis is accepted, one could proceed with the usual ANOVA F -tests.

Duck (1965) and Hays and Feldt (1970) showed that the most general condition under which univariate F -tests remain valid is that

$$\mathbf{C} \mathbf{C}^T = \sigma^2 \mathbf{I}, \quad (6.88)$$

where $\mathbf{C}$ is a $(p-1) \times p$ matrix whose rows are orthonormal contrasts (orthogonal contrasts that have been normalized to unit length). We can construct $\mathbf{C}$ by choosing any $p-1$ orthogonal contrasts among the means $\mu_1, \mu_2, \dots, \mu_p$ of the repeated measure factor and dividing each contrast $\sum_{i=1}^p c_i \mu_i$ by $\sqrt{\sum_{i=1}^p c_i^2}$. (This matrix $\mathbf{C}$ is different from $\mathbf{C}$ used in Section 6.8 and its the remainder of Section 6.9, whose rows are contrasts that are not normalized to unit length.) It can be shown that (6.87) is a special case of (6.88). The condition (6.88) is sometimes referred to as *sphericity*, although this term can also refer to the covariance pattern $\mathbf{I} = \sigma^2 \mathbf{I}$ on the untransformed y_{ij} (see Section 7.2.2).

A simple way to test the hypothesis that (5.88) holds is to transform the data by $\mathbf{y}_{ij} = \mathbf{C}\mathbf{y}_{ij}$ and test $\mathbf{H}_0: \Sigma_c = \sigma^2\mathbf{I}$, as in Section 7.2.2, using $\mathbf{C}\mathbf{y}_{ij}\mathbf{C}'$ in place of $\mathbf{y}_{ij}\mathbf{y}_{ij}' = \mathbf{E}/v_j$.

Thus one procedure for repeated measures designs is to make a preliminary test for (5.87) or (5.88) and, if the hypothesis is accepted, use univariate F -tests, as in Table 6.13. Fienberg (1958) investigated the use of larger α -values with a preliminary test of structure of the covariance matrix, as in (5.88). He concludes that using $\alpha = .40$ sufficiently controls the problem of falsely accepting sphericity so as to justify the use of a preliminary test.

If the univariate test for the repeated measures factor A is appropriate, it is more powerful because it has more degrees of freedom for error than the corresponding multivariate test. However, even small departures from (5.88) seriously inflate the Type I error rate of the univariate test for factor A (Dun 1954; Davidian 1972; Dells 1981). Because such departures can be easily missed in a preliminary test, Dells (1981) concludes that "On the whole, the ordinary F tests have nothing to recommend them" (p. 248) and emphasized that "there is no justification for employing ordinary univariate F tests for repeated measures treatment contrasts" (p. 254).

Another approach to analysis of repeated measures designs is to adjust the univariate F -test for the amount of departure from sphericity. Dun (1954) and Greenhouse and Geisser (1959) showed that when $\Sigma_c \neq \sigma^2\mathbf{I}$, an approximate F -test for effects involving the repeated measures is obtained by reducing the degrees of freedom for both numerator and denominator by a factor of

$$f = \frac{1 + (\mathbf{J} - \mathbf{J}\mathbf{J}'/p)^2}{(p-1) + (\mathbf{J} - \mathbf{J}\mathbf{J}'/p)^2} \quad (5.89)$$

where $\mathbf{J}$ is a $p \times p$ matrix of 1's defined in (2.12). For example, in Table 6.10 the F -value for the $B \times A$ interaction would be compared to F_{α} with $rv(1 - 1)(p - 1)$ and $rv(s - 1)(p - 1)$ degrees of freedom. An estimate $\hat{f}$ can be obtained by substituting $\hat{\Sigma} = \mathbf{R}/v_j$ in (5.89). Greenhouse and Geisser (1959) showed that $\hat{f}$ and $\hat{f}$ vary between $1/(p - 1)$ and 1, with $\hat{f} = 1$ when sphericity holds and $\hat{f} \geq 1/(p - 1)$ for other values of $\hat{\Sigma}$. Thus $\hat{f}$ is a measure of nonsphericity. For a conservative test, Greenhouse and Geisser recommend dividing numerator and denominator degrees of freedom by $\hat{p} - 1$. Hays and Feldt (1976) provided an improved estimator of $\hat{f}$.

The behavior of the approximate univariate F -test with degrees of freedom adjusted by $\hat{f}$ has been investigated by Collier et al. (1987), Hays (1978), Davidian (1972), Rogan, Kowtman, and Menckens (1978), and Maxwell and Arvey (1982). In these studies, the true α -level turned out to be close to the nominal α , and the power was close to that of the multivariate test. However, since the $\hat{f}$ -adjusted F -test is only approximate and has no power advantage over the exact multivariate test, there appears to be no compelling reason to use it. The only case in which we need to fall back on a univariate test is when there are insufficient degrees of freedom to perform a multivariate test, that is, when $p > v_j$.

In Sections 8.3.2-8.3.6, we discuss the multivariate approach to repeated measures. We will cover several models, beginning with the simple one-sample design.

5.5.1 One-Sample Repeated Measures Model

We illustrate some of the procedures in this section with $p = 4$. A one-sample design with four repeated measures on n subjects would appear as in Table 5.11. There is a superficial resemblance to a univariate randomized block design. However, in the repeated measures design, the observations y_{11} , y_{12} , y_{13} , and y_{14} are correlated because they are measured on the same subject (experimental unit), whereas in a randomized block design y_{11} , y_{21} , y_{31} , and y_{41} would be measured on four different experimental units. Thus we have a single sample of n observation vectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$.

To test for significance of factor A , we compare the means of the four variables in $\mathbf{y}_i$.

$$\mathbf{y}'(\mathbf{y}_i) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix}.$$

The hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, which can be expressed as $H_0: \boldsymbol{\mu}_1 - \mu_2 = \mu_1 - \mu_3 = \mu_2 - \mu_3 = \mu_2 - \mu_4 = 0$ or $\mathbf{C}_1\boldsymbol{\mu} = \mathbf{0}$, where

$$\mathbf{C}_1 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Test: $H_0: \mathbf{C}_1\boldsymbol{\mu} = \mathbf{0}$ for a general value of p (p repeated measures on subjects), we calculate $\bar{\mathbf{y}}$ and $\mathbf{S}$ from $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ and extend $\mathbf{C}_1$ to $p - 1$ rows. Then when H_0 is true, $\mathbf{C}_1\bar{\mathbf{y}}$ is $(p-1) \times 1$, $\mathbf{C}_1\mathbf{S}\mathbf{C}_1'$ is $(p-1) \times (p-1)$, and

$$\mathbf{T}^2 = n(\mathbf{C}_1\bar{\mathbf{y}})'(\mathbf{C}_1\mathbf{S}\mathbf{C}_1')^{-1}\mathbf{C}_1\bar{\mathbf{y}} \quad (5.36)$$

is distributed as $F_{p-1, p-1}^1$. We reject $H_0: \mathbf{C}_1\boldsymbol{\mu} = \mathbf{0}$ if $\mathbf{T}^2 \geq F_{p-1, p-1}^1$. Note that the dimension is $p-1$ because $\mathbf{C}_1\bar{\mathbf{y}}$ is $(p-1) \times 1$ [see (5.33)].

The multivariate approach involving transformed observations $\mathbf{z}_i = \mathbf{C}_1\mathbf{y}_i$ was first suggested by Hot (1928) and has been discussed further by Williams (1970).

Table 5.11. Data Layout for a Single-Sample Repeated Measures Design

Subjects	Factor A (Repeated Measures)				
	A_1	A_2	A_3	A_4	
$\mathbf{y}_1$	y_{11}	y_{12}	y_{13}	y_{14}	$= \mathbf{y}_1$
$\mathbf{y}_2$	y_{21}	y_{22}	y_{23}	y_{24}	$= \mathbf{y}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}_n$	y_{n1}	y_{n2}	y_{n3}	y_{n4}	$= \mathbf{y}_n$

and Morrison (1972). Note that in $C_2\beta$ (for $p = 4$), we work with contrasts on the elements $\bar{y}_1, \bar{y}_2, \bar{y}_3$, and $\bar{y}_4$ within the vector

$$\bar{y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \bar{y}_4 \end{pmatrix},$$

as opposed to the contrasts involving comparisons of several mean vectors themselves, as, for example, in Section 8.3.2.

The hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ can also be expressed as $M_0: \mu_1 - \mu_2 = \mu_2 - \mu_3 = \mu_3 - \mu_4 = \mu_4 - \mu_1 = 0$, or $C_2\beta = 0$, where

$$C_2 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

The matrix C_1 can be obtained from C_2 by simple row operations, for example, subtracting the second row from the first and the third row from the second. Hence, $C_1 = AC_2$, where

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

In fact, $M_0: \mu_1 = \mu_2 = \cdots = \mu_p$ can be expressed as $C\beta = 0$ for any full-rank $(p - 1) \times p$ matrix C such that $C\mathbf{1} = 0$, and the same value of T^2 in (8.93) will result. The contrasts in C can be either linearly independent or orthogonal.

The hypothesis $M_0: \mu_1 = \mu_2 = \cdots = \mu_p = \mu$, say, can also be expressed as

$$M_0: \mu = \mu_0$$

where $\mathbf{j} = (1, 1, \dots, 1)'$. The maximum likelihood estimate of μ is

$$\hat{\mu} = \frac{\mathbf{y}'\mathbf{y}^{-1}\mathbf{j}}{\mathbf{j}'\mathbf{y}^{-1}\mathbf{j}}. \quad (8.94)$$

The likelihood ratio test of M_0 is a function of

$$\mathbf{y}'\mathbf{y}^{-1}\mathbf{j} = \frac{(\mathbf{j}'\mathbf{y}^{-1}\mathbf{j})^2}{\mathbf{j}'\mathbf{y}^{-1}\mathbf{j}}.$$

Williams (1970) showed that for any $(p - 1) \times p$ matrix C of rank $p - 1$ such that $C\mathbf{1} = 0$,

$$\mathbf{y}'\mathbf{S}^{-1}\mathbf{y} = \frac{\mathbf{y}'\mathbf{S}^{-1}\mathbf{y}'}{\mathbf{1}'\mathbf{S}^{-1}\mathbf{1}} = (\mathbf{C}\mathbf{Y})'(\mathbf{C}\mathbf{M}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{Y})_0$$

and thus the F^2 -test in (6.9) is equivalent to the likelihood ratio test.

Example 6.8.1. The data in Table 6.11 were given by Cochran and Cox (1957, p. 130). As arranged by Tukey (1958), the observations constitute a one-sample repeated measures design with two within-subjects factors. Factor A is a comparison of two tasks; factor B is a comparison of two types of calculators. The measurements are speed of calculations.

To test the hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, we use the contrast matrix

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

where the first row compares the two levels of A , the second row compares the two levels of B , and the third row corresponds to the $A:B$ interaction. From the five-column vectors in Table 6.12, we obtain

$$\mathbf{y} = \begin{pmatrix} 23.2 \\ 15.6 \\ 20.0 \\ 11.6 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 51.7 & 29.8 & 9.2 & 7.4 \\ 29.8 & 46.8 & 16.2 & -8.7 \\ 9.2 & 16.2 & 8.3 & -10.3 \\ 7.4 & -8.7 & -10.3 & 24.3 \end{pmatrix}.$$

For the overall test of equality of means, we have, by (6.9),

$$F^2 = n(\mathbf{C}\mathbf{Y})'(\mathbf{C}\mathbf{M}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{Y}) = 29.798 = F_{3,20,0.01}^2 = 114.938.$$

Since the F^2 -test is not significant, we would ordinarily not proceed with tests based on the individual rows of $\mathbf{C}$. We will do so, however, for illustrative purposes. (Note that the F^2 -test has very low power in this case, because $n - 1 = 4$ is very small.)

To test A , B , and $A:B$, we test each row of $\mathbf{C}$, where $F^2 = n(\mathbf{y}'\mathbf{P})(\mathbf{y}'\mathbf{S}_0\mathbf{y})^{-1}\mathbf{y}'\mathbf{y}$ is the square of the t -statistic

Table 6.12. Calculator-Speed Data

Subjects	A_1		A_2	
	B_1	B_2	B_1	B_2
S_1	39	25	21	14
S_2	22	15	22	9
S_3	28	15	18	17
S_4	12	7	16	16
S_5	25	26	23	8

$$q_i = \frac{\sum_{j=1}^p \frac{(\bar{y}_{ij} - \bar{y}_{i.})^2}{n_j}}{s_i^2(\bar{R}_i)} \quad i = 1, 2, 3,$$

where $\bar{y}_{ij}$ is the i th row of $\bar{Y}$.

The three results are as follows:

$$\text{Factor A:} \quad q_1 = 3.459 < \text{tost}_4 = 3.778,$$

$$\text{Factor B:} \quad q_2 = 5.247 > \text{tost}_4 = 4.604,$$

$$\text{Interaction AB:} \quad q_3 = -1.93.$$

Thus only the main-effect for B is significant. Note that in all but one case in Table 6.11, the value for $\bar{R}_i$ is greater than that for $\bar{R}_j$. $\square$

6.5.2 k -Sample Repeated Measures Model

We turn now to the k -sample repeated measures design depicted in Table 6.3. As noted in Section 6.5.1, the multivariate approach to this repeated measures design uses the one-way MANOVA model $\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\epsilon}_{ij}$. From the k groups of n observations (each each), we calculate $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \dots, \bar{\mathbf{y}}_k$ and the error matrix $\mathbf{E}$.

The layout in Table 6.9 is similar to that of a k -sample profile analysis in Section 6.3. To test (the within-subjects) factor A , we need to compare the means of the variables $y_1, y_2, \dots, y_p$ within p overlapped across the levels of factor B . The p variables correspond to the levels of factor A . In the model $\mathbf{y}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\epsilon}_{ij}$, the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k$ correspond to the levels of factor B and are estimated by $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \dots, \bar{\mathbf{y}}_k$. To compare the means of $y_1, y_2, \dots, y_p$ overlapped across the levels of B , we use $\bar{\boldsymbol{\mu}} = \sum_{i=1}^k \boldsymbol{\mu}_i / k$, which is estimated by $\bar{\mathbf{y}}_{..} = \sum_{i=1}^k \bar{\mathbf{y}}_i / k$. The hypothesis $H_0: \bar{\mathbf{y}}_1 = \bar{\mathbf{y}}_2 = \dots = \bar{\mathbf{y}}_k$ (comparing the means of $y_1, y_2, \dots, y_p$ for factor A) can be expressed using contrasts:

$$H_0: \mathbf{C}\bar{\mathbf{y}}_{..} = \mathbf{0}, \quad (6.32)$$

where $\mathbf{C}$ is any $(p-1) \times p$ full-rank contrast matrix with $\mathbf{C}\mathbf{1} = \mathbf{0}$. This is equivalent to the “Thomson” test of profile analysis, the third test in Section 6.3. Under H_0 , the vector $\mathbf{C}\bar{\mathbf{y}}_{..}$ is distributed as $N_{p-1}(\mathbf{0}, \mathbf{C}\bar{\Sigma}\mathbf{C}^T/N)$, where $N = \sum_{i=1}^k n_i$ for an unbalanced design and $N = kn$ in the balanced case. We can, therefore, test H_0 with

$$T^2 = N(\mathbf{C}\bar{\mathbf{y}}_{..})^T(\mathbf{C}\bar{\Sigma}_p\mathbf{C}^T)^{-1}(\mathbf{C}\bar{\mathbf{y}}_{..}), \quad (6.33)$$

where $\bar{\Sigma}_p = \mathbf{E}/n_p$. The T^2 -statistic in (6.33) is distributed as $F_{p-1, N-p}^{(p)}$ when H_0 is true, where $n_p = N - k$ [see (6.30) and the comments following]. Note that the dimension of T^2 is $p - 1$, because $\mathbf{C}\bar{\mathbf{y}}_{..}$ is $(p - 1) \times 1$.

For the grouping of between-subjects factor B , we wish to compare the means for the k levels of B . The mean response for the i th level of B (averaged over the levels of A) is $\sum_{j=1}^p \mu_{ij}/p = \bar{\mu}_{B_i}$. The hypothesis can be expressed as

$$H_0: \bar{\mu}_{B_1} = \bar{\mu}_{B_2} = \cdots = \bar{\mu}_{B_k} \quad (5.94)$$

which is analogous to (5.80), the "levels" hypothesis in profile analysis. This is easily tested by calculating a univariate F -statistic for a one-way ANOVA on $\mu_{ij} = \bar{Y}_{ij}$, $i = 1, 2, \dots, k$, $j = 1, 2, \dots, p$. There is a μ_{ij} corresponding to each subject, X_{ij} . The observation vector for each subject is thus solved to a singular vector observation, and we have a one-way ANOVA comparing the means $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \dots, \bar{Y}_{k\cdot}$. (Note that $\bar{Y}_{k\cdot}/p$ is an average over the p levels of A .)

The AB interaction hypothesis is equivalent to the parallelism hypothesis in profile analysis [see (5.76)],

$$H_0: C\mu_{1j} = C\mu_{2j} = \cdots = C\mu_{kj} \quad (5.95)$$

In other words, differences or contrasts among the levels of factor A are the same across all levels of factor B . This is easily tested by performing a one-way MANOVA on $\alpha_{ij} = C\mu_{ij}$ or directly by

$$L = \frac{(CKL')}{(K)(E + H)C'} \quad (5.96)$$

[see (5.78)], which is distributed as $F_{(p-1)(k-1)}$ with $v_1 = k - 1$ and $v_2 = B - k$, that is, $v_2 = \sum_{j=1}^p (n_j - 1)$ for the unbalanced model or $v_2 = k(n - 1)$ in the balanced model.

5.5.4 Computation of Repeated Measures Tests

Some statistical software packages have automated repeated measures procedures that are easily implemented. However, if one is unsure as to how the resulting tests correspond to the tests in Section 5.5.3, there are two ways to obtain the tests directly. One approach is to calculate (5.95) and (5.96) outright using a matrix manipulation system. We would need to have available the $\mathbf{H}$ and $\mathbf{H}$ matrices of a one-way MANOVA using n -data layout as in Table 5.3.

The second approach uses simple data transformations available in virtually all programs. To test (5.95) for factor A , we would transform each $\mu_{ij} = C\mu_{ij}$ by using the rows of C . For example, if

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

then each $y' = (y_1, y_2, y_3, \dots, y_p)$ becomes $x' = (y_1 - y_3, y_2 - y_3, y_3 - y_3, \dots, y_p - y_3)$. We then test $H_0: \mu_3 = 0$ using a one-sample T^2 on all N of the $\tilde{y}_i$'s.

$$T^2 = N\tilde{y}'\tilde{S}_y^{-1}\tilde{y}$$

where $\tilde{y} = \sum_{i=1}^N \tilde{y}_i$, $\tilde{y}' = \sum_{i=1}^N \tilde{y}_i'/N$, and $\tilde{S}_y = \tilde{S}_{yy}/N$ is the pooled covariance matrix. Reject H_0 if $T^2 = T_{\alpha, p-1, N}^2$.

To test (5.24) for factor B , we sum the components of each observation vector to obtain $y_{1+} = \sum_{j=1}^p y_{1j} = y_{11} + y_{12} + \dots + y_{1p}$, and compute the means $\bar{y}_{1+}, \bar{y}_{2+}, \dots, \bar{y}_{q+}$ by an F -test, as in one-way MANOVA.

To test the interaction hypothesis (5.25), we transform each y_{ij} to $\tilde{y}_{ij} = Cy_{ij}$ using the rows of C , as before. Note that $\tilde{y}_{ij}$ is $(p-1) \times 1$. We then do a one-way MANOVA, using $\tilde{y}_{ij}$ to obtain

$$L = \frac{|E_L|}{|E_L + H_L|}, \quad (5.27)$$

5.5.2 Repeated Measures with Two Within-Subjects Factors and One Between-Subjects Factor

The repeated measures model with two within-subjects factors A and B and one between-subjects factor C corresponds to a one-way MANOVA design in which each observation vector includes measurements on a two-way factorial arrangement of treatments. Thus, each subject receives all treatment combinations of the two factors A and B . As usual, the sequence of administration of treatment combinations should be randomized for each subject. A design of this type is illustrated in Table 5.13.

Each p_{ij} in Table 5.13 has nine elements, consisting of responses to the nine treatment combinations $A_1B_1, A_1B_2, \dots, A_3B_3$. We are interested in the same hypotheses as in a univariate split-plot design, but we use a multivariate approach to allow for correlated y 's. The model for the observation vectors in the one-way MANOVA model

$$y_{ij} = \mu + \tau_i + \alpha_{ij} = \mu_0 + \alpha_{ij},$$

where τ_i is the C effect.

To test factors A , B , and AB in Table 5.13, we use contrasts in the y 's. As an example of contrast matrices, consider

$$A = \begin{pmatrix} 2 & 2 & 2 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix}, \quad (5.28)$$

$$B = \begin{pmatrix} 2 & -1 & -1 & 2 & -1 & -1 & 2 & -1 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \end{pmatrix}, \quad (5.29)$$

Table 6.13. Data Layout for Repeated Measures with Two Within-Subjects Factors and One Between-Subjects Factor

Between-Subjects Factor	Subjects	Within-Subjects Factors								
		A_1			A_2			A_3		
		B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
C_1	S_{11}	X_{111}	X_{112}	X_{113}	X_{121}	X_{122}	X_{131}	X_{132}	X_{133}	$X_{134} = X_{135}$
	S_{12}							F_{13}		
	$\vdots$							$\vdots$		
	S_{1m}							S_{1m3}		
C_2	S_{21}							F_{23}		
	S_{22}							F_{23}		
	$\vdots$							$\vdots$		
	S_{2m}							S_{2m3}		
C_3	S_{31}							F_{33}		
	S_{32}							F_{33}		
	$\vdots$							$\vdots$		
	S_{3m}							S_{3m3}		

$$G = \begin{pmatrix} 4 & -2 & -2 & -2 & 1 & 1 & -2 & 1 & 1 \\ 8 & 2 & -2 & 0 & -1 & 1 & 0 & -1 & 1 \\ 8 & 8 & 0 & 2 & -1 & -1 & -2 & 1 & 1 \\ 8 & 8 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}. \quad (6.100)$$

The rows of A are orthogonal contrasts with two components:

$$A_1 \text{ vs. } A_2 \text{ and } A_3,$$

$$A_2 \text{ vs. } A_3.$$

Similarly, B compares

$$B_1 \text{ vs. } B_2 \text{ and } B_3,$$

$$B_2 \text{ vs. } B_3.$$

(Other orthogonal (or linearly independent) contrasts could be used for A and B . The matrix G is for the AB interactions and is obtained from products of the corresponding elements of the rows of A and the rows of B .)

As before, we define $\bar{y}_{\cdot} = \sum_{ij} \bar{y}_{ij}/N$, $\bar{y}_{0j} = B_j/\nu_j$, and $\bar{y}_{\cdot} = \sum_{\cdot} \bar{y}_{\cdot}$. If there were k levels of C in Table 6.13 with mean vectors $\mu_{10}, \mu_{20}, \dots, \mu_{k0}$, then $\bar{y}_{\cdot} = \sum_{i=1}^k \mu_i/k$, and the A main effect corresponding to $A_2 : A_3\bar{y}_{\cdot} = 0$ could be tested

with

$$T^2 = N(\mathbf{BY})'(\mathbf{BY}_B\mathbf{B}')^{-1}(\mathbf{BY}_-), \quad (5.10)$$

which is distributed as T^2_{a, r_B} under H_0 , where $r_B = \sum_{i=1}^I (n_i - 1)$. The dimension is 2, corresponding to the two rows of $\mathbf{B}$.

Similarly, to test $H_0: \mathbf{B}\boldsymbol{\beta}_- = \mathbf{0}$ and $H_0: \mathbf{G}\boldsymbol{\beta}_- = \mathbf{0}$ for the B main effect and the AB interaction, respectively, we have

$$T^2 = N(\mathbf{BY})'(\mathbf{BY}_B\mathbf{B}')^{-1}(\mathbf{BY}_-), \quad (5.11)$$

$$T^2 = N(\mathbf{BY})'(\mathbf{BY}_B\mathbf{G}')^{-1}(\mathbf{BY}_-), \quad (5.12)$$

which are distributed as T^2_{a, r_B} and $T^2_{a, r_{BG}}$, respectively. In general, if factor A has a levels and factor B has b levels, then $\mathbf{A}$ has $a - 1$ rows, $\mathbf{B}$ has $b - 1$ rows, and $\mathbf{G}$ has $(a - 1)(b - 1)$ rows. The T^2 -statistics are then distributed as T^2_{a-1, r_B} , $T^2_{a-1, r_{BG}}$, and $T^2_{a-1, (a-1)r_B}$, respectively.

Factors A , B , and AB can be tested with Wilks' Λ as well as T^2 . Define $\mathbf{H}' = N\mathbf{Y}'\mathbf{Y}'$ from the partitioning $\sum_{i=1}^I \mathbf{Y}_i\mathbf{Y}_i' = \mathbf{K} + \mathbf{H} + N\mathbf{Y}'\mathbf{Y}'$. This can be used to test $H_0: \boldsymbol{\beta}_- = \mathbf{0}$ (not usually a hypothesis of interest) by means of

$$\Lambda = \frac{|\mathbf{H}|}{|\mathbf{K} + \mathbf{H}|}, \quad (5.13)$$

which is Λ_{a, r_B} if H_0 is true. Then the hypothesis of interest, $H_0: \mathbf{A}\boldsymbol{\beta}_- = \mathbf{0}$ for factor A , can be tested with

$$\Lambda = \frac{|\mathbf{A}\mathbf{H}\mathbf{A}'|}{|\mathbf{A}(\mathbf{K} + \mathbf{H})\mathbf{A}'|}, \quad (5.14)$$

which is distributed as Λ_{a-1, r_B} when H_0 is true, where a is the number of levels of factor A . There are similar expressions for testing factors B and AB . Note that the dimension of $\mathbf{A}$ is (b, K) ($a - 1$, because $\mathbf{A}\mathbf{H}\mathbf{A}'$ is $(a - 1) \times (a - 1)$).

The T^2 and Wilks' Λ expressions in (5.11) and (5.14) are related by

$$\Lambda = \frac{r_B}{r_B + T^2}, \quad (5.15)$$

$$T^2 = r_B \frac{1 - \Lambda}{\Lambda}. \quad (5.16)$$

We can establish (5.15) as follows. By (5.28),

$$\mathbf{U}^{(B)} = \sum_{i=1}^I \mathbf{U}_i = \mathbf{u}[(\mathbf{A}\mathbf{H}\mathbf{A}')^{-1}(\mathbf{A}\mathbf{H}'\mathbf{A}')]'$$

$$\begin{aligned}
&= \text{tr}[(\mathbf{A}\mathbf{E}\mathbf{E}\mathbf{E})^{-1}\mathbf{A}\mathbf{N}\mathbf{N}']\mathbf{y}'\mathbf{A}\mathbf{y} \\
&= N + \mathbf{y}'(\mathbf{A}\mathbf{E}\mathbf{A}')^{-1}\mathbf{A}\mathbf{y}_-(\mathbf{A}\mathbf{y}_-') \\
&= N + \mathbf{y}'_-(\mathbf{A}\mathbf{y}_-')(\mathbf{A}\mathbf{E}\mathbf{A}')^{-1}\mathbf{A}\mathbf{y}_- \\
&= \frac{N}{r_B}(\mathbf{A}\mathbf{y}_-')(\mathbf{A}\mathbf{E}_B\mathbf{E})^{-1}\mathbf{A}\mathbf{y}_- \\
&= \frac{\mathbf{y}_-'\mathbf{y}_-}{r_B}.
\end{aligned}$$

Since $\text{rank}(\mathbf{E}\mathbf{E}') = 1$, only λ_1 is nonzero, and

$$\mathbf{E}^{-1}\mathbf{E} = \sum_{i=1}^n \lambda_i^{-1} \mathbf{e}_i = \lambda_1^{-1}.$$

By (5.14),

$$\lambda = \frac{1}{1 + r_A} = \frac{1}{1 + \mathbf{E}^{-1}\mathbf{E}} = \frac{1}{1 + \mathbf{E}^{-1}/r_B},$$

which is the same as (5.100).

Factor C is tested exactly as factor B in Section 5.3.3. The hypothesis is

$$H_0: \bar{\mathbf{f}}_{p_1} = \bar{\mathbf{f}}_{p_2} = \cdots = \bar{\mathbf{f}}_{p_3},$$

as in (5.94), and we perform a multivariate F -test on $\mathbf{y}_{ij} = \bar{\mathbf{f}}\mathbf{y}_{ij}$ in a one-way ANOVA layout.

The AC , BC , and ABC interactions are tested as follows.

AC Interaction

The AC interaction hypothesis is

$$H_0: \mathbf{A}\mathbf{p}_1 = \mathbf{A}\mathbf{p}_2 = \cdots = \mathbf{A}\mathbf{p}_3,$$

which states that centroids in factor A are the same across all levels of factor C . This can be tested by

$$A = \frac{(\mathbf{A}\mathbf{E}\mathbf{A})}{(\mathbf{A}\mathbf{E} + \mathbf{E}\mathbf{A})}, \quad (5.105)$$

which is distributed as λ_1/r_{A+r_B} , where $r = 1 = 2$ is the number of rows of $\mathbf{A}$, and r_A and r_B are from the multivariate one-way model. Alternatively, the test can be carried out by transforming $\mathbf{y}_{ij}$ to $\mathbf{A}\mathbf{y}_{ij}$ and doing a one-way MANOVA on $\mathbf{A}\mathbf{y}_{ij}$.

BC Interaction

The *BC* interaction hypothesis,

$$\beta_{10} : \beta_{11} = \beta_{12} = \cdots = \beta_{1b},$$

is tested by

$$A = \frac{(\mathbf{BDB})}{(\mathbf{BDB} + \mathbf{I} \cdot \mathbf{B})},$$

which is $A_{L_{b-1}, b}$, where $b - 1 = 2$. H_0 can also be tested by doing MANOVA on $\mathbf{r}_1 = \mathbf{B}\mathbf{r}_b$.

ABC Interaction

The *ABC* interaction hypothesis,

$$\beta_0 : \beta_{10} = \beta_{20} = \cdots = \beta_{c0},$$

is tested by

$$A = \frac{(\mathbf{CDB})}{(\mathbf{CDB} + \mathbf{I} \cdot \mathbf{C})},$$

which is $A_{L_{c-1}, c}$, or by doing MANOVA on $\mathbf{r}_c = \mathbf{C}\mathbf{r}_b$. In this case the dimension is $bc - 1(3 - 1) = 4$.

The preceding tests for *AC*, *BC*, or *ABC* can be also carried out with the other three MANOVA test statistics using eigenvalues of the appropriate matrices. For example, for *AC* we would use $(\mathbf{ABN})^{-1}(\mathbf{ABN})$.

Example 6.5.5. The data in Table 6.14 represent a repeated measures design with two within-subjects factors and one between-subjects factor (Thom, 1980). Since *A* and *B* have three levels each, as in the illustration in this section, we will use the $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ matrices in (5.54), (5.55), and (5.56). The fixed $\mathbf{H}$ matrices are 3×3 and will not be shown. The overall mean vector is given by

$$\bar{\mathbf{y}}_c = (46.45, 38.25, 31.70, 34.85, 45.40, 40.15, 34.55, 36.80, 38.15).$$

By (5.54), the test for factor *A* is

$$\begin{aligned} F^2 &= N(\mathbf{A}\bar{\mathbf{y}}_c)'(\mathbf{ABN})^{-1}(\mathbf{A}\bar{\mathbf{y}}_c) \\ &= 20(-28, 15.85) \begin{pmatrix} 2158.4 & 138.6 \\ 138.6 & 456.9 \end{pmatrix}^{-1} \begin{pmatrix} -20 \\ 15.85 \end{pmatrix} \\ &= 8.645 > F_{26, 2, 38}^2 = 3.686. \end{aligned}$$

Table 6.1.4. Data From a Repeated Measures Experiment with Two Within-Subjects Factors and One Between-Subjects Factor

Between-Subjects Factor	Subjects	Within-Subjects Factors								
		A_1			A_2			A_3		
		B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
C_1	S_{11}	28	31	28	32	42	37	32	32	32
	S_{12}	37	48	28	41	36	48	38	40	41
	S_{13}	37	31	25	27	28	30	31	30	34
	S_{14}	42	40	38	37	36	36	38	37	33
	S_{15}	57	45	32	27	31	25	36	29	34
	S_{16}	39	35	38	44	34	43	33	29	38
	S_{17}	43	32	28	33	46	44	42	37	31
	S_{18}	38	34	34	39	43	39	33	39	42
	S_{19}	48	32	23	37	31	39	27	28	30
	$S_{1,10}$	39	32	24	30	28	31	28	29	32
C_2	S_{21}	47	35	29	34	36	29	33	28	37
	S_{22}	50	45	32	40	46	47	46	50	54
	S_{23}	38	38	33	38	42	43	48	48	49
	S_{24}	48	31	41	34	47	40	33	32	30
	S_{25}	37	36	23	40	43	40	34	40	46
	S_{26}	29	46	37	45	32	44	36	46	32
	S_{27}	67	50	33	47	46	46	33	41	50
	S_{28}	49	35	27	32	36	35	33	35	32
	S_{29}	64	39	33	56	62	51	46	40	43
	$S_{2,10}$	48	38	34	41	47	42	37	41	46

For factor B , we use (5.302) to obtain

$$\begin{aligned}
 F^B &= N(\mathbf{SS}_B)'(\mathbf{SS}_B\mathbf{R}^2)^{-1}(\mathbf{SS}_B) \\
 &= 26(7.15, 18.55) \begin{pmatrix} 365.7 & 94.0 \\ 94.0 & 29.8 \end{pmatrix}^{-1} \begin{pmatrix} 7.15 \\ 18.55 \end{pmatrix} \\
 &= 13.416 = F_{2,9,0.05}^B = 11.943.
 \end{aligned}$$

By (5.163), the test for the AB interaction is given by

$$\begin{aligned}
 F^B &= N(\mathbf{R}_B)'(\mathbf{R}_B\mathbf{R}_A\mathbf{R}^2)^{-1}(\mathbf{R}_B) \\
 &= 61.825 = F_{2,9,0.05}^{AB} = 21.697.
 \end{aligned}$$

To test factor C , we carry out a one-way ANOVA on $\mathbf{z}_C = \int \mathbf{z}_B/\mathbf{R}$.

Source	Sum-of-Squares	df	Mean Square	F
Between	3641.22	1	3641.22	8.34
Error	6408.98	18	356.05	

The observed F , 8.34, has a p -value of .0093 and is therefore significant.

The AC interaction is tested by $F(5,18)$ as

$$\begin{aligned} F &= \frac{(\mathbf{A}\mathbf{B}\mathbf{B})}{(\mathbf{A}\mathbf{B} + \mathbf{B}\mathbf{B})} = \frac{3.098 \times 10^6}{1.002 \times 10^6} \\ &= 3.089 > F_{.05,15,18} = 2.703. \end{aligned}$$

For the BC interaction, we have

$$\begin{aligned} F &= \frac{(\mathbf{B}\mathbf{B}\mathbf{B})}{(\mathbf{B}\mathbf{B} + \mathbf{B}\mathbf{B})} = \frac{4.059 \times 10^6}{4.179 \times 10^6} \\ &= .9718 < F_{.05,15,18} = 2.703. \end{aligned}$$

For ABC , we obtain

$$\begin{aligned} F &= \frac{(\mathbf{B}\mathbf{B}\mathbf{B})}{(\mathbf{B}\mathbf{B} + \mathbf{B}\mathbf{B})} = \frac{2.845 \times 10^6}{2.917 \times 10^6} \\ &= .9759 < F_{.05,15,18} = 2.703. \end{aligned}$$

In summary, factors A , B , and C and the AB interaction are significant. $\square$

6.5.6 Repeated Measures with Two Within-Subjects Factors and Two Between-Subjects Factors

In this section we consider a balanced two-way MANOVA design in which each observation vector arises from a two-way factorial arrangement of treatments. This is illustrated in Table 6.15 for a balanced design with three levels of all factors. Each $\mathbf{y}_{ijk}$ has nine elements, consisting of responses to the nine treatment combinations $A_1B_1, A_2B_1, \dots, A_3B_3$ (see Table 6.17).

To test A , B , and AB , we can use the same contrast matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{G}$ as in (6.95)–(6.100). We define a grand mean vector $\bar{\mathbf{y}}_{..} = \sum_{ijk} \mathbf{y}_{ijk}/N$, where N is the total number of observation vectors in this illustration, $N = 27$. In general, $N = cab$, where c and a are the number of levels of factors C and D and b is the number of replications in each cell in the illustration, $b = 3$. The test statistics for A , B , and AB are as follows, where $\mathbf{B}_0 = \mathbf{K}/n_T$ and the $\mathbf{B}$ matrix is obtained from the two-way MANOVA with $n_T = ca(b-1)$ degrees of freedom.

Table 6.18. Data Layout for Repeated Measures with Two Within-Subjects Factors and One Between-Subjects Factor

Between-Subjects Factors			Within-Subjects Factors								
<i>C</i>	<i>D</i>	Subject	<i>A</i> ₁			<i>A</i> ₂			<i>A</i> ₃		
			<i>a</i> ₁₁	<i>a</i> ₁₂	<i>a</i> ₁₃	<i>a</i> ₂₁	<i>a</i> ₂₂	<i>a</i> ₂₃	<i>a</i> ₃₁	<i>a</i> ₃₂	<i>a</i> ₃₃
<i>C</i> ₁	<i>A</i> ₁	<i>S</i> ₁₀₁							<i>F</i> ₁₀₁		
		<i>S</i> ₁₀₂							<i>F</i> ₁₀₂		
		<i>S</i> ₁₀₃							<i>F</i> ₁₀₃		
	<i>A</i> ₂	<i>S</i> ₂₀₁							<i>F</i> ₂₀₁		
		<i>S</i> ₂₀₂							<i>F</i> ₂₀₂		
		<i>S</i> ₂₀₃							<i>F</i> ₂₀₃		
	<i>A</i> ₃	<i>S</i> ₃₀₁							<i>F</i> ₃₀₁		
		<i>S</i> ₃₀₂							<i>F</i> ₃₀₂		
		<i>S</i> ₃₀₃							<i>F</i> ₃₀₃		
<i>C</i> ₂	<i>A</i> ₁	<i>S</i> ₄₀₁							<i>F</i> ₄₀₁		
		<i>S</i> ₄₀₂							<i>F</i> ₄₀₂		
		<i>S</i> ₄₀₃							<i>F</i> ₄₀₃		
	<i>A</i> ₂	<i>S</i> ₅₀₁							<i>F</i> ₅₀₁		
		<i>S</i> ₅₀₂							<i>F</i> ₅₀₂		
<i>C</i> ₃	<i>A</i> ₁	⋮							⋮		
		⋮							⋮		
	<i>A</i> ₃	<i>S</i> ₆₀₁							<i>F</i> ₆₀₁		

Factor *A*

$$T^2 = N(\mathbf{R}\mathbf{I}_p - \mathbf{I})'(\mathbf{R}\mathbf{S}_p\mathbf{R})^{-1}(\mathbf{R}\mathbf{I}_p - \mathbf{I})$$

is distributed as $F_{p-1, Np}^2$ Factor *B*

$$T^2 = N(\mathbf{R}\mathbf{I}_p - \mathbf{I})'(\mathbf{R}\mathbf{S}_p\mathbf{R})^{-1}(\mathbf{R}\mathbf{I}_p - \mathbf{I})$$

is distributed as $F_{p-1, Np}^2$ *AB Interaction*

$$T^2 = N(\mathbf{R}\mathbf{I}_p - \mathbf{I})'(\mathbf{R}\mathbf{S}_p\mathbf{R})^{-1}(\mathbf{R}\mathbf{I}_p - \mathbf{I})$$

is distributed as $F_{(p-1)(p-1), Np}^2$

To test factors C , D , and CD , we transform to $z_{ijk} = \bar{y}_{ijk}$ and carry out univariate F -tests in a two-way ANOVA design.

To test factors AC , AD , and ACD , we perform a two-way MANOVA on Ag_{ijk} . Then, the C main effect on Ag_{ijk} compares the levels of C on Ag_{ijk} , which is an effective description of the AC interaction. Similarly, the D main effect on Ag_{ijk} yields the AD interaction, and the CD interaction on Ag_{ijk} gives the ACD interaction.

To test factors BC , BD , and BCD , we carry out a two-way MANOVA on Bg_{ijk} . The C main effect on Bg_{ijk} gives the BC interaction, the D main effect on Bg_{ijk} yields the BD interaction, and the CD interaction on Bg_{ijk} corresponds to the BCD interaction.

Finally, to test factors ABC , ABD , and $ABCD$, we perform a two-way MANOVA on Cg_{ijk} . Then the C main effect on Cg_{ijk} gives the ABC interaction, the D main effect on Cg_{ijk} yields the ABD interaction, and the CD interaction on Cg_{ijk} corresponds to the $ABCD$ interaction.

6.8.7 Additional Topics

Wang (1983) and Tien (1980) give a method for obtaining univariate mixed-model sums of squares from the multivariate B and H matrices. Cepani et al. (1985) consider repeated measures experiments with missing data. Pedersen (1988) discusses the planning of repeated measures designs, emphasizing such aspects as determining the length of treatment period, eliminating carry-over effects, the nature of pre- and posttreatment, the nature of a response to a treatment, treatment sequences, and the choice of a model. Worselt (1985) discusses sample size requirements to achieve a given power level in repeated measures designs. Paul (1986) presents a model that accommodates both within- and between-subjects covariates in repeated measures designs. Jensen (1982) compares the efficiency and robustness of various procedures.

A multivariate or multiresponse repeated-measures design will result if more than one variable is measured on each subject at each treatment combination. Such designs are discussed by Tien (1980), Keimel (1982), Wang (1983), and Thomas (1983). Buck (1975) refers to observations of this type as doubly multivariate data.

6.10 GROWTH CURVES

When the subject responds to a treatment or stimulus at successive time periods, the pattern of responses is often referred to as a growth curve. As in repeated measures experiments, subjects are usually human or animal. We consider estimators and testing hypotheses about the form of the response curve for a single sample in Section 6.10.1 and extend to growth curves for several samples in Section 6.10.2.

6.10.1 Growth Curve for One Sample

The data layout for a single sample growth curve experiment is analogous to Table 5.11, with the levels of factor A representing time periods. Thus we have

a sample of n observation vectors $y_1, y_2, \dots, y_n$, for which we compute $\bar{y}$ and S . The usual approach is to approximate the shape of the growth curve by a polynomial function of time. If the time points are equally spaced, we can use orthogonal polynomials. This approach will be described first, followed by a method suitable for unequal time intervals.

Orthogonal polynomials are special contrasts that are often used in testing for linear, quadratic, cubic, and higher order trends in quantitative factors. For a more complete description and derivation see Ottman (1992, pp. 194–207), Morrison (1983, pp. 182–198), or Rencher (2000, pp. 323–331). Here we give only a heuristic introduction to the use of these contrasts.

Suppose we administer a drug to some subjects and measure a certain reaction at 3-min intervals. Let $\mu_1, \mu_2, \mu_3, \mu_4$, and μ_5 designate the average responses at 0, 3, 6, 9, and 12 min, respectively. To test the hypothesis that there are no trends in the μ_i 's, we could test $H_0: \mu_1 = \mu_2 = \dots = \mu_5$ or $H_0: C\mu = 0$ using the contrast matrix

$$C = \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 2 & -1 & -2 & -1 & 2 \\ -1 & 2 & 0 & -2 & 1 \\ 1 & -4 & 6 & -4 & 1 \end{pmatrix} \quad (6.108)$$

in $F^2 = n(C'VC)^{-1}C'\bar{y}$, as in (6.90). The four rows of C are orthogonal polynomials that test for linear, quadratic, cubic, and quartic trends in the means. As noted in Section 6.1.1, any set of orthogonal contrasts in C will give the same value of F^2 to test $H_0: \mu_1 = \mu_2 = \dots = \mu_5$. However, in this case we will be interested in using a subset of the rows of C to determine the shape of the response curve.

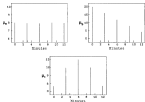
Table A.15 (Kleinbaum, Kupper, and Muller 1988) gives orthogonal polynomials for $p = 3, 4, \dots, 10$. The $p - 1$ entries for each value of p constitute the matrix C . Some software programs will generate these automatically.

As with all orthogonal contrasts, the rows of C in (6.108) sum to zero and are mutually orthogonal. It is also apparent that the coefficients in each row increase and decrease in conformity with the desired pattern. Thus the entries in the first row, $(-2, -1, 0, 1, 2)$, increase steadily in a straightline trend. The values in the second row dip down and back up in a quadratic-type bend. The third-row entries increase, decrease, then increase in a cubic pattern with two-bends. The fourth row bends three times in a quartic curve.

To further illustrate how the orthogonal polynomials pinpoint trends in the means when testing $H_0: C\mu = 0$, consider the three different patterns for μ depicted in Figure 6-4, where $\mu_0' = (0, 0, 0, 0, 0)$, $\mu_1' = (20, 16, 12, 8, 4)$, and $\mu_2' = (5, 12, 18, 12, 5)$. Let us denote the rows of C in (6.108) as c_1', c_2', c_3' , and c_4' . It is clear that $c_i'\mu_0 = 0$ for $i = 1, 2, 3, 4$; that is, when $H_0: \mu_1 = \dots = \mu_5$ is true, all four comparisons confirm it. If μ has the pattern μ_1 , only $c_1'\mu_1$ is nonzero. The other rows are not sensitive to a linear pattern. We illustrate this for c_1' and c_2' :

$$c_1'\mu_1 = (-2)(20) + (-1)(16) + (0)(12) + (1)(8) + (2)(4) = -44,$$

$$c_2'\mu_1 = 2(20) - 1(16) - 2(12) - 1(8) + 2(4) = 0.$$

Figure 6.4. Three different patterns for μ .

For μ , only $d_{j\mu}$ is nonzero. For example,

$$d_{1\mu} = -2(3) - 12 + 12 + 2(3) = 0,$$

$$d_{2\mu} = 2(3) - 12 - 2(3) - 12 + 2(3) = -18.$$

Thus each orthogonal polynomial independently detects the type of curvature it is designed for and ignores other types. Of course real curves generally exhibit a mixture of more than one type of curvature, and in practice more than one orthogonal polynomial contrast may be significant.

To test hypotheses about the shape of the curve, we therefore use the appropriate rows of C in (5.10). Suppose we suspected a priori that there would be a combined linear and quadratic trend. Then we would partition C as follows:

$$C_1 = \begin{pmatrix} -3 & -1 & 0 & 1 & 3 \\ 2 & -1 & -2 & -1 & 2 \end{pmatrix},$$

$$C_2 = \begin{pmatrix} -1 & 3 & 6 & -2 & 1 \\ 1 & -4 & 6 & -4 & 1 \end{pmatrix}.$$

We would test $H_0: C_1\mu = 0$ by

$$T^2 = n(C_1'Y'YC_1)^{-1}C_1'Y,$$

which is distributed as $T_{2, n-1}^2$, where 2 is the number of rows of C_1 , n is the number of subjects in the sample, and $\bar{Y}$ and $\bar{S}$ are the mean vector and covariance matrix for

the sample. Similarly, $H_0: C_2\mu = 0$ is tested by

$$F^2 = s(C_2g)'(C_2gS_1^{-1})^{-1}(C_2g),$$

which is $F_{2, p-2}^2$. In this case we might expect the first to reject M_0 and the second to accept M_0 .

If we have no a priori expectations as to the shape of the curve, we would proceed as follows. Test the overall hypothesis $H_0: C\mu = 0$, and if M_0 is rejected, test each of the rows of C separately to test $M_{0i}: \mu_{ji} = 0, i = 1, 2, 3, 4$. The respective test statistics are

$$F^i = ss(C_i g)'(S_1 C_i')^{-1}(C_i g),$$

which is $F_{1, p-1}^2$ and

$$\lambda_i = \frac{r_i^2}{\sum_{j=1}^4 \lambda_j}, \quad i = 1, 2, 3, 4,$$

each of which is distributed as $\chi_{1, p-1}^2$ (see Example 6.9.2).

In a case where p is large so that μ has a large number of levels, say 10 or more, we would likely want to stop testing after the first few or few rows of C and test the remaining rows in one group. However, for larger values of p , most tables of orthogonal polynomials give only the first few rows and omit those corresponding to higher degrees of curvature. We can find a matrix whose rows are orthogonal to the rows of a given matrix as follows. Suppose $p = 11$ so that C is 10×11 and C_1 contains the first five orthogonal polynomials. Then a matrix C_2 , with rows orthogonal to those of C_1 , can be obtained by selecting five linearly independent rows of

$$B = I - C_1'(C_1 C_1')^{-1}C_1, \quad (6.11.6)$$

whose rows can easily be shown to be orthogonal to those of C_1 . The matrix B is not full rank, and some care must be exercised in choosing linearly independent rows. However, if an incorrect choice of C_2 is made, the computer algorithm should indicate this as it attempts to invert $C_2 g S_1^{-1}$ in $F^2 = s(C_2 g)'(C_2 g S_1^{-1})^{-1}(C_2 g)$.

Alternatively, to check for significant curvature beyond the rows of C_1 without finding C_2 , we can use the test for additional information in a subset of variables in Section 5.8. We need not find C_2 in order to find the overall F^2 , since, as noted in Section 6.9.2, any full rank $(p-1) \times p$ matrix C such that $C\mathbf{1} = 0$ will give the same value for the overall F^2 -test of $H_0: C\mu = 0$. We can conveniently use a simple contrast matrix such as

$$C = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

is

$$T^2 = s(CY^T(CSY)^{-1}CY), \quad (8.11.1)$$

which is $F_{p-1, p-1}^2$. Let p_1 be the number of orthogonal polynomials in C_1 and p_2 be the number of rows of C_2 if it were available; that is $p_1 + p_2 = p - 1$. Then the test statistic for the p_1 orthogonal polynomials in C_1 is

$$T_1^2 = s(C_1\bar{y}(C_1SY)^{-1}C_1\bar{y}), \quad (8.11.2)$$

which is $F_{p_1, p-1}^2$. We wish to compare T_1^2 in (8.11.2) to T^2 in (8.11.1) to check for significant curvature beyond the rows of C_1 . However, the test for additional information is a subset of variables in Section 5.8 was for the two-sample case. We can adapt (5.29) for use with the one-sample case, as follows. The test for significance of any curvature remaining after that accounted for in C_1 is made by comparing

$$W = p_1 - 1 - \frac{T^2 - T_1^2}{s - 1 + T_1^2}$$

with the critical value $F_{p_1, p-1-p_1}^1$.

We now describe an approach that can be used when the time points are not equally spaced. It may also be of interest in the equal-time-increment case because it provides an estimate of the response function.

Suppose we observe the response of the subject at p time points $t_1, t_2, \dots, t_p$ and that the average response μ at any time point t is a polynomial in t of degree $d \leq p$:

$$\mu = \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_d t^d.$$

This holds for each point t_i and the corresponding average response μ_i . Thus our hypothesis becomes

$$H_0: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 t_1 + \beta_2 t_1^2 + \dots + \beta_d t_1^d \\ \beta_0 + \beta_1 t_2 + \beta_2 t_2^2 + \dots + \beta_d t_2^d \\ \vdots \\ \beta_0 + \beta_1 t_p + \beta_2 t_p^2 + \dots + \beta_d t_p^d \end{pmatrix}, \quad (8.11.3)$$

which can be expressed in matrix notation as

$$H_0: \mu = A\beta, \quad (8.11.4)$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^p \\ 1 & t_2 & t_2^2 & \cdots & t_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^p \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}.$$

In practice, it may be useful to transform the t_i 's by subtracting the mean or the smallest value in order to reduce their size for computational purposes.

The following method of testing H_0 is due to Rao (1958, 1973). The model $\mu = \mathbf{A}\boldsymbol{\beta}$ is similar to a regression model $E(y) = \mathbf{X}\boldsymbol{\beta}$ (see Section 10.2.1). However, in this case, we have $\text{cov}(y) = \mathbf{I}$ rather than $\sigma^2\mathbf{I}$, as in the standard regression assumption. In place of the usual regression approach of seeking $\hat{\boldsymbol{\beta}}$ to minimize $\text{SSE} = (y - \mathbf{X}\hat{\boldsymbol{\beta}})'(y - \mathbf{X}\hat{\boldsymbol{\beta}})$ (see (10.4) and (10.5)), we use a standardized distance as in (3.83), $(y - \mathbf{A}\hat{\boldsymbol{\beta}})'(\mathbf{S}^{-1}(y - \mathbf{A}\hat{\boldsymbol{\beta}}))$. The value of $\hat{\boldsymbol{\beta}}$ that minimizes $(y - \mathbf{A}\hat{\boldsymbol{\beta}})'(\mathbf{S}^{-1}(y - \mathbf{A}\hat{\boldsymbol{\beta}}))$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}y \quad (6.115)$$

[see Renscher (2008, Section 7.8.1.3)], and $M_{\text{DF}} = \mu = \mathbf{A}\hat{\boldsymbol{\beta}}$ can be tested by

$$F^2 = n(y' - \mathbf{A}\hat{\boldsymbol{\beta}}'(\mathbf{S}^{-1}(y - \mathbf{A}\hat{\boldsymbol{\beta}}))), \quad (6.116)$$

which is distributed as $F_{p-k, k+1, n-k}$. The dimension of F^2 is reduced from p to $p - k + 1$ because $k + 1$ parameters have been estimated in $\hat{\boldsymbol{\beta}}$. The F^2 -statistic in (6.116) is usually given in the equivalent form

$$F^2 = n(y'\mathbf{S}^{-1}y - y'\mathbf{S}^{-1}\mathbf{A}\hat{\boldsymbol{\beta}}). \quad (6.117)$$

The mean response at the i th time point,

$$\begin{aligned} \mu_i &= \mu_0 + \mu_1 t_i + \mu_2 t_i^2 + \cdots + \mu_p t_i^p \\ &= (\zeta_1, \zeta_2, \zeta_3^2, \dots, \zeta_1^p)\boldsymbol{\beta} = \mathbf{a}_i'\boldsymbol{\beta}, \end{aligned}$$

can be estimated by

$$\hat{\mu}_i = \mathbf{a}_i'\hat{\boldsymbol{\beta}}. \quad (6.118)$$

Simultaneous confidence intervals for all possible $\mathbf{a}'\boldsymbol{\beta}$ are given by

$$\mathbf{a}'\hat{\boldsymbol{\beta}} \pm \frac{T_\alpha}{\sqrt{n}} \sqrt{\mathbf{a}'(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{a} \left(1 + \frac{F^2}{n-k-1}\right)}, \quad (6.119)$$

where $T_\alpha = \sqrt{F_{\alpha, k+1, n-k}^2}$ is from Table A.7 and F^2 is given by (6.115) or (6.117).

The interval in (5.118) for $\alpha'\beta$ includes, of course, $\alpha'\beta$ for the p rows of $\mathbf{A}$, that is, confidence intervals for the p time points. If $\alpha'\beta, j = 1, 2, \dots, p$, are the only values of interest, we can shorten the intervals in (5.118) by using a Bonferroni coefficient $t_{\alpha/2p}$ in place of T_{α} :

$$\alpha'\beta \pm \frac{t_{\alpha/2p}}{\sqrt{p}} \left\{ \alpha'(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A} \left(1 + \frac{p^2}{n-1} \right) \right\}, \quad (6.120)$$

where $t_{\alpha/2p} = t_{\alpha/2p, n-1}$. Bonferroni critical values $t_{\alpha/2p, n-1}$ are given in Table A.8. See procedures 2 and 3 in Section 5.5 for additional comments on the use of $t_{\alpha/2p}$ and T_{α} .

Example 6.10.1. Porttuff and Roy (1964) reported measurements in a dental study on boys and girls from ages 8 to 14. The data are given in Table 6.10.

To illustrate the methods of this section, we use the data for the boys alone. In Example 6.10.2 we will compare the growth curves of the boys with those of the girls. We first test the overall hypothesis $H_0: C\mu = 0$, where C contains orthogonal polynomials for linear, quadratic, and cubic effects:

$$C = \begin{pmatrix} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & -3 & 1 \end{pmatrix}. \quad (6.121)$$

Table 6.10. Dental Measurements

Subject	Girls' Ages in Years				Subject	Boys' Ages in Years			
	8	10	12	14		8	10	12	14
1	21.0	26.8	28.5	32.0	1	26.0	32.0	29.8	34.0
2	21.0	26.5	26.0	25.5	2	21.5	22.5	23.8	26.5
3	20.5	24.8	26.5	26.0	3	25.0	22.5	24.8	27.5
4	18.5	24.5	25.0	26.5	4	25.5	27.5	26.5	27.0
5	21.5	23.8	22.5	23.5	5	20.0	18.5	22.5	26.0
6	20.0	21.8	21.8	22.8	6	26.8	28.5	27.8	26.8
7	21.8	22.8	20.8	23.0	7	22.0	22.0	24.5	26.8
8	23.0	23.8	20.5	24.0	8	26.0	21.5	24.5	28.8
9	20.0	21.8	22.0	21.5	9	21.0	20.5	18.8	20.0
10	19.5	19.8	19.8	19.5	10	27.5	26.0	30.8	31.5
11	24.5	25.8	26.8	26.0	11	25.0	23.0	23.5	25.0
					12	21.5	23.5	24.8	26.0
					13	17.0	14.8	26.8	26.8
					14	22.8	23.8	23.8	26.0
					15	26.0	24.5	26.8	30.0
					16	22.0	21.5	23.5	25.0

From the 1st observation vector we obtain

$$\bar{\mathbf{y}} = \begin{pmatrix} 21.88 \\ 21.81 \\ 21.72 \\ 21.47 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 8.02 & 2.79 & 3.87 & 1.81 \\ 2.79 & 4.56 & 2.13 & 2.81 \\ 3.87 & 2.13 & 7.03 & 3.24 \\ 1.81 & 2.81 & 3.24 & 4.39 \end{pmatrix}.$$

To test $H_0: C\boldsymbol{\mu} = \mathbf{0}$, we calculate

$$\Gamma^2 = n(\mathbf{C}\bar{\mathbf{y}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{y}}) = 17.971,$$

which exceeds $F_{20,23,41}^1 = 15.581$. We now test $H_0: c_j\boldsymbol{\mu} = \mathbf{0}$ for each row of $\mathbf{C}$ to determine the shape of the growth curve. For the linear trend, using the first row of $\mathbf{C}$, we obtain

$$t_1 = \frac{c_1'\bar{\mathbf{y}}}{\sqrt{c_1'\mathbf{S}c_1/n}} = 2.772 < t_{20,23,41} = 2.847.$$

The test of significance of the quadratic component, using the second row yields

$$t_2 = \frac{c_2'\bar{\mathbf{y}}}{\sqrt{c_2'\mathbf{S}c_2/n}} = 1.379 < t_{20,23,41} = 2.131.$$

To test for a cubic trend, we use the third row of $\mathbf{C}$:

$$t_3 = \frac{c_3'\bar{\mathbf{y}}}{\sqrt{c_3'\mathbf{S}c_3/n}} = -2.51 > -t_{20,23,41} = -2.134.$$

Thus only the linear trend is needed to describe the growth curve.

To model the curve for each variable, we use (8.115),

$$\begin{aligned} \mu_x &= \beta_0 + \beta_1 x, & x &= 1, 2, 3, 4, \quad \text{or} \\ \boldsymbol{\mu} &= \mathbf{A}\boldsymbol{\beta}, \end{aligned}$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

The values in the second column of $\mathbf{A}$ are obtained as $t = \text{age} - 11$. By (8.115), we obtain

$$\hat{\beta} = (A'S^{-1}A)^{-1}A'S^{-1}Y = \begin{pmatrix} 25.002 \\ .834 \end{pmatrix},$$

and our prediction equation is

$$\begin{aligned}\hat{\mu} &= 25.002 + .834t = 25.002 + .834(\text{age} - 11) \\ &= 15.828 + .834(\text{age}).\end{aligned}\quad (3)$$

6.3.2 Growth Curves for Several Samples

For the case of several samples or groups, the data layout would be similar to that in Table 6.5, where the p levels of factor A represent time points. Assuming the time points are equally spaced, we can use orthogonal polynomials in the $(p-1) \times p$ contrast matrix C and express the basic hypothesis in the form $H_0: C\bar{\mu} = \mathbf{0}$, where $\bar{\mu} = \sum_{i=1}^p \mathbf{a}_i/\mathbf{1}'$. This is equivalent to $H_0: \bar{\mu}_1 = \bar{\mu}_2 = \bar{\mu}_3 = \cdots = \bar{\mu}_p$, which compares the means of the p time points averaged across groups. As in Section 6.3.1, let us denote the sample mean vectors for the k groups as $\bar{\mathbf{y}}_{1k}, \bar{\mathbf{y}}_{2k}, \dots, \bar{\mathbf{y}}_{pk}$, with grand mean $\bar{\mathbf{y}}_{..}$ and pooled covariance matrix $\mathbf{S}_{p2} = \mathbf{R}_2/v_2$. For the overall test of $H_0: C\bar{\mu} = \mathbf{0}$ we use the test statistic

$$T^2 = N(\mathbf{C}\bar{\mathbf{y}}_{..})'(\mathbf{C}\mathbf{S}_{p2}\mathbf{C}')^{-1}(\mathbf{C}\bar{\mathbf{y}}_{..}), \quad (6.122)$$

which is $T_{p-1, N-p}^2$ as in (5.37), where $N = \sum_{i=1}^p n_i$ for unbalanced data or $N = k$ for balanced data. The corresponding degrees of freedom for error is given by $v_2 = N - k$ or $v_2 = k(\alpha - 1)$. A test that the average growth curve (averaged over groups) has a particular form can be tested with C_1 , containing a subset of the rows of C :

$$T^2 = N(\mathbf{C}_1\bar{\mathbf{y}}_{..})'(\mathbf{C}_1\mathbf{S}_{p2}\mathbf{C}_1')^{-1}(\mathbf{C}_1\bar{\mathbf{y}}_{..}), \quad (6.123)$$

which is distributed as $F_{p_1, N-p}^2$, where p_1 is the number of rows in C_1 .

The growth curves for the k groups can be compared by the interaction or population test of Section 6.4.3 using either C or C_1 . We do a one-way MANOVA on $C\mathbf{y}_{ik}$ or $C_1\mathbf{y}_{ik}$, or equivalently calculate by (6.95)

$$A = \frac{|\mathbf{C}\mathbf{B}\mathbf{C}'|}{|\mathbf{C}(\mathbf{B} + \mathbf{B}\mathbf{C}'\mathbf{C})|} \quad \text{or} \quad A = \frac{|\mathbf{C}_1\mathbf{B}\mathbf{C}_1'|}{|\mathbf{C}_1(\mathbf{B} + \mathbf{B}\mathbf{C}_1'\mathbf{C}_1)|} \quad (6.124)$$

which are distributed as $A_{p-p_1, N-p_1}$ and $A_{p_1, N-p_1}$, respectively.

Example 6.3.2. In Example 6.1.1, we found a linear trend for the growth curve for dental measurements of boys in Table 6.10. We now consider the growth curve for the combined group and also compare the girls' group with the boys' group.

The two sample sizes are unequal and we use (5.37) to calculate the $\mathbf{B}$ matrix for the two groups.

$$\mathbf{E} = \begin{pmatrix} 115.39 & 67.88 & 91.76 & 67.76 \\ 67.88 & 105.76 & 72.86 & 82.71 \\ 97.76 & 72.86 & 161.59 & 103.17 \\ 67.76 & 82.71 & 103.17 & 124.64 \end{pmatrix},$$

from which we obtain $\hat{\mathbf{S}}_{\hat{\mu}} = \mathbf{E}/v_{\hat{\mu}}$. Using the $\hat{\mathbf{C}}$ matrix in (5.111), we can test the basic hypothesis of equal means for the combined samples, $H_0: \mathbf{C}\hat{\mu} = \mathbf{0}$, using (5.122):

$$\begin{aligned} F^1 &= N(\mathbf{C}\hat{\mathbf{y}}_1)'(\mathbf{C}\hat{\mathbf{S}}_{\hat{\mu}}\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{y}}_1) \\ &= 118.522 = T_{2,3,4,11}^2 = 25.538. \end{aligned}$$

To test for a linear trend, we use the first row of $\mathbf{C}$ in (5.123):

$$\begin{aligned} F^2 &= N(\mathbf{c}_1'\hat{\mathbf{y}}_1)'(\mathbf{c}_1'\hat{\mathbf{S}}_{\hat{\mu}}\mathbf{c}_1)^{-1}(\mathbf{c}_1'\hat{\mathbf{y}}_1) \\ &= 39.445 = T_{2,3,4,11}^2 = 1.776. \end{aligned}$$

This is, of course, the square of a t -statistic, but in the F^2 form it can readily be compared with the preceding T^2 using all three rows of $\mathbf{C}$. The linear trend is seen to dominate the relationship among the means.

We now compare the growth curves of the two groups using (5.124). For $\mathbf{C}$, we obtain

$$\begin{aligned} A &= \frac{|\mathbf{C}\mathbf{B}\mathbf{C}'|}{|\mathbf{C}|(\mathbf{K} + \mathbf{B}\mathbf{B}')|} = \frac{1.1696 \times 10^6}{1.9025 \times 10^6} \\ &= .776 = A_{(2,3,4,1,1,2)} = .717. \end{aligned}$$

For the linear trend, we have

$$\begin{aligned} A &= \frac{|\mathbf{c}_1'\mathbf{B}\mathbf{c}_1|}{|\mathbf{c}_1'(\mathbf{K} + \mathbf{B}\mathbf{B}')|} = \frac{1184.2}{1427.9} \\ &= .829 = A_{(2,3,4,1,1,2)} = .855. \end{aligned}$$

Thus the overall comparison does not reach significance, but the more specific comparison of linear trends does give a significant result. $\square$

6.6.3 Additional Topics

Lukacs and Brown (1981) presented methods of analyzing growth curves based on univariate linear models. Snoco (1972) and Simon, Acosta, and Gilman (1979) proposed the use of eigenvalues and eigenvectors of covariates derived from residuals after fitting the model. If one of the eigenvalues is dominant, certain simplifications result. Brown (1980) discussed a similar simplification for the two-group case. Gelman (1980) and Fourn (1973, 1977) gave the Bayesian approach to growth curves, including estimation and prediction. Zar (1979a, b) provided a randomization test requiring fewer assumptions than normal-based tests.

6.11 TESTS ON A SUBVECTOR

6.11.1 Test for Additional Information

In Section 5.8, we considered tests of significance of the additional information in a subvector when comparing two groups. We now extend these concepts to several groups and use similar notation.

Let $\mathbf{y}$ be a $p \times 1$ vector of measurements and $\mathbf{x}$ be a $q \times 1$ vector measured in addition to $\mathbf{y}$. We are interested in determining whether $\mathbf{x}$ makes a significant contribution to the test of $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ above and beyond $\mathbf{y}$. Another way to phrase the question is, Can the separation of groups achieved by $\mathbf{x}$ be predicted from the separation achieved by $\mathbf{y}$? It is not necessary, of course, that $\mathbf{x}$ represent new variables. It may be that $\mathbf{y}$ is a partitioning of the present variables, and we wish to know if the variables in $\mathbf{x}$ can be deleted because they do not contribute to rejecting H_0 .

We consider here only the one-way MANOVA, but the results could be extended to higher order designs, where various possibilities arise. In a two-way context, for example, it may happen that $\mathbf{x}$ contributes nothing to the A main effect but does contribute significantly to the B main effect.

It is assumed that we have k samples,

$$\left\{ \begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix} \right\}, \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, p,$$

from which we calculate

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{yy} & \mathbf{K}_{yx} \\ \mathbf{K}_{xy} & \mathbf{K}_{xx} \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} \mathbf{H}_{yy} & \mathbf{H}_{yx} \\ \mathbf{H}_{xy} & \mathbf{H}_{xx} \end{pmatrix},$$

where $\mathbf{K}$ and $\mathbf{H}$ are $(p+q) \times (p+q)$ and $\mathbf{K}_{yy}$ and $\mathbf{H}_{yy}$ are $p \times p$.

Then

$$\Lambda(\mathbf{y}, \mathbf{x}) = \frac{|\mathbf{K}|}{|\mathbf{K} + \mathbf{H}|} \quad (6.125)$$

is distributed as $\Lambda_{p, q, p, q}$ and tests the significance of group separation using the full vector $\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix}$. In the balanced one-way model, the degrees of freedom are $v_1 = k-1$ and $v_2 = k(p-1)$. To test group separation using the reduced vector $\mathbf{y}$, we can compute

$$\Lambda(\mathbf{y}) = \frac{|\mathbf{K}_{yy}|}{|\mathbf{K}_{yy} + \mathbf{H}_{yy}|}, \quad (6.126)$$

which is distributed as $\Lambda_{p, 0, p, 0}$.

To test the hypothesis that the extra variables in $\mathbf{x}$ do not contribute anything significant to separating the groups beyond the information already available in $\mathbf{y}$,

we calculate

$$A(x) = \frac{A(x, y, z)}{A(y, z)}, \quad (6.125)$$

which is distributed as $A_{q, n_2, n_2 - p}$. Note that the dimension of $A(x, y, z)$ is q , the number of x 's. The error degrees of freedom, $n_2 - p$, has been adjusted for the p y 's. Thus to test for the contribution of additional variation to separation of groups, we take the ratio of 'Wilks' Λ for the full set of variables in (6.123) to 'Wilks' Λ for the reduced set in (6.126). If the addition of q makes $A(x, y, z)$ sufficiently smaller than $A(y, z)$, then $A(x, y, z)$ in (6.127) will be small enough to reject the hypothesis.

If we are interested in the effect of adding a single x , then $q = 1$, and (6.125) becomes

$$A(x, y_1, \dots, y_p) = \frac{A(y_1, \dots, y_p, x)}{A(y_1, \dots, y_p)}, \quad (6.128)$$

which is distributed as $A_{1, n_2, n_2 - p}$. In this test we are inquiring whether x reduces the overall Λ by a significant amount. With a dimension of 1, the Λ -statistic in (6.128) has an exact F -transformation from Table 6.1,

$$F = \frac{1 - \Lambda}{\Lambda} \frac{n_2 - p}{n_2}, \quad (6.129)$$

which is distributed as $F_{n_2, n_2 - p}$. The statistic (6.128) is often referred to as a *partial* Λ -statistic; correspondingly, (6.129) is called a *partial* F -statistic.

In (6.128) and (6.129), we have a test of the significance of a variable in the presence of the other variables. For a breakdown of precisely how the contribution of a variable depends on the other variables, see Rencher (1995; 1998, Section 4.1.6).

We can rewrite (6.128) as

$$A(y_1, \dots, y_p, x) = A(y_1, \dots, y_p) W(y_1, \dots, y_p) \leq A(y_1, \dots, y_p), \quad (6.130)$$

which shows that 'Wilks' Λ can only decrease with an additional variable.

Example 6.3.1. We use the rockfish data of Table 6.2 to illustrate tests on subvariables. From Example 6.1.7, we have, for all four variables, $A(y_1, y_2, y_3, y_4) = .1540$. For the first two variables, we obtain $A(y_1, y_2) = .6890$. Thus to test the significance of y_3 and y_4 adjusted for y_1 and y_2 , we have by (6.127),

$$A(y_3, y_4 | y_1, y_2) = \frac{A(y_1, y_2, y_3, y_4)}{A(y_1, y_2)} = \frac{.1540}{.6890} = .2235,$$

which is less than the critical value $A_{.05, 1, 1, 40} = .839$.

Similarly, the test for y_4 adjusted for y_1 , y_2 , and y_3 is given by (6.128) as

$$\begin{aligned} A(y_4|y_1, y_2, y_3) &= \frac{A(y_1, y_2, y_3, y_4)}{A(y_1, y_2, y_3)} = \frac{.1540}{.3885} \\ &= .3966 = A_{.05, 1, 3, 3} = .759. \end{aligned}$$

For each of the other variables, we have a similar test:

$$y_1: A(y_1|y_2, y_3, y_4) = \frac{.1540}{.2541} = .6018 = A_{.05, 1, 3, 3} = .759,$$

$$y_2: A(y_2|y_1, y_3, y_4) = \frac{.1540}{.1922} = .8014 = A_{.05, 1, 3, 3} = .759,$$

$$y_3: A(y_3|y_1, y_2, y_4) = \frac{.1540}{.1599} = .9638 > A_{.05, 1, 3, 3} = .759.$$

Thus the two variables y_3 and y_4 , either individually or together, contribute a significant amount to separation of the six groups. $\square$

6.11.2 Stepwise Selection of Variables

If there are no variables for which we have a priori interest in testing for significance, we can do a data-directed search for the variables that best separate the groups. Such a strategy is often called *stepwise discriminant analysis*, although it could more aptly be called *stepwise MANOVA*. The procedure appears in many software packages.

We first describe an approach that is usually called *forward selection*. At the first step calculate $A(y_i)$ for each individual variable and choose the one with minimum $A(y_i)$ for maximum associated F . At the second step calculate $A(y_i|y_1)$ for each of the $p - 1$ variables not entered at the first step, where y_1 indicates the first variable entered. For the second variable we choose the one with minimum $A(y_i|y_1)$ for maximum associated partial F , that is, the variable that adds the maximum separation to the one entered at step 1. Denote the variable entered at step 2 by y_2 . At the third step calculate $A(y_i|y_1, y_2)$ for each of the $p - 2$ remaining variables and choose the one that minimizes $A(y_i|y_1, y_2)$ for maximum associated partial F . Continue this process until the F falls below some predetermined threshold value, say, F_{th} .

A *backward procedure* follows a similar sequence, except that after a variable has entered, the variables previously selected are reexamined to see if each still contributes a significant amount. The variable with smallest partial F will be removed if the partial F is less than a second threshold value, F_{th2} . If F_{th2} is the same as F_{th} , there is a very small possibility that the procedure will cycle continuously without stopping. This possibility can be eliminated by using a value of F_{th2} slightly less than F_{th} . For an illustration of the stepwise procedure, see Example 6.3.

PROBLEMS

6.1 Verify the computational forms given in (6.3) and (6.5); that is, show that

$$(a) \sum_{i=1}^n (Y_i - \bar{Y}_1)^2 = \sum_{i=1}^n Y_i^2 - \sum_{i=1}^n Y_i^2/n,$$

$$(b) \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2/n - n\bar{Y}^2/n.$$

6.2 Show that Wilks' A can be expressed in terms of the eigenvalues of $E^{-1}H$ as in (6.14).

6.3 Show that the eigenvalues of $E^{-1}H$ are the same as those of $(E^{1/2})^{-1}HE^{1/2}E_1^{-1}$, as noted in Section 6.1.4, where $E^{1/2}$ is the square root matrix defined in (2.11.2).

6.4 Show that F_2 in (6.27) is the same as F_1 in (6.25).

6.5 Show that if $p \geq n_2$, then F_2 in (6.31) is the same as F_1 in (6.30).

6.6 Show that if there is only one nonzero eigenvalue λ_1 , then $E^{1/2}$, $V^{1/2}$, and A can be expressed in terms of E , as in (6.34), (6.36), and (6.38).

6.7 Show that (5.16), (5.18), and (5.19), which relate T^2 to A , $V^{1/2}$, and E , follow from (6.34)–(6.36) and (6.38), $V^{1/2} = T^2/(n_1 + n_2 - 2)$.

6.8 Verify the computational forms of H and E in (6.31) and (6.32), that is, show that

$$(a) \sum_{i=1}^n x_i(\bar{Y}_1 - \bar{Y})(\bar{Y}_1 - \bar{Y})' = \sum_{i=1}^n x_i Y_i Y_i' / n_1 - n_1 \bar{Y}' \bar{Y} / n_1,$$

$$(b) \sum_{i=1}^n \sum_{j=1}^n (Y_{ij} - \bar{Y}_1)(Y_{ij} - \bar{Y}_1)' = \sum_{i=1}^n \sum_{j=1}^n Y_{ij} Y_{ij}' - \sum_{i=1}^n n_2 \bar{Y}_1' \bar{Y}_1 / n_1.$$

6.9 Show that for two groups, $H = \sum_{i=1}^n x_i(\bar{Y}_1 - \bar{Y})(\bar{Y}_1 - \bar{Y})'$ can be expressed as $H = [n_1 n_2 / (n_1 + n_2)] (\bar{Y}_1 - \bar{Y}_2)(\bar{Y}_1 - \bar{Y}_2)'$, thus verifying (6.38). Note that

$$\bar{Y}_1 = \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{n_1 + n_2}.$$

6.10 Show that θ can be expressed as $\theta = \text{SSB}(x_1)/\text{SSB}(x_2) + \text{SSB}(x_2)$ as in (6.42).

6.11 Show that

$$\prod_{i=1}^p \frac{1}{1 + \lambda_i} = \prod_{i=1}^p (1 - c_i^2),$$

as in (6.45), where $c_i^2 = \lambda_i / (1 + \lambda_i)$.

6.12 Show that the F -approximation based on A_F in (6.50) reduces to (6.56) if $\lambda_1 = V^{1/2} \lambda_1$, as in (6.49).

6.13 Show that if $i = 1$, A_{1E} in (6.51) reduces to (6.48).

6.14 Show that the F -approximation denoted by F_2 in (6.51) is equivalent to (6.52).

6.15 Show that $\text{acc}(\hat{\theta}) = \frac{1}{2} \sum_{i=1}^p c_i^2$ as in (6.64).

- 6.16 If $\mathbf{B}_0 = \mathbf{C}\mathbf{Y}_{(p)}$, where $\mathbf{C}$ is $(p-1) \times p$, show that $\mathbf{B}_0 = \mathbf{C}\mathbf{B}\mathbf{C}'$ and $\mathbf{B}_0 = \mathbf{C}\mathbf{B}\mathbf{C}'$, as used in (5.19).
- 6.17 Why do $\mathbf{C}$ and $\mathbf{C}'$ not “cancel out” of Wilks’ A in (5.20)?
- 6.18 Show that under H_{00} and H_{01} , $\mathbf{C}\mathbf{Y}_{(p)} \sim N_{p-1}(\mathbf{0}, \mathbf{C}\mathbf{B}\mathbf{C}'/\lambda_1)$, as noted preceding (5.19).
- 6.19 Show that $\mathbf{T}^2 = \det(\mathbf{Y}'/\mathbf{C}\mathbf{B}\mathbf{C}'/\lambda_1)^{-1}\mathbf{C}\mathbf{Y}_{(p)}$ in (5.19) is distributed as $\Gamma_{p-1, \lambda_1}^2$.
- 6.20 For τ defined by (5.19), show that $\tau = 1$ when $\mathbf{X} = \mathbf{a}^2\mathbf{I}$.
- 6.21 Give a justification of the Wilks’ A test of $H_0: \boldsymbol{\mu}_1 = \mathbf{0}$ in (5.20).
- 6.22 Provide an alternative derivation of (5.19), $A = \mathbf{y}_{(p)}'\mathbf{y}_{(p)} + \mathbf{T}^2$, starting with (5.18).
- 6.23 Obtain $\mathbf{T}^2$ in terms of A in (5.19), starting with (5.20).
- 6.24 Show that the rows of $\mathbf{C}_1$ are orthogonal to those of $\mathbf{B} = \mathbf{I} - \mathbf{C}_1\mathbf{C}_1'\mathbf{C}_1^{-1}\mathbf{C}_1$ in (5.19).
- 6.25 Show that $\hat{\boldsymbol{\beta}}$ in (5.15) minimizes $(\mathbf{Y} - \mathbf{A}\hat{\boldsymbol{\beta}})'\mathbf{B}^{-1}(\mathbf{Y} - \mathbf{A}\hat{\boldsymbol{\beta}})$.
- 6.26 Show that $\mathbf{T}^2$ in (5.17) is equivalent to $\mathbf{T}^2$ in (5.19).
- 6.27 Bates, Tuck, and Hunter (1958) compared judges’ scores on fish prepared by three methods. Twelve fish were cooked by each method, and several judges tasted fish samples and rated each on five variables: $y_1 = \text{taste}$, $y_2 = \text{flavor}$, $y_3 = \text{texture}$, and $y_4 = \text{moisture}$. The data are in Table 6.27. Each entry is an average score for the judges on that fish.
- (a) Compare the three methods using all four MANOVA tests.

Table 6.27. Judges’ Scores on Fish Prepared by Three Methods

Method 1				Method 2				Method 3			
y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
9.4	8.8	8.9	8.7	9.8	9.3	9.3	8.9	4.8	5.0	4.9	7.0
8.2	8.2	8.8	9.8	4.8	4.9	4.2	3.6	3.4	3.0	4.0	6.4
6.1	5.8	6.8	7.8	3.8	4.8	4.4	3.0	4.8	3.1	3.9	6.5
4.8	5.8	4.8	5.8	4.8	5.1	4.8	3.8	3.7	3.2	4.4	6.4
5.8	5.7	5.8	6.5	5.6	5.4	5.1	6.1	4.2	4.6	5.3	6.3
5.7	6.1	6.8	6.6	6.8	5.5	5.7	6.0	6.0	5.3	5.8	6.4
6.8	6.8	5.8	6.8	5.2	4.8	5.4	6.0	5.1	5.2	6.2	6.5
4.8	5.8	4.8	5.8	5.3	5.3	5.8	6.4	4.8	4.6	5.7	5.7
8.7	8.4	4.8	8.8	8.4	6.3	5.7	6.0	8.3	8.4	6.8	6.8
5.8	5.2	5.4	5.8	6.1	6.8	6.1	6.2	4.6	4.4	5.7	5.8
5.8	6.1	5.2	6.4	6.2	5.7	5.8	6.0	4.3	4.0	5.0	5.9
5.9	5.9	5.8	6.8	5.1	4.8	5.3	4.8	4.4	4.2	5.5	5.5

Source: Bates, Tuck, and Hunter (1958, p. 8).

- (b) Compute the following measures of multivariate association from Section 5.1.8: η^2_{ij} , η^2_{ij} , A_{ij} , A_{2ij} , A_{3j} .
- (c) Based on the eigenvalues, is the essential dimensionality of the space containing the mean vectors equal to 1 or 2?
- (d) Using contrasts, test the following two comparisons of methods: 1 and 2 vs. 3, and 1 vs. 2.
- (e) If any of the four tests in (c) is significant, run an ANCOVA F -test on each y_i and determine the discriminant function for $i = 1, 2$ (Section 6.4).
- (f) Test the significance of y_1 and y_2 adjusted for y_3 and y_4 .
- (g) Test the significance of each variable adjusted for the other three.

6.28 Table 6.18, from Kozia, Matuszko, and Biegaj (1994), gives data from a two-way (fixed-effects) MANOVA on maple leaves showing the results of five vari-

Table 6.18. Maple-leaf Data

S	F		x_1	x_2	x_3	x_4	S	F		x_1	x_2	x_3	x_4
1	1	1	50.5	4.5	364	295	3	1	1	48.3	5.4	41.3	280
		2	49.5	4.5	365	302			2	48.5	5.9	41.4	284
		3	49.9	5.5	37.5	318			3	48.5	5.5	40.5	286
		4	49.0	5.8	36.1	345			4	48.6	5.1	40.9	288
		5	49.4	6.8	36.8	329			5	48.6	5.5	41.1	294
1	2	1	50.5	4.5	374	279	3	2	1	48.8	5.6	40.8	283
		2	50.4	4.8	366	299			2	48.4	5.9	41.4	288
		3	49.0	5.1	38.7	295			3	48.5	5.7	40.6	284
		4	50.9	5.8	37.5	296			4	48.4	5.7	40.4	286
		5	50.5	4.8	37.0	308			5	48.5	4.1	40.1	284
1	3	1	50.4	5.1	36.7	299	3	3	1	48.6	5.7	41.3	285
		2	49.2	5.5	37.0	315			2	48.7	5.5	40.6	284
		3	49.7	6.4	37.4	304			3	48.7	5.8	40.7	277
		4	49.8	7.1	37.0	302			4	48.4	5.8	41.8	299
		5	49.1	7.8	36.9	298			5	48.6	5.4	41.4	283
2	1	1	49.7	5.4	39.5	271	4	1	1	48.8	5.4	40.4	285
		2	49.1	5.4	39.2	284			2	48.5	5.5	40.8	287
		3	48.4	5.4	39.0	288			3	48.5	5.5	40.9	284
		4	48.2	5.5	39.0	296			4	48.9	5.5	40.5	298
		5	48.2	5.8	39.0	279			5	48.3	5.1	40.1	287
2	2	1	49.6	6.8	39.1	288	4	2	1	48.6	5.8	40.7	288
		2	49.0	6.8	38.6	284			2	48.8	5.7	40.8	289
		3	49.7	6.8	38.8	287			3	48.1	5.7	40.3	285
		4	49.5	7.1	38.0	289			4	48.8	5.8	40.3	285
		5	49.5	6.8	38.5	288			5	48.6	5.9	40.1	285
2	3	1	49.8	5.7	40.5	282	4	3	1	48.3	5.7	40.3	285
		2	49.2	6.1	40.1	284			2	48.5	5.7	40.8	286
		3	49.5	6.8	40.0	291			3	48.7	5.8	40.7	288
		4	49.2	6.8	40.0	289			4	48.9	5.5	40.1	288
		5	48.1	5.4	39.7	285			5	48.8	5.4	40.7	286

where y_0 = yield coefficient, y_1 = specific leaf area (SLA)-coefficient, y_2 = total yield, and y_3 = average SLA. The factors are sowing date (2) and variety (3).

- Test for main effects and interaction using all four MANOVA statistics.
- In previous experiments, the second variety gave higher yields. Compare variety 2 with varieties 1 and 3 by means of a test on a contrast.
- Test linear, quadratic, and cubic contrasts for sowing date. (Interpretations of these for mean vectors is not as straightforward as for univariate means.)
- Repeat of the tests in part (a) rejects H_0 , carry out ANOVA F-tests on the four variables.
- Test for significance of y_2 and y_3 adjusted for y_0 and y_1 in main effects and interaction.
- Test the significance of each variable adjusted for the other three in main effects and interaction.

6.29 The bar steel data in Table 6.6 were analyzed in Example 5.3.2 as a two-way fixed-effects design. Consider lubricants to be random so that we have a mixed model. Test for main effects and interaction.

6.30 In Table 6.19, we have a comparison of four reagents (Bretzick 1979). The first reagent is the one presently in use and the other three are less expensive.

Table 6.19. Mixed Data

Subject	Reagent 1			Reagent 2			Reagent 3			Reagent 4		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
1	8.0	1.96	12.5	6.8	1.95	12.7	7.8	5.88	13.0	7.8	5.87	13.2
2	4.0	5.77	16.8	4.2	5.38	17.2	4.0	5.59	17.1	4.8	5.58	17.5
3	6.5	5.47	17.1	6.5	5.36	17.8	6.8	5.59	17.1	6.0	5.44	17.4
4	9.4	5.36	18.2	9.4	5.36	18.7	9.4	5.17	18.7	9.0	5.18	18.7
5	8.2	5.36	17.8	8.0	5.13	17.8	8.1	5.15	17.4	7.8	5.12	17.8
6	11.0	4.87	19.2	10.7	4.80	19.7	10.8	4.50	19.6	10.2	4.58	19.7
7	6.8	5.20	18.2	6.8	5.16	18.5	6.8	5.13	18.8	6.7	5.13	18.8
8	9.0	4.85	19.7	9.0	4.59	19.8	8.8	4.58	19.0	8.6	4.55	19.1
9	6.1	5.21	18.5	6.0	5.16	18.8	6.0	5.14	18.8	6.0	5.18	18.8
10	6.4	5.11	18.8	6.4	5.11	18.4	6.4	5.11	18.4	6.3	5.07	18.5
11	5.6	4.47	18.2	5.6	4.48	18.6	5.3	4.48	18.6	5.3	4.44	18.7
12	8.2	5.22	18.8	8.2	5.14	18.8	8.8	5.11	18.8	7.8	5.18	18.8
13	5.7	5.10	18.8	5.6	5.08	19.3	5.5	5.02	18.4	5.4	5.05	18.5
14	9.8	5.25	18.1	9.8	5.15	18.8	8.1	5.13	18.8	9.4	5.18	18.6
15	5.9	5.28	18.8	5.8	5.25	18.4	5.7	5.26	18.4	5.6	5.29	18.2
16	6.6	4.65	18.8	6.6	4.58	18.2	6.5	4.58	18.1	6.4	4.57	18.2
17	5.7	4.45	19.5	5.5	4.34	19.8	5.5	4.30	19.9	5.4	4.32	19.8
18	6.7	4.88	18.1	6.5	4.82	18.4	6.5	4.52	18.6	6.3	4.81	18.5
19	6.8	4.87	18.8	6.8	4.87	18.8	6.8	4.88	18.6	6.8	4.88	18.8
20	9.6	5.44	17.8	9.5	5.38	17.8	8.3	5.80	17.4	9.2	5.68	17.8

Table 6.20. Wear of Coated Fabric over Three Periods (mg)

Surface Treatment	Filter	Proportion of Filler								
		P_1 (25%)			P_2 (50%)			P_3 (75%)		
		F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
S_1	F_1	104	102	144	155	177	171	265	282	207
		208	188	168	241	220	204	290	283	194
		269	177	95	224	123	79	243	177	100
S_2	F_1	197	105	85	243	123	110	226	123	73
		155	109	154	198	197	175	235	223	166
		173	152	141	177	195	157	220	230	181
	F_2	137	92	77	125	54	76	133	75	52
	F_3	150	92	85	98	85	48	132	105	57

also reports that we wish to compare with the first. All four reports are used with a blood sample from each patient. The three variables measured for each report are y_1 = white blood count, y_2 = red blood count, and y_3 = hemoglobin count.

(a) Analyze as a randomized block design with subjects as blocks.

(b) Compare the first report with the other three using a contrast.

6.21 The data in Table 6.20, from Box (1976), show the amount of fabric wear (y_1 , y_2 , and y_3) in three successive periods: (1) the first 1000 revolutions, (2) the second 1000 revolutions, and (3) the third 1000 revolutions of the abrasive wheel. There were three factors: type of abrasive surface, type of filter, and proportion of filler. There were two replications. Carry out a three-way MANOVA, testing for main effects and interactions. (Ignore the repeated measure aspects of the data.)

6.22 The fabric wear data in Table 6.20 can be considered to be a growth curve model, with the three periods (y_1 , y_2 , y_3) representing repeated measurements on the same specimen. We thus have one within-subjects factor, to which we should assign polynomial contrasts ($-1, 0, 1$) and ($-1, 1, -1$), and a three-way between-subjects classification. Test for period and the interaction of period with the between-subjects factors and interactions.

6.23 Carry out a profile analysis on the fish data in Table 6.13, testing for parallelism, equal levels, and flatness.

6.24 Rao (1948) measured the weight of cork borings taken from the north (N), east (E), south (S), and west (W) directions of 28 trees. The data are given in Table 6.21. It is of interest to compare the bark thickness (and hence weight) in the four directions. This can be done by analyzing the data as a one-sample repeated measures design. Since the primary comparison of interest is north and south vs. east and west, use the contrast matrix

Table 8.21. Weights of Cork Straps (g) in Four Directions for 20 Trees

Tree	N	E	S	W	Tree	N	E	S	W
1	71	66	76	77	15	81	74	100	75
2	60	53	65	65	16	56	68	47	50
3	66	57	64	68	17	76	63	70	61
4	63	58	56	60	18	81	80	68	78
5	52	52	58	66	19	78	53	67	60
6	50	55	54	55	20	46	68	57	68
7	59	54	54	57	21	59	55	54	57
8	45	43	54	55	22	52	56	50	50
9	57	46	54	55	23	66	56	67	54
10	53	54	57	56	24	55	57	48	56
11	52	56	54	56	25	59	56	56	51
12	63	43	74	66	26	66	54	57	60
13	54	46	65	50	27	43	57	59	50
14	47	51	52	46	28	46	54	55	46

$$C = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

(a) Test $H_0: \mu_N = \mu_E = \mu_S = \mu_W$ using the matrix C .

(b) If the test in (a) rejects H_0 , test each row of C .

8.25 Analyze the glucose data in Table 5.8 as a one-sample repeated measures design with two within-subjects factors. Factor A is a comparison of fasting test vs. 1 hour posttest. The three levels of factor B are y_1 (and u_1), y_2 (and u_2), and y_3 (and u_3).

8.26 Table 8.22 gives survival times for cancer patients (Cameron and Finkel, 1978; see also Anderson and Hartzberg 1985, pp. 283–286). The factors in this two-way design are gender (1 = male, 2 = female) and type of cancer (1 = stomach, 2 = breast, 3 = colon, 4 = rectum, 5 = bladder, 6 = kidney). The variables (repeated measures) are y_1 = survival time (days) of patient

Table 8.22. Survival Times for Cancer Patients

Type of Cancer	Gender	Age	y_1	y_2	y_3	y_4
1	2	61	128	264	128	88
1	1	69	42	62	12	68
1	2	62	25	149	19	36
1	2	65	45	18	45	52
1	1	65	415	188	167	64
1	1	70	81	143	26	28
1	1	76	1112	39	128	53

Table 4.12. (Continued)

Type of Error	Gender	Age	F ₁	F ₂	F ₃	F ₄
1	1	54	46	209	40	51
1	1	62	104	83	60	18
1	1	66	146	304	120	52
1	1	67	140	209	160	28
1	2	50	196	190	150	55
2	1	54	61	72	76	10
2	1	54	404	154	405	18
2	1	66	20	64	16	20
2	1	62	480	58	490	58
2	2	48	146	48	87	11
2	2	66	166	112	115	49
2	1	70	68	115	50	38
2	1	67	64	90	50	24
2	1	71	115	50	115	18
2	1	50	151	260	50	24
2	1	70	166	116	150	20
2	1	70	37	67	37	37
2	1	66	113	60	205	30
2	1	70	188	100	188	27
2	1	66	71	115	30	30
2	1	70	145	160	24	65
3	2	70	146	200	105	18
3	2	58	177	400	50	30
3	1	48	160	400	160	65
3	1	66	643	268	1267	17
3	2	70	180	204	188	37
3	2	66	177	144	208	16
3	1	50	169	640	200	25
3	2	70	415	304	126	21
3	1	66	406	140	90	17
3	2	70	365	641	365	40
3	2	50	641	272	641	40
3	2	70	171	37	660	28
3	1	66	169	160	166	31
3	2	66	104	114	60	20
3	1	77	20	669	20	33
3	1	66	260	100	270	60
4	2	50	165	410	60	30
4	2	75	470	60	208	18
4	2	57	673	551	457	60
4	1	66	113	140	80	11
4	1	66	162	106	120	20
4	1	50	141	665	166	60
4	1	50	173	607	750	60
5	1	60	426	404	260	20

Table 6.22. (Continued)

Type of Cancer	Gender	Age	y_1	y_2	y_3	y_4
3	2	70	508	494	505	32
4	2	77	41	231	37	21
3	2	72	278	498	109	18
3	1	44	348	433	37	11
6	2	71	205	102	8	81
6	2	65	334	177	55	47
6	2	59	504	147	150	35
6	1	55	246	288	54	34
6	1	57	470	289	240	19
6	1	74	111	983	126	27
6	1	69	1493	1876	45	13

tested with y_1 -score measured from date of first hospital admission, y_2 = mean survival time for the patient's 10 matched controls (treated with second), y_3 = survival time after immediate treatment (control) and y_4 = mean survival time after all treatment ceased for the patient's 10 matched controls. Analysis is a repeated measures design with one within-subjects factor (y_1, y_2, y_3, y_4) and a two-way (unbalanced) design between subjects. Since the two-way classification of subjects is unbalanced, you will need to use a program that allows for this or delete some observations to achieve a balanced design.

6.27 Analyze the mean/force data of Table 5.4 as a one-sample growth curve design.

- (a) Using a matrix C of orthogonal polynomial contrasts, test the hypothesis of overall equality of means, $M_0: C\mu = 0$.

Table 6.23. Weights of 12 Male Mice Measured at Successive Intervals of 3 Days over 21 Days from Birth to Weaning

Mouse	Day 3	Day 6	Day 9	Day 12	Day 15	Day 18	Day 21
1	1.98	.988	.621	.825	1.078	1.150	1.191
2	28.8	.804	.968	1.729	.819	.850	1.004
3	26.1	.964	.649	1.900	.783	.879	.613
4	.288	.419	.649	.850	1.081	1.028	1.089
5	1.93	.962	.928	.590	.641	.680	.731
6	.288	.904	.982	.510	.637	.760	.888
7	.282	.390	.498	.650	.783	.858	.910
8	1.98	.550	.518	.680	.819	.879	.924
9	28.9	.959	.578	.690	.789	.822	.913
10	.229	.400	.949	.690	.796	.828	.836
11	.224	.981	.377	.736	.869	.909	.994
12	1.87	.329	.441	.528	.589	.621	.756
13	.278	.411	.888	.770	.888	1.008	1.103

- (b4) If the overall hypothesis in (a) is rejected, find the degree of growth curve by testing each row of C .
- 6.28** Table 6.23 contains the weights of 12 male mice measured every 3 days from birth to weaning. The data set was reported and analyzed by Williams and Ironman (1981) and by Ironman and Williams (1983) and has been further analyzed by Rao (1984, 1987) and by Lee (1988). Analyze as a one-sample growth curve design.
- (a) Using a matrix C of orthogonal polynomial contrasts, test the hypothesis of overall equality of means, $H_0: C\mu = 0$.
- (b4) If the overall hypothesis in (a) is rejected, find the degree of growth curve by testing each row of C .
- 6.29** In Table 6.24, we have measurements of proportions of albumin at four time points on three groups of trout (Brenschamp and Ilović 1974).
- (a) Using a matrix C of orthogonal contrasts, test the hypothesis of overall equality of means, $H_0: C\mu = 0$, for the combined samples, as in Section 6.18.2.
- (b4) If the overall hypothesis is rejected, find the degree of growth curve for the combined samples by testing each row of C .
- (c1) Compare the three groups using the entire matrix C .
- (b) Compare the three groups using each row of C .

Table 6.23. Measurements of Trout

Group	Time Point			
	1	2	3	4
1	.257	.288	.328	.358
1	.266	.282	.345	.404
1	.256	.303	.326	.381
1	.270	.436	.288	.281
2	.342	.388	.275	.270
2	.280	.228	.344	.281
2	.238	.261	.279	.224
2	.258	.243	.308	.254
3	.270	.278	.250	.295
3	.246	.282	.279	.302
3	.280	.271	.258	.284
3	.290	.261	.244	.264

- 6.40** Table 6.25 contains weight gains for three groups of rats (Bee 1959).

The variables are y_{ij} = gain in i th week, $j = 1, 2, 3, 4$.

Three groups are 1 = controls, 2 = Bayonin added to drinking water, and 3 = thioacetol added to drinking water.

Table 6.25. Weekly Output in Weight for 27 Rats

Row	Group 1				Group 2					Group 3				
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
1	28	28	23	33	31	26	36	25	35	18	25	23	11	9
2	33	30	23	31	32	17	19	28	28	19	21	21	10	11
3	25	34	33	41	32	19	30	40	38	28	26	21	6	27
4	38	33	29	33	34	28	31	30	29	21	29	12	11	11
5	33	33	17	30	19	15	29	20	24	22	24	26	22	17
6	34	32	29	33	38	26	34	19	24	23	24	17	8	19
7	26	24	36	31	37	18	35	33	33	28	22	17	8	5
8	28	21	19	24						25	11	24	21	24
9	19	23	22	28						26	15	17	12	17
10	25	28	29	30						27	19	17	15	18

- Using a matrix C of orthogonal contrasts, test the hypothesis of overall equality of means, $H_0: C\bar{\mu}_i = \mathbf{0}$, for the combined samples, as in Section 6.10.2.
- If the overall hypothesis is rejected, find the degree of growth curve for the combined samples by testing each row of C .
- Compare the three groups using the entire matrix C .
- Compare the three groups using each row of C .

6.41 Table 6.26 contains measurements of coronary sinus potassium at 2-min intervals after coronary occlusion on four groups of dogs (Diaz and Allen 1959). The groups are 1 = control dogs, 2 = dogs with extrinsic vagal denervation 3 wk prior to coronary occlusion, 3 = dogs with extrinsic vagal denervation immediately prior to coronary occlusion, and 4 = dogs with bilateral thoracic sympathectomy and atellectomy 3 wk prior to coronary occlusion.

Table 6.26. Coronary Sinus Potassium Measured at 2-min Intervals on Dogs

Group	Time						
	1	3	5	7	9	11	13
1	4.0	4.3	4.1	3.6	3.6	3.8	3.1
1	4.2	4.3	3.7	3.7	4.0	3.9	3.2
1	4.0	4.2	4.3	4.3	4.0	3.8	3.4
1	4.2	4.4	4.4	4.9	5.0	5.0	4.8
1	4.6	4.4	5.3	5.6	5.9	5.9	5.3
1	5.1	5.6	4.9	5.1	5.5	4.2	4.0
1	5.7	5.9	5.9	4.8	5.2	5.4	4.2
1	4.9	4.2	4.4	5.1	5.6	5.4	4.7
1	4.8	4.8	4.4	4.6	5.4	5.9	5.6
2	3.6	3.4	3.9	3.1	3.1	3.7	3.3

Table 6.26. (Continued)

Group	Time						
	1	2	3	7	9	11	15
1	30	3.1	3.2	3.6	3.3	3.8	3.8
1	30	3.2	3.6	3.6	3.1	3.2	3.1
1	3.1	3.2	3.2	3.2	3.5	3.1	3.1
1	3.6	3.3	4.8	3.9	3.5	3.5	3.4
1	3.0	3.6	3.2	3.1	3.0	3.8	3.8
1	3.3	3.3	3.3	3.4	3.6	3.1	3.1
1	4.2	4.8	4.2	4.1	4.2	4.8	4.8
1	4.1	4.2	4.3	4.3	4.2	4.8	4.2
1	4.5	4.4	4.3	4.5	5.3	4.4	4.4
2	3.2	3.3	3.8	3.8	4.4	4.2	3.7
2	3.5	3.4	3.4	3.7	3.7	3.8	3.7
2	3.1	3.3	3.2	3.1	3.2	3.1	3.1
2	3.6	3.4	3.8	4.6	4.9	3.2	4.4
2	4.3	4.3	3.4	3.7	4.9	4.8	4.8
2	3.7	4.8	4.4	4.2	4.6	4.8	3.4
2	3.5	3.8	3.8	3.4	4.9	5.3	3.8
2	3.9	4.8	4.3	3.8	5.4	4.4	3.9
4	3.1	3.5	3.5	3.2	3.0	3.8	3.2
4	3.5	3.2	3.6	3.7	3.7	4.2	4.4
4	3.5	3.8	4.3	4.3	5.0	3.4	3.3
4	3.4	3.4	3.8	3.3	3.4	3.2	3.4
4	3.7	3.8	4.2	4.3	3.6	3.8	3.7
4	4.0	4.8	4.8	4.9	3.4	3.8	4.8
4	4.2	3.9	4.5	4.7	3.9	3.8	3.7
4	4.1	4.3	3.7	4.6	4.1	4.6	4.7
4	3.5	3.6	3.6	4.1	4.6	4.9	5.6

(a) Using a matrix C of orthogonal contrasts, test the hypothesis of overall equality of means, $H_0: C\mu = 0$, for the combined samples, as in Section 6.10.2.

(b) If the overall hypothesis is rejected, find the degree of growth curve for the combined samples by testing each row of C .

(c) Compare the four groups using the entire matrix C .

(d) Compare the four groups using each row of C .

6.43 Table 6.27 contains blood pressure measurements at intervals after inducing a heart attack for four groups of rats; group 1 is the controls and groups 2–4 have been exposed to halothane concentrations of .25%, .50%, 1.0%, respectively (Crepas et al., 1983).

(a) Find the degree of growth curve for the combined sample using the methods in (5.1.13)–(5.1.18).

Table 6.21. Blood Pressure Data

Group	Number of Minutes after Laparal					
	1	5	10	15	50	60
1	102.5	106.5	102.5	102.5	107.5	107.5
1	92.5	102.5	103.0	100.0	100.0	117.5
1	102.5	105.0	105.0	102.5	100.0	100.0
1	102.5	107.5	107.5	102.5	90.0	112.5
1	106.0	106.0	105.0	105.0	102.5	100.0
1	97.5	97.5	86.0	82.0	82.0	100.0
1	96.0	76.0	83.0	83.0	92.5	97.5
2	105.0	105.0	107.5	107.5	102.5	107.5
2	105.0	105.0	106.0	106.0	107.5	105.0
2	95.0	90.0	95.0	90.0	100.0	107.5
2	97.5	95.5	95.0	90.0	105.0	90.0
2	96.0	87.5	97.5	97.0	100.0	95.0
2	97.5	92.0	97.5	94.0	96.0	97.5
2	107.5	107.5	103.0	106.0	103.0	112.0
2	102.5	106.0	85.0	86.0	107.5	97.5
2	107.5	107.5	102.5	102.5	102.5	97.5
3	97.5	100.5	94.5	102.5	102.5	107.5
3	100.0	105.0	105.0	105.0	100.0	100.0
3	95.0	97.0	90.0	100.0	100.0	100.0
3	85.0	92.0	92.5	92.0	90.0	100.0
3	92.0	77.0	79.0	69.0	69.0	75.0
3	92.0	79.0	103.0	106.0	100.0	100.0
4	76.0	87.5	87.5	77.5	77.5	77.5
4	45.0	37.5	45.0	45.0	47.5	45.0
4	52.5	22.5	90.0	85.0	80.0	85.5
4	100.0	100.0	100.0	100.0	97.5	92.5
4	105.0	100.0	100.0	100.0	105.0	105.0
4	97.5	97.5	97.5	105.0	95.0	92.5
4	99.0	103.0	106.0	103.0	103.0	117.0
4	72.0	87.0	69.0	87.0	92.0	82.0
4	105.0	105.0	105.0	105.0	102.5	100.0

(b) Repeat (a) for group 1.

(c) Repeat (a) for groups 2–4 combined.

6.42 Table 6.28, from Zerbe (1979a), compares 13 control and 20 obese patients on a glucose tolerance test using plasma inorganic phosphate. Derive the observations corresponding to $\frac{1}{2}$ and $1\frac{1}{2}$ hours so that the time points are equally spaced.

(a) For the control group, use orthogonal polynomials to find the degree of growth curve.

(b) Repeat (a) for the obese group.

Table A.29. Plasma-free gastric Phosphatase-activity

Patient	Normalised Kinase Challenge						
	0	1	2	3	4	5	6
Control							
1	4.5	3.3	3.8	2.8	2.2	2.3	4.8*
2	3.7	2.6	2.4	1.9	2.0	3.2	3.9
3	4.0	4.1	3.1	2.0	2.0	3.1	3.8
4	3.0	3.0	3.2	2.8	2.0	3.9	4.0
5	4.1	3.8	2.1	3.8	3.6	3.4	3.8
6	3.8	2.2	2.8	2.8	3.8	3.6	3.8
7	3.8	3.0	2.4	2.5	3.1	3.4	3.7
8	4.4	3.9	2.8	2.1	3.0	3.8	3.9
9	3.0	4.0	3.4	3.4	3.5	3.0	4.3
10	3.7	3.1	2.9	2.2	1.5	2.3	2.8
11	3.7	2.6	2.4	2.0	2.0	2.2	3.9
12	4.4	3.7	3.0	3.2	3.7	4.3	3.8
13	4.7	3.1	3.2	3.8	3.2	4.2	3.7
Glucose							
1	4.5	3.3	3.8	2.8	2.2	2.3	3.8*
2	3.0	4.8	4.1	3.7	3.7	4.1	4.7
3	4.0	4.4	3.8	3.9	3.7	4.2	4.8
4	4.5	3.9	3.1	3.1	3.1	3.1	4.0
5	3.1	3.1	3.0	2.8	2.8	3.9	2.7
6	4.8	3.0	2.9	2.8	2.2	3.1	3.9
7	3.7	3.1	3.9	2.8	2.0	3.6	4.3
8	3.4	4.7	3.9	4.1	2.8	3.7	3.9
9	3.0	2.9	2.9	2.9	2.1	2.8	3.8
10	4.9	3.0	4.0	3.7	3.7	4.1	4.9
11	4.8	4.3	4.7	4.0	4.7	3.7	3.8
12	4.4	4.2	4.2	3.4	3.5	3.4	3.9
13	4.0	4.3	4.8	4.8	3.5	4.1	4.5
14	3.1	4.1	4.8	4.1	3.4	4.3	4.4
15	4.8	4.4	4.4	4.4	4.1	4.0	3.8
16	4.2	3.9	3.8	3.8	3.0	3.1	3.8
17	6.0	6.1	3.2	4.1	4.0	3.8	4.8
18	3.0	3.4	3.0	2.8	2.1	2.4	3.5
19	4.5	4.0	3.0	3.9	2.4	2.3	3.9
20	4.0	4.4	3.8	3.8	3.8	3.0	3.8

*The similarity in the data for patient 1 in the control group and patient 1 in the glucose group is coincidental.

Table 8.28. Mandibular Measurements

Group	Subject	Anterior Treatment								
		1			2			3		
		y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
1	1	117.0	117.3	118.3	99.0	99.0	99.0	18.3	18.3	19.3
	2	109.0	109.3	110.3	99.0	99.3	99.3	20.3	20.3	20.3
	3	117.0	120.0	120.5	99.0	99.5	99.0	21.5	21.5	21.5
	4	121.0	126.0	127.8	97.5	97.5	97.5	23.8	23.8	24.5
	5	105.0	106.3	109.5	99.5	99.5	99.5	24.5	24.5	24.5
	6	125.0	126.0	127.8	97.5	99.5	97.5	22.8	22.8	22.0
	7	130.3	131.8	134.8	98.8	99.8	97.0	23.8	23.8	25.0
	8	126.3	126.8	128.8	99.0	97.0	97.0	20.8	20.8	20.0
	9	113.0	116.3	118.8	99.0	99.0	99.0	23.8	23.8	24.8
2	1	129.0	129.0	129.5	97.0	97.5	99.0	24.8	24.8	24.0
	2	116.5	120.0	120.5	97.5	99.0	99.0	29.5	29.5	29.5
	3	121.3	121.8	127.8	94.5	97.8	99.0	26.9	27.8	27.0
	4	109.8	112.0	114.8	94.8	98.8	97.0	18.8	18.8	19.0
	5	133.0	136.0	137.8	92.8	93.8	93.8	34.8	34.8	34.8
	6	120.0	124.3	126.8	92.5	99.0	99.0	26.8	26.8	26.0
	7	129.3	131.3	134.3	95.8	98.0	98.0	18.3	18.3	18.3
	8	121.0	124.0	125.5	94.5	95.5	99.0	18.5	18.5	19.5
	9	125.0	127.0	128.8	95.5	96.5	97.0	21.5	21.5	21.0

(a) Plot the degree of growth curves for the combined groups, and compare the growth curves of the two groups.

8.44 Consider the complete data from Table 8.28 including the observations corresponding to $\frac{1}{2}$ and $1\frac{1}{2}$ years. Use the methods in 85.1.13–85.1.18 for unequally spaced time points to analyze each group separately and the combined groups.

8.45 Table 8.28 contains mandible measurements (Timm 1980). There were two groups of subjects. Each subject was measured at three time points: y_1 , y_2 , and y_3 for each of three types of anterior treatment. Analyze as a repeated measures design with two within-subjects factors and one between-subjects factor. Use linear and quadratic contrasts for time (growth curve).

CHAPTER 7

Tests on Covariance Matrices

7.1 INTRODUCTION

We now consider tests of hypotheses involving the variance-covariance structure. These tests are often carried out to check assumptions pertaining to other tests. In Sections 7.2–7.4, we cover three basic types of hypotheses: (I) the covariance matrix has a particular structure, (II) two or more covariance matrices are equal, and (III) certain elements of the covariance matrix are zero, thus implying independence of the corresponding (multivariate normal) random variables. In most cases we use the likelihood ratio approach (Section 5.4.3). The resulting test statistics often involve the ratio of the determinants of the sample covariance matrix under the null hypothesis and under the alternative hypothesis.

7.2 TESTING A SPECIFIED PATTERN FOR Σ

In this section, the discussion is in terms of a sample covariance matrix $\mathbf{S}$ from a single sample. However, the tests can be applied to a sample covariance matrix $\hat{\mathbf{S}}_p = \mathbf{E}/v$ obtained by pooling across several samples. To allow for either possibility, the degrees-of-freedom parameter has been indicated by v . For a single sample, $v = n - 1$; for a pooled covariance matrix, $v = \sum_{i=1}^g (n_i - 1) = \sum_{i=1}^g n_i - g = N - J$.

7.2.1 Testing $H_0: \Sigma = \Sigma_0$

We begin with the basic hypothesis $H_0: \Sigma = \Sigma_0$ vs. $H_1: \Sigma \neq \Sigma_0$. The hypothesized covariance matrix Σ_0 is a target value for Σ or a nominal value from previous experience. Note that Σ_0 is completely specified in H_0 , whereas μ is not specified.

To test H_0 , we obtain a random sample of n observation vectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ from $N_p(\mu, \Sigma)$ and calculate $\mathbf{S}$. To see if $\mathbf{S}$ is significantly different from Σ_0 , we use the following test statistic, which is a modification of the likelihood ratio (Sec-

from 5.4.34:

$$u = v \left\{ \ln(\Sigma_0) - \ln(S) + \ln(S \Sigma_0^{-1}) - p \right\}, \quad (7.1)$$

where v represents the degrees of freedom of S (parameters at the beginning of Section 7.2), $\ln$ is the natural logarithm (base e), and $\ln$ is the trace of a matrix (Section 2.8). Note that, if $S = \Sigma_0$, then $u = 0$; otherwise u increases with the "distance" between S and Σ_0 (see 7.4) and the comment following).

When v is large, the statistic in (7.1) is approximately distributed as $\chi^2(p, p+1)/2N_0$ is true. For moderate size v ,

$$u' = \left[1 - \frac{1}{2v-2} \left(2p+1 - \frac{2}{p+1} \right) \right] u \quad (7.2)$$

is a better approximation to the $\chi^2(p, p+1)$ distribution. We reject H_0 if u or u' is greater than $\chi^2_{\alpha, \frac{1}{2}(p(p+1))}$. Note that the degrees of freedom for the χ^2 -statistic, $\frac{1}{2}p(p+1)$, is the number of distinct parameters in Σ .

We can express u in terms of the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$ of $S \Sigma_0^{-1}$ by noting that $\ln(S \Sigma_0^{-1})$ and $\ln(\Sigma_0) - \ln(S)$ become

$$\begin{aligned} \ln(S \Sigma_0^{-1}) &= \sum_{j=1}^p \lambda_j && \text{[Eq. (2.307)]} \\ \ln(\Sigma_0) - \ln(S) &= -\ln(\Sigma_0)^{-1} - \ln(S) \\ &= -\ln(S \Sigma_0^{-1}) && \text{[Eq. (2.49) and (2.51)]} \\ &= -\ln \left(\prod_{j=1}^p \lambda_j \right) && \text{[Eq. (2.308)]} \end{aligned} \quad (7.3)$$

from which (7.1) can be written as

$$\begin{aligned} u &= v \left[-\ln \left(\prod_{j=1}^p \lambda_j \right) + \sum_{j=1}^p \lambda_j - p \right] \\ &= v \left[\sum_{j=1}^p (\lambda_j - \ln \lambda_j) - p \right]. \end{aligned} \quad (7.4)$$

A plot of $y = x - \ln x$ will show that $x - \ln x \geq 1$ for all $x > 0$, with equality holding only for $x = 1$. Thus, $\sum_{j=1}^p (\lambda_j - \ln \lambda_j) \geq p$ and $u \geq 0$.

The hypothesis that the variables are independent and have unit variances,

$$R_0: \Sigma = I,$$

can be tested by simply setting $\Sigma_0 = I$ in (7.1).

7.3.1 Testing Sphericity

The hypothesis that the variables $y_1, y_2, \dots, y_p$ in $\mathbf{y}$ are independent and have the same variance can be expressed as $H_0: \Sigma = \sigma^2 \mathbf{I}$ versus $H_1: \Sigma \neq \sigma^2 \mathbf{I}$, where σ^2 is the unknown common variance. This hypothesis is of interest in repeated measures (see Section 6.3.1). Under H_0 , the ellipsoid $(\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) = c^2$ reduces to $(\mathbf{y} - \boldsymbol{\mu})' (\mathbf{y} - \boldsymbol{\mu}) = \sigma^2 c^2$, the equation of a sphere, hence the term sphericity is applied to the covariance structure $\Sigma = \sigma^2 \mathbf{I}$. Another sphericity hypothesis of interest in repeated measures is $H_0: \mathbf{C} \Sigma \mathbf{C}' = \sigma^2 \mathbf{I}$, where $\mathbf{C}$ is any full-rank $(p-1) \times p$ matrix of orthogonal contrasts (see Section 6.3.1).

For a random sample $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ from $N_p(\boldsymbol{\mu}, \Sigma)$, the likelihood ratio for testing $H_0: \Sigma = \sigma^2 \mathbf{I}$ is

$$LR = \left[\frac{|\mathbf{S}|}{(1/n) |\mathbf{I}| p^p} \right]^{n/2}. \quad (7.5)$$

In some cases that we have considered previously, the likelihood ratio is a simple function of a test statistic such as F , F' , Wilks' Λ , and so on. However, LR in (7.5) does not reduce to a standard statistic, and we resort to an approximation for its distribution. It has been shown that for a general likelihood ratio statistic LR,

$$-2 \ln LR \text{ is approximately } \chi_r^2 \quad (7.6)$$

for large n , where r is the total number of parameters minus the number estimated under the restrictions imposed by H_0 .

For the likelihood ratio statistic in (7.5), we obtain

$$-2 \ln LR = -n \ln \left[\frac{|\mathbf{S}|}{(1/n) |\mathbf{I}| p^p} \right] = -n \ln a,$$

where

$$a = (LR)^{2/n} = \frac{p^p |\mathbf{S}|}{(1/n) p^p}. \quad (7.7)$$

By (2.18T) and (2.18B), a becomes

$$a = \frac{p^p \prod_{i=1}^p \lambda_i}{(\sum_{i=1}^p \lambda_i)^p}, \quad (7.8)$$

where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigenvalues of $\mathbf{S}$. An improvement over $-n \ln a$ is given by

$$a' = - \left(r - \frac{2p^2 + p + 2}{np} \right) \ln a, \quad (7.9)$$

where r is the degrees of freedom for S (see comments at the beginning of Section 7.2). The statistic χ^2 has an approximate χ^2 -distribution with $\frac{1}{2}p(p+1)-1$ degrees of freedom. We reject H_0 if $\chi^2 \geq \chi^2_{\alpha, \frac{1}{2}p(p+1)-1}$. As noted before, the degrees of freedom in the χ^2 -approximation is equal to the total number of parameters minus the number of parameters estimated under H_0 . The number of parameters in Σ is $p + p_0^2 = \frac{1}{2}p(p+1)$, and the loss of 1 degree of freedom is due to estimation of σ^2 .

We see from (7.8) and (7.9) that if the sample $\mathbf{z}_i$'s are all equal, $\alpha = 1$ and $\alpha' = 0$. Hence, this statistic also tests the hypothesis of equality of the population eigenvalues.

To test $H_0: \mathbf{C}\Sigma\mathbf{C}' = \sigma^2\mathbf{I}$, use $\mathbf{C}\Sigma\mathbf{C}'$ in place of S in (7.7) and use $p-1$ in place of p in (7.7)–(7.9) and in the degrees-of-freedom for χ^2 .

The likelihood ratio (7.5) was first given by Mauchly (1940), and his name is often associated with this test. Nagelesheim and Pillai (1973) gave the exact distribution of α and provided a table for $p = 4, 5, \dots, 20$. Mardia (1976) showed that α can be obtained by a union-intersection approach (Section 6.1.4).

Example 7.2.2. We use the pebax watch-belt in Table 3.3 to illustrate tests of sphericity. The five variables appear to be commensurate, and the hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_5$ may be of interest. We would expect the variables to be correlated, and H_0 would ordinarily be tested using a multivariate approach, as in Sections 5.9.1 and 6.9.2. However, if $\Sigma = \sigma^2\mathbf{I}$ or $\mathbf{C}\Sigma\mathbf{C}' = \sigma^2\mathbf{I}$, then the hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_5$ can be tested with a univariate ANOVA F -test (see Section 6.9.1).

We first test $H_0: \Sigma = \sigma^2\mathbf{I}$. The sample covariance matrix S was obtained in Example 3.3.1. By (7.7),

$$\alpha = \frac{p^2 \mathbf{1}' S \mathbf{1}}{64 \mathbf{1}' S \mathbf{1}} = \frac{3^2 (27, 236, 546)}{(281, 691)^2} = .0082.$$

Then by (7.9), with $\alpha = .01$ and $p = 5$, we have

$$\chi^2 = - \left(\alpha - 1 + \frac{2p^2 + p + 2}{6p} \right) \ln \alpha = 26.177.$$

The approximate χ^2 -test has $\frac{1}{2}p(p+1)-1 = 14$ degrees of freedom. We therefore compute $\chi^2 = 26.177$ with $\chi^2_{.01, 14} = 23.68$ and reject $H_0: \Sigma = \sigma^2\mathbf{I}$.

To test $H_0: \mathbf{C}\Sigma\mathbf{C}' = \sigma^2\mathbf{I}$, we use the following matrix of orthonormal contrasts:

$$\mathbf{C} = \begin{pmatrix} 4/\sqrt{35} & -1/\sqrt{35} & -1/\sqrt{35} & -1/\sqrt{35} & -1/\sqrt{35} \\ 0 & 3/\sqrt{15} & -1/\sqrt{15} & -1/\sqrt{15} & -1/\sqrt{15} \\ 0 & 0 & 2/\sqrt{6} & -1/\sqrt{6} & -1/\sqrt{6} \\ 0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}.$$

Then using CSC* in place of S and with $p - 1 = 4$ for the four rows of C, we obtain

$$u = \frac{(p-1)s^{p-1}(\text{CSC}^*)}{[u(\text{CSC}^*)]^p} = \frac{4^4(14409.8)}{(93.04)^4} = .488,$$

$$s^2 = 6.176.$$

For degrees of freedom, we now have $\frac{1}{2}p(p-1) = 9$, and the critical value is $\chi^2_{.05,9} = 16.92$. Hence, we do not reject $H_0: \text{CSC}^* = s^2\mathbf{I}$, and a univariate F -test of $H_0: \mu_1 = \mu_2 = \cdots = \mu_5$ may be justified. $\square$

7.2.3 Testing $H_0: \Sigma = \sigma^2(\mathbf{I} - \rho\mathbf{J} + \rho\mathbf{J})$

In Section 6.3.1, it was noted that univariate ANOVA remains valid if

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \vdots & & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix} \quad (7.10)$$

$$= \sigma^2(\mathbf{I} - \rho\mathbf{J} + \rho\mathbf{J}), \quad (7.11)$$

where $\mathbf{J}$ is a square matrix of 1's, as defined in (2.1.2), and ρ is the population correlation between any two variables. This pattern of equal variances and equal covariances in Σ is variously referred to as *spherically composed symmetry*, or the *exchangeable correlation model*.

We now consider the hypothesis that (7.11) holds:

$$H_0: \Sigma = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho \\ \vdots & \vdots & & \vdots \\ \sigma^2\rho & \sigma^2\rho & \cdots & \sigma^2 \end{pmatrix}.$$

From a sample, we obtain the sample covariance matrix S. Estimates of σ^2 and $\sigma^2\rho$ under H_0 are given by

$$s^2 = \frac{1}{p} \sum_{j=1}^p v_{jj} \quad \text{and} \quad s^2r = \frac{1}{p(p-1)} \sum_{j \neq k} v_{jk}, \quad (7.12)$$

respectively, where v_{jj} and v_{jk} are from S. Thus s^2 is an average of the variances on the diagonal of S, and s^2r is an average of the off-diagonal covariances in S. An estimate of ρ can be obtained as $r = s^2r/s^2$. Using s^2 and s^2r in (7.12), the estimate of Σ under H_0 is then

$$\mathbf{S}_0 = \begin{pmatrix} \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\ \vdots & \vdots & & \vdots \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2(\mathbf{I} - \rho\mathbf{I} + \rho\mathbf{J}), \quad (7.13)$$

To compare $\mathbf{S}$ and $\mathbf{S}_0$, we use the following function of the likelihood ratio

$$u = \frac{|\mathbf{S}|}{|\mathbf{S}_0|}, \quad (7.14)$$

which can be expressed in the alternative form

$$u = \frac{|\mathbf{S}|}{\sigma^2 p(1 - \rho)^{p-1} + (\rho - 1)^2}. \quad (7.15)$$

By analogy with (7.5), the test statistic is given by

$$u' = - \left[r - \frac{p(p+1)^2(2p-3)}{8(p-1)(p^2+p-4)} \right] \ln u, \quad (7.16)$$

where r is the degree of freedom of $\mathbf{S}$ (see comments at the beginning of Section 7.2). The statistic u' is approximately $p^{-1/2} \left\{ \frac{1}{2} p(p+1-2) \right\}$, and we reject H_0 if $u' > p^{-1/2} \left\{ \frac{1}{2} p(p+1-2) \right\}$. Note that 2 degrees of freedom are lost due to estimation of σ^2 and ρ .

An alternative approximate test that is more precise when p is large and r is relatively small is given by

$$F = \frac{-(r_1 - 10r_2 - r_1)r}{r_1 r_2} \ln u,$$

where

$$\begin{aligned} r_1 &= \frac{p(p+1)^2(2p-3)}{8(p-1)(p^2+p-4)}, & r_2 &= \frac{p(p^2-1)(p+2)}{8r_1(p^2+p-4)}, \\ r_3 &= \frac{1}{2} p(p+1) - 2, & r_4 &= \frac{r_1 + 2}{r_3 - r_1}. \end{aligned}$$

We reject $H_0: \Sigma = \sigma^2(\mathbf{I} - \rho\mathbf{I} + \rho\mathbf{J})$ if $F > F_{\alpha, r_1, r_2}$.

Example 7.2.3. To illustrate this test, we use the cork data of Table 6.21. In Problem 6.34, a comparison is made of average thickness, and hence weight, in the four directions. A standard ANOVA approach to this repeated measures design would be valid if (7.13) holds. To check this assumption, we test $H_0: \Sigma = \sigma^2(\mathbf{I} - \rho\mathbf{I} + \rho\mathbf{J})$.

The sample covariance matrix is given by

$$\mathbf{S} = \begin{pmatrix} 290.41 & 221.75 & 288.44 & 226.27 \\ 221.75 & 218.00 & 229.05 & 171.37 \\ 288.44 & 229.05 & 350.00 & 230.54 \\ 226.27 & 171.37 & 230.54 & 226.80 \end{pmatrix}.$$

From which we obtain

$$\begin{aligned} \bar{\mathbf{y}} &= (25.61, 3.561, 28, \\ s^2 &= \frac{1}{p} \sum_{j=1}^p s_{jj} = 271.586, \\ s_{12} &= \frac{1}{p(p-1)} \sum_{j=2}^p s_{1j} = 200.072, & r = \frac{s_{12}}{s^2} = \frac{200.072}{271.586} = .736. \end{aligned}$$

From (7.15) and (7.16), we now have

$$\begin{aligned} n &= \frac{28.66(7.583, 28)}{(271.586(p-1) - .858(7.5)) + (1 \times .858)} = 401, \\ n' &= - \left[r - \frac{p(p+1)(2p-3)}{6(p-1)(p^2+p-4)} \right] \ln n \\ &= - \left[.736 - \frac{(1 \times 28)(7.5)}{(6 \times 28)(35)} \right] .734 = 18.511. \end{aligned}$$

Since $18.511 > z_{1-\alpha/2}^1 = 15.5$, we reject H_0 and conclude that Σ does not have the pattern in (7.10). $\square$

7.3 TESTS COMPARING COVARIANCE MATRICES

An assumption for F^2 or MANOVA tests comparing two or more mean vectors is that the corresponding population-covariance matrices are equal, $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$. Under this assumption, the sample covariance matrices $S_1, S_2, \dots, S_k$ reflect a common population Σ and are therefore pooled to obtain an estimate of Σ . If $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ is not true, large differences in $S_1, S_2, \dots, S_k$ may possibly lead to rejection of $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$. However, the F^2 and MANOVA tests are fairly robust to heterogeneity of covariance matrices as long as the sample sizes are large and equal. For other cases it is useful to have available a test of equality of covariance matrices. We begin with a review of the univariate case.

7.3.1 Univariate Tests of Equality of Variances

The two-sample univariate hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$ is tested with

$$F = \frac{s_1^2}{s_2^2}, \quad (7.17)$$

where s_1^2 and s_2^2 are the variances of the two samples. If H_0 is true (and assuming normality), f is distributed as F_{v_1, v_2} , where v_1 and v_2 are the degrees of freedom of s_1^2 and s_2^2 typically, $v_1 = 1$ and $v_2 = 1$. Note that s_1^2 and s_2^2 must be independent, which will hold if the two samples are independent.

For the several-sample case, various procedures have been proposed. We present Bartlett's (1937) test of homogeneity of variances because it has been extended to the multivariate case. To test

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2,$$

we calculate

$$\begin{aligned}x &= 1 + \frac{1}{k(k-1)} \left[\sum_{i=1}^k \frac{1}{v_i} - \frac{1}{\sum_{i=1}^k v_i} \right], & s^2 &= \frac{\sum_{i=1}^k v_i s_i^2}{\sum_{i=1}^k v_i}, \\w &= \left(\sum_{i=1}^k v_i \right) \ln s^2 - \sum_{i=1}^k v_i \ln s_i^2,\end{aligned}$$

where $s_1^2, s_2^2, \dots, s_k^2$ are independent sample variances with $v_1, v_2, \dots, v_k$ degrees of freedom, respectively. Then

$$\frac{w}{x} \quad \text{is approximately} \quad \chi_{k-1}^2.$$

We reject H_0 if $w/x > \chi_{k-1, \alpha}^2$.

For an F -approximation, we use v and m and calculate in addition

$$d_1 = k - 1, \quad d_2 = \frac{k+1}{k^2 - 1}, \quad F = \frac{w}{x - d + \frac{1}{2}d_2}.$$

Then

$$F = \frac{w/m}{d_1(k-m)} \quad \text{is approximately} \quad F_{d_1, d_2}.$$

We reject H_0 if $F > F_{\alpha}$.

Note that an assumption for either form of the preceding test is independence of $s_1^2, s_2^2, \dots, s_k^2$, which will hold for random samples from k distinct populations. This test would therefore be inappropriate for comparing $s_{1j}, s_{2j}, \dots, s_{kj}$ from the diagonal of $\mathbf{S}$, since the s_{ij} 's are correlated.

7.3.2 Multivariate Tests of Equality of Covariance Matrices

For k multivariate populations, the hypothesis of equality of covariance matrices is

$$H_0: \Sigma_1 = \Sigma_2 = \cdots = \Sigma_k. \quad (7.18)$$

The test of $H_0: \Sigma_1 = \Sigma_2$ for two groups is treated as a special case by setting $k = 2$. There is no exact test of $H_0: \Sigma_1 = \Sigma_2$ as there is in the unidimensional case (see (7.1.7)). We assume independent samples of size $n_1, n_2, \dots, n_k$ from multivariate normal distributions, as in an unbalanced one-way MANOVA, for example. To make the test, we calculate

$$M = \frac{(\mathbf{S}_1)^{n_1/2} (\mathbf{S}_2)^{n_2/2} \cdots (\mathbf{S}_k)^{n_k/2}}{(\mathbf{S}_{\text{pooled}})^{(n_1+n_2+\cdots+n_k)/2}}, \quad (7.1.8)$$

in which $n_i = n_i - 1$, $\mathbf{S}_i$ is the covariance matrix of the i th sample, and $\mathbf{S}_{\text{pooled}}$ is the pooled sample covariance matrix

$$\mathbf{S}_{\text{pooled}} = \frac{\sum_{i=1}^k n_i \mathbf{S}_i}{\sum_{i=1}^k n_i} = \frac{\mathbf{R}}{n_E}, \quad (7.1.9)$$

where $\mathbf{R}$ is given by (6.1.6) and $n_E = \sum_{i=1}^k n_i = \sum_{i=1}^k n_i - k$. It is clear that we must have every $n_i > p$, otherwise $(\mathbf{S}_i) = 0$ for some i , and M would be zero. Exact upper percentage points of $-2 \ln M = n_E \ln (\mathbf{S}_{\text{pooled}}) - \sum_{i=1}^k n_i \ln (\mathbf{S}_i)$ for the special case of $v_1 = v_2 = \cdots = v_k = v$ are given in Table A.14 for $p = 2, 3, 4, 5$ and $k = 2, 3, \dots, 10$ (Law, Chiang, and Kishorwala 1977). We can easily modify (7.1.8) and (7.1.9) to compare covariance matrices for the cells of a two-way model by using $n_E = n_{E1} - 1$.

The statistic M is a modification of the likelihood ratio and varies between 0 and 1, with values near 1 favoring H_0 in (7.1.8) and values near 0 leading to rejections of H_0 . It is not immediately obvious that M in (7.1.8) behaves in this way, and we offer the following heuristic argument. First we note that (7.1.8) can be expressed as

$$M = \left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_{\text{pooled}}|} \right)^{n_1/2} \left(\frac{|\mathbf{S}_2|}{|\mathbf{S}_{\text{pooled}}|} \right)^{n_2/2} \cdots \left(\frac{|\mathbf{S}_k|}{|\mathbf{S}_{\text{pooled}}|} \right)^{n_k/2}. \quad (7.1.10)$$

If $\mathbf{S}_1 = \mathbf{S}_2 = \cdots = \mathbf{S}_k = \mathbf{S}_{\text{pooled}}$, then $M = 1$. As the disparity among $\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_k$ increases, M approaches zero. To see this, note that the determinant of the pooled covariance matrix, $|\mathbf{S}_{\text{pooled}}|$, lies somewhere near the "middle" of the $|\mathbf{S}_i|$'s and that as a set of variables $x_1, x_2, \dots, x_n$ increases in spread, $x_{\text{max}}/x_{\text{min}}$ reduces the product more than $x_{\text{max}}/x_{\text{min}}$ increases it, where x_{min} and x_{max} are the minimum and maximum values, respectively. We illustrate this with the two sets of numbers, 4, 5, 6 and 1, 5, 9, which have the same mean but different spread. If we assume $v_1 = v_2 = v_3 = v$, then for the first set,

$$M_1 = \left[\left(\frac{4}{5} \right) \left(\frac{5}{5} \right) \left(\frac{6}{5} \right) \right]^{3/2} = (1.8)(1)(1.2)^{3/2} = (.86)^{3/2}$$

and for the second set,

$$M_2 = \left[\binom{1}{3} \binom{3}{3} \binom{3}{3} \right]^{-1} = (3, 2)(1)(3, 3)(1)^{-1} = 1/36r^2.$$

In M_2 , the smallest value, .2, follows the product proportionally more than the largest value, 1.8, increases it. Another illustration is found in Problem T.9.

Box (1949, 1953) gave χ^2 - and F -approximations for the distribution of M . Either of these approximate tests is referred to as Box's W -test. For the χ^2 -approximation, calculate

$$v_1 = \left[\sum_{i=1}^k \frac{1}{n_i} - \frac{1}{\sum_{i=1}^k n_i} \right] \left[\frac{2p^2 + 3p - 1}{4(p + 1)(p - 1)} \right]. \quad (7.22)$$

Then

$$u = -2(1 - v_1)/\ln W \quad \text{is approximately} \quad \chi^2 \left[\frac{1}{2}(k - 1)p(p + 1) \right], \quad (7.23)$$

where W is defined in (7.18), and

$$\ln M = \frac{1}{2} \sum_{i=1}^k n_i \ln |B_i| - \frac{1}{2} \left(\sum_{i=1}^k n_i \right) \ln |B_{\text{ref}}|. \quad (7.24)$$

We reject H_0 if $u > \chi^2_{\alpha}$. If $n_1 = n_2 = \cdots = n_k = n$, then v_1 simplifies to

$$v_1 = \frac{(k + 1)(2p^2 + 3p - 1)}{6k(p + 1)}. \quad (7.25)$$

To justify the degrees of freedom of the χ^2 -approximation, note that the total number of parameters estimated under H_0 is $k[\frac{1}{2}p(p + 1)]$, whereas under H_1 we estimate only Σ , which has $p + \binom{p}{2} = \frac{1}{2}p(p + 1)$ parameters. The difference is $(k - 1)[\frac{1}{2}p(p + 1)]$. The quantity $k[\frac{1}{2}p(p + 1)]$ arises from the assumption that all $\Sigma_i, i = 1, 2, \dots, k$, are different. Technically, H_1 can be stated as $\Sigma_i \neq \Sigma_j$ for some $i \neq j$. However, the most general case is all Σ_i different, and the distribution of M is derived accordingly.

For the F -approximation, we use v_1 from (7.22) and calculate, additionally,

$$v_2 = \frac{(p + 1)(p + 2)}{6(k - 1)} \left[\sum_{i=1}^k \frac{1}{n_i^2} - \frac{1}{\left(\sum_{i=1}^k n_i \right)^2} \right], \quad (7.26)$$

$$v_3 = \frac{1}{2}(k - 1)p(p + 1), \quad v_4 = \frac{v_1 + 2}{(v_1 + v_2)^2},$$

$$h_1 = \frac{1 - v_1 - \ln v_3}{v_1}, \quad h_2 = \frac{1 - v_1 + 2 \ln v_2}{v_1}.$$

$$E\tau_1 > c_1^2,$$

$$F = -2\tau_1 \ln M \quad \text{is approximately} \quad F_{0, 2g} \quad (7.27)$$

$$E\tau_1 < c_1^2,$$

$$F = -\frac{2\tau_1\tau_2 \ln M}{\tau_1(1) + 2\tau_2 \ln M} \quad \text{is approximately} \quad F_{0, 2g} \quad (7.28)$$

In either case, we reject M_0 if $F > F_{\alpha}$. If $\tau_1 = \tau_2 = \dots = \tau_g = \tau$, then τ_1 simplifies as in (7.27) and τ_2 simplifies as

$$\tau_2 = \frac{(g-1)(g+2)g^2 + 4 + 1)}{2g^2g^2} \quad (7.29)$$

Here's M and is calculated routinely in many computer programs for MANOVA. However, Olson (1974) showed that the M -test with equal τ_i may detect some forms of heterogeneity that have only minor effects on the MANOVA tests. The test is also sensitive to some forms of nonnormality. For example, it is sensitive to kurtosis for which the MANOVA tests are rather robust. Thus the M -test may signal covariance heterogeneity in some cases where it is not damaging to the MANOVA tests. Hence we may not wish to automatically rule out standard MANOVA tests if the M -test leads to rejection of M_0 . Olson showed that the skewness and kurtosis statistics $\beta_{1,p}$ and $\beta_{2,p}$ (see Section 4.4.2) have similar shortcomings.

Example 7.3.1. We test the hypothesis $H_0: \Sigma_1 = \Sigma_2$ for the psychological data of Table 5.1. The covariance matrices S_1 , S_2 , and S_p were given in Example 5.4.1. Using these, we obtain, by (7.24),

$$\begin{aligned} \ln M &= \frac{1}{2}(\tau_1 \ln |S_1| + \tau_2 \ln |S_2|) - \frac{1}{2}(\tau_1 + \tau_2) \ln |S_p| \\ &= \frac{1}{2}[(31) \ln(7917.7) + (31) \ln(38958.1)] \\ &\quad - \frac{1}{2}(31 + 31) \ln(71315.2) = -1.2603. \end{aligned}$$

For an exact test, we compare

$$-2 \ln M = 2.521$$

with 3.74, its critical value from Table A.34.

For the g^2 -approximation, we compare, by (7.25) and (7.26),

$$\begin{aligned} \tau_1 &= \frac{(2 + 14(2g^2) + 3g) - 11}{6(2 \times 31)(4 + 1)} = .0093, \\ \tau &= -(3) - c_1(\ln M) = 15.331 < g_{M, 0}^2 = 18.503. \end{aligned}$$

For an approximate F -test, we first calculate the following:

$$c_1 = \frac{(2 - 1)(4 + 2)}{6(2 - 1)} \left[\frac{1}{3(2)} + \frac{1}{3(2)} - \frac{1}{(3) + 3(2)} \right] = .003469,$$

$$d_1 = \frac{1}{2} (2 - 1)(4)(4 + 1) = 10,$$

$$e_1 = \frac{10 + 2}{(.003469 - .000335)} = 38379.7,$$

$$f_1 = \frac{1 - .003469 - .98183777}{10} = .0150,$$

$$g_1 = \frac{1 - .003469 + 2(18377.7)}{18377.7} = 5.0046 \times 10^{-2}.$$

Since $c_1 = .003469 = c_1^2 = .003469$, we use (1.27) to obtain

$$F = -2g_1 \ln M = 1.354 = F_{.05, 10, 10} = 1.83.$$

Thus all three tests accept H_0 . □

5.4 TESTS OF INDEPENDENCE

5.4.1 Independence of Two Subvectors

Suppose the observation vector is partitioned into two subvectors of interest, which we label $\mathbf{y}$ and $\mathbf{x}$, as in Section 3.3.1, where $\mathbf{y}$ is $p \times 1$ and $\mathbf{x}$ is $q \times 1$. By (3.46), the corresponding partitioning of the population covariance matrix is

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix},$$

with analogous partitioning of $\mathbf{S}$ and $\mathbf{R}$ as in (3.47):

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{pmatrix}.$$

The hypothesis of independence of $\mathbf{y}$ and $\mathbf{x}$ can be expressed as

$$H_0: \Sigma = \begin{pmatrix} \Sigma_{yy} & \mathbf{0} \\ \mathbf{0} & \Sigma_{xx} \end{pmatrix} \quad \text{or} \quad H_0: \Sigma_{yx} = \mathbf{0}.$$

Thus independence of $\mathbf{y}$ and $\mathbf{x}$ means that every variable in $\mathbf{y}$ is independent of every variable in $\mathbf{x}$. Note that there is no restriction on Σ_{yy} or Σ_{xx} .

The likelihood ratio test statistic for $H_0: \Sigma_{yz} = \mathbf{O}$ is given by

$$A = \frac{|\mathbf{S}|}{|\mathbf{S}_{yz}| |\mathbf{S}_{xz}|} = \frac{|\mathbf{R}|}{|\mathbf{R}_{yz}| |\mathbf{R}_{xz}|}, \quad (7.28)$$

which is distributed as $\lambda_{p,q,p+q-1-p}$. We reject H_0 if $A \leq A_{\alpha}$. We thus have an exact test for $H_0: \Sigma_{yz} = \mathbf{O}$. Critical values for Wilks' A are given in Table A.9 using $v_1 = p$ and $v_2 = n - 1 - q$. The test statistic in (7.30) is equivalent (when H_0 is true) to the A -statistic (10.55) in Section 10.5.1 for testing the significance of the regression of $\mathbf{y}$ on $\mathbf{x}$.

By the symmetry of

$$\frac{|\mathbf{R}|}{|\mathbf{R}_{yz}| |\mathbf{R}_{xz}|} = \frac{|\mathbf{R}|}{|\mathbf{R}_{xz}| |\mathbf{R}_{yz}|},$$

A in (7.28) is also distributed as $\lambda_{q,p,p+q-1-p}$. This is equivalent to property 3 in Section 6.1.3.

Note that $|\mathbf{R}_{yz}| |\mathbf{R}_{xz}|$ in (7.28) is an estimate of $|\hat{\Sigma}_{yz}| |\hat{\Sigma}_{xz}|$, while by (2.81) is the determinant of $\hat{\Sigma}$ when $\Sigma_{yz} = \mathbf{O}$. Thus Wilks' A compares an estimate of $\hat{\Sigma}$ without restrictions to an estimate of $\hat{\Sigma}$ under $H_0: \Sigma_{yz} = \mathbf{O}$. We can see intuitively that $|\mathbf{R}| = |\mathbf{R}_{yz}| |\mathbf{R}_{xz}|$ by noting from (2.84) that $|\mathbf{R}| = |\mathbf{R}_{xz}| |\mathbf{R}_{yz} - \mathbf{R}_{xz} \mathbf{R}_{zz}^{-1} \mathbf{R}_{zy}|$, and since $\mathbf{R}_{zz} \mathbf{R}_{zz}^{-1} \mathbf{R}_{zy}$ is positive definite, $|\mathbf{R}_{yz} - \mathbf{R}_{xz} \mathbf{R}_{zz}^{-1} \mathbf{R}_{zy}| < |\mathbf{R}_{yz}|$. This can be illustrated for the case $p = q = 1$:

$$|\mathbf{R}| = \begin{vmatrix} r_{11}^2 & r_{12} \\ r_{12} & r_{22}^2 \end{vmatrix} = r_{12}^2 r_{22}^2 - (r_{12})^2 = r_{12}^2 r_{22}^2.$$

As r_{22}^2 increases, $|\mathbf{R}|$ decreases.

Wilks' A in (7.28) can be expressed in terms of eigenvalues:

$$A = \prod_{i=1}^r (1 - \epsilon_i^2), \quad (7.29)$$

where $r = \min(p, q)$ and the ϵ_i^2 's are the nonzero eigenvalues of $\mathbf{S}_{xz}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1} \mathbf{S}_{yx}$. We could also use $\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$, since the (nonzero) eigenvalues of $\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$ are the same as those of $\mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1} \mathbf{S}_{yx}$ (these two matrices are of the form $\mathbf{AB}$ and $\mathbf{BA}$, see Section 2.11.2). The number of nonzero eigenvalues is $r = \min(p, q)$, since r is the rank of both $\mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1} \mathbf{S}_{yx}$ and $\mathbf{S}_{yy}^{-1} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$. The eigenvalues are designated ϵ_i^2 because they are the squared canonical correlations between $\mathbf{y}$ and $\mathbf{x}$ (see Chapter 11). In the special case $p = 1$, (7.29) becomes

$$A = 1 - \epsilon_1^2 = 1 - R^2,$$

where R^2 is the square of the multiple correlation between $\mathbf{y}$ and $(x_1, x_2, \dots, x_p)$.

The other test statistics, $U^{(0)}$, $F^{(0)}$, and Roy's δ , can also be defined in terms of the σ_j^2 's (see Section 11.4.1).

Example 7.4.1. Consider the diabetes data in Table 3.4. There is a natural partitioning in the variables, with y_1 and y_2 of minor interest and x_1 , x_2 , and x_3 of major interest. We test independence of the y 's and the x 's, that is, $H_0: \Sigma_{yx} = \mathbf{0}$. From Example 3.8.1, the partitioned-covariance matrix is

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} = \begin{pmatrix} \begin{array}{cc|ccc} .0362 & .2690 & .7873 & -.2138 & 3.158 \\ .2690 & 79.36 & 35.23 & -23.96 & -20.94 \end{array} \\ \hline \begin{array}{cc|ccc} .7873 & 28.23 & 11.66 & 356.7 & 308.4 \\ -.2138 & -23.96 & 356.7 & 2582 & 1143 \\ 3.158 & -20.94 & 308.4 & 1143 & 2138 \end{array} \end{pmatrix}.$$

To make the test, we compute

$$\Lambda = \frac{|\Sigma|}{|\Sigma_{yy}||\Sigma_{xx}|} = \frac{3.158 \times 10^6}{(1.082)(3.572) \times 10^7} = .334 < \Lambda_{.05, 2, 42, 42} = .736.$$

Thus we reject the hypothesis of independence. Note the use of 42 in $\Lambda_{.05, 2, 42, 42}$ in place of $n - 1 - q = 48 - 1 - 5 = 42$. This is a conservative approach that allows the use of a table value without interpolation. $\square$

7.4.2 Independence of Several Subvectors

Let there be k sets of variables so that $\mathbf{y}$ and $\mathbf{z}$ are partitioned as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_k \end{pmatrix} \quad \text{and} \quad \mathbf{z} = \begin{pmatrix} \mathbf{z}_{11} & \mathbf{z}_{12} & \cdots & \mathbf{z}_{1k} \\ \mathbf{z}_{21} & \mathbf{z}_{22} & \cdots & \mathbf{z}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_{k1} & \mathbf{z}_{k2} & \cdots & \mathbf{z}_{kk} \end{pmatrix},$$

with p_i variables in $\mathbf{y}_i$, where $p_1 + p_2 + \cdots + p_k = p$. Note that $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k$ represents a partitioning of $\mathbf{y}$, not a random sample of independent vectors. The hypothesis that the subvectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k$ are mutually independent can be expressed as $H_0: \Sigma_{y_j y_i} = \mathbf{0}$ for all $i \neq j$, or

$$H_0: \Sigma = \begin{pmatrix} \Sigma_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_{kk} \end{pmatrix}. \quad (7.32)$$

The likelihood ratio statistic is

$$w = \frac{|\mathbf{S}|}{|\mathbf{S}_{11}||\mathbf{S}_{22}|\cdots|\mathbf{S}_{kk}|} \quad (7.33)$$

$$= \frac{|\mathbf{R}|}{|\mathbf{R}_{11}||\mathbf{R}_{22}|\cdots|\mathbf{R}_{kk}|}, \quad (7.34)$$

where $\mathbf{S}$ and $\mathbf{R}$ are obtained from a random sample of n observations and are partitioned as $\mathbf{S}$ above, conforming to $p_1, p_2, \dots, p_k$. Note that the denominator of (7.33) is the determinant of $\mathbf{S}$ restricted by $\mathbf{R}_0$, that is, with $S_{ij} = 0$ for all $i \neq j$ [see (2.31)]. The statistic w does not have Wilks' Λ -distribution as it does in (7.36) when $k = 2$, but a good g^2 -approximation to its distribution is given by

$$w' = -w \ln w, \quad (7.35)$$

where

$$w = 1 - \frac{1}{12f^2}(\lambda_1 + \lambda_2), \quad (7.36)$$

$$f = \frac{1}{2}p_2, \quad \lambda_1 = p^2 - \sum_{i=1}^p p_i^2, \quad \lambda_2 = p^2 - \sum_{i=1}^p p_i^3,$$

and v is the degrees of freedom of $\mathbf{S}$ or $\mathbf{R}$ (see comments at the beginning of Section 7.2). We reject the independence hypothesis if $w' > \chi_{\alpha, f}^2$.

The degrees of freedom, $f = \frac{1}{2}p_2$, arise from the following consideration. The number of parameters in Σ unrestricted by the hypothesis is $\frac{1}{2}p(p+1)$. Under the hypothesis (7.32), the number of parameters in each Σ_{α} is $\frac{1}{2}p_{\alpha}(p_{\alpha}+1)$, for a total of $\frac{1}{2}\sum_{\alpha=1}^k p_{\alpha}(p_{\alpha}+1)$. The difference is

$$\begin{aligned} f &= \frac{1}{2}p(p+1) - \frac{1}{2}\sum_{\alpha=1}^k p_{\alpha}(p_{\alpha}+1) = \frac{1}{2}\left(p^2 + p - \sum_{\alpha} p_{\alpha}^2 - \sum_{\alpha} p_{\alpha}\right) \\ &= \frac{1}{2}\left(p^2 + p - \sum_{\alpha} p_{\alpha}^2 - p\right) = \frac{1}{2}\left(p^2 - \sum_{\alpha} p_{\alpha}^2\right) = \frac{p_2}{2}. \end{aligned}$$

Example 7.4.1. For 30 brands of Japanese-Grande wine, Sotani et al. (1980) studied the relationship between

$$y_1 = \text{taste},$$

$$y_2 = \text{color},$$

Table 3.1. Taste Measurements

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1.0	.8	4.05	1.59	.85	3.8	3.67	5.00	16.59	122.0
.1	.2	3.83	1.50	.80	.4	3.63	4.50	15.80	95.0
.8	.8	4.26	1.66	.82	-0.6	3.68	4.66	16.86	109.0
.7	.7	4.39	1.63	.87	-1.6	3.69	5.00	17.49	106.0
-1.1	-1.1	4.35	1.59	.87	-0.9	3.67	4.20	15.49	100.0
.4	.3	4.03	1.68	.83	-0.5	3.63	5.00	16.78	120.0
.2	-0.3	4.26	1.64	1.09	-1.7	3.25	4.15	15.80	170.0
.3	-1.1	4.32	1.65	.83	-5.0	4.06	5.65	16.76	144.0
.7	.4	4.23	1.74	.83	-1.5	3.40	4.25	16.62	155.0
.5	-1.1	4.37	1.72	.83	-1.2	3.62	4.54	16.76	121.0
-1.1	.1	4.45	1.79	1.19	-0.9	3.69	3.90	16.76	105.0
.5	-0.5	4.45	1.69	.86	-0.9	3.32	4.69	15.49	126.0
.3	.8	4.25	1.55	.83	-5.0	3.46	4.54	15.55	125.0
.6	.2	4.25	1.60	.86	3.8	3.13	5.65	15.69	126.0
.0	-0.3	4.09	1.68	.86	.8	3.67	4.50	15.86	99.0
-1.2	-0.2	4.22	1.64	.86	-0.2	3.69	4.69	16.37	100.8
.0	-0.2	4.16	1.58	.86	3.8	3.62	5.60	15.33	111.0
.1	.1	4.28	1.80	.79	-4.8	3.64	6.09	15.73	126.0
-1.1	-0.2	4.32	1.54	.83	-0.9	3.37	4.60	16.69	109.0
.0	.1	4.32	1.69	.84	-0.1	3.72	4.60	16.93	111.0
.8	.3	4.36	1.50	.83	-1.3	3.66	5.00	15.36	96.0
.8	.3	4.09	1.80	1.04	.4	3.60	5.68	15.76	107.0
.4	.7	4.39	1.60	.79	-7.0	4.33	6.09	15.89	106.0
.6	-0.3	4.33	1.80	.83	.8	3.36	4.60	15.49	49.5
-0.3	-0.3	4.25	1.77	1.02	.3	3.64	4.15	16.65	104.0
-1.2	.8	3.85	1.56	.25	3.8	3.67	4.50	15.86	56.5
.3	-1.1	4.35	1.60	.86	-0.5	3.40	4.12	15.86	154.0
.1	.4	4.35	1.77	1.06	-4.5	3.69	5.00	16.76	106.2
.4	.3	4.16	1.61	.83	-0.1	3.63	4.68	15.76	100.0
-0.8	-0.3	3.89	1.80	.86	.7	3.63	4.29	15.73	46.0

and

$$\begin{aligned}
 x_1 &= \text{pH}, & x_6 &= \text{direct reducing sugar}, \\
 x_2 &= \text{acidity 1}, & x_7 &= \text{total sugar}, \\
 x_3 &= \text{acidity 2}, & x_8 &= \text{alcohol}, \\
 x_4 &= \text{water matter}, & x_9 &= \text{ferment-alkalies}.
 \end{aligned}$$

The data are in Table 3.1.

We test independence of the following four subsets of variables:

$$\{x_1, x_2\}, \{x_1, x_3, x_4, x_5\}, \{x_1, x_2, x_3\}, \{x_7, x_8\}.$$

The sample covariance matrix is

$$S = \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{pmatrix} = \begin{pmatrix} \begin{array}{cc|cc|cc|cc} .16 & .16 & .01 & .008 & .02 & -.04 & .02 & .01 & -.02 & 1.44 \\ .16 & .16 & -.01 & .008 & .02 & -.04 & .02 & .01 & -.02 & 1.69 \end{array} \\ \hline \begin{array}{cc|cc|cc} .04 & -.01 & .03 & .004 & .03 & -.11 & -.10 & -.03 & -.10 & 4.43 \\ .004 & .009 & .000 & .014 & .009 & -.043 & -.000 & .0004 & .008 & 1.11 \\ .02 & .02 & .03 & .020 & .07 & -.10 & -.10 & -.03 & .02 & 9.03 \end{array} \\ \hline \begin{array}{cc|cc|cc} -.04 & -.16 & -.10 & -.003 & -.10 & 4.68 & -.10 & -.07 & -.10 & -28.11 \\ .02 & .03 & -.03 & -.009 & -.03 & -.19 & 1.1 & .19 & .02 & -4.26 \\ .04 & .08 & -.03 & .0004 & -.03 & -.07 & 1.1 & .26 & .10 & -1.07 \end{array} \\ \hline \begin{array}{cc|cc} -.02 & .04 & -.10 & .006 & .05 & -.12 & .05 & .13 & .02 & 6.71 \\ 1.44 & 1.69 & 4.43 & 1.11 & 9.03 & -28.11 & -4.26 & -1.07 & 6.70 & 19.11 \end{array} \end{pmatrix},$$

where S_{11} is 2×2 , S_{22} is 3×3 , S_{33} is 3×3 , and S_{44} is 2×2 . We first obtain

$$\begin{aligned} n^2 &= \frac{100}{|S_{11}| |S_{22}| |S_{33}| |S_{44}|} \\ &= \frac{2.025 \times 10^{-7}}{(.00218)(.0000458)(.00012)(.000504)} = 19.827. \end{aligned}$$

For the χ^2 -approximation, we calculate

$$\begin{aligned} a_1 &= p^2 - \sum_{i=1}^n p_i^2 = 10^2 - (2^2 + 3^2 + 3^2 + 2^2) = 74, \\ a_2 &= p^3 - \sum_{i=1}^n p_i^3 = 999, \quad f = \frac{1}{2}a_1 = 37, \\ r &= 1 - \frac{1}{12} \frac{a_2}{f^2} = 1 - \frac{2(999) + 3(74)}{12(37)^2} = .89. \end{aligned}$$

Then,

$$n^2 = -rc \ln n = -(20)(.89) \ln(19.827) = 108.122,$$

which exceeds $\chi^2_{108, .95} = 69.15$, and we reject the hypothesis of independence of the four subtests. $\square$

7.4.3 Test for Independence of All Variables

If all $\rho_{jk} = 0$ in the hypothesis (7.32) in Section 7.4.2, we have the special case in which all the variables are mutually independent, $\alpha_{jk} = 0$ for $j \neq k$, or

$$H_0: \Sigma = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{pmatrix}.$$

There is no restriction on the σ_{jj} 's. With $\sigma_{jk} = 0$ for all $j \neq k$, the corresponding ρ_{jk} 's are also 0, and an equivalent form of the hypothesis is $\mathbf{R}_0: \mathbf{P}_p = \mathbf{I}$, where $\mathbf{P}_p$ is the population correlation matrix defined in (5.37).

Since all $\rho_{jk} = 0$, the statistics (7.33) and (7.34) reduce to

$$u = \frac{|\mathbf{R}|}{r_1 r_2 \cdots r_{p-1}} = |\mathbf{R}|, \quad (7.37)$$

and (7.33) becomes

$$u' = -\{v - \frac{1}{2}(2p + 3)\} \ln u, \quad (7.38)$$

which has an approximate χ^2 -distribution, where v is the degrees of freedom of $\mathbf{R}$ or $\mathbf{R}$ (see a comment at the beginning of Section 7.2) and $f = \frac{1}{2}p(p-1)$ is the degrees of freedom of χ^2 . We reject H_0 if $u' > \chi^2_{\alpha, f}$. Exact percentage points of u' for selected values of n and p are given in Table A.15 (Mantel and Haensler 1959). Percentage points for the limiting χ^2 -distribution are also given for comparison.

Note that $|\mathbf{R}|$ in (7.37) varies between 0 and 1. If the variables were uncorrelated in the sample, we would have $|\mathbf{R}| = 0$ and $|\mathbf{R}| = 1$. On the other hand, if two or more variables were linearly related, $\mathbf{R}$ would not be full rank and we would have $|\mathbf{R}| = 0$. If the variables were highly correlated, $|\mathbf{R}|$ would be close to 0, if the correlations were all small, $|\mathbf{R}|$ would be close to 1. This can be illustrated for $p = 2$

$$|\mathbf{R}| = \begin{vmatrix} 1 & r \\ r & 1 \end{vmatrix} = 1 - r^2.$$

Example 7.4.3. To test the hypothesis $H_0: \sigma_{jk} = 0, j \neq k$, for the probe word data from Table 3.5, we calculate

$$\mathbf{R} = \begin{pmatrix} 1.000 & .604 & .757 & .578 & .415 \\ .604 & 1.000 & .347 & .756 & .548 \\ .757 & .347 & 1.000 & .689 & .892 \\ .578 & .756 & .689 & 1.000 & .528 \\ .415 & .548 & .892 & .528 & 1.000 \end{pmatrix}.$$

Then by (3.37) and (3.38),

$$\begin{aligned} s &= (\mathbf{B}) = .5408, \\ s' &= - \left[s - 1 - \frac{1}{2} (2p + 5) \right] \ln s = 23.81. \end{aligned}$$

The exact χ^2 critical value for s' from Table A.15 is 23.78, and we therefore reject H_0 . The approximate g^2 critical value for s' is $\chi^2_{23.78, .05} = 23.71$, with which we also reject H_0 . $\square$

PROBLEMS

- 7.1 Show that if $\mathbf{S} = \mathbf{I}_n$ in (3.1), then $s = 0$.
- 7.2 Verify (7.16), that is, show that $\ln(\mathbf{I}_n) = \ln(\mathbf{S}) = -\ln(\mathbf{S}\mathbf{I}_n^{-1})$.
- 7.3 Verify (7.16), that is, show that $-\ln\left[\prod_{i=1}^p \lambda_i\right] + \sum_{i=1}^p \lambda_i = \sum_{i=1}^p (\lambda_i - \ln \lambda_i)$.
- 7.4 Show that the likelihood ratio for $H_0: \mathbf{\Sigma} = \sigma^2 \mathbf{I}$ is given by (7.2), i.e. $LR = |\mathbf{S}|/(\det \mathbf{S})/p^{p/2} T^{-p/2}$.
- 7.5 Show that $s = 1$ and $s' = 0$ if all the λ_i 's are equal, as noted in Section 7.2.1, where s is given by (3.8) and s' by (7.9).
- 7.6 Show that the covariance matrix in (3.10) can be written in the form $\sigma^2(1 - r)\mathbf{I} + r\mathbf{J}$, as given in (7.13).
- 7.7 Obtain (7.15) from (3.14) as follows:
 - (a) Show that the $p \times p$ matrix $\mathbf{J}$ has a single nonzero eigenvalue equal to p and corresponding eigenvector proportional to $\mathbf{j}$.
 - (b) Show that $\mathbf{S}_0 = r^2\mathbf{j}(1 - r)\mathbf{j}' + r\mathbf{J}$ in (7.13) can be written in the form $\mathbf{S}_0 = r^2(1 - r)\mathbf{J} + \frac{r}{1-r}\mathbf{J}$.
 - (c) Use Section 2.11.2 and (2.106) to obtain (7.15).
- 7.8 Show that M' in (3.18) can be expressed in the form given in (7.21).
- 7.9 (a) Calculate M as given in (7.21) for

$$\mathbf{S}_1 = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}, \quad \mathbf{S}_2 = \begin{pmatrix} 4 & 3 \\ 3 & 8 \end{pmatrix}.$$

Assume $v_1 = v_2 = 5$.

- (b) Calculate M for

$$\mathbf{S}_1 = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}, \quad \mathbf{S}_2 = \begin{pmatrix} 10 & 18 \\ 18 & 26 \end{pmatrix}.$$

Assume $v_1 = v_2 = 5$.

In (b), $\mathbf{S}_1$ and $\mathbf{S}_2$ differ more than in (a) and M is accordingly much smaller. This illustrates the comments following (7.21).

7.00 Obtain (7.31), $A = [\sum_{i=1}^p c_i - c_i^2]$, by using (2.64) to write (8) in the form $\mathbf{B} = \mathbf{B}_{\text{col}}(\mathbf{B}_{\text{col}} - \mathbf{B}_{\text{col}}\mathbf{B}_{\text{col}}^{-1}\mathbf{B}_{\text{col}})$.

7.01 Show that the forms of $\mathbf{u}$ in (7.33) and (7.34) reduce to (7.37) when all $p_i = 1$.

7.02 Show that when all $p_i = 1$, ϵ in (7.36) reduces to $1 - (2p + S_0)/6\pi$, so that (7.35) becomes (7.38).

7.03 Give a justification for the degrees of freedom $f = \frac{1}{2}n(p-1)$ for the approximate χ^2 test statistic χ^2 in (7.38).

7.04 In Example 5.2.2, we assumed that, for the height and weight data of Table 3.1, the population covariance matrix is

$$\Sigma = \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}.$$

Test this as a hypothesis using (7.1).

7.05 Test $H_0: \Sigma = \sigma^2\mathbf{I}$ and $H_0: \text{COC} = \sigma^2\mathbf{I}$ for the calculator speed data of Table 5.11.

7.06 Test $H_0: \Sigma = \sigma^2\mathbf{I}$ and $H_0: \text{COC} = \sigma^2\mathbf{I}$ for the rumor-bone data of Table 3.5.

7.07 Test $H_0: \Sigma = \sigma^2\mathbf{I}$ and $H_0: \text{COC} = \sigma^2\mathbf{I}$ for the rock data of Table 5.1.

7.08 Test $H_0: \Sigma = \sigma^2(1 - \rho)\mathbf{I} + \rho\mathbf{J}$ for the pocket word data in Table 3.5. Use both χ^2 and F -approximations.

7.09 Test $H_0: \Sigma = \sigma^2(1 - \rho)\mathbf{I} + \rho\mathbf{J}$ for the calculator speed data in Table 5.11. Use both χ^2 and F -approximations.

7.10 Test $H_0: \Sigma = \sigma^2(1 - \rho)\mathbf{I} + \rho\mathbf{J}$ for the rumor-bone data in Table 3.5. Use both χ^2 and F -approximations.

7.11 Test $H_0: \Sigma_0 = \Sigma_2$ for the brother data of Table 5.5. Use an exact critical value from Table A.14 as well as χ^2 and F -approximations.

7.12 Test $H_0: \Sigma_0 = \Sigma_2$ for the engineer data of Table 5.6. Use an exact critical value from Table A.14 as well as χ^2 and F -approximations.

7.13 Test $H_0: \Sigma_0 = \Sigma_2$ for the orthography data of Table 5.5. Use an exact critical value from Table A.14 as well as χ^2 and F -approximations.

7.14 Test $H_0: \Sigma_0 = \Sigma_2$ for the optical data of Table 5.5. Use an exact critical value from Table A.14 as well as χ^2 and F -approximations.

7.15 Test $H_0: \Sigma_0 = \Sigma_2 = \Sigma_3$ for the fish data of Table 5.17. Use an exact critical value from Table A.14 as well as χ^2 and F -approximations.

7.16 Test $H_0: \Sigma_0 = \Sigma_2 = \Sigma_3 = \dots = \Sigma_9$ for the mankind data in Table 5.1. Use an exact critical value from Table A.14 as well as χ^2 and F -approximations.

7.17 Test $H_0: \Sigma_0 = \Sigma_{12} = \Sigma_{13} = \dots = \Sigma_{12}$ for the wrap-bone data in Table 5.18. Use both χ^2 and F -approximations.

- 7.28 Test independence of (y_1, y_2) and (x_1, x_2) for the urea data in Table 3.7.
- 7.29 Test independence of (y_1, y_2, y_3) and (x_1, x_2, x_3) for the glucose data in Table 3.8.
- 7.30 Test independence of (y_1, y_2) and $(x_1, x_2, \dots, x_6)$ for the Swedes data of Table 7.1.
- 7.31 The data in Table 7.2 relate temperature, humidity, and evaporation (summary of R. J. Freund). The variables are

- y_1 = maximum daily air temperature,
 y_2 = minimum daily air temperature,
 y_3 = integrated area under daily air temperature curve, that is,
 a measure of average air temperature,
 y_4 = maximum daily soil temperature,
 y_5 = minimum daily soil temperature,
 y_6 = integrated area under soil temperature curve,
 y_7 = maximum daily relative humidity,
 y_8 = minimum daily relative humidity,
 y_9 = integrated area under daily humidity curve,
 x_{10} = total wind, measured in miles per day,
 x_{11} = evaporation.

Test independence of the following five groups of variables: (y_1, y_2, y_3) , (y_4, y_5, y_6) , (y_7, y_8, y_9) , x_{10} and x_{11} .

- 7.32 Test for independence of all the variables for the calcium data of Table 3.3.
- 7.33 Test for independence of all the variables for the calculator speed data of Table 6.12.
- 7.34 Test for independence of all the variables for the rumen bone data of Table 3.6.
- 7.35 Test for independence of all the variables for the oak data of Table 5.21.

Table 1.1. Temperatures, Boundaries, and Exposures

X	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T ₇	T ₈	T ₉	T ₁₀	T ₁₁
66	65	147	85	59	151	85	48	548	275	38	38
66	65	149	86	61	150	84	58	545	548	34	34
79	66	142	83	64	152	84	41	588	538	33	33
81	67	147	83	69	150	84	58	406	282	36	36
86	68	167	88	69	160	83	66	379	531	41	41
76	69	151	77	67	167	86	73	478	448	4	4
75	69	151	76	69	159	86	72	452	294	5	5
75	69	154	64	68	159	85	79	454	513	28	28
86	68	161	86	71	165	85	63	450	493	31	31
86	71	169	94	76	208	83	56	406	684	56	56
88	73	176	94	76	208	84	69	508	618	43	43
89	74	187	94	76	211	84	51	383	528	47	47
88	72	171	94	73	211	86	54	458	663	43	43
58	72	171	95	76	201	85	51	392	467	45	45
81	69	154	87	68	167	85	51	448	184	31	31
79	68	149	85	68	162	85	59	456	173	38	38
86	69	168	87	64	175	85	42	502	175	38	38
86	70	168	87	68	177	84	44	505	76	39	39
82	70	168	88	76	169	83	48	506	73	24	24
77	67	147	83	66	179	87	68	431	183	58	58
87	67	166	95	67	196	86	64	379	76	37	37
89	69	171	95	72	199	84	48	399	248	58	58
89	72	188	94	72	204	85	48	504	193	56	56
95	72	186	95	73	201	84	47	585	488	54	54
95	74	188	95	73	208	85	47	585	539	44	44
86	73	189	94	73	208	86	45	525	173	41	41
83	74	183	94	73	214	83	58	506	238	43	43
83	74	186	95	76	213	86	45	585	118	42	42
96	75	188	95	71	207	83	48	568	83	58	58
95	76	202	95	69	202	83	59	559	269	48	48
86	75	175	96	69	175	84	58	459	128	37	37
81	71	176	94	69	168	84	44	450	435	58	58
88	72	179	86	76	189	83	58	566	415	35	35
89	72	179	94	71	213	88	46	585	588	43	43
81	72	182	96	73	208	83	45	584	193	44	44
92	74	186	97	73	215	86	46	589	195	41	41
86	75	182	96	69	198	85	56	580	215	49	49
96	75	185	95	67	196	87	54	554	185	53	53
95	76	188	94	75	211	83	45	564	466	55	55
88	74	188	95	73	196	85	52	456	589	51	51
88	74	178	96	74	197	85	61	447	363	5	5
81	73	179	94	76	205	84	45	580	279	44	44
92	72	188	94	71	209	86	44	379	188	44	44
92	73	189	96	72	208	83	45	372	189	46	46
86	75	184	95	71	208	83	45	375	184	47	47
95	76	202	96	71	208	84	48	508	559	56	56

Discriminant Analysis: Description of Group Separation

8.1 INTRODUCTION

We use the term *group* to represent either a population or a sample from the population. There are two major objectives in separation of groups:

1. **Description of group separation**, in which linear functions of the variables (*discriminant functions*) are used to describe or elucidate the differences between two or more groups. The goals of descriptive discriminant analysis include identifying the relative contribution of the p variables to separation of the groups and finding the optimal plane on which the points can be projected to best illustrate the configuration of the groups.
2. **Prediction or allocation of observations to groups**, in which linear or quadratic functions of the variables (*classification functions*) are employed to assign an individual sampling unit to one of the groups. The measured values in the observation vector for an individual or object are evaluated by the classification functions to find the group to which that individual most likely belongs.

For consistency, we will use the term *discriminant analysis* only in connection with objective 1. We will refer to all aspects of objective 2 as *classification analysis*, which is the subject of Chapter 9. Unfortunately, there is no general agreement with regard to usage of the terms *discriminant analysis* and *discriminant functions*. Many writers, perhaps the majority, use the term *discriminant analysis* in connection with the second objective, prediction or allocation. The linear functions contributing to the first objective, description of group separation, are often referred to as *canonical variates* or *discriminant coordinates*. To avoid confusion, we prefer to reserve the term *canonical* for canonical correlation analysis in Chapter 11.

Discriminant functions are linear combinations of variables that best separate groups. They were introduced in Section 3.3 for two groups and in Sections 6.3.4 and 6.4 for several groups. In those sections, interest was centered on follow-up to Hotelling's T^2 -tests and MANOVA tests. In this chapter, we further develop these useful multivariate tools.

8.2 THE DISCRIMINANT FUNCTION FOR TWO GROUPS

We assumed that the two populations to be compared have the same covariance matrix Σ but distinct mean vectors μ_1 and μ_2 . We work with samples $\mathbf{y}_{11}, \mathbf{y}_{12}, \dots, \mathbf{y}_{1n_1}$ and $\mathbf{y}_{21}, \mathbf{y}_{22}, \dots, \mathbf{y}_{2n_2}$ from the two populations. As usual, each vector $\mathbf{y}_{ij}$ consists of measurements on p variables. The discriminant function is the linear combination of these p variables that maximizes the distance between the two (transformed) group mean vectors. A linear combination $\mathbf{z} = \mathbf{a}'\mathbf{y}$ transforms each observation vector to a scalar

$$z_{1i} = \mathbf{a}'\mathbf{y}_{1i} = a_1 y_{1i1} + a_2 y_{1i2} + \dots + a_p y_{1ip}, \quad i = 1, 2, \dots, n_1$$

$$z_{2i} = \mathbf{a}'\mathbf{y}_{2i} = a_1 y_{2i1} + a_2 y_{2i2} + \dots + a_p y_{2ip}, \quad i = 1, 2, \dots, n_2$$

Hence for $n_1 + n_2$ observations vectors in the two samples,

$$\begin{array}{cc} \mathbf{y}_{11} & \mathbf{y}_{12} \\ \mathbf{y}_{13} & \mathbf{y}_{14} \\ \vdots & \vdots \\ \mathbf{y}_{1n_1} & \mathbf{y}_{1n_2} \end{array}$$

are transformed to scalars,

$$\begin{array}{cc} z_{11} & z_{12} \\ z_{13} & z_{14} \\ \vdots & \vdots \\ z_{1n_1} & z_{1n_2} \end{array}$$

We find the means $\bar{z}_1 = \sum_{i=1}^{n_1} z_{1i}/n_1 = \mathbf{a}'\bar{\mathbf{y}}_1$ and $\bar{z}_2 = \mathbf{a}'\bar{\mathbf{y}}_2$ by (3.34), where $\bar{\mathbf{y}}_1 = \sum_{i=1}^{n_1} \mathbf{y}_{1i}/n_1$ and $\bar{\mathbf{y}}_2 = \sum_{i=1}^{n_2} \mathbf{y}_{2i}/n_2$. We then wish to find the vector $\mathbf{a}$ that maximizes the standardized difference $(\bar{z}_1 - \bar{z}_2)/s_1$. Since $\bar{z}_1 - \bar{z}_2/s_1$ can be negative, we use the squared distance $(\bar{z}_1 - \bar{z}_2)^2/s_1^2$, which, by (3.50) and (3.52), can be expressed as

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_1^2} = \frac{[\mathbf{a}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)]^2}{\mathbf{a}'\Sigma_0\mathbf{a}} \quad (8.1)$$

The maximum of (8.1) occurs when

$$\mathbf{a} = \Sigma_0^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2), \quad (8.2)$$

or when $\mathbf{a}$ is any multiple of $\Sigma_0^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$. Thus the maximizing vector $\mathbf{a}$ is not unique. However, its "direction" is unique; that is, the relative values or ratios of $a_1, a_2, \dots, a_p$ are unique, and $\mathbf{z} = \mathbf{a}'\mathbf{y}$ projects points $\mathbf{y}$ onto the line on which

$(\bar{Q}_1 - \bar{Y}_1)^2/\sigma_1^2$ is maximised. Note that in order for Σ_{μ}^{-1} to exist, we must have $\sigma_1 + \sigma_2 - 2 > 0$.

The optimum direction given by $\mathbf{a} = \Sigma_{\mu}^{-1}(\bar{Q}_1 - \bar{Y}_1)$ is effectively parallel to the line joining $\bar{Y}_1$ and $\bar{Y}_2$, because the squared distance $(\bar{Q}_1 - \bar{Y}_1)^2/\sigma_1^2$ is equivalent to the standardised distance between $\bar{Y}_1$ and $\bar{Y}_2$. This can be seen by substituting (8.2) into (8.1) to obtain

$$\frac{(\bar{Q}_1 - \bar{Y}_1)^2}{\sigma_1^2} = (\bar{Q}_1 - \bar{Y}_2)' \Sigma_{\mu}^{-1} (\bar{Q}_1 - \bar{Y}_2) \quad (8.3)$$

for $z = \mathbf{a}'\mathbf{y}$ with $\mathbf{a} = \Sigma_{\mu}^{-1}(\bar{Q}_1 - \bar{Y}_1)$. Since $\mathbf{a}' = (\bar{Q}_1 - \bar{Y}_2)'\Sigma_{\mu}^{-1}$, we can write (8.3) as $(\bar{Q}_1 - \bar{Y}_1)^2/\sigma_1^2 = \mathbf{a}'(\bar{Q}_1 - \bar{Y}_2)$, and any other direction than that represented by $\mathbf{a} = \Sigma_{\mu}^{-1}(\bar{Q}_1 - \bar{Y}_1)$ would yield a smaller difference between $\mathbf{a}'\bar{Y}_1$ and $\mathbf{a}'\bar{Y}_2$ (see Section 3.5).

Figure 8.1 illustrates the separation of two bivariate normal ($p = 2$) groups along the single dimension represented by the discriminant function $z = \mathbf{a}'\mathbf{y}$, where $\mathbf{a}$ is given by (8.2). In this illustration the population covariance matrices are equal. The linear combinations $z_1 = \mathbf{a}'\mathbf{y}_{11} = a_1y_{11} + a_2y_{12}$ and $z_2 = \mathbf{a}'\mathbf{y}_{21} = a_1y_{21} + a_2y_{22}$ project the points $\mathbf{y}_{11}$ and $\mathbf{y}_{21}$ onto the line of optimum separation of the two groups. Since the two variables y_1 and y_2 are bivariate normal, a linear combination $z = a_1y_1 + a_2y_2 = \mathbf{a}'\mathbf{y}$ is univariate normal (see property 1a in Section 6.2). We have therefore indicated this by two univariate normal densities along the line representing z .

The point where the line joining the points of intersection of the two ellipses intersects the discriminant function line z is the point of maximum separation (minimum



Figure 8.1. Two-group discriminant analysis.



Figure 8.2. Separation achieved by the discriminant function.

overlap of points projected onto the line. If the two populations are multivariate normal with common covariance matrix Σ , as illustrated in Figure 8.1, it can be shown that all possible group separation is expressed in a single new dimension.

In Figure 8.2, we illustrate the optimum separation achieved by the discriminant function. Projection in another direction—denoted by x' —gives a smaller standardized distance between the transformed means $\bar{x}'_1$ and $\bar{x}'_2$, and also a larger overlap between the projected points.

Example 8.1. Samples of steel produced at two different rolling temperatures are compared in Table 8.1 (Kassam and Lomax 1989a). The variables are y_1 = yield

Table 8.1. Yield Point and Ultimate Strength of Steel Produced at Two Rolling Temperatures

Temperature 1		Temperature 2	
y_1	y_2	y_1	y_2
33	68	35	57
36	65	36	59
35	64	39	59
38	63	39	60
40	65	41	63
		43	63
		44	58

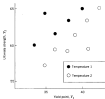


Figure 8.3. Ultimate strength and fatigue for steel tested at two temperatures.

point and y_2 = ultimate strength. From the data, we calculate

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 36.4 \\ 82.6 \end{pmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 38.8 \\ 80.4 \end{pmatrix}, \quad \mathbf{S}_{p2} = \begin{pmatrix} 7.92 & 3.68 \\ 3.68 & 6.29 \end{pmatrix}.$$

A plot of the data appears in Figure 8.3. We see that if the points were projected on either the y_1 or the y_2 axis, there would be considerable overlap. In fact, when the two groups are compared by means of a t -statistic for each variable separately, both t 's are nonsignificant.

$$t_1 = \frac{\bar{T}_{11} - \bar{T}_{12}}{\sqrt{s_{12}^2(1/n_1 + 1/n_2)}} = -1.88,$$

$$t_2 = \frac{\bar{T}_{21} - \bar{T}_{22}}{\sqrt{s_{22}^2(1/n_1 + 1/n_2)}} = 1.48.$$

However, it is clear in Figure 8.3 that the two groups can be separated. If they are projected in an appropriate direction, as in Figure 8.4, there will be no overlap. The single dimension onto which the points would be projected is the discriminant function

$$z = \mathbf{a}'\mathbf{y} = a_1 y_1 + a_2 y_2 = -1.655y_1 + 1.85y_2.$$

Table 3.2. Discriminant Function ($a = 1.439y_1 + 1.829y_2$)
Evaluated for Data in Table 3.1

Temperature 1	Temperature 2
55.26	46.56
51.50	48.57
56.50	45.00
51.88	47.00
51.44	47.66
	48.69
	48.48

where $\mathbf{a}$ is obtained as

$$\mathbf{a} = \mathbf{S}_p^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = \begin{pmatrix} -1.659 \\ 1.829 \end{pmatrix}.$$

The values of the projected points are found by calculating $\hat{y}$ for each observation vector $\mathbf{p}$ in the two groups. The results are given in Table 3.2, where the separation provided by the discriminant function is clearly evident. $\square$

3.2 RELATIONSHIP BETWEEN TWO-GROUP DISCRIMINANT ANALYSIS AND MULTIPLE REGRESSION

The natural connection between multiple regression and two-group discriminant analysis was introduced as a computational device in Section 2.5.2. The roles of independent and dependent variables are reversed in the two models. The dependent variables (y 's) of discriminant analysis become the independent variables in regression.

Let w be a grouping variable identifying groups 1 and 2, such that $\bar{w} = 0$ and define $\mathbf{b} = (b_1, b_2, \dots, b_p)'$ as the vector of regression coefficients when w is fit to the y 's. Then by (2.21), $\mathbf{b}$ is proportional to the discriminant function coefficient vector $\mathbf{a} = \mathbf{S}_p^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$:

$$\mathbf{b} = \frac{n_2 \bar{\mathbf{y}}_1}{(n_1 + n_2)(n_1 + n_2 - 2 + T^2)} \mathbf{a}, \quad (3.4)$$

where $T^2 = [n_1 n_2 / (n_1 + n_2)] (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$ as in (3.9). From (3.20) the squared multiple correlation R^2 is related to T^2 by

$$R^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{b} = \frac{T^2}{n_1 + n_2 - 2 + T^2}. \quad (3.5)$$

Thruston statistic (5.35) for the hypothesis that q of the $p + q$ variables are redundant for separating the groups can also be obtained in terms of regression by (5.31) as

$$F = \frac{s_1 + s_2 - p - q - 1}{q} \frac{R_{p+q}^2 - R_p^2}{1 - R_{p+q}^2}, \quad (5.36)$$

where R_{p+q}^2 and R_p^2 are from regressions with $p + q$ and p variables, respectively.

The link between two-group discriminant analysis and multiple regression was first noted by Fisher (1936). Flury and Riedwyl (1985) give further insights into the relationship.

Example 5.3. In Example 5.1.2, the psychological data of Table 5.1 were used in an illustration of the regression approach to computation of $\mathbf{a}$ and F^2 . We use the same data to obtain $\mathbf{b}$ and R^2 from $\mathbf{a}$ and F^2 .

From the results of Examples 5.1.2 and 5.3, we have

$$F^2 = 87.6815, \\ \mathbf{a} = \begin{pmatrix} -.5164 \\ -.2033 \\ .4660 \\ -.3897 \end{pmatrix}.$$

To find $\mathbf{b}$ from $\mathbf{a}$ and F^2 , we use (5.4):

$$\mathbf{b} = \frac{(32)(32)}{(32 + 32)(32 + 32 - 2 + 87.6815)} \mathbf{a} = \begin{pmatrix} .093 \\ -.028 \\ .047 \\ -.093 \end{pmatrix}.$$

To find R^2 , we use (5.5):

$$R^2 = (\mathbf{G}_1 - F_0)\mathbf{b} = \begin{pmatrix} 1.625 \\ 2.080 \\ 10.531 \\ -.812 \end{pmatrix}' \begin{pmatrix} .093 \\ -.028 \\ .047 \\ -.093 \end{pmatrix} = .641.$$

We can also use the relationship with F^2 in (5.3):

$$R^2 = \frac{F^2}{s_1 + s_2 - 2 + F^2} = \frac{87.6815}{32 + 32 - 2 + 87.6815} = .641. \quad \square$$

5.4 DISCRIMINANT ANALYSIS FOR SEVERAL GROUPS

5.4.1 Discriminant Functions

In discriminant analysis for several groups, we are concerned with finding linear combinations of variables that best separate the g groups of multivariate observations. Discriminant analysis for several groups may serve any one of various purposes:

1. Examine group separation in a two-dimensional plot. When there are more than two groups, it requires more than one discriminant function to describe group separation. If the points in the p -dimensional space are projected onto a two-dimensional space represented by the first two discriminant functions, we obtain the best possible view of how the groups are separated.
2. Find a subset of the original variables that separates the groups almost as well as the original set. This topic was introduced in Section 5.11.2.
3. Rank the variables in terms of their relative contribution to group separation. This use for discriminant functions has been mentioned in Sections 5.3, 5.3.4, 5.3.8, and 5.4. In Section 5.5, we discuss standardized discriminant function coefficients that provide a more valid comparison of the variables.
4. Interpret the new dimensions represented by the discriminant functions.
5. Follow up on fixed-effects MANCOV.

Purposes 3 and 4 are closely related. Any of the first four can be used to accomplish purpose 5. Methods of achieving these five goals of discriminant analysis are discussed in subsequent sections. In the present section we review discriminant functions for the several-group case and discuss attendant assumptions. For alternative estimators of discriminant functions that may be useful in the presence of multicollinearity or outliers, see Rencher (1995, Section 3.11).

For J groups (samples) with n_j observations in the j th group, we transform each observation vector $\mathbf{x}_{ij}$ to obtain $\mathbf{z}_{ij} = \mathbf{w}'\mathbf{x}_{ij}$, $i = 1, 2, \dots, n_j$; $j = 1, 2, \dots, J$, and find the means $\bar{\mathbf{z}}_j = \mathbf{w}'\bar{\mathbf{x}}_j$, where $\bar{\mathbf{x}}_j = \sum_{i=1}^{n_j} \mathbf{x}_{ij}/n_j$. As in the two-group case, we seek the vector $\mathbf{w}$ that maximally separates $\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2, \dots, \bar{\mathbf{z}}_J$. To express separation among $\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2, \dots, \bar{\mathbf{z}}_J$, we extend the separation criterion (5.2) to the J -group case. Since $\mathbf{w}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{w}$, we can express (5.2) in the form

$$\frac{(\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2)^2}{s_j^2} = \frac{[\mathbf{w}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{w}'\mathbf{S}_{p2}\mathbf{w}} = \frac{\mathbf{w}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{w}}{\mathbf{w}'\mathbf{S}_{p2}\mathbf{w}} \quad (5.7)$$

To extend (5.7) to J groups, we use the $\mathbf{B}$ matrix from MANCOV in place of $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'$ [see (5.38)] and $\mathbf{E}$ in place of $\mathbf{S}_{p2}$ to obtain

$$h = \frac{\mathbf{w}'\mathbf{B}\mathbf{w}}{\mathbf{w}'\mathbf{E}\mathbf{w}} \quad (5.8)$$

which can also be expressed as

$$\lambda = \frac{SSB(\bar{x})}{SSB(\bar{y})} \quad (8.8)$$

where $SSB(\bar{x})$ and $SSB(\bar{y})$ are the between and within sums of squares for $\bar{x}$ [see (6.42)].

We can write (8.8) in the form

$$\begin{aligned} \mathbf{a}'\mathbf{H}\mathbf{a} &= \lambda\mathbf{a}'\mathbf{K}\mathbf{a}, \\ \mathbf{a}'(\mathbf{H}\mathbf{a} - \lambda\mathbf{K}\mathbf{a}) &= 0. \end{aligned} \quad (8.9)$$

We maximize values of λ and $\mathbf{a}$ that are solutions of (8.9) to a search for the value of $\mathbf{a}$ that results in maximum λ . The solution $\mathbf{a}' = \mathbf{0}$ is not permissible because it gives $\lambda = 0/0$ in (8.8). Other solutions are found from

$$\mathbf{H}\mathbf{a} - \lambda\mathbf{K}\mathbf{a} = \mathbf{0}, \quad (8.10)$$

which can be written in the form

$$(\mathbf{K}^{-1}\mathbf{H} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}. \quad (8.11)$$

The solutions of (8.11) are the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_d$ and associated eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_d$ of $\mathbf{K}^{-1}\mathbf{H}$. As in previous discussions of eigenvalues, we consider them to be ranked $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$. The number of (nonzero) eigenvalues r is the rank of $\mathbf{H}$, which can be found as the smaller of $d - 1$ or p . Thus the largest eigenvalue λ_1 is the maximum value of $\lambda = \mathbf{a}'\mathbf{H}\mathbf{a}/\mathbf{a}'\mathbf{K}\mathbf{a}$ in (8.8), and the coefficient vector that produces the maximum is the corresponding eigenvector $\mathbf{a}_1$. (This can be verified using calculus.) Hence the discriminant function that maximally separates the means is $\zeta_1 = \mathbf{a}_1'\mathbf{x}$, that is, ζ_1 represents the dimension or direction that maximally separates the means.

From the r eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_r$ of $\mathbf{K}^{-1}\mathbf{H}$ corresponding to $\lambda_1, \lambda_2, \dots, \lambda_r$, we obtain r discriminant functions $\zeta_1 = \mathbf{a}_1'\mathbf{x}, \zeta_2 = \mathbf{a}_2'\mathbf{x}, \dots, \zeta_r = \mathbf{a}_r'\mathbf{x}$, which show the dimensions or directions of differences among $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \dots, \bar{\mathbf{y}}_J$. These discriminant functions are uncorrelated, but they are not orthogonal ($\mathbf{a}_i'\mathbf{a}_j = 0$ for $i \neq j$) because $\mathbf{K}^{-1}\mathbf{H}$ is not symmetric [see Rencher (1995, pp. 204–204)]. Note that the numbering $\zeta_1, \zeta_2, \dots, \zeta_r$ corresponds to the eigenvalues, not to the d groups as was done earlier in this section.

The relative importance of each discriminant function ζ_i can be assessed by considering its eigenvalue as a proportion of the total:

$$\frac{\lambda_i}{\sum_{j=1}^r \lambda_j}. \quad (8.12)$$

By this criterion, two or three discriminant functions will often suffice to describe the group differences. The discriminant functions associated with small eigenvalues can

be neglected. A test of significance for each discriminant function is also available (see Section 9.6).

The matrix $\mathbf{E}^{-1}\mathbf{H}$ is not symmetric. Many algorithms for computation of eigenvalues and eigenvectors accept only symmetric matrices. In Section 5.1.4, it was shown that the eigenvalues of the symmetric matrix $(\mathbf{U}^{-1})'\mathbf{H}\mathbf{U}^{-1}$ are the same as those of $\mathbf{E}^{-1}\mathbf{H}$, where $\mathbf{E} = \mathbf{U}\mathbf{U}'$ is the Cholesky factorization of $\mathbf{E}$. However, an adjustment is needed for the eigenvectors. If $\mathbf{h}$ is an eigenvector of $(\mathbf{U}^{-1})'\mathbf{H}\mathbf{U}^{-1}$, then $\mathbf{u} = \mathbf{U}^{-1}\mathbf{h}$ is an eigenvector of $\mathbf{E}^{-1}\mathbf{H}$.

The preceding discussion was presented in terms of unequal sample sizes $n_1, n_2, \dots, n_k$. In applications, this situation is common and can be handled with no difficulty. Ideally, the smallest n_i should exceed the number of variables, p . This is not required mathematically but will lead to more stable discriminant functions.

Example 8.4.1. The data in Table 8.3 were collected by G. R. Borge and H. M. Fisher (Brigham Young University) as part of a preliminary study of a possible link between football helmet design and neck injuries.

Six head measurements were made on each subject. There were 30 subjects in each of three groups: high school football players (group 1), college football players (group 2), and midget football players (group 3). The six variables are

WIDTH = head width at widest dimension,

CIRCUM = head circumference,

FRONT = front-to-back measurement at eye level,

DEPTH = eye-to-top-of-head measurement,

EARSD = ear-to-ear measurement,

JAW = jaw width.

The eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ are $\lambda_1 = 1.9178$ and $\lambda_2 = .1158$. The corresponding eigenvectors are

$$\mathbf{u}_1 = \begin{pmatrix} -.548 \\ .384 \\ .386 \\ .647 \\ .584 \\ .829 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -1.487 \\ .081 \\ .028 \\ -.540 \\ .384 \\ 1.528 \end{pmatrix}.$$

The first eigenvalue accounts for a substantial proportion of the total

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.9178}{1.9178 + .1158} = .94.$$

Thus the mean vectors lie largely in one dimension, and one discriminant function suffices to describe most of the separation among the three groups. $\square$

Table A.3. Head Measurements for Three Groups

Group	WEAM	CTACTUM	POSTIO	STYLOS	BARBES	LAB
1	12.5	37.2	19.5	12.5	14.8	31.8
1	12.5	38.4	20.8	120	18.8	32.8
1	14.8	38.8	18.8	180	18.8	32.8
1	18.8	38.4	28.8	18.8	18.8	32.8
1	14.5	38.4	28.8	180	15.5	32.8
1	18.8	40.8	20.8	120	18.8	32.8
1	15.8	38.4	19.5	18.5	15.5	32.8
1	15.8	38.4	20.8	180	14.8	32.8
1	15.5	38.7	28.5	15.5	14.5	32.5
1	15.5	38.7	28.5	150	15.8	32.8
1	18.8	37.2	18.8	140	14.8	31.8
1	15.5	38.7	20.8	180	18.8	32.8
1	18.8	37.2	18.8	140	18.8	32.8
1	15.5	42.2	20.5	140	18.8	32.8
1	15.5	37.2	18.5	18.5	15.8	32.8
1	14.8	40.8	28.8	150	15.8	32.8
1	14.5	38.4	28.8	120	14.5	32.8
1	15.8	38.8	18.8	150	14.8	32.8
1	18.8	38.7	28.8	12.8	18.8	32.8
1	18.8	37.2	18.8	120	18.8	31.8
1	15.8	38.8	18.8	120	18.8	32.8
1	15.5	38.8	18.5	14.5	14.5	32.8
1	17.5	43.5	20.5	140	15.5	32.5
1	15.5	37.2	18.8	150	15.5	32.5
1	15.5	40.8	28.5	120	15.8	32.5
1	15.5	40.8	20.8	14.5	15.5	32.5
1	18.8	43.8	20.8	14.8	18.8	32.8
1	14.5	38.4	28.5	180	18.8	38.5
1	15.5	38.8	28.8	18.5	18.8	32.8
1	18.8	40.8	28.8	12.5	14.5	32.5
2	15.5	40.8	20.1	103	18.4	32.4
2	15.4	38.7	28.8	128	14.5	31.5
2	15.1	38.7	28.2	11.4	14.1	32.1
2	14.8	38.8	18.8	11.0	15.4	31.8
2	18.8	38.8	28.1	98	11.1	31.7
2	18.2	37.8	14.8	98	12.8	31.4
2	15.8	38.8	28.8	102	12.8	31.8
2	18.8	38.8	28.1	88	12.8	32.8
2	15.5	37.8	18.8	10.5	13.8	31.8
2	15.8	38.5	18.8	104	14.5	32.8
2	15.5	37.2	28.8	11.2	15.4	32.4
2	15.5	38.8	18.8	92	12.8	31.2
2	18.7	37.8	18.8	11.8	12.8	32.8
2	18.8	37.8	28.8	102	12.7	32.8
2	14.8	34.8	18.5	11.2	18.8	31.2

Table 8.3. (Continued)

Group	WORM	CRUCUM	POCTO	RYEBE	BARBE	DAVE
2	16.5	29.8	28.2	9.6	64.5	32.2
2	15.9	26.1	18.8	9.8	10.8	32.6
2	18.2	31.0	18.8	18.1	18.2	31.8
2	14.8	31.6	18.3	12.0	10.8	31.8
2	15.5	26.3	28.8	9.9	10.4	31.3
2	13.2	23.8	18.3	9.9	18.4	31.8
2	15.0	26.5	19.3	9.1	10.8	31.7
2	15.0	26.8	28.2	8.6	64.2	31.5
2	15.8	23.5	19.2	8.2	10.8	32.6
2	14.8	27.0	28.2	9.8	10.8	30.5
2	18.2	18.4	18.1	9.8	10.8	31.2
2	15.8	28.8	21.8	8.8	10.5	31.8
2	13.3	27.3	28.1	9.8	18.1	32.3
2	18.5	28.0	19.5	9.0	10.8	32.3
2	17.0	32.8	21.5	18.0	10.8	32.8
3	14.9	26.5	28.4	7.4	10.8	32.0
3	15.4	27.5	19.5	18.5	10.8	31.5
3	15.0	31.4	18.2	9.7	10.5	31.8
3	14.8	18.0	18.8	8.8	10.8	31.8
3	18.2	18.3	18.3	11.8	18.8	31.8
3	14.8	28.0	18.8	18.0	10.4	31.3
3	15.8	26.7	18.7	18.8	10.8	32.6
3	14.7	23.8	18.7	11.1	10.8	31.5
3	15.5	26.5	19.4	11.5	10.4	31.8
3	18.1	68.0	28.5	18.8	10.7	32.2
3	15.2	27.8	19.8	18.4	10.5	31.4
3	18.1	18.0	18.4	18.0	10.1	30.4
3	15.8	39.8	28.5	12.0	10.8	31.3
3	18.1	27.7	18.7	18.0	10.8	31.3
3	15.7	28.5	28.7	11.0	10.8	31.3
3	15.0	26.8	19.6	18.5	10.5	32.1
3	15.0	26.8	19.5	9.9	64.8	32.1
3	15.2	28.0	28.4	11.0	10.1	31.7
3	18.8	19.3	18.8	12.1	18.8	32.1
3	18.8	18.2	18.7	11.7	10.8	32.1
3	15.8	27.3	18.4	11.8	18.7	31.8
3	18.8	27.2	18.8	18.8	10.8	32.0
3	15.4	27.0	18.8	11.0	18.8	31.4
3	18.8	28.2	28.8	18.4	10.8	32.2
3	15.4	27.6	19.6	18.2	10.8	31.7
3	15.8	68.3	28.8	12.4	10.4	32.1
3	15.4	23.8	18.8	18.7	18.2	30.8
3	18.8	18.4	18.8	18.1	18.8	31.7
3	13.7	28.8	28.4	12.1	10.8	32.7
3	17.0	31.7	28.7	11.8	10.5	32.3

8.4.1 A Measure of Association for Discriminant Functions

Measures of association between the dependent variable $y_1, y_2, \dots, y_p$ and the independent grouping variable(s) associated with $\mathbf{g}_i$, $i = 1, 2, \dots, J$, were presented in Section 6.1.8. These measures attempt to answer the questions, How well do the variables separate the groups? It was noted that Ray's statistic R serves as an R^2 -like measure of association, since it is the ratio of between to total sum of squares for the first discriminant function, $\tau_1 = \mathbf{a}_1'\mathbf{x}$:

$$\eta_1^2 = R^2 = \frac{\lambda_1}{1 + \lambda_1} = \frac{\text{SSB}(\tau_1)}{\text{SSB}(\tau_1) + \text{SSB}(\tau_2)}.$$

[see (6.42) and (6.31). Another interpretation of η_1^2 is the maximum squared correlation between the first discriminant function and the best linear combination of the $k = 1$ (dummy) group membership variables [see a comment following (6.48) in Section 6.1.8]. Dummy variables were defined in the first two paragraphs of Section 6.1.8. The maximum correlation is called the (first) canonical correlation (see Chapter 11). The squared canonical correlation can be calculated for each discriminant function:

$$\tau_i^2 = \frac{\lambda_i}{1 + \lambda_i}, \quad i = 1, 2, \dots, p. \quad (8.14)$$

The average squared canonical correlation was used as a measure of association in (6.49).

Example 8.4.1. For the football data of Table 8.3, we obtain the squared canonical correlation between each of the two discriminant functions and the grouping variable,

$$\begin{aligned} \tau_1^2 &= \frac{\lambda_1}{1 + \lambda_1} = \frac{1.9778}{1 + 1.9778} = .687, \\ \tau_2^2 &= \frac{\lambda_2}{1 + \lambda_2} = \frac{.1138}{1 + .1138} = .104. \end{aligned} \quad \square$$

8.5 STANDARDIZED DISCRIMINANT FUNCTIONS

In Section 5.5, it was noted that in the two-group case the relative contribution of the y 's to separation of the two groups can best be assessed by comparing the coefficients a_r , $r = 1, 2, \dots, p$, in the discriminant function

$$Z = \mathbf{a}'\mathbf{x} = a_1y_1 + a_2y_2 + \cdots + a_py_p.$$

Similar comments appeared in Section 6.1.4, 6.1.8, and 6.4 concerning the use of discriminant functions to assess contributions of the y 's to separation of several groups.

However, such comparisons are informative only if the y 's are commensurate, that is, measured on the same scale and with comparable variances. If the y 's are not commensurate, we need coefficients a_i^* that are applicable to standardized variables.

Consider the case of two groups. For the i th observation vector $\mathbf{y}_{1i}$ of $\mathbf{y}_{1n}$ in group 1 or 2, we can express the discriminant function in terms of standardized variables as

$$z_{1i} = a_1^* \frac{y_{1i1} - \bar{y}_{11}}{s_1} + a_2^* \frac{y_{1i2} - \bar{y}_{12}}{s_2} + \cdots + a_p^* \frac{y_{1ip} - \bar{y}_{1p}}{s_p}, \quad (3.15)$$

$$i = 1, 2, \dots, n_1,$$

$$z_{2i} = a_1^* \frac{y_{2i1} - \bar{y}_{21}}{s_1} + a_2^* \frac{y_{2i2} - \bar{y}_{22}}{s_2} + \cdots + a_p^* \frac{y_{2ip} - \bar{y}_{2p}}{s_p},$$

$$i = 1, 2, \dots, n_2,$$

where $\bar{\mathbf{y}}_1 = (\bar{y}_{11}, \bar{y}_{12}, \dots, \bar{y}_{1p})$ and $\bar{\mathbf{y}}_2 = (\bar{y}_{21}, \bar{y}_{22}, \dots, \bar{y}_{2p})$ are the mean vectors for the two groups, and s_r is the within-sample standard deviation of the r th variable, obtained as the square root of the r th diagonal element of $\mathbf{S}_{D_r}$. Clearly, these standardized coefficients must be of the form

$$a_r^* = a_r/s_r, \quad r = 1, 2, \dots, p. \quad (3.16)$$

In vector form, this becomes

$$\mathbf{a}^* = (\det \mathbf{S}_{D_r})^{1/2} \mathbf{a}. \quad (3.17)$$

For the several-group case, we can standardize the discriminant functions in an analogous fashion. If we denote the r th coefficient in the m th discriminant function by a_{mr} , $m = 1, 2, \dots, n$, $r = 1, 2, \dots, p$, then the standardized form is

$$a_{mr}^* = a_{mr}/s_r,$$

where s_r is the within-group standard deviation obtained from the diagonal of $\mathbf{S}_{D_r} = \mathbf{E}/v_r$. Note that a_{mr}^* has two subscripts because there are several discriminant functions, whereas a_r^* in (3.15) has only one subscript because there is one discriminant function for two groups.

Alternatively, since the m th eigenvector is unique only up to multiplication by a scalar, we can simplify the standardization by using

$$a_{mr}^* = \sqrt{v_r} a_{mr}, \quad r = 1, 2, \dots, p,$$

where v_r is the r th diagonal element of $\mathbf{E}$. For further discussion of the use of standardized discriminant function coefficients to gauge the relative contribution of the variables to group separation, see Section 5.3.1 [see also Rencher and Sprott (1990) and Rencher (1995, Section 3.4)].

Example 8.5. In Example 8.4.1, we obtained the discriminant function coefficient vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ for the football data of Table 8.3. Since $\lambda_1/(\lambda_1 + \lambda_2) = .98$, we concentrate on $\mathbf{a}_1$. To standardize $\mathbf{a}_1$, we modify it with sample standard deviations of the variables. The pooled covariance matrix is given by

$$\mathbf{S}_{12} = \frac{\mathbf{E}}{87} = \begin{pmatrix} .828 & .378 & .158 & .084 & .125 & .228 \\ .378 & 3.364 & 1.828 & .888 & .568 & .803 \\ .158 & 1.828 & .548 & .877 & .128 & .139 \\ .084 & .888 & .877 & 1.132 & .345 & .034 \\ .128 & .568 & .128 & .345 & .618 & .089 \\ .228 & .803 & .139 & .034 & .089 & .276 \end{pmatrix}.$$

Using the square roots of the diagonal elements of $\mathbf{S}_{12}$, we obtain

$$\mathbf{a}_1^* = \begin{pmatrix} \sqrt{.828}(-.948) \\ \sqrt{3.364}(.004) \\ \vdots \\ \sqrt{.276}(.828) \end{pmatrix} = \begin{pmatrix} -.811 \\ .007 \\ .005 \\ .719 \\ .387 \\ .508 \end{pmatrix}.$$

Thus the fourth, first, sixth, and fifth variables contribute most to separating the groups, in that order. The second and third variables are not useful (in the presence of the others) in distinguishing groups. $\square$

8.6 TESTS OF SIGNIFICANCE

In order to test hypotheses, we need the assumption of multivariate normality. This is not explicitly required for the development of discriminant functions.

8.6.1 Tests for the Two-Group Case

By 8.5) we see that the separation of transformed means, $(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^T \mathbf{a}_1^*$, achieved by the discriminant function $z = \mathbf{a}_1^T \mathbf{y}$ is equivalent to the standardized distance between the mean vectors $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$. This standardized distance is proportional to the two-group T^2 in (8.51) in Section 5.4.1. Hence the discriminant function coefficient vector $\mathbf{a}$ is significantly different from $\mathbf{0}$ if T^2 is significant. More formally, if the population discriminant function coefficient vector is expressed as $\mathbf{a} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, then $H_0: \mathbf{a} = \mathbf{0}$ is equivalent to $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

To test the significance of a subset of the discriminant function coefficients, we can use the test of the corresponding subset of $\mathbf{y}$'s given in Section 5.8. To test the hypothesis that the population discriminant function has a specified form $\mathbf{a}_0^T \mathbf{y}$, see Banerjee (1998, Section 5.5.1).

8.4.2 Tests for the Several-Group Case

In Section 8.4.1 we noted that the discriminant criterion $\hat{\lambda} = \mathbf{a}'\mathbf{H}\mathbf{a}/\mathbf{a}'\mathbf{a}$ is maximized by λ_1 , the largest eigenvalue of $\mathbf{B}^{-1}\mathbf{B}$, and that the remaining eigenvalues $\lambda_2, \dots, \lambda_k$ correspond to other discriminant dimensions. These eigenvalues are the same as those in the Wilks' Λ -test in (8.14) for significant differences among mean vectors.

$$\Lambda_1 = \prod_{j=2}^k \frac{1}{1 + \lambda_j}, \quad (8.18)$$

which is distributed as $\Lambda_{k,1,1,1,1,1}$, where $R = \sum_{i=1}^k n_i$ for an unbalanced design or $R = N$ in the balanced case. Since Λ_1 is small if one or more λ_j 's are large, Wilks' Λ tests for significance of the eigenvalues and thereby for the discriminant functions. The k eigenvalues represent k dimensions of separation of the mean vectors $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \dots, \bar{\mathbf{y}}_g$. We are interested in which, if any, of these dimensions are significant. In the context of discriminant functions, Wilks' Λ is more useful than the other three MANOVA test statistics, because it can be used on a subset of eigenvalues, as we see shortly.

In addition to the exact test provided by the critical values for Λ found in Table A(9), we may use the χ^2 -approximation for Λ_1 given in (8.18), with $\nu_1 = R - k = \sum_{i=1}^k n_i - k$ and $\nu_2 = k - 1$:

$$\begin{aligned} V_1 &= -\left[\nu_1 - \frac{1}{2}(\nu_2 + \nu_1 + 1)\right] \ln \Lambda_1 \\ &= -\left[\nu_1 - \frac{1}{2}(\nu_2 + \nu_1 + 1)\right] \ln \prod_{j=2}^k \frac{1}{1 + \lambda_j} \\ &= \left[\nu_1 - \frac{1}{2}(\nu_2 + \nu_1 + 1)\right] \sum_{j=2}^k \ln(1 + \lambda_j), \end{aligned} \quad (8.19)$$

which is approximately χ^2 with $\nu_2(k - 1)$ degrees of freedom. This test statistic V_1 and its approximation (8.19) test the significance of all of $\lambda_2, \lambda_3, \dots, \lambda_k$. If the test leads to rejection of H_0 , we conclude that at least one of the λ_j 's is significantly different from zero, and therefore there is at least one dimension of separation of mean vectors. Since λ_1 is the largest, we are only sure of its significance, along with that of $\lambda_1 = \mathbf{a}'\mathbf{g}$.

To test the significance of $\lambda_2, \lambda_3, \dots, \lambda_k$, we delete λ_1 from Wilks' Λ and the associated χ^2 -approximation to obtain

$$\Lambda_2 = \prod_{j=3}^k \frac{1}{1 + \lambda_j}, \quad (8.20)$$

$$V_1 = -\left[N-1-\frac{1}{2}(p+k)\right]\ln \lambda_1 = \left[N-1-\frac{1}{2}(p+k)\right]\sum_{i=1}^k \ln(1+\lambda_i), \quad (8.20)$$

which is approximately χ^2 with $(p-1)(k-2)$ degrees of freedom. If this test leads to rejection of H_0 , we conclude that at least λ_1 is significant along with the associated discriminant function $\tau_1 = u_1^T x$. We can continue in this fashion, testing each λ_i in turn until a test fails to reject H_0 . (To compensate for making several tests, an adjustment to the α -level of each test could be made as in procedure 2, Section 5.5.) The test statistic at the m th step is

$$\lambda_m = \prod_{i=m}^k \frac{1}{1+\lambda_i}, \quad (8.21)$$

which is distributed as $\lambda_{p-m+1, k-m, N-m+1}$. The statistic

$$\begin{aligned} V_m &= -\left[N-1-\frac{1}{2}(p+k)\right]\ln \lambda_m \\ &= \left[N-1-\frac{1}{2}(p+k)\right]\sum_{i=m}^k \ln(1+\lambda_i) \end{aligned} \quad (8.22)$$

has an approximate χ^2 -distribution with $(p-m+1)(k-m)$ degrees of freedom. In some cases, more λ 's will be statistically significant than the number would seem to be of practical importance. If $\lambda_i/\sum_{j=1}^k \lambda_j$ is small, the associated discriminant function may not be of interest, even if it is significant.

We can also use an F -approximation for each λ_i . For

$$\lambda_1 = \prod_{i=1}^k \frac{1}{1+\lambda_i},$$

we use (5.12), with $v_0 = N-k$ and $v_2 = k-1$:

$$F = \frac{1 - \lambda_1^{1/v_0} c_0}{\lambda_1^{1/v_0} \frac{c_0}{c_2}}, \quad (8.23)$$

where

$$\begin{aligned} c &= \sqrt{\frac{p^2(k-1)^2-4}{p^2+(k-1)^2-2}}, & w &= N-1-\frac{1}{2}(p+k), \\ d_1 &= p(k-1), & d_2 &= w - \frac{1}{2}(p(k-1)+1). \end{aligned}$$

For

$$\lambda_m = \prod_{i=m}^k \frac{1}{1+\lambda_i}, \quad m = 2, 3, \dots, k,$$

we get

$$\rho = \frac{1 - A_{m+1}^{1/2} \cdot dR_1}{A_m^{1/2} \cdot dR_1} \quad (8.25)$$

with $p = m + 1$ and $k = m$ in place of p and $k - 1$:

$$r = \sqrt{\frac{(p - m + 1)^2(k - m)^2 - 4}{(p - m + 1)^2 + (k - m)^2 - 5}},$$

$$w = N - 1 - \frac{1}{2}(p + k),$$

$$dR_1 = (p - m + 1)(k - m),$$

$$dR_2 = w - \frac{1}{2}(p - m + 1)(k - m) - 1.$$

Example 8.6.2. We test the significance of the two discriminant functions obtained in Example 8.4.1 for the football data. For the overall test we have, by (8.18),

$$\lambda_0 = \prod_{i=1}^2 \frac{1}{1 + \lambda_i} = \frac{1}{1 + 1.7878} \frac{1}{1 + .1158} = .307.$$

With $p = 6$, $k = 3$, and $N - k = 37$, the critical value from Table A.8 is $\lambda_{.05,6,3,36} = .302$. By (8.19), the χ^2 -approximation is

$$\begin{aligned} \overline{V}_1 &= -\left[N - 1 - \frac{1}{2}(p + k)\right] \ln \lambda_1 \\ &= -\left[36 - 1 - \frac{1}{2}(6 + 3)\right] \ln(.307) = 10.75, \end{aligned}$$

which exceeds the critical value $\chi_{.05,11}^2 = 19.117$. Thus at least the first discriminant function is significant.

To test the second discriminant function, we have, by (8.20),

$$\lambda_2 = \frac{1}{1 + .1158} = .896.$$

With $m = 2$, the (conservative) critical value is $\lambda_{.05,2,2,36} = .867$. Since this is close to $\lambda = .896$, we interpolate in Table A.8 to obtain $\lambda_{.05,2,2,36} = .875$. By (8.21), the χ^2 -approximation is

$$\overline{V}_2 = -\left[N - 1 - \frac{1}{2}(p + k)\right] \ln \lambda_2$$

$$= - \left[90 - 1 - \frac{1}{2}(8 + 3) \right] \ln \frac{1}{1 + .1259} = 8.27 + 2 \chi^2_{80.0} = 11.078.$$

For the F -approximation for A_0 , we obtain by (8.24)

$$\begin{aligned} t &= \sqrt{\frac{p^2 q^2 - 1}{p^2 + (q-1)^2 - 2}} = \sqrt{\frac{.5^2 .5^2 - 1}{.5^2 + .5^2 - 2}} = 2, \\ w &= N - 1 - \frac{1}{2}(p + q) = 90 - 1 - \frac{1}{2}(8 + 3) = 84.5, \\ d_0 &= p(q-1) = 6(2) = 12, \\ d_1 &= wt - \frac{1}{2}(p(q-1) - 2) = (84.5)(2) - \frac{1}{2}(6(2) - 2) = 169, \\ F &= \frac{1 - .5^{\frac{1}{2}}}{.5^{\frac{1}{2}}} \frac{d_0}{d_1} = \frac{1 - .50^{\frac{1}{2}}}{.50^{\frac{1}{2}}} \frac{169}{12} = 10.894. \end{aligned}$$

The p -value for $F = 10.894$ is less than .001. For the F -approximation for A_1 , we reduce p and q by 1 and obtain by (8.25)

$$\begin{aligned} t &= \sqrt{\frac{p^2 q^2 - 1}{p^2 + (q-1)^2 - 2}} = 1, \quad w = 90 - 1 - \frac{1}{2}(p + q) = 84.5, \\ d_0 &= 5(1) = 5, \quad d_1 = 84.5(1) - \frac{1}{2}(5(1) - 2) = 83, \\ F &= \frac{1 - .5}{.5} \frac{d_0}{d_1} = \frac{1 - .50}{.50} \frac{5}{83} = 1.514. \end{aligned}$$

The p -value for $F = 1.514$ is .979. Thus only the first discriminant function significantly separates groups. The exact test using λ_1 appears to be somewhat closer to rejection than are the approximate tests. $\square$

8.7 INTERPRETATION OF DISCRIMINANT FUNCTIONS

There is a close correspondence between interpreting discriminant functions and determining the contribution of each variable, and we shall not always make a distinction. In interpretation, the signs of the coefficients are taken into account; in ascertaining the contribution, the signs are ignored, and the coefficients are ranked in absolute value. (We discuss this distinction further in Section 8.7.1.) We are more commonly interested in assessing the contribution of the variables than in interpreting the function.

In the next three sections, we review three common approaches to assessing the contribution of each variable (in the presence of the other variables) to separating the

groups. The three methods are (1) examine the standardized discriminant function coefficients, (2) calculate a partial F -test for each variable, and (3) calculate a correlation between each variable and the discriminant function. The third method is the most widely recommended, but we note in Section 8.3.3 that it is the least useful.

8.3.1 Standardized Coefficients

To offset differing scales among the variables, the discriminant function coefficients can be standardized using (8.16) or (8.17), in which the coefficients have been adjusted so that they apply to standardized variables. For the observations in the first of two groups, for example, we have by (8.15),

$$z_{1i} = a_1' \frac{y_{1i1} - \bar{y}_{11}}{s_{y1}} + a_2' \frac{y_{1i2} - \bar{y}_{12}}{s_{y2}} + \cdots + a_p' \frac{y_{1ip} - \bar{y}_{1p}}{s_{yp}}, \\ i = 1, 2, \dots, n_1.$$

The standardized variables $(y_{1ir} - \bar{y}_{1r})/s_{yr}$ are scale free, and the standardized coefficients $a_r' = a_r/s_{yr}$, $r = 1, 2, \dots, p$, therefore correctly reflect the joint contribution of the variables to the discriminant function z as it maximally separates the groups. For the case of several groups, each discriminant function coefficient vector $\mathbf{a} = (a_1, a_2, \dots, a_p)'$ is an eigenvector of $\mathbf{B}^{-1}\mathbf{H}$, and as such, it takes into account the sample correlations among the variables as well as the influence of each variable in the presence of the others.

As noted in Section 8.3, this standardization is carried out for each of the r discriminant functions. Typically, each will have a different interpretation; that is, the pattern of the coefficients a_r' will vary from one function to another.

The absolute values of the coefficients can be used to rank the variables in order of their contribution to separating the groups. If we wish to go further and interpret or "name" a discriminant function, the signs can be taken into account. Thus, for example, $z_1 = .8y_1 - .5y_2 + .3y_3$ has a different meaning than does $z_2 = .8y_1 + .8y_2 + .5y_3$, since z_1 depends on the difference between y_1 and y_2 , whereas z_2 is related to the sum of y_1 and y_2 .

The discriminant function is subject to the same limitations as other linear combinations such as a regression equation; for example, (1) the coefficient for a variable may change notably if variables are added or deleted, and (2) the coefficients may not be stable from sample to sample unless the sample size is large relative to the number of variables. With regard to limitation 1, we note that the coefficients reflect the contribution of each variable in the presence of the particular variables at hand. This is, in fact, what we want the coefficients to do. As to limitation 2, for processing of a substantial number of variables is not "free." More stable estimates will be obtained from 50 observations on 2 variables than from 50 observations on 20 variables. In other words, if R^2/p is too small, the variables that rank high in one sample may emerge as less important in another sample.

8.3.1 Partial F -Values

For any variable y_i , we can calculate a partial F -test showing the significance of y_i after adjusting for the other variables, that is, the separation provided by y_i in addition to that due to the other variables. After computing the partial F for each variable, the variables can then be ranked.

In the case of two groups, the partial F is given by (8.25) as

$$F = (r - p + 1) \frac{T_p^2 - T_{p-1}^2}{r + T_{p-1}^2}, \quad (8.25)$$

where T_p^2 is the two-sample Hotelling T^2 with all p variables, T_{p-1}^2 is the T^2 statistic with all variables except y_i , and $r = n_1 + n_2 - 1$. The F -statistic in (8.25) is distributed as $F_{1, r-p+1}$.

For the several-group case, the partial A for y_i adjusted for the other $p - 1$ variables is given by (8.26) as

$$A(y_i | y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_p) = \frac{\Lambda_p}{\Lambda_{p-1}}, \quad (8.26)$$

where Λ_p is Wilks' Λ for all p variables and Λ_{p-1} involves all variables except y_i . The corresponding partial F is given by (8.27) as

$$F = \frac{1 - \Lambda}{\Lambda} \frac{r_p - p + 1}{r_p}, \quad (8.27)$$

where Λ is defined in (8.22), $r_p = R - k$, and $r_p = k - 1$. The partial A -statistic in (8.27) is distributed as $A_{1, r_p, r_p - p + 1}$, and the partial F in (8.27) is distributed as $F_{1, r_p - p + 1}$.

The partial F -values in (8.25) and (8.27) are not associated with a single dimension of group separation, as are the standardized discriminant function coefficients. For example, y_2 will have a different contribution to each of the r discriminant functions, but the partial F for y_2 constitutes an overall index of the contribution of y_2 to group separation taking into account all dimensions. However, the partial F -values will often rank the variables in the same order as the standardized coefficients for the first discriminant functions, especially if $\lambda_1 / \sum_{j=2}^p \lambda_j$ is very large so that the first function accounts for most of the available separation.

A partial index of association for y_i similar to the overall measure for all p 's given in (8.41), $\eta_{y_i}^2 = 1 - \Lambda_i$, can be defined by

$$R_i^2 = 1 - \Lambda_i, \quad i = 1, 2, \dots, p, \quad (8.28)$$

where Λ_i is the partial Λ in (8.27) for y_i . This partial R^2 is a measure of association between the grouping variables and y_i after adjusting for the other $p - 1$ y 's.

5.3.3 Correlations between Variables and Discriminant Functions

Many textbooks and research papers assert that the best measure of variable importance is the correlation between each variable and a discriminant function, $r_{y_{j|k}}$. It is claimed that these correlations are more informative than standardized coefficients with respect to the joint contribution of the variables to the discriminant functions. The correlations are often referred to as loadings or structure coefficients and are routinely provided in many major programs. However, Rencher (1995, 1992b, 1993, Section 5.7) has shown that the correlations in question show the contribution of each variable in a univariate context rather than in a multivariate one. The correlations actually reproduce the t or F for each variable, and consequently they show only how each variable by itself separates the groups, ignoring the presence of the other variables. Hence these correlations provide no information about how the variables contribute jointly to separation of the groups. They become misleading if used for interpretation of discriminant functions.

Upon reflection, we could have anticipated this failure of the correlations to provide multivariate information. The objection to standardized coefficients is based on the argument that they are “variable” because they change if some variables are deleted and others added. However, we actually want them to behave this way, so as to reflect the mutual influence of the variables on each other. In a multivariate analysis, interest is centered on the joint performance of the set of variables at hand. To ask for the contribution of each variable independent of all other variables is to request a univariate index that ignores the other variables. The correlations $r_{y_{j|k}}$ are stable and do not change when variables are added or deleted; this should be a clear signal that they are univariate in nature. There is no middle ground between the univariate and multivariate notions.

5.3.4 Rotation

Rotation of the discriminant function coefficients is sometimes recommended. This is a procedure (see Section 13.5) that attempts to produce a pattern with (absolute values of) coefficients close to 0 or 1. Discriminant functions with such coefficients might be easier to interpret, but they have two deficiencies: they no longer maximize group separation and they are correlated.

Accordingly, for interpretation of discriminant functions we recommend standardized coefficients rather than correlations or rotated coefficients.

5.4 SCATTER PLOTS

One benefit of the dimension reduction effected by discriminant analysis is the potential for plotting. It was noted in Section 5.1 that the number of large eigenvalues of $\mathbf{E}^{-1}\mathbf{W}$ reflects the dimensionality of the space occupied by the mean vectors. In many data sets, the first few discriminant functions account for most of $k_1 + k_2 + \cdots + k_p$, and consequently the pattern of the mean vectors can be effectively portrayed in a

two-dimensional plot. If the essential dimensionality is greater than 2, there may be some distortion in intersoup configuration in a two-dimensional plot, that is, some groups that overlap in two dimensions may be well separated in a third dimension.

To plot the first two discriminant functions for the individual observation vectors $\mathbf{p}_{ij}$, simply calculate $z_{1ij} = \mathbf{a}_1^T \mathbf{p}_{ij}$ and $z_{2ij} = \mathbf{a}_2^T \mathbf{p}_{ij}$ for $i = 1, 2, \dots, R$; $j = 1, 2, \dots, N_i$, and plot a scatter plot of the $R = \sum_i N_i$ values of

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^T \mathbf{p}_i \\ \mathbf{a}_2^T \mathbf{p}_i \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} \mathbf{p}_i = \mathbf{A} \mathbf{p}_i. \quad (8.30)$$

The transformed mean vectors,

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} \bar{\mathbf{p}}_i = \mathbf{A} \bar{\mathbf{p}}_i, \quad i = 1, 2, \dots, d. \quad (8.31)$$

should be plotted along with the individual values, $\mathbf{z}_{ij}$. In some cases, a plot would show only the transformed mean vectors $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_d$. For confidence regions for $\mathbf{p}_{i0} = \mathbf{A} \mathbf{p}_{i0}$, see Rousseeu (1988, Section 5.4).

We note that the eigenvalues of $\mathbf{B}^{-1} \mathbf{H}$ reveal the dimensionality of the mean vectors, not of the individual points. The dimensionality of the individual observations is p , although the essential dimensionality may be less because the variables are correlated. (The dimensionality of the observation vectors is the concern of principal components; see Chapter 12.) If $r = 2$, for example, so that the mean vectors occupy only two dimensions, the individual observation vectors ordinarily lie in more than two dimensions, and their inclusion in a plot constitutes a projection onto the two-dimensional plane of the mean vectors.

It was noted in Section 8.6.1 that the discriminant functions are uncorrelated but not orthogonal. Thus the angle between $\mathbf{a}_1$ and $\mathbf{a}_2$ as given by (3.14) is not 90° (that is, $\mathbf{a}_1^T \mathbf{a}_2 \neq 0$). In practice, however, the usual procedure is to plot discriminant functions on a rectangular coordinate system. The resulting distortion is generally not serious.

Example 8.8. Figure 8.4 contains a scatterplot of (z_1, z_2) for the observations in the football data of Table 8.3. Each observation in group 1 is denoted by a square, observations in group 2 are denoted by circles, and observations in group 3 are indicated by triangles. We see that the first discriminant function z_1 (the horizontal direction) effectively separates group 1 from groups 2 and 3, whereas the second discriminant function z_2 (the vertical direction) is less successful in separating group 2 from group 3.

The group mean vectors are indicated by solid circles. They are almost collinear, as we would expect since $\lambda_2 = 1.92$ dominates $\lambda_1 = .12$. $\square$

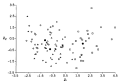


Figure 8.4. Scatter plot of discriminant function values for the football data of Table 8.1.

8.5 STEPWISE SELECTION OF VARIABLES

In many applications, a large number of dependent variables is available and the experimenter would like to discard those that are redundant (in the presence of the other variables) for separating the groups. Our discussion is limited to procedures that delete or add variables one at a time. We emphasize that we are selecting *dependent* variables (y 's), and therefore the basic model (one-way MANOVA) does not change. In subset selection in regression, on the other hand, we select *independent* variables with a consequent alteration of the model.

A *forward* selection method was discussed in Section 5.11.2. We begin with a single variable, the one that maximally separates the groups by itself. Then the variable entered at each step is the one that maximizes the partial F -statistic based on Wilks' Λ , thus obtaining the maximal additional separation of groups alone and beyond the separations already attained by the other variables. Since we choose the variable with maximum partial F at each step, the proportion of these maximum F 's that exceed F_α is greater than α . This bias is discussed in Rencher and Lauryssen (1985) and Rencher (1995, Section 5.10).

Backward elimination is a similar operation in which we begin with all the variables and then at each step, the variable that contributes least is deleted, as indicated by the partial F .

Stepwise selection is a combination of the forward and backward approaches. Variables are added one at a time, and at each step, the variables are examined to see if any variable that entered earlier has become redundant in the presence of recently added variables. The procedure stops when the largest partial F among the variables available for entry fails to exceed a preset threshold value. The stepwise procedure has long been popular with practitioners. Some detail about the steps in this procedure was given in Section 5.11.2.

All the preceding procedures are routinely referred to as *stepwise discriminant analysis*. However, as noted in Section 6.11.2, we are actually doing stepwise MANOVA. No discriminant functions are calculated in the selection process. After the subset selection is completed, we can calculate discriminant functions for the selected variables. We could also use the variables in a classification analysis, as described in Chapter 9.

Example 8.8. We use the Kodak data of Table 8.6 to illustrate the stepwise procedure outlined in this section and in Section 6.11.2. At the first step, we carry out a univariate F using ordinary ANOVA for each variable to determine which variable best separates the three groups by itself.

Variable	F	p -Value
WDEM	2.558	.0884
CRACKUM	8.246	.0000
HAETE	5.688	.0045
ETYRED	58.162	1.11×10^{-12}
ENRRED	22.427	1.40×10^{-5}
APR	4.911	.0197

Thus ETYRED is the first variable to “enter.” The Wilks Λ value equivalent to $F = 58.162$ is $\Lambda(y_1) = .4179$ (see Table 8.1) with $p = 1$. At the second step we calculate a partial Λ and accompanying partial F using (8.27) and (8.28):

$$\Lambda(y_1, y_2) = \frac{\Lambda(y_0, y_2)}{\Lambda(y_0)},$$

$$F = \frac{1 - \Lambda(y_1, y_2) / y_2 - 1}{\Lambda(y_1, y_2) / y_2},$$

where y_1 indicates the variable selected at step 1 (ETYRED) and y_2 represents each of the five variables to be examined at step 2. The results are

Variable	Partial Λ	Partial F	p -Value
WDEM	.9155	2.764	.0997
CRACKUM	.8607	38.2	.0000
HAETE	.8946	208	.0001
ENRRED	.8525	2.163	.1205
APR	.8548	2.872	.1022

The variable WDEM would enter at this step, since it has the largest partial F . With a p -value of .0997, entering this variable may be questionable, but we will continue the procedure for illustrative purposes. We must check to see if ETYRED is still significant now that WDEM has entered. The partial Λ and F for ETYRED adjusted for WDEM

are $A = .456$ and $F = 58.47$. Thus EYEDH stays in. The overall Wilks' λ for EYEDH and WDEM is $\lambda(y_1, y_2) = .4656$.

At step 3 we check each of the four remaining variables for possible entry using

$$A(y_1, y_2, y_3) = \frac{A(y_1, y_2, y_3)}{A(y_1, y_2)},$$

$$F = \frac{1 - A(y_1, y_2, y_3) \, n_g - 2}{A(y_1, y_2, y_3) \, n_g},$$

where $y_1 = \text{EYEDH}$, $y_2 = \text{WDEM}$, and y_3 represents each of the other four variables. The results are

Variable	Partial λ	Partial F	p -Value
CHCTUM	.6774	.941	.3795
FRCTYB	.6748	1.008	.3161
RAMBBI	.6282	5.290	.0441
LYR	.6181	7.794	.0006

The indicated variable for entry at this step is LYR. To determine whether one of the first two should be removed after LYR has entered, we calculate the partial A and F for each, adjusted for the other two:

Variable	Partial λ	Partial F	p -Value
WDEM	.6307	6.767	.0003
EYEDH	.4634	49.211	6.33×10^{-10}

Thus both previously entered variables remain in the model. The overall Wilks' λ for EYEDH, WDEM, and LYR is $\lambda(y_1, y_2, y_3) = .3385$.

At step 4 there are three candidate variables for entry. The partial A - and F -statistics are

$$A(y_1, y_2, y_3, y_4) = \frac{A(y_1, y_2, y_3, y_4)}{A(y_1, y_2, y_3)},$$

$$F = \frac{1 - A(y_1, y_2, y_3, y_4) \, n_g - 3}{A(y_1, y_2, y_3, y_4) \, n_g},$$

where y_1 , y_2 , and y_3 are the three variables already entered and y_4 represents each of the other three remaining variables. The results are

Variable	Partial λ	Partial F	p -Value
CHCTUM	.6987	.873	.3482
FRCTYB	.6933	.189	.6282
RAMBBI	.6080	4.207	.0373

Hence LARED enters at this step, and we check to see if any of the three previously entered variables has now become redundant. The partial A and partial F for each of these three are

Variable	Partial A	Partial F	p -Value
WIDM	.7669	11.237	4.79×10^{-12}
EYED3	.8719	28.588	3.59×10^{-17}
LQR	.8758	8.894	.0033

Consequently, all three variables are retained. The overall Wilks' A for all four variables is now $A(Y_1, Y_2, Y_3, Y_4) = .9551$.

At step 5, the partial A - and F -values are

Variable	Partial A	Partial F	p -Value
CHOCUM	.9999	.003	.9571
PRCCE	.9999	.004	.9999

Thus no more variables will enter.

We summarize the selection process as follows:

Step	Variable Entered	Overall A	Partial A	Partial F	p -Value
1	EYED3	.4379	.4379	98.162	1.13×10^{-19}
2	WIDM	.4004	.8393	2.994	.0869
3	LQR	.3583	.8431	7.791	.0068
4	LARED	.9551	.9880	4.257	.0179

□

PROBLEMS

- 8.1 Show that if $\mathbf{u} = S_{Y_1}^{-1}(\mathbf{Y}_1 - \mathbf{Y}_2)$ is substituted into $\mathbf{u}'(\mathbf{Y}_1 - \mathbf{Y}_2)I^2/\mathbf{u}'\mathbf{S}_u\mathbf{u}$, the result is (8.3).
- 8.2 Verify (8.4) for the relationship between $\mathbf{b}$ and $\mathbf{a}$.
- 8.3 Verify the relationship between R^2 and F^2 shown in (8.5).
- 8.4 Show that $\mathbf{u}'(\mathbf{Y}_1 - \mathbf{Y}_2)I^2 = \mathbf{u}'(\mathbf{Y}_1 - \mathbf{Y}_2)(\mathbf{Y}_1 - \mathbf{Y}_2)'\mathbf{u}$ as in (8.7).
- 8.5 Show that $\mathbf{H}\mathbf{a} - \lambda\mathbf{H}\mathbf{a} = \mathbf{0}$ can be written in the form $(\mathbf{H}^{-1}\mathbf{H} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}$, as in (8.12).
- 8.6 Verify (8.16) by substituting $\mathbf{a}_i' = \lambda_i\mathbf{a}_i$ into (8.15) to obtain $\lambda_1 = \mathbf{a}_1'\mathbf{y}_{1.2} + \mathbf{a}_2'\mathbf{y}_{2.2} + \cdots + \mathbf{a}_p'\mathbf{y}_{p.2} - \mathbf{a}_1'\mathbf{y}_1$.
- 8.7 For the psychological data in Table 2.1, the discriminant function coefficient vector was given in Example 5.5.

- (a) Find the standardized coefficients.
- (b) Calculate t -tests for the individual variables.
- (c) Compare the results of (a) and (b) as to the contribution of the variables to separation of the two groups.
- (d) Find the partial F for each variable, as in (8.26), and compare with the standardized coefficients.

8.8 Using the beetle data of Table 5.5, do the following:

- (a) Find the discriminant function coefficient vector.
- (b) Find the standardized coefficients.
- (c) Calculate t -tests for individual variables.
- (d) Compare the results of (b) and (c) with the contribution of each variable to separation of the groups.
- (e) Find the partial F for each variable, as in (8.26). Do the partial F 's rank the variables in the same order of importance as the standardized coefficients?

8.9 Using the dystrophy data of Table 5.1, do the following:

- (a) Find the discriminant function coefficient vector.
- (b) Find the standardized coefficients.
- (c) Calculate t -tests for individual variables.
- (d) Compare the results of (b) and (c) with the contribution of each variable to separation of the groups.
- (e) Find the partial F for each variable, as in (8.26). Do the partial F 's rank the variables in the same order of importance as the standardized coefficients?

8.10 For the eyelid data of Table 5.8, do the following:

- (a) Find the discriminant function coefficient vector.
- (b) Find the standardized coefficients.
- (c) Calculate t -tests for individual variables.
- (d) Compare the results of (b) and (c) with the contribution of each variable to separation of the groups.
- (e) Find the partial F for each variable, as in (8.26). Do the partial F 's rank the variables in the same order of importance as the standardized coefficients?

8.11 Using the fish data in Table 6.17, do the following:

- (a) Find the eigenvectors of $\mathbf{B}^{-1}\mathbf{B}$.
- (b) Carry out tests of significance for the discriminant functions and find the relative importance of each as in (8.18), $\lambda_j / \sum_{j=1}^p \lambda_j$. Do these two procedures agree as to the number of important discriminant functions?
- (c) Find the standardized coefficients and comment on the contribution of the variables to separation of groups.
- (d) Find the partial F for each variable, as in (8.26). Do they rank the variables in the same order as the standardized coefficients for the first discriminant function?

- (ii) Plot the first two discriminant functions for each observation and for the mean vectors.

8.12 For the roostock data of Table 6.2, do the following:

- (a) Find the eigenvalues and eigenvectors of $\mathbf{B}^{-1}\mathbf{B}$.
- (b) Carry out tests of significance for the discriminant functions and find the relative importance of each, as in (8.13), $k_i / \sum_j k_j$. Do these two procedures agree with the number of important discriminant functions?
- (c) Find the standardized coefficients and comment on the contribution of the variables to separation of groups.
- (d) Find the partial F for each variable, as in (8.28). Do they rank the variables in the same order as the standardized coefficients for the first discriminant function?
- (e) Plot the first two discriminant functions for each observation and for the mean vectors.

8.13 Carry out a stepwise selection of variables on the roostock data of Table 6.2.

8.14 Carry out a stepwise selection of variables on the engineot data of Table 6.6.

8.15 Carry out a stepwise selection of variables on the fish data of Table 6.13.

Classification Analysis: Allocation of Observations to Groups

9.1 INTRODUCTION

The *descriptive* aspect of discriminant analysis, in which group separations are ascertained by means of discriminant functions, was covered in Chapter 8. We turn now to allocation of observations to groups, which is the *predictive* aspect of discriminant analysis. We prefer to call this classification analysis to clearly distinguish it from the descriptive aspect. However, classification is often referred to simply as discriminant analysis. In engineering and computer science, classification is usually called pattern recognition. Some writers use the term classification analysis to describe cluster analysis, in which the observations are clustered according to variable values rather than into predefined groups (see Chapter 14).

In classification, a sampling unit (subject or object) whose group membership is unknown is assigned to a group on the basis of the vector of p measured values, $\mathbf{y}$, associated with the unit. To classify the unit, we must have available a previously obtained sample of observation vectors from each group. Then one approach is to compare $\mathbf{y}$ with the mean vectors $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \dots, \bar{\mathbf{y}}_k$ of the k samples and assign the unit to the group whose $\bar{\mathbf{y}}_i$ is closest to $\mathbf{y}$.

In this chapter, the term groups may refer to either the k samples or the k populations from which they were taken. It should be clear from the context which of the two uses is intended in every case.

We give some examples to illustrate the classification technique.

1. A university admissions committee wants to classify applicants as likely to succeed or likely to fail. The variables available are high school grades in various subject areas, standardized test scores, rating of high school, number of advanced placement courses, etc.
2. A psychiatrist gives a battery of diagnostic tests in order to assign a patient to the appropriate mental illness category.
3. A college student takes aptitude and interest tests in order to determine which vocational available or her profile best matches.

4. African, or "killer," bees cannot be distinguished visually from ordinary domestic honey bees. The variables based on chromatograph peaks can be used to readily identify them (Lawrie and Carlson 1987).
5. The Air Force wishes to classify each applicant into the training program where he or she has the most potential.
6. Timbers of the Federalist Papers were claimed by both Madison and Hamilton. Can we identify authorship by measuring frequencies of word usage (Mosteller and Wallace 1964)?
7. Variables such as availability of fingerprints, availability of eye witnesses, and time until police arrive can be used to classify burglaries into solvable and unsolvable.
8. One approach to speech recognition by computer consists of an attempt to identify phonemes based on the energy levels in speech waves.
9. A number of variables are measured at five weather stations. Based on these variables, we wish to predict the cooling at a particular airport in 2 hours. The cooling categories are closed, low instrument, high instrument, low open, and high open (Ladenheusch 1978, p. 2).

9.2 CLASSIFICATION INTO TWO GROUPS

In the case of two populations, we have a sampling unit (subject or object) to be classified into one of two populations. The information we have available consists of the p variables in the observation vector $\mathbf{y}$ measured on the sampling unit. In the first illustration in Section 9.1, for example, we have an applicant with high school grades and various test scores recorded in $\mathbf{y}$. We do not know if the applicant will succeed or fail at the university, but we have data on previous students at the university for whom it is now known whether they succeeded or failed. By comparing $\mathbf{y}$ with $\bar{\mathbf{y}}_1$ for those who succeeded and $\bar{\mathbf{y}}_2$ for those who failed, we attempt to predict the group to which the applicant will eventually belong.

When there are two populations, we can use a classification procedure due to Fisher (1936). The principal assumption for Fisher's procedure is that the two populations have the same covariance matrix ($\Sigma_1 = \Sigma_2$). Normality is not required. We obtain a sample from each of the two populations and compute $\bar{\mathbf{y}}_1$, $\bar{\mathbf{y}}_2$, and $\mathbf{S}_p$. A simple procedure for classification can be based on the discriminant function,

$$z = \mathbf{w}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_p^{-1}\mathbf{y} \quad (9.1)$$

(see Sections 5.3, 5.6, 8.2, and 8.5), where $\mathbf{y}$ is the vector of measurements on a new sampling unit that we wish to classify into one of the two groups (populations). For convenience we speak of classifying $\mathbf{y}$ rather than classifying the subject or object associated with $\mathbf{y}$.

To determine whether $\mathbf{y}$ is closer to $\bar{\mathbf{y}}_1$ or $\bar{\mathbf{y}}_2$, we check to see if z in (9.1) is closer to the transformed mean $\bar{z}_1$ or to $\bar{z}_2$. We evaluate (9.1) for each observation

$\mathbf{p}_1$, from the first sample and about $\mathbf{z}_{10}, \mathbf{z}_{11}, \dots, \mathbf{z}_{1n_1}$, from which, by (3.54), $\bar{\mathbf{z}}_1 = \sum_{i=1}^{n_1} \mathbf{z}_{1i}/n_1 = \mathbf{w}'\bar{\mathbf{y}}_1 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2/\mathbf{S}_p^{-1}\bar{\mathbf{y}}_2)$. Similarly, $\bar{\mathbf{z}}_2 = \mathbf{w}'\bar{\mathbf{y}}_2$. Denote the two groups by G_1 and G_2 . Fisher's (1936) linear classification procedure assigns $\mathbf{z}$ to G_1 if $\mathbf{z} = \mathbf{w}'\mathbf{y}$ is closer to $\bar{\mathbf{z}}_1$ than to $\bar{\mathbf{z}}_2$ and assigns $\mathbf{z}$ to G_2 if $\mathbf{z}$ is closer to $\bar{\mathbf{z}}_2$. This is illustrated in Figure 8.1.

For the configuration in Figure 8.1, we see that $\mathbf{z}$ is closer to $\bar{\mathbf{z}}_1$ if

$$\mathbf{z} > \frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2). \quad (8.2)$$

This is true in general because $\bar{\mathbf{z}}_1$ is always greater than $\bar{\mathbf{z}}_2$, which can easily be shown as follows:

$$\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2 = \mathbf{w}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2/\mathbf{S}_p^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)) > 0, \quad (8.3)$$

because $\mathbf{S}_p^{-1}$ is positive definite. Thus $\bar{\mathbf{z}}_1 > \bar{\mathbf{z}}_2$. [It is more of the form $\mathbf{w}' = (\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}_1/\mathbf{S}_p^{-1})$, then $\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2$ would be positive.] Since $\frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2)$ is the midpoint, $\mathbf{z} > \frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2)$ implies that $\mathbf{z}$ is closer to $\bar{\mathbf{z}}_1$. By (8.2) the distance from $\bar{\mathbf{z}}_1$ to $\bar{\mathbf{z}}_2$ is the same as that from $\bar{\mathbf{y}}_1$ to $\bar{\mathbf{y}}_2$.

To express the classification rule in terms of $\mathbf{y}$, we first write $\frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2)$ in the form

$$\frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2) = \frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2/\mathbf{S}_p^{-1}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)). \quad (8.4)$$

Then the classification rule becomes: Assign $\mathbf{y}$ to G_1 if

$$\mathbf{w}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2/\mathbf{S}_p^{-1})\mathbf{y} > \frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2/\mathbf{S}_p^{-1})(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2). \quad (8.5)$$



Figure 8.1. Fisher's procedure for classification into two groups.

and assign y to G_1 if

$$a'y = \bar{Q}_1 - \bar{S}_1/\bar{S}_D^2 > -\frac{1}{2}(\bar{Q}_1 - \bar{S}_1/\bar{S}_D^2)(\bar{Q}_1 + \bar{Q}_2). \quad (9.6)$$

This linear classification rule employs the same discriminant function $z = a'y$ used in Section 8.7 in connection with descriptive equations of groups. Thus in the two-group case, the discriminant function serves as a linear classification function as well. However, in the several-group case in Section 8.3, we use classification functions that are different from the descriptive discriminant functions in Section 8.4.

Fisher's CPD approach using (9.5) and (9.6) is essentially nonparametric because no distributional assumptions were made. However, if the two populations are normal with equal covariance matrices, then this method is (asymptotically) optimal; that is, the probability of misclassification is minimized [see comments following (9.8)].

If prior probabilities p_1 and p_2 are known for the two populations, the classification rule can be modified to exploit this additional information. We define the prior probabilities as follows: p_1 is the proportion of observations in G_1 and p_2 is the proportion in G_2 , where $p_2 = 1 - p_1$. For example, suppose that at a certain university 70% of entering freshmen ultimately graduate. Then $p_1 = .7$ and $p_2 = .3$.

In order to use the prior probabilities, the density functions for the two populations, $f(y|G_1)$ and $f(y|G_2)$, must also be known. Then the optimal classification rule (Wald, 1950) that minimizes the probability of misclassification is: Assign y to G_1 if

$$p_1 f(y|G_1) > p_2 f(y|G_2) \quad (9.7)$$

and to G_2 otherwise. Note that $f(y|G_1)$ is a convenient notation for the density when sampling from the population represented by G_1 . It does not represent a conditional distribution in the usual sense (Section 4.2).

Assuming that the two densities are multivariate normal with equal covariance matrices, namely, $f(y|G_1) = N_p(\mu_1, \Sigma)$ and $f(y|G_2) = N_p(\mu_2, \Sigma)$, then from (9.7) we obtain the following rule (with estimates in place of μ_1 , μ_2 , and Σ): Assign y to G_1 if

$$a'y = \bar{Q}_1 - \bar{S}_1/\bar{S}_D^2 > -\frac{1}{2}(\bar{Q}_1 - \bar{S}_1/\bar{S}_D^2)(\bar{Q}_1 + \bar{Q}_2) + \ln\left(\frac{p_1}{p_2}\right) \quad (9.8)$$

and to G_2 otherwise [see Kancher (1998, p. 285)]. Because we have substituted estimates for the parameters, the rule in (9.8) is no longer optimal, as in (9.7). However, it is asymptotically optimal (approaches optimality as the sample size increases).

If $p_1 = p_2$, the normal-based classification rule in (9.8) becomes the same as Fisher's procedure given in (9.5) and (9.6). Thus Fisher's rule, which is not based on a normality assumption, has optimal properties when the data come from multivariate normal populations with $\Sigma_1 = \Sigma_2$ and $p_1 = p_2$. [For the case when $\Sigma_1 \neq \Sigma_2$, see Kancher (1998, Section 8.2.2).] Hence, even though Fisher's method is nonparametric, it works better for normally distributed populations or other populations with

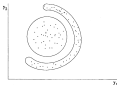


Figure 5.2. Two populations with nonlinear separation.

linear trends. For example, suppose two populations have 550 contours, as in Figure 5.2. If the points are projected in any direction onto a straight line, there will be almost total overlap. A linear discriminant procedure will not successfully separate the two populations.

Example 5.2. For the psychological data of Table 5.1, $\bar{y}_1$, $\bar{y}_2$, and $\bar{S}_p$ were obtained in Example 5.4.1. The discriminant function coefficients were obtained in Example 5.5 as $\mathbf{a}' = (.5154, -.2032, .4660, -.5897)$. For G_1 (the male group), we find

$$\begin{aligned}T_1 &= \mathbf{a}'\bar{\mathbf{y}}_1 = .5154(15.87) - .2032(15.81) + .4660(17.38) - .5897(12.79) \\&= 10.5427.\end{aligned}$$

Similarly, for G_2 (the female group), $T_2 = \mathbf{a}'\bar{\mathbf{y}}_2 = 4.4426$. Thus we assign an observation vector $\mathbf{y}$ to G_1 if

$$z = \mathbf{a}'\mathbf{y} = \frac{1}{2}(T_1 + T_2) = 7.4927$$

and assign $\mathbf{y}$ to G_2 if $z < 7.4927$.

There are no new observations available, so we will illustrate the procedure by classifying two of the observations in G_1 . For $\mathbf{y}_{11}' = (15, 17, 16, 14)$, the first observation in G_1 , we have $z_{11} = \mathbf{a}'\mathbf{y}_{11} = .5154(15) - .2032(17) + .4660(16) - .5897(14) = 11.0498$, which is greater than 7.4927, and $\mathbf{y}_{11}$ would be correctly classified as belonging to G_1 . For $\mathbf{y}_{14}' = (15, 12, 15, 18)$, the fourth observation in G_1 , we find $z_{14} = 3.9835$, which would misclassify $\mathbf{y}_{14}$ into G_2 . $\square$

9.3 CLASSIFICATION INTO SEVERAL GROUPS

In this section we discuss classification rules for several groups. As in the two-group case, we use a sample from each of the k groups to find the sample mean vectors $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \dots, \bar{\mathbf{y}}_k$. For a vector $\mathbf{y}$ whose group membership is unknown, one approach is to use a distance function to find the mean vector that $\mathbf{y}$ is closest to and assign $\mathbf{y}$ to the corresponding group.

9.3.1 Equal Population Covariance Matrices: Linear Classification Functions

In this section, we assume $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k$. We can estimate the common population covariance matrix by a pooled sample covariance matrix

$$\mathbf{S}_0 = \frac{1}{N - k} \sum_{i=1}^k (n_i - 1)\mathbf{S}_i = \frac{\mathbf{E}}{N - k},$$

where n_i and $\mathbf{S}_i$ are the sample size and covariance matrix of the i th group, $\mathbf{E}$ is the error matrix from one-way MANOVA, and $N = \sum_{i=1}^k n_i$. We compare $\mathbf{y}$ to each $\bar{\mathbf{y}}_i$, $i = 1, 2, \dots, k$, by the distance function

$$D_i^2(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}}_i)' \mathbf{S}_0^{-1} (\mathbf{y} - \bar{\mathbf{y}}_i) \quad (9.9)$$

and assign $\mathbf{y}$ to the group for which $D_i^2(\mathbf{y})$ is smallest.

We can obtain a linear classification rule by expanding (9.9):

$$\begin{aligned} D_i^2(\mathbf{y}) &= \mathbf{y}' \mathbf{S}_0^{-1} \mathbf{y} - \mathbf{y}' \mathbf{S}_0^{-1} \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_i' \mathbf{S}_0^{-1} \mathbf{y} + \bar{\mathbf{y}}_i' \mathbf{S}_0^{-1} \bar{\mathbf{y}}_i \\ &= \mathbf{y}' \mathbf{S}_0^{-1} \mathbf{y} - 2\bar{\mathbf{y}}_i' \mathbf{S}_0^{-1} \mathbf{y} + \bar{\mathbf{y}}_i' \mathbf{S}_0^{-1} \bar{\mathbf{y}}_i. \end{aligned}$$

The term $\mathbf{y}' \mathbf{S}_0^{-1} \mathbf{y}$ on the right can be neglected since it is not a function of i and, consequently, does not change from group to group. The second term is a linear function of $\mathbf{y}$, and the third does not involve $\mathbf{y}$. We thus delete $\mathbf{y}' \mathbf{S}_0^{-1} \mathbf{y}$ and obtain a linear classification function, which we denote by $L_i(\mathbf{y})$. If we multiply by $-\frac{1}{2}$ to agree with the rule based on the normal distribution and prior probabilities given in (8.12), our linear classification rule becomes: Assign $\mathbf{y}$ to the group for which

$$L_i(\mathbf{y}) = \bar{\mathbf{y}}_i' \mathbf{S}_0^{-1} \mathbf{y} - \frac{1}{2} \bar{\mathbf{y}}_i' \mathbf{S}_0^{-1} \bar{\mathbf{y}}_i, \quad i = 1, 2, \dots, k \quad (9.10)$$

is a maximum (we reversed the sign when multiplying by $-\frac{1}{2}$). To highlight the linearity of (9.10) as a function of $\mathbf{y}$, we can express it as

$$L_i(\mathbf{y}) = c_1^i \mathbf{y} + c_{10} = c_{11}^i y_1 + c_{12}^i y_2 + \dots + c_{1p}^i y_p + c_{10},$$

where $c_1^i = \bar{\mathbf{y}}_i' \mathbf{S}_0^{-1}$ and $c_{10} = -\frac{1}{2} \bar{\mathbf{y}}_i' \mathbf{S}_0^{-1} \bar{\mathbf{y}}_i$. To assign $\mathbf{y}$ to a group using this procedure, we calculate c_1^i and c_{10} for each of the k groups, evaluate $L_i(\mathbf{y})$, $i = 1, 2, \dots, k$, and

assign $\mathbf{y}$ to the group for which $L_i(\mathbf{y})$ is largest. This will be the same group for which $L_i^0(\mathbf{y})$ in (9.9) is smallest, that is, the group whose mean vector $\bar{\mathbf{y}}_i$ is closest to $\mathbf{y}$.

For the case of several groups, the optimal rule in (9.7) amounts to

$$\text{Assign } \mathbf{y} \text{ to the group for which } p_i f(\mathbf{y}|\Sigma_i) \text{ is maximum.} \quad (9.10)$$

With this rule, the probability of misclassification is minimized. If we assume normality with equal covariance matrices and with prior probabilities of group membership, $p_1, p_2, \dots, p_k$, then $f(\mathbf{y}|\Sigma_i) = N_{\mathbf{y}}(\mathbf{y}|\mu_i, \Sigma)$, and the rule in (9.10) becomes (with estimates in place of parameters) Calculate

$$L_i(\mathbf{y}) = \ln p_i + \mathbf{y}' \mathbf{R}_i^{-1} \mathbf{y} = \frac{1}{2} \mathbf{y}' \mathbf{R}_i^{-1} \mathbf{y}, \quad i = 1, 2, \dots, k \quad (9.12)$$

and assign $\mathbf{y}$ to the group with maximum value of $L_i(\mathbf{y})$. Note that if $p_1 = p_2 = \dots = p_k$, then (9.12), which optimizes the classification rate for the normal distribution, reduces to (9.10), which was based on the heuristic approach of minimizing the distance of $\mathbf{y}$ to $\bar{\mathbf{y}}_i$.

The linear functions $L_i(\mathbf{y})$ defined in (9.12) are called *linear classification functions* (many writers refer to them as *linear discriminant functions*). They are different from the linear discriminant functions in Sections 8.1.4, 8.4, and 8.4.1, whose coefficients are eigenvectors of $\mathbf{K}^{-1}\mathbf{B}$. In fact, there will be k classification functions and $r = \min(k, d - 1)$ discriminant functions, where k is the number of groups and p is the number of variables. In many cases we do not need all r discriminant functions to effectively detect the group differences, whereas all k classification functions must be used in assigning observations to groups.

Example 9.3.1. For the fueloil data of Table 8.3, the mean vectors for the three groups are as follows:

$$\bar{\mathbf{y}}_1 = (25.2, 26.0, 26.1, 15.1, 14.5, 12.3),$$

$$\bar{\mathbf{y}}_2 = (25.6, 27.6, 26.8, 15.1, 15.5, 11.8),$$

$$\bar{\mathbf{y}}_3 = (25.6, 27.8, 26.8, 15.8, 15.5, 11.8).$$

Using these values of $\bar{\mathbf{y}}_i$ and the pooled covariance matrix $\mathbf{R}_{p0}$ given in Example 8.5, the linear classification functions (9.12) become

$$L_1(\mathbf{y}) = 7.6y_0 + 13.3y_1 + 4.7y_2 - 1.2y_3 + 16.6y_4 + 8.1y_5 - 641.1,$$

$$L_2(\mathbf{y}) = 10.2y_1 + 13.3y_2 + 4.7y_3 - 3.6y_4 + 18.2y_5 + 6.1y_6 - 808.6,$$

$$L_3(\mathbf{y}) = 10.8y_1 + 13.3y_2 + 4.1y_3 - 2.7y_4 + 18.1y_5 + 5.2y_6 - 816.6.$$

We note that y_2 and y_3 have essentially the same coefficients in all three functions and hence do not contribute to classification of $\mathbf{y}$. These same two variables were eliminated in the stepwise discriminant analysis in Example 8.9.

We illustrate the use of these linear functions for the first and third observations in group 1. For the first observation, y_{11} , we obtain

$$\begin{aligned}L_1(y_{11}) &= 3.6(13.5) + 13.3(57.2) + 4.3(18.5) - 1.1(12.5) + 14.6(14.0) \\&\quad + 8.2(11.0) - 646.1 = 582.124, \\L_2(y_{11}) &= 10.2(13.5) + 13.3(57.2) + 4.1(19.5) - 3.4(12.5) + 15.3(14.0) \\&\quad + 6.1(11.0) - 608.0 = 579.099, \\L_3(y_{11}) &= 10.9(13.5) + 13.3(57.2) + 4.1(19.5) - 2.7(12.5) + 15.3(14.0) \\&\quad + 5.2(11.0) - 614.6 = 578.768.\end{aligned}$$

We classify y_{11} into group 1 since $L_1(y_{11}) = 582.1$ exceeds $L_2(y_{11})$ and $L_3(y_{11})$. For the third observation in group 1, y_{13} , we obtain

$$L_1(y_{13}) = 581.054, \quad L_2(y_{13}) = 579.290, \quad L_3(y_{13}) = 588.121.$$

This observation is misclassified into group 2 since $L_2(y_{13}) = 579.290$ exceeds $L_1(y_{13})$ and $L_3(y_{13})$. $\square$

9.3.1 Unequal Population Covariance Matrices: Quadratic Classification Functions

The linear classification functions in Section 9.3.1 are based on the assumption $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$. The resulting classification rules are sensitive to heterogeneity of covariance matrices. Observations tend to be classified too frequently into groups whose covariance matrices have larger variances on the diagonal. Thus the population covariance matrices should not be assumed to be equal if there is reason to suspect otherwise.

If $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ does not hold, the classification rules can easily be altered to preserve optimality of classification rates. In place of (9.9), we consider

$$D_i^2(y) = y - \mathbf{f}_i'(\mathbf{S}_i^{-1}(y - \mathbf{f}_i)), \quad i = 1, 2, \dots, k, \quad (9.12)$$

where $\mathbf{S}_i$ is the sample covariance matrix for the i th group. As before, we would assign y to the group for which $D_i^2(y)$ is smallest. With $\mathbf{S}_i$ in place of $\mathbf{S}_{kp}$, (9.12) cannot be reduced to a linear function of y as in (9.10) but remains a quadratic function. Hence rules based on $\mathbf{S}_i$ are called *quadratic classification rules*.

If we assume normality with unequal covariance matrices and with prior probabilities $p_1, p_2, \dots, p_k$, then $f(y|\mathbf{G}_i) = K_i(p_i, \mathbf{S}_i)$, and the optimal rule in (9.11) based on $p_i f(y|\mathbf{G}_i)$ becomes: Assign y to the group for which

$$Q_i(y) = \ln p_i - \frac{1}{2} \ln |\mathbf{S}_i| - \frac{1}{2} (y - \mathbf{f}_i'(\mathbf{S}_i^{-1}(y - \mathbf{f}_i)) \quad (9.14)$$

is maximum. If $p_1 = p_2 = \cdots = p_k$ (or if the p_i 's are unknown, the term $\ln p_i$ is deleted).

In order to use a quadratic classification rule based on $\mathbf{S}_i$, each n_i must be greater than p so that $\mathbf{S}_i^{-1}$ will exist. This restriction does not apply to linear classification rules based on $\mathbf{S}_{ij}$. Once more parameters are estimated with quadratic classification functions, larger values of the n_i 's are needed for stability of estimates. Note the distinction between p , the number of variables, and p_i , the prior probability for the i th group.

5.4 ESTIMATING MISCLASSIFICATION RATES

In Chapter 3, we assessed the effectiveness of the discriminant functions in group separation by the use of significance tests or by examining $\lambda_1/\sum_{i=1}^p \lambda_i$. To judge the ability of classification procedures to predict group membership, we usually use the probability of misclassification, which is known as the error rate. We could also use its complement, the correct classification rate.

A simple estimate of the error rate can be obtained by trying out the classification procedure on the same data set that has been used to compute the classification functions. This method is commonly referred to as *resubstitution*. Each observation vector $\mathbf{y}_{ij}$ is substituted in the classification functions and assigned to a group. We then count the number of correct classifications and the number of misclassifications. The proportion of misclassifications resulting from resubstitution is called the *apparent error rate*. The results can be conveniently displayed in a classification table or confusion matrix, such as Table 5.1 for two groups.

Among the n_1 observations in G_1 , n_{11} are correctly classified into G_1 , and n_{12} are misclassified into G_2 , where $n_1 = n_{11} + n_{12}$. Similarly, of the n_2 observations in G_2 , n_{21} are misclassified into G_1 , and n_{22} are correctly classified into G_2 , where $n_2 = n_{21} + n_{22}$. Thus

$$\begin{aligned}\text{Apparent error rate} &= \frac{n_{12} + n_{21}}{n_1 + n_2} \\ &= \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}.\end{aligned}\quad (5.15)$$

Similarly, we can define

$$\text{Apparent correct classification rate} = \frac{n_{11} + n_{22}}{n_1 + n_2}, \quad (5.16)$$

Table 5.1. Classification Table for Two Groups

Actual Group	Number of Observations	Predicted Group	
		1	2
1	n_1	n_{11}	n_{12}
2	n_2	n_{21}	n_{22}

Table 9.1. Classification Table for the Psychological Data of Table 8.1

Actual Group	Number of Observations	Predicted Group	
		1	2
Male	32	28	4
Female	32	4	28

Clearly,

$$\text{Apparent error rate} = 1 - \text{apparent correct classification rate}.$$

The method of resubstitution can be readily extended to the case of several groups.

The apparent error rate is easily obtained and is routinely provided by most classification software programs. It is an estimate of the probability that our classification functions based on the present sample will misclassify a future observation. This probability is called the *actual error rate*. Unfortunately, the apparent error rate underestimates the actual error rate because the data set used to compute the classification functions is also used to evaluate them. The classification functions are optimized for the particular sample and may be capitalizing on chance to some degree, especially for small samples. For other estimates of error rates, see Rencher (1995, Section 6.4). In Section 9.5 we consider some approaches to reducing the bias in the apparent error rate.

Example 9.4(a). We use the psychological data of Table 8.1 to illustrate the apparent error rate obtained by the resubstitution method for two groups. The hypothesis $H_0: \Sigma_1 = \Sigma_2$ was not rejected in Example 7.3.1, and we therefore classify each of the 64 observations using the linear classification procedure obtained in Example 9.2. Classify as G_1 if $\hat{y} = 1.4921$ and as G_2 otherwise. The resulting classification table is given in Table 9.1. By (9-15),

$$\text{Apparent error rate} = \frac{n_{21} + n_{12}}{n_1 + n_2} = \frac{4 + 4}{32 + 32} = .125. \quad \square$$

Example 9.4(b). We use the football data of Table 8.3 to illustrate the use of the resubstitution method for estimating the error rate in the case of more than two groups. The sample covariance matrices for the three groups are almost significantly different, and we will use both linear and quadratic classification functions.

The linear classification functions $J_1(x)$ from (9-18) were given in Example 9.3.1 for the football data. Using these, we classify each of the 99 observations. The results are shown in Table 9.2.

An examination of this data set in Example 8.2 showed that groups 2 and 3 are harder to separate than 1 and 2 or 1 and 3. This pattern is reflected here in the misclas-

Table 8.3. Classification Table for the Football Data of Table 8.2 Using Linear Classification Functions

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	30	26	1	3
2	30	1	26	3
3	30	2	8	20

$$\text{Apparent correct classification rate} = \frac{26 + 26 + 20}{90} = .833$$

$$\text{Apparent error rate} = 1 - .833 = .167$$

Table 8.4. Classification Table for the Football Data of Table 8.2 Using Quadratic Classification Functions

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	30	27	1	2
2	30	2	26	2
3	30	1	4	25

$$\text{Apparent correct classification rate} = \frac{27 + 26 + 25}{90} = .889$$

$$\text{Apparent error rate} = 1 - .889 = .111$$

sifications. Only 4 of the observation vectors in group 1 are misclassified, whereas 16 observations in each of groups 2 and 3 are misclassified.

Using the quadratic classification functions $\hat{Q}(\mathbf{x})$, $i = 1, 2, 3$, in (8.14) and assuming $\mu_1 = \mu_2 = \mu_3$, we obtain the classification results in Table 8.4. There is some improvement in the apparent error rate using quadratic classification functions. $\square$

8.3 IMPROVED ESTIMATES OF ERROR RATES

For large samples, the apparent error rate has only a small amount of bias for estimating the actual error rate and can be used with little concern. For small samples, however, it is overly optimistic (biased downwards), as noted before. We discuss two techniques for reducing the bias in the apparent error rate, that is, increasing the apparent error rate to a more realistic level.

9.5.1 Partitioning the Sample

One way to avoid bias is to split the sample into two parts, a training sample used to construct the classification rule and a validation sample used to evaluate it. With the training sample, we calculate linear or quadratic classification functions. We then submit each observation vector in the validation sample to the classification functions obtained from the training sample. Since these observations are not used in calculating the classification functions, the resulting error rate is unbiased. To increase the information gained, we could also reverse the roles of the two samples so that the classification functions are obtained from the validation sample and evaluated on the training sample. The two estimates of error could then be averaged.

Partitioning the sample has at least two disadvantages:

1. It requires large samples that may not be available.
2. It does not evaluate the classification functions we will use in practice. The estimate of error based on half the sample may vary considerably from that based on the entire sample. We prefer to use all or almost all the data to construct the classification functions so as to minimize the variance of our error rate estimate.

9.5.2 Holdout Method

The holdout method is an improved version of the sample-splitting procedure in Section 9.5.1. In the holdout procedure, all but one observation is used to compute the classification rule, and this rule is then used to classify the omitted observation. We repeat this procedure for each observation, so that, in a sample of size $N = \sum_{i=1}^J n_i$, each observation is classified by a function based on the other $N - 1$ observations. The computational load is increased because N distinct classification procedures have to be constructed. The holdout procedure is also referred to as the *leave-one-out* method or as *cross validation*. Note that this procedure is used to estimate error rates. The actual classification rule for future observations would be based on all N observations.

Example 9.5.1. We use the football data of Table 9.3 to illustrate the holdout method for estimating the error rate. Each of the 50 observations is classified by linear classification functions based on the other 49 observations. To begin the procedure, the first observation in group 1 ($\mathbf{p}_{11}$) is held out and the linear classification functions $L_i(\mathbf{p})$, $i = 1, 2, 3$, in (9.18) are calculated using the remaining 49 observations in group 1 and the 50 observations in groups 2 and 3. The observation $\mathbf{p}_{11}$ is now classified using $L_1(\mathbf{p})$, $L_2(\mathbf{p})$, and $L_3(\mathbf{p})$. Then $\mathbf{p}_{11}$ is reinserted in group 1, and $\mathbf{p}_{12}$ is held out. The functions $L_i(\mathbf{p})$, $L_j(\mathbf{p})$, and $L_k(\mathbf{p})$ are recomputed, and $\mathbf{p}_{12}$ is then classified. This procedure is followed for each of the 50 observations, and the results are in Table 9.5.

Table 8.8. Classification Table for the Football Data of Table 8.3 Using the Wilcoxon Method Based on Linear Classification Functions

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	30	26	0	4
2	30	1	18	11
3	30	2	9	19

$$\text{Correct classification rate} = \frac{26 + 18 + 19}{90} = .700$$

$$\text{Error rate} = 1 - .700 = .300$$

As expected, the incident error rate has increased somewhat from the apparent error rate based on resubstitution in Tables 8.3 and 8.4 in Example 8.4(b). An error rate of .300 is a (less optimistic, more realistic) estimate of what the classification functions can do with future samples. (2)

8.6 SUBJECT SELECTION

The experimenter often has available a large number of variables and wishes to keep any that might aid in predicting group membership but, at the same time, to delete any superfluous variables that do not contribute to allocations. A reduction in the number of redundant variables may in fact lead to improved error rates. As an additional consideration, there is an increase in robustness to nonnormality of linear and quadratic classification functions as p (the number of variables) decreases.

The majority of selection schemes for classification analysis are based on stepwise discriminant analysis or a similar approach (Section 8.5). One finds the subset of variables that best separates groups using Wilks' Λ , for example, and then uses those variables to construct classification functions. Most of the major statistical software packages offer this method. When the "best" subset is selected in this way, an optimistic bias in error rates is introduced. For a discussion of this bias, see Rencher (1995a, 1998, Section 8.7).

Another link between separation and classification is the use of error rates in an informal stopping rule in a stepwise discriminant analysis. Thus, for example, if a subset of 5 variables out of 18 gives a misclassification rate of 35% compared to 50% for the full set of variables, we may decide that the 5 variables are adequate for separating the groups. We could try several subsets of decreasing sizes to see when the error rate begins to escalate noticeably.

Example 8.6(a). In Example 8.5, a stepwise discriminant analysis based on a partial Wilks' Λ (or partial F') was carried out for the football data of Table 8.3. Four variables were selected: CYED, WDBI, JGR, and LARBI. These same four vari-

Table 9.6. Classification Table for the Football Data of Table 9.3 Using Linear Classification Functions Based on Four Variables Chosen by Stepwise Selection

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	30	28	1	1
2	30	1	28	1
3	30	2	8	20

ables are indicated by the coefficients in the linear classification functions in Example 9.3.1. We now use these four variables to classify the observations using the method of substitution to obtain the apparent error rate.

The linear classification functions (9.10) are

$$\begin{aligned}\text{Group 1: } L_1(\mathbf{y}) &= 5(\mathbf{S}_1^{-1}\mathbf{y} - \frac{1}{2}\mathbf{S}_1^{-1}\bar{\mathbf{y}}_1) \\ &= 39.67y_1 + 4.13y_2 + 17.67y_3 + 28.44y_4 - 425.90, \\ \text{Group 2: } L_2(\mathbf{y}) &= 21.12y_1 + 1.86y_2 + 16.24y_3 + 18.36y_4 - 362.15, \\ \text{Group 3: } L_3(\mathbf{y}) &= 21.87y_1 + 2.67y_2 + 16.12y_3 + 12.46y_4 - 369.03.\end{aligned}$$

When each observation vector is classified using these linear functions, we obtain the classification results in Table 9.6.

Table 9.6 is identical to Table 9.5 in Example 9.4.6a, where all six variables were used. Thus the four selected variables can classify the sample as well as all six variables classify it. $\square$

Example 9.4.6b. We illustrate the use of error rates as an informal stopping rule in a stepwise discriminant analysis. Fifteen teacher and pupil behaviors were observed during 5-min intervals of reading instruction in elementary school classrooms (Kendall, Wachman, and Young 1970). The observations were recorded in rate of occurrences per minute for each variable. The variables were the following:

Teacher Behaviors

1. *Explains*—Explains task to learner.
2. *Models*—Models the task response for the learner.
3. *Questions*—Asks a question to elicit a task response.
4. *Directs*—Gives a direct signal to elicit a task response.
5. *Controls*—Controls management behavior with direction statements or gestures.
6. *Positive*—Gives a positive (affirmative) statement or gesture.
7. *Negative*—Gives a negative statement or gesture.

Pupil Behaviors

- 8. *Overl delayed*—An overl learner response to task signals that cannot be judged correct or incorrect until later.
- 9. *Correct*—A correct learner response with relationship to task signals.
- 10. *Incorrect*—An incorrect learner response with relationship to task signals.
- 11. *No response*—Learner gives no response with relationship to task signals.
- 12. *Ask*—Learner asks a question about the task.
- 13. *Statement*—Learner gives a positive statement or gesture not related to the task.
- 14. *Inappropriate*—Learner gives inappropriate management behavior.
- 15. *Other*—Other learner than one being observed gives responses as teacher directs task signals.

The teachers were grouped into four categories:

Group 1: Outstanding teachers,

Group 2: Poor teachers,

Group 3: First-year teachers,

Group 4: Teacher aides.

The sample sizes in groups 1–4 were 61, 61, 31, and 41, respectively. Because of the large values of R and p ($R = 121$, $p = 121$), the data are not given here.

The stepwise discriminant analysis was run several times with different threshold F -to-enter values in order to select subsets with different sizes. A classification analysis based on substitution was carried out with each of the resulting subsets of variables. In Table 9.3, we compare the overall 'Wilks' A and the apparent correct classification rate.

According to the correct classification rate, we would choose to stop at five variables because of the abrupt change from 5 to 4. On the other hand, the changes in 'Wilks' A are more gradual, and no clear stopping point is indicated. $\square$

Table 9.3. Stepwise Selection Statistics for the Teacher Data

Number of Variables	Overall 'Wilks' A	Percentage of Correct Classification
13	.102	77.4
10	.129	75.4
9	.170	75.3
8	.182	76.8
7	.199	76.8
6	.211	76.1
5	.231	76.6
4	.256	65.6

9.2 NONPARAMETRIC PROCEDURES

We have previously discussed both parametric and nonparametric procedures. Welch's optimal rule in (9.7) and (9.11) is parametric, whereas Fisher's linear classification rule for two groups as given in (9.5) and (9.6) is essentially nonparametric, since no distributional assumptions were involved in its derivation. However, Fisher's procedure also turns out to be equivalent to the optimal normal-based approach in (9.8). Nonparametric procedures for estimating error rate include the substitution and holdout methods. In the next three sections we discuss three additional nonparametric classification procedures.

9.2.1 Multinomial Data

We now consider data in which an observation vector consists of responses on each of several categorical variables. The various combinations of categories constitute the possible outcomes of a multinomial random variable. For example, consider the following four categorical variables: gender (male or female), political party (Republican, Democrat, other), size of city of residence (under 10,000, between 10,000 and 100,000, over 100,000), and education (less than high-school graduation, high school graduate, college graduate, advanced degree). An observation vector might be (2, 1, 3, 4), that is, a female Republican who lives in a city of over 100,000 and is a college graduate. The total number of possible outcomes in this multinomial distribution is the product of the number of states of the individual variables: $2 \times 3 \times 3 \times 4 = 72$. We will use this example to illustrate classification procedures for multinomial data. Suppose we are attempting to predict whether or not a person will vote. Then there are two groups, G_1 and G_2 , and we assign a person to one of the groups after observing which of the 72 possible outcomes he or she gives.

Welch's (1939) optimum rule given in (9.7) can be written as: Assign $\mathbf{y}$ to G_1 if

$$\frac{f(\mathbf{y}|G_1)}{f(\mathbf{y}|G_2)} \leq \frac{\pi_2}{\pi_1} \quad (9.17)$$

and to G_2 otherwise. In our categorical example, $f(\mathbf{y}|G_1)$ is represented by q_{1i} , $i = 1, 2, \dots, 72$, and $f(\mathbf{y}|G_2)$ becomes q_{2i} , $i = 1, 2, \dots, 72$, where q_{1i} is the probability that a person in group 1 will give the i th outcome, with an analogous definition for q_{2i} . In terms of these multinomial probabilities, the classification rule in (9.17) becomes: If a person gives the i th outcome, assign him or her to G_1 if

$$\frac{q_{1i}}{q_{2i}} \leq \frac{\pi_2}{\pi_1} \quad (9.18)$$

and to G_2 otherwise. If the probabilities q_{1i} and q_{2i} were known, it would be easy to check (9.18) for each i and partition the 72 possible outcomes into two subsets, those for which the person would be assigned to G_1 and those corresponding to G_2 .

The values of q_{1i} and q_{2i} are usually unknown and must be estimated from a sample. Let n_{1i} and n_{2i} be the numbers of persons in groups 1 and 2 who give the

i th outcome, $i = 1, 2, \dots, 71$. Then we estimate p_{1i} and p_{2i} by

$$\hat{p}_{1i} = \frac{n_{1i}}{N_1} \quad \text{and} \quad \hat{p}_{2i} = \frac{n_{2i}}{N_2}, \quad i = 1, 2, \dots, 71, \quad (9.18)$$

where $N_1 = \sum_i n_{1i}$ and $N_2 = \sum_i n_{2i}$. However, a large sample size would be required for stable estimates; in any given example, some of the n 's may be zero.

Multinomial data can also be classified by ordinary linear classification functions. We must distinguish between ordered and unordered categories. If all the variables have ordered categories, the data can be submitted directly to an ordinary classification program. In the preceding example, city size and education are variables of this type. It is customary to assign ordered categories ranked values such as 1, 2, 3, 4. It has been shown that linear classification functions perform reasonably well on ordered-discrete data of this type [see Lachenbruch (1975, p. 45), Tinbergen et al. (1987), and Gilbert (1989)].

Unordered categorical variables cannot be handled this same way. Thus the political party variable in the preceding example should not be ranked 1, 2, 3 and entered into the computation of the classification functions. However, an unordered categorical variable with k categories can be replaced by $k - 1$ dummy variables (see Sections 6.1.8 and 11.3.2) for use with linear classification functions. For example, the political preference variable with three categories can be converted to two dummy variables as follows:

$$x_1 = \begin{cases} 1 & \text{if Republican,} \\ 0 & \text{otherwise,} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if Democrat,} \\ 0 & \text{otherwise.} \end{cases}$$

Thus the (x_1, x_2) pair takes the value (1, 0) for a Republican, (0, 1) for a Democrat, and (0, 0) for other. Many software programs will create dummy variables automatically. Note that if a subset selection program is used, the dummy variables for a given categorical variable must be kept together; that is, they must all be included in the chosen subset or all excluded, because all are necessary to describe the categorical variable.

In some cases, such as in medical data collection, there is a mixture of continuous and categorical variables. Various approaches to classification with such data have been discussed by Kruskowski (1973, 1975, 1977, 1979, 1980), Lachenbruch and Gelfand (1979), Tu and Hsu (1982), and Bayne et al. (1983). See Kinsch (1988, Section 6.6) for a discussion of logistic and probit classification, which are useful for certain types of continuous and discrete data that are not normally distributed.

9.3.2 Classification Based on Density Estimators

In (9.8), (9.12), and (9.14) we have linear and quadratic classification rules based on the multivariate normal density and prior probabilities. These normal-based rules arise from Wald's optimal rule that assigns y to the group for which $p_i f(y|G_i)$ is maximum. If the form of $f(y|G_i)$ is nonnormal and unknown, the density can be

estimated directly from the data. The approach we describe is known as the *kernel estimator*.

We first describe the kernel method for a univariate continuous random variable y . Suppose y has density $f(y)$, which we wish to estimate using a sample $y_1, y_2, \dots, y_n$. A simple estimate of $f(y_0)$ for an arbitrary point y_0 can be based on the proportion of points in the interval $(y_0 - h, y_0 + h)$. If the number of points in the interval is denoted by $N(y_0)$, then the proportion $N(y_0)/n$ is an estimate of $P(y_0 - h < y < y_0 + h)$, which is approximately equal to $2hf(y_0)$. Thus we estimate $f(y_0)$ by

$$\hat{f}(y_0) = \frac{N(y_0)}{2hn}. \quad (9.20)$$

We can express $\hat{f}(y_0)$ as a function of all y_i in the sample by defining

$$K(u) = \begin{cases} \frac{1}{2} & \text{for } |u| \leq 1, \\ 0 & \text{for } |u| > 1, \end{cases} \quad (9.21)$$

so that $N(y_0) = 2 \sum_{i=1}^n K(y_0 - y_i/h)$, and (9.20) becomes

$$\hat{f}(y_0) = \frac{1}{hn} \sum_{i=1}^n K\left(\frac{y_0 - y_i}{h}\right). \quad (9.22)$$

The function $K(u)$ is called the *kernel*. In (9.22), $K(y_0 - y_i)/h$ is $\frac{1}{2}$ for any point y_i in the interval $(y_0 - h, y_0 + h)$ and is zero for points outside the interval. Points in the interval add $1/2hn$ to the density and points outside the interval contribute nothing.

Kernel estimators were first proposed by Rosenblatt (1956) and Parzen (1962). A good review of nonparametric density estimation including kernel estimators has been given by Silverman (1986), who noted that classification analysis provided the initial motivation for the development of density estimation.

The kernel defined by (9.21) is rectangular, and the graph of $\hat{f}(y_0)$ plotted as a function of y_0 will be a step-function, since there will be a jump (or drop) whenever y_0 is a distance h from one of the y_i 's. (A moving average has a similar property.) To obtain a smooth estimator of $f(y)$, we must choose a smooth kernel. Two possibilities are

$$K(u) = \frac{1}{\pi} \frac{\sin^2 \pi u}{u^2}, \quad (9.23)$$

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad (9.24)$$

which have the property that all n sample points $y_1, y_2, \dots, y_n$ contribute to $\hat{f}(y_0)$, with the closest points weighted heavier than the more distant points. Even though $K(u)$ in (9.24) has the form of the normal distribution, this does not imply any

assumption about the density $f(y)$. We have used the normal density function because it is symmetric and unimodal. Other density functions could be used as kernels.

Caozoulin (1966) provided kernel estimates for multivariate density functions; see also Scott (1982). If $y_0 = (y_{01}, y_{02}, \dots, y_{0p})$ is an arbitrary point whose density we wish to estimate, then the estimator of (9.22) is

$$\hat{f}(y_0) = \frac{1}{nh_1h_2 \cdots h_p} \sum_{i=1}^n K\left(\frac{y_{01} - y_{i1}}{h_1}, \dots, \frac{y_{0p} - y_{ip}}{h_p}\right). \quad (9.23)$$

An estimate $\hat{f}(y_0)$ based on a multivariate normal kernel is given by

$$\hat{f}(y_0) = \frac{1}{nh_1h_2 \cdots h_p(2\pi)^{p/2}} \sum_{i=1}^n e^{-1/2(y_0 - y_i)'S_0^{-1}(y_0 - y_i)/h^2}, \quad (9.24)$$

where $h_1 = h_2 = \cdots = h_p = h$ and S_0 is the pooled covariance matrix from the k groups in the sample. The covariance matrix S_0 could be replaced by other forms. Two examples are (1) S_0 for the k th group and (2) a diagonal matrix.

The choice of the smoothing parameter h is critical in a kernel density estimator. The size of h determines how much each y_i contributes to $\hat{f}(y_0)$. If h is too small, $\hat{f}(y_0)$ has a peak at each y_i , and if h is too large, $\hat{f}(y_0)$ is almost uniform (overly smoothed). Therefore, the value chosen for h must depend on the sample size n to avoid too much or too little smoothing; the larger the sample size, the smaller h should be. In practice, we could try several values of h and check the resulting error rates from the classification analysis.

To use the kernel method of density estimation in classification, we can apply it to each group to obtain $\hat{f}(y_0|G_1), \hat{f}(y_0|G_2), \dots, \hat{f}(y_0|G_k)$, where y_0 is the vector of measurements for an individual of unknown group membership. The classification is then to choose $\hat{f}(y_0|G_i)$ for which

$$P_i \hat{f}(y_0|G_i) \text{ is maximum.} \quad (9.25)$$

Haberman, Hermann, and Van den Broek (1974) proposed a forward selection method for classification based on density estimation. Wipulan (1972) and Haberman, Hermann, and Kemmer (1976) found that the size of the h_i 's is more important than the shape of the kernel. The choice of h was investigated by Pao (1982) in a superior mode. Romano, Haberman, and Hermann (1980) compared linear, quadratic, and kernel classification methods for two groups and reported that for multivariate normal data with equal covariance matrices, the linear classifications were clearly superior. For some cases with departure from these assumptions, the kernel methods gave better results.

Example 9.3.2. We illustrate the density estimation method of classification for the football data of Table 8.3. We use the multivariate normal kernel estimation in (9.24) with $h = 1$ to obtain $\hat{f}(y_0|G_i), i = 1, 2, 3$, for the three groups. Using $p_1 =$

$p_1 = p_2$, the rule in (9.27) becomes: Assign p_0 to the group for which $\hat{f}(p_0|G_i)$ is greatest. To obtain an apparent error rate, we follow this procedure for each of the 90 observations and obtain the classification results in Table 9.8.

Applying a likelihood method in which the observation p_0 being classified is excluded from computation of $\hat{f}(p_0|G_1)$, $\hat{f}(p_0|G_2)$, and $\hat{f}(p_0|G_3)$, we obtain the classification results in Table 9.9. As expected, the likelihood error rate has increased somewhat from the apparent error rate in Table 9.8. $\square$

9.5.3 Nearest Neighbor Classification Rule

The earliest nonparametric classification method was the nearest neighbor rule of Fix and Hodges (1951), also known as the 1 nearest neighbor rule. The procedure is conceptually simple. We compute the distance from an observation p_0 to all other points p_j using the distance function

$$(p_0 - p_j)'(S_0^{-1})(p_0 - p_j), \quad j \neq 0.$$

Table 9.8. Classification Table for the Football Data of Table 8.5 Using the Density Estimation Method of Classification with Multivariate Normal Kernel

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	30	28	1	1
2	30	8	12	10
3	30	8	3	19

$$\text{Apparent correct classification rate} = \frac{28 + 11 + 17}{90} = .711$$

$$\text{Apparent error rate} = 1 - .711 = .289$$

Table 9.9. Classification Table for the Football Data of Table 8.5 Using the Likelihood Method Based on Density Estimation

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	30	28	1	1
2	30	8	18	20
3	30	1	3	26

$$\text{Correct classification rate} = \frac{28 + 10 + 16}{90} = .667$$

$$\text{Error rate} = .71 = .667 = .333$$

To classify y_i into one of two groups, the k points nearest to y_i are examined, and if the majority of the k points belong to G_1 , assign y_i to G_1 ; otherwise assign y_i to G_2 . If we denote the number of points from G_1 as k_1 , with the remaining k_2 points from G_2 , where $k = k_1 + k_2$, then the rule can be expressed as: Assign y_i to G_1 if

$$k_1 > k_2 \quad (9.28)$$

and to G_2 otherwise. If the sample sizes n_1 and n_2 differ, we may wish to use proportions in place of counts. Assign y_i to G_1 if

$$\frac{k_1}{n_1} > \frac{k_2}{n_2} \quad (9.29)$$

A further refinement can be made by taking into account prior probabilities. Assign y_i to G_1 if

$$\frac{k_1/n_1}{k_2/n_2} > \frac{p_2}{p_1} \quad (9.30)$$

These rules are easily extended to more than two groups. For example, (9.29) becomes: Assign the observation to the group that has the highest proportion k_i/n_i , where k_i is the number of observations from G_i among the k nearest neighbors of the observation in question.

A decision must be made as to the value of k . Loftsgaarden and Quesenberry (1965) suggest choosing k near $\sqrt{n_i}$ for a typical n_i . In practice, one could try several values of k and use the one with the best error rate.

Reviews and extensions of the nearest neighbor method have been given by Hart (1968), Dales (1972), Hand and Dababneh (1978), Chittamuru Chavala and Krishna (1979), Rogers and Wagner (1979), and Brown and Kaplowitz (1979).

Example 9.5.3. We use the football data of Table 9.3 to illustrate the k nearest neighbor method of estimating error rate, with $k = 5$. Since $n_1 = n_2 = n_3 = 30$ and the p_i 's are also assumed to be equal, we simply examine the five points closest to a

Table 9.1b. Classification Table for the Football Data of Table 9.3 Using the k Nearest Neighbor Method with $k = 5$

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	30	26	4	0
2	30	1	18	9
3	30	1	4	25

$$\text{Correct classification rate} = \frac{26 + 18 + 25}{93} = .867$$

$$(\text{Error rate} = 1 - .867 = .133)$$

point y and classify y into the group that has the most points among the five points. If there is a tie, we do not classify the point. For example, if the numbers from G_1 , G_2 , and G_3 were 1, 2, and 1, respectively, then we do not assign y to either G_1 or G_2 .

For each point y_{ij} , $i = 1, 2, 3$, $j = 1, 2, \dots, 20$, we find the five nearest neighbors and classify the point accordingly. Table 9.18 gives the classification results. As can be seen, there were 5 observations in group 1 that were not classified because of ties, 1 in group 2, and 3 in group 3. This left a total of 85 observations classified. $\square$

PROBLEMS

- 9.1 Show that if $T_1 = n\bar{y}_1$, $i = 1, 2, \dots, n_0$, and $T_2 = n\bar{y}_2$, $i = 1, 2, \dots, n_1$, where τ is the discriminant function defined in (9.1), then $T_1 - T_2 = (\bar{y}_1 - \bar{y}_2)/S_{p-1}^2(\bar{y}_1 + \bar{y}_2) - \tau$ is in (τ, τ_0) .
- 9.2 With $\tau = n\bar{y}$ as in (9.1) and $T_1 = n\bar{y}_1$, $T_2 = n\bar{y}_2$, show that $\frac{1}{2}(T_1 + T_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)/S_{p-1}^2(\bar{y}_1 + \bar{y}_2) - \tau$ is in (τ, τ_0) .
- 9.3 Obtain the normal-based classification rule in (9.8).
- 9.4 Derive the linear classification rule in (9.13).
- 9.5 Derive the quadratic classification function in (9.14).
- 9.6 Do a classification analysis on the beetle data in Table 5.5 as follows:
 - (a) Find the classification function $\tau = (\bar{y}_1 - \bar{y}_2)/S_{p-1}^2\bar{y}$ and the cutoff point $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$.
 - (b) Find the classification table using the linear classification function in part (a).
 - (c) Find the classification table using the nearest neighbor method.
- 9.7 Do a classification analysis on the hydrophy data of Table 5.7 as follows:
 - (a) Find the classification function $\tau = (\bar{y}_1 - \bar{y}_2)/S_{p-1}^2\bar{y}$ and the cutoff point $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$.
 - (b) Find the classification table using the linear classification function in part (a).
 - (c) Repeat part (b) using p_1 and p_2 proportional to sample sizes.
- 9.8 Do a classification analysis on the cyclical data of Table 5.8 as follows:
 - (a) Find the classification function $\tau = (\bar{y}_1 - \bar{y}_2)/S_{p-1}^2\bar{y}$ and the cutoff point $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$.
 - (b) Find the classification table using the linear classification function in part (a).
 - (c) Find the classification table using the holdout method.
 - (d) Find the classification table using a kernel density estimator method.

- 8.9** Using the engine data of Table 5.5, carry out a classification analysis as follows:
- Find the classification table using the linear classification function.
 - Carry out a stepwise least-squares selection of variables (see Problem 8.14).
 - Find the classification table for the variables selected in part (b).
- 8.10** Do a classification analysis on the fish data in Table 5.17 as follows. Assume $P_1 = P_2 = P_3$.
- Find the linear classification functions.
 - Find the classification table using the linear classification functions in part (a) (assuming $\Sigma_1 = \Sigma_2 = \Sigma_3$).
 - Find the classification table using quadratic classification functions (assuming population covariance matrices are not equal).
 - Find the classification table using linear classification functions and the holdout method.
 - Find the classification table using a nearest neighbor method.
- 8.11** Do a classification analysis on the rocket data of Table 5.2 as follows:
- Find the linear classification functions.
 - Find the classification table using the linear classification functions in part (a) (assuming $\Sigma_1 = \Sigma_2 = \Sigma_3$).
 - Find the classification table using quadratic classification functions (assuming population covariance matrices are not equal).
 - Find the classification table using the nearest neighbor method.
 - Find the classification table using a kernel density estimator method.

Multivariate Regression

10.1 INTRODUCTION

In this chapter, we consider the linear relationship between one or more y 's (the dependent or response variables) and one or more x 's (the independent or predictor variables). We will use a linear model to relate the y 's to the x 's and will be concerned with estimation and testing of the parameters in the model. One aspect of interest will be choosing which variables to include in the model if this is not already known.

We can distinguish three cases according to the number of variables:

1. **Simple linear regression:** one y and one x . For example, suppose we wish to predict college grade point average (GPA) based on an applicant's high school GPA.
2. **Multiple linear regression:** one y and several x 's. We could attempt to improve our prediction of college GPA by using more than one independent variable, for example, high school GPA, standardized test scores (such as ACT or SAT), or rating of high school.
3. **Multivariate multiple linear regression:** several y 's and several x 's. In the preceding illustration, we may wish to predict several y 's (such as number of years of college the person will complete or GPA in the sciences, arts, and humanities). As another example, suppose the Air Force wishes to predict several measures of pilot efficiency. These response variables could be regressed against independent variables (such as math and science skills, reaction time, eyesight acuity, and manual dexterity).

To further distinguish case 2 from case 3, we could designate case 2 as *univariate multiple regression* because there is only one y . Thus in case 3, *multivariate* indicates that there are several y 's and *multiple* implies several x 's. The term *multivariate regression* usually refers to case 3.

There are two basic types of independent variables, *fixed* and *random*. In the preceding illustrations, all x 's are random variables and are therefore not under the control of the researcher. A person is chosen at random, and all the y 's and x 's are

measured, or observed, for that person. In some experimental situations, the x 's are fixed, that is, under the control of the experimenter. For example, a researcher may wish to relate yield per acre and nutritional value to level of application of various chemical fertilizers. The experimenter can choose the amount of chemicals to be applied and then observe the changes in the yield and nutritional responses.

In order to provide a solid base for multivariate multiple regression, we review several aspects of multiple regression with fixed x 's in Sections 18.2. The random- x case for multiple regression is discussed briefly in Section 18.3.

18.2 MULTIPLE REGRESSION: FIXED x 's

18.2.1 Model for Fixed x 's

In the fixed- x regression model, we express each y in a sample of n observations as a linear function of the x 's plus a random error, ϵ :

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_q x_{1q} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_q x_{2q} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_q x_{nq} + \epsilon_n. \end{aligned} \quad (18.1)$$

The number of x 's is denoted by q . The β 's in (18.1) are called regression coefficients. Additional assumptions that accompany the equations of the model are as follows:

1. $E(\epsilon_i) = 0$ for all $i = 1, 2, \dots, n$.
2. $\text{var}(\epsilon_i) = \sigma^2$ for all $i = 1, 2, \dots, n$.
3. $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

Assumption 1 states that the model is linear and that no additional terms are needed to predict y ; all remaining variation in y is purely random and unpredictable. Thus if $E(\epsilon_i) = 0$ and the x 's are fixed, then $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq}$, and the mean of y is expressible in terms of these q x 's with no others needed. In assumption 2, the variance of each y_i is the same, which also implies that $\text{var}(y_i) = \sigma^2$, since the x 's are fixed. Assumption 3 imposes the condition that the error terms be uncorrelated, from which it follows that the y 's are also uncorrelated, that is, $\text{cov}(y_i, y_j) = 0$.

Thus the three assumptions can be restated in terms of y as follows:

1. $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq}$, $i = 1, 2, \dots, n$.
2. $\text{var}(y_i) = \sigma^2$, $i = 1, 2, \dots, n$.
3. $\text{cov}(y_i, y_j) = 0$, for all $i \neq j$.

by $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, then $\hat{E}(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}$. However, $\hat{E}(y_i)$ is usually designated $\hat{y}_i$. Thus $\hat{y}_i$ estimates $E(y_i)$, not y_i . We now consider the least squares estimates of the β 's.

The *least squares* estimates of $\beta_0, \beta_1, \dots, \beta_p$ minimize the sum of squares of deviations of the n observed y 's from their "modeled" values, that is, from their values $\hat{y}_i$ predicted by the model. Thus we seek $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ that minimize

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2. \end{aligned} \quad (10.4)$$

The value of $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)'$ that minimizes SSE in (10.4) is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (10.5)$$

In (10.5), we assume that $\mathbf{X}'\mathbf{X}$ is nonsingular. This will ordinarily hold if $n > p + 1$ and no x_j is a linear combination of other x 's.

In expression (10.5), we use a characteristic pattern similar to that for $\hat{\beta}_1$ in simple linear regression given in (3.11), $\hat{\beta}_1 = s_{yz}/s_x^2$. The product $\mathbf{X}'\mathbf{y}$ can be used to compute the covariances of the x 's with y . The product $\mathbf{X}'\mathbf{X}$ can be used to obtain the covariance matrix of the x 's, which includes the variances and covariances of the x 's [see the comment following (10.16) about variances and covariances involving the fixed x 's]. Since $\mathbf{X}'\mathbf{X}$ is typically not diagonal, each $\hat{\beta}_j$ depends on s_{yz} and s_x^2 , as well as the relationship of x_j to the other x 's.

We now demonstrate algebraically that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (10.5) minimizes SSE. This can also be done readily with calculus. If we designate the i th row of $\mathbf{X}$ as $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$, we can write (10.4) as

$$\text{SSE} = \sum_{i=1}^n (y_i - \mathbf{x}_i\hat{\beta})^2.$$

The quantity $y_i - \mathbf{x}_i\hat{\beta}$ is the i th element of the vector $\mathbf{y} - \mathbf{X}\hat{\beta}$. Hence, by (2.33),

$$\text{SSE} = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}). \quad (10.6)$$

Let $\mathbf{h}$ be an alternative estimate that may lead to a smaller value of SSE than does $\hat{\beta}$. We add $\mathbf{X}(\hat{\beta} - \mathbf{h})$ to one of the addends SSE,

$$\text{SSE} = (\mathbf{y} - \mathbf{X}\hat{\beta}) + \mathbf{X}(\hat{\beta} - \mathbf{h}) \quad (\mathbf{y} - \mathbf{X}\hat{\beta}) + \mathbf{X}(\hat{\beta} - \mathbf{h}).$$

We now expand this using the two terms $y - X_0\hat{\beta}$ and $X_0(\hat{\beta} - \beta)$ to obtain

$$\begin{aligned} SSE &= (y - X_0\hat{\beta})'(y - X_0\hat{\beta}) + 2X_0(\hat{\beta} - \beta)'(X_0\hat{\beta} - y) + 2X_0(\hat{\beta} - \beta)'(y - X_0\hat{\beta}) \\ &= (y - X_0\hat{\beta})'(y - X_0\hat{\beta}) + (\hat{\beta} - \beta)'X'X_0(\hat{\beta} - \beta) + 2(\hat{\beta} - \beta)'X'(y - X_0\hat{\beta}) \\ &= (y - X_0\hat{\beta})'(y - X_0\hat{\beta}) + (\hat{\beta} - \beta)'X'X_0(\hat{\beta} - \beta) + 2(\hat{\beta} - \beta)'(X'y - X'X_0\hat{\beta}). \end{aligned}$$

The third term vanishes if we substitute $\hat{\beta} = (X'X)^{-1}X'y$ into $X'X_0\hat{\beta}$. The second term is a positive definite quadratic form, and SSE is therefore minimized when $\hat{\beta} = \beta$. Thus, no value of $\hat{\beta}$ can reduce SSE from the value given by $\hat{\beta}$. For a review of properties of $\hat{\beta}$ and an alternative derivation of $\hat{\beta}$ based on the assumption that y is normally distributed, see Rousseeuw (1978, Chapter 7; 2000, Chapter 7).

10.1.3 An Estimator for σ^2

It can be shown that

$$E(SSE) = \sigma^2(n - (q + 1)) = \sigma^2(n - p - 1). \quad (10.7)$$

We can therefore obtain an unbiased estimator of σ^2 as

$$s^2 = \frac{SSE}{n - p - 1} = \frac{1}{n - p - 1}(y - X_0\hat{\beta})'(y - X_0\hat{\beta}). \quad (10.8)$$

We can also express SSE in the form

$$SSE = y'y - \hat{\beta}'X'y, \quad (10.9)$$

and we note that there are n terms in $y'y$ and $q + 1$ terms in $\hat{\beta}'X'y$. The difference is the denominator of s^2 in (10.8). Thus the degrees of freedom (denominator) for SSE are reduced by $q + 1$.

The need for an adjustment of $q + 1$ to the degrees of freedom of SSE can be illustrated with a simple random sample of a random variable y from a population with mean μ and variance σ^2 . The sum of squares $\sum_{i=1}^n (y_i - \mu)^2$ has n degrees of freedom, whereas $\sum_{i=1}^n (y_i - \bar{y})^2$ has $n - 1$. It is intuitively clear that

$$E\left[\sum_{i=1}^n (y_i - \mu)^2\right] = E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right]$$

because $\bar{y}$ fits the sample better than μ , which is the mean of the population but not of the sample. Thus (squared) deviations from $\bar{y}$ will tend to be smaller than deviations from μ . In fact, it is easily shown that

$$\begin{aligned}\sum_{i=1}^n (y_i - \mu)^2 &= \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \mu)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2,\end{aligned}\quad (10.10)$$

whence

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \mu)^2 - n(\bar{y} - \mu)^2.$$

Thus $\sum (y_i - \bar{y})^2$ is expressible as a sum of n squares minus 1 square, which corresponds to $n - 1$ degrees of freedom. More formally, we have

$$E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] = n\sigma^2 - \frac{\sigma^2}{n} = (n-1)\sigma^2.$$

10.1.4 The Model Corrected for Means

It is sometimes convenient to "center" the x 's by subtracting their means. $T_1 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)/n$, $T_2 = \sum_{i=1}^n (x_{i2} - \bar{x}_2)/n$, and so on. $T_1, T_2, \dots, T_p$ are the means of the columns of $\mathbf{X}$ in (10.11). In terms of centered x 's, the model for each y_i in (10.1) becomes

$$y_i = \alpha + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_p(x_{ip} - \bar{x}_p) + \varepsilon_i, \quad (10.13)$$

where

$$\alpha = \beta_0 + \beta_1\bar{x}_1 + \beta_2\bar{x}_2 + \cdots + \beta_p\bar{x}_p. \quad (10.12)$$

To estimate

$$\beta_1 = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

we use the centered x 's in the matrix

$$\mathbf{X}_0 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix} = \begin{pmatrix} (x_1 - \bar{x})' \\ (x_2 - \bar{x})' \\ \vdots \\ (x_n - \bar{x})' \end{pmatrix}, \quad (10.14)$$

where $x'_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $\bar{x}' = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$. Then, by analogy to (10.5), the least squares estimate of β_1 is

$$\hat{\beta}_1 = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{y}. \quad (10.15)$$

If $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is evaluated at $x_1 = \bar{x}_1, x_2 = \bar{x}_2, \dots, x_p = \bar{x}_p$, the result is the same as $\bar{y}$ in (10.12). Thus, we estimate α by $\bar{y}$:

$$\hat{\alpha} = \bar{y}.$$

In other words, if the origin of the x 's is shifted to $\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)'$, then the intercept of the fitted model is $\bar{y}$. With $\hat{\alpha} = \bar{y}$, we obtain

$$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 - \cdots - \hat{\beta}_p \bar{x}_p = \bar{y} - \hat{\beta}'_1 \bar{x} \quad (10.13)$$

as an estimate of β_0 in (10.12). Together, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in (10.13) and (10.14) are the same as the usual least squares estimator $\hat{\beta} = (X'X)^{-1}X'y$ in (10.8).

We can express $\hat{\beta}_1$ in (10.14) in terms of sample variances and covariances. The overall sample-covariance matrix of y and the x 's is

$$S = \begin{pmatrix} s_{yy} & s_{y1} & s_{y2} & \cdots & s_{yp} \\ s_{y1} & s_{11} & s_{12} & \cdots & s_{1p} \\ s_{y2} & s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{yp} & s_{1p} & s_{2p} & \cdots & s_{pp} \end{pmatrix} = \begin{pmatrix} s_{yy} & s'_{yx} \\ s_{yx} & S_{xx} \end{pmatrix}, \quad (10.16)$$

where s_{yy} is the variance of y , s_{y1} is the covariance of y and x_1 , s_{12} is the variance of x_1 , s_{12} is the covariance of x_1 and x_2 , and $s'_{yx} = (s_{y1}, s_{y2}, \dots, s_{yp})$. These sample variances and covariances are mathematically equivalent to analogous formulas (3.23) and (3.26) for random variables, where the sample variances and covariances were estimates of population variances and covariances. However, here the x 's are considered to be constants that remain fixed from sample to sample, and a formula such as $s_{11} = \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 / (n - 1)$ summarizes the spread in the n values of x_1 but does not estimate a population variance.

To express $\hat{\beta}_1$ in terms of S_{xx} and s'_{yx} in (10.16), we note first that the diagonal elements of $S_y S_y$ are corrected sums of squares. For example, in the second diagonal position, we have

$$\sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 = (n - 1)s_{22}.$$

The off-diagonal elements of $S_y S_y$ are analogous corrected sums of products; for example, the element in the $(1, 2)$ position is

$$\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = (n - 1)s_{12}.$$

Thus

$$\frac{1}{n-1} S_y S_y = S_{yy}. \quad (10.17)$$

Similarly,

$$\frac{1}{n-1}\mathbf{X}_2\mathbf{y} = \mathbf{u}_{2y}, \quad (10.18)$$

even though $\mathbf{y}$ has not been centered. The second element of $\mathbf{X}_2\mathbf{y}$, for example, is $\sum_i (x_{2i} - \bar{x}_2)y_i$, which is equal to $(n-1)u_{2y}$:

$$\begin{aligned} (n-1)u_{2y} &= \sum_i (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \\ &= \sum_i (x_{2i} - \bar{x}_2)y_i - \sum_i (x_{2i} - \bar{x}_2)\bar{y} \\ &= \sum_i (x_{2i} - \bar{x}_2)y_i, \end{aligned}$$

since

$$\sum_i (x_{2i} - \bar{x}_2)\bar{y} = 0. \quad (10.19)$$

Now, multiplying and dividing by $n-1$ in (10.14), we obtain

$$\begin{aligned} \hat{\beta}_1 &= (n-1)(\mathbf{X}_1'\mathbf{X}_1)^{-1}\frac{\mathbf{X}_1'\mathbf{y}}{n-1} = \left(\frac{\mathbf{X}_1'\mathbf{X}_1}{n-1}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{y}}{n-1} \\ &= \mathbf{S}_{11}^{-1}\mathbf{u}_{1y} \quad \text{(by (10.17) and (10.18)).} \end{aligned} \quad (10.20)$$

and substituting this in (10.15) gives

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1'\bar{\mathbf{x}} = \bar{y} - \mathbf{c}_{1y}'\mathbf{S}_{11}^{-1}\bar{\mathbf{x}}. \quad (10.21)$$

10.1.5 Hypothesis Tests

In this section, we review two basic tests on the β 's. For other tests and confidence intervals, see Raudenbush (1998, Section 7.2.4; 2008, Sections 8.4–8.7). In order to obtain F -tests, we assume that $\mathbf{y}$ is $N_1(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

10.1.5a Test of Overall Regression

The overall regression hypothesis that none of the x 's predict y can be expressed as $H_0: \beta_1 = 0$, since $\beta_1' = (\beta_{11}, \beta_{12}, \dots, \beta_{1q})$. We do not include $\beta_0 = 0$ in the hypothesis so as not to restrict y to have an intercept of zero.

We can write SSE = $\mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$ in (10.9) in the form

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\bar{\mathbf{y}} + \hat{\beta}'\mathbf{X}'\mathbf{y} + \hat{\beta}'\mathbf{X}'\mathbf{y}, \quad (10.22)$$

which partitions $y'y$ into a part due to β and a part due to deviations from the fitted model.

To correct y for its mean and thereby avoid inclusion of $\beta_0 = 0$, we subtract ny^2 from both sides of (18.12) to obtain

$$\begin{aligned} y'y - ny^2 &= (y'y - \beta'X'y) + (\beta'X'y - ny^2) \\ &= SSR + SSR_0 \end{aligned} \quad (18.23)$$

where $y'y - ny^2 = \sum (y_i - \bar{y})^2$ is the total sum of squares adjusted for the mean and $SSR = \beta'X'y - ny^2$ is the overall regression sum of squares adjusted for the intercept.

We can test $H_0: \beta_k = 0$ by means of

$$F = \frac{SSR_k/q}{SSR_0/(n - q - 1)}, \quad (18.24)$$

which is distributed as $F_{q, n-q-1}$ when $H_0: \beta_k = 0$ is true. We reject H_0 if $F > F_{\alpha, q, n-q-1}$.

18.2.14 Test on a Subset of the β 's

In an attempt to simplify the model, we may wish to test the hypothesis that some of the β 's are zero. For example, in the model

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1^2 + \beta_4x_2^2 + \beta_5x_1x_2 + \varepsilon,$$

we may be interested in the hypothesis $H_0: \beta_3 = \beta_4 = \beta_5 = 0$. If H_0 is true, the model is linear in x_1 and x_2 . In other cases, we may want to ascertain whether a single β_j can be deleted.

For convenience of exposition, let the β 's that are candidates for deletion be rearranged to appear last in β and denote this subset of β 's by β_k , where k reminds us that these β 's are to be deleted if $H_0: \beta_k = 0$ is accepted. Let the subset to be retained in the reduced model be denoted by β_r . Thus β is partitioned into

$$\beta = \begin{pmatrix} \beta_r \\ \beta_k \end{pmatrix}.$$

Let k designate the number of parameters in β_k . Then there are $q + 1 - k$ parameters in β_r .

To test the hypothesis $H_0: \beta_k = 0$, we fit the full model containing all the β 's in β and then fit the reduced model containing only the β 's in β_r . Let X_r be the columns of X corresponding to β_r . Then the reduced model can be written as

$$y = X_r\beta_r + \varepsilon, \quad (18.25)$$

and β_0 is estimated by $\hat{\beta}_0 = (X'X_0)^{-1}X'_0y$. To compare the fit of the full model and the reduced model, we calculate

$$\hat{\mu}'Xy - \hat{\mu}'_0X_0y, \quad (10.26)$$

where $\hat{\mu}'Xy$ is the regression sum of squares from the full model and $\hat{\mu}'_0X_0y$ is the regression sum of squares for the reduced model. The difference in (10.26) shows what β_0 contributes "above and beyond" β_1 . We can test $H_0: \beta_0 = 0$ with an F -statistic:

$$F = \frac{(\hat{\mu}'Xy - \hat{\mu}'_0X_0y)/q}{(y'y - \hat{\mu}'Xy)/(n - q - 1)} \quad (10.27)$$

$$= \frac{(SSR_f - SSR_0)/q}{SSR_0/(n - q - 1)} = \frac{SSR}{SSR_0}, \quad (10.28)$$

where $SSR_f = \hat{\mu}'Xy$ and $SSR_0 = \hat{\mu}'_0X_0y$. The F -statistic in (10.27) and (10.28) is distributed as $F_{q, n-q-1}$ if H_0 is true. We reject H_0 if $F > F_{\alpha, q, n-q-1}$.

The test in (10.27) is easy to carry out in practice. We fit the full model and obtain the regression and error sums of squares $\hat{\mu}'Xy$ and $y'y - \hat{\mu}'Xy$, respectively. We then fit the reduced model and obtain its regression sum of squares $\hat{\mu}'_0X_0y$ to be subtracted from $\hat{\mu}'Xy$. If a software package gives the regression sums of squares in corrected form, this can readily be used to obtain $\hat{\mu}'Xy - \hat{\mu}'_0X_0y$, since

$$\hat{\mu}'Xy - ny^2 = (\hat{\mu}'Xy - y^2) - (\hat{\mu}'_0X_0y - y^2) = \hat{\mu}'Xy - \hat{\mu}'_0X_0y.$$

Alternatively, we can obtain $\hat{\mu}'Xy - \hat{\mu}'_0X_0y$ as the difference between error sums of squares for the two models:

$$\begin{aligned} SSR - SSR_f &= y'y - \hat{\mu}'Xy - (y'y - \hat{\mu}'_0X_0y) \\ &= \hat{\mu}'Xy - \hat{\mu}'_0X_0y. \end{aligned}$$

A test for an individual β_j above and beyond the other β 's is readily obtained using (10.27). To test $H_0: \beta_j = 0$, we arrange β_j last in β ,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_j \end{pmatrix},$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_{q-1})'$ contains all the β 's except β_j . By (10.27), the test statistic is

$$F = \frac{\hat{\mu}'Xy - \hat{\mu}'_jX_jy}{SSR_j/(n - q - 1)}, \quad (10.29)$$

which is $F_{1, n-q-1}$. Note that $h = 1$. The test of $H_0: \beta_j = 0$ by the F -statistic in (10.28) is called a *partial F-test*. A detailed breakdown of the effect of each variable in the presence of the others is given by Kennedy (1993, 2008), Section 10.5.

Since the F -statistic in (10.28) has 1 and $n - q - 1$ degrees of freedom, it is the square of a t -statistic. The t -statistic equivalent to (10.28) is

$$t = \frac{\hat{\beta}_j}{s_y \sqrt{g_{jj}}},$$

where g_{jj} is the j th diagonal element of $(X'X)^{-1}$ and $s = \sqrt{\text{SSE}_T/(n - q - 1)}$ (Kennedy 2008, Section 8.5.1).

10.2.6 R^2 in Fixed- α Regression

The proportion of the (explained) total variation in the y 's that can be attributed to regression on the x 's is denoted by R^2 :

$$\begin{aligned} R^2 &= \frac{\text{regression sum of squares}}{\text{total sum of squares}} \\ &= \frac{\hat{\beta}'X'y - n\bar{y}^2}{y'y - n\bar{y}^2}. \end{aligned} \quad (10.29)$$

The ratio R^2 is called the *coefficient of multiple determination*, or more commonly the *squared multiple correlation*. The *multiple correlation* R is defined as the positive square root of R^2 .

The F -test for overall regression in (10.26) can be expressed in terms of R^2 as

$$F = \frac{n - q - 1}{q} \frac{R^2}{1 - R^2}. \quad (10.30)$$

For the reduced model (10.25), R^2 can be written as

$$R_1^2 = \frac{\hat{\beta}_1'X_1y - n\bar{y}^2}{y'y - n\bar{y}^2}. \quad (10.31)$$

Then in terms of R^2 and R_1^2 , the full and reduced model test in (10.27) for $H_0: \beta_h = 0$ becomes

$$F = \frac{(R^2 - R_1^2)/h}{(1 - R_1^2)/(n - q - 1)} \quad (10.32)$$

[see (11.36)].

We can express R^2 in terms of sample variances, covariances, and correlations:

$$R^2 = \frac{\hat{y}_m'X_1'X_1\hat{y}_m}{y'y} = \hat{y}_{1m}'R_{11}^{-1}\hat{y}_{1m}, \quad (10.33)$$

where r_{xy} , s_{yy} , and S_{xx} are defined in (18.35) and r_{yx} and $\mathbf{R}_{xx}$ are from an analogous partitioning of the sample correlation matrix of y and the x 's:

$$\mathbf{R} = \left(\begin{array}{c|cccc} 1 & r_{y1} & r_{y2} & \cdots & r_{yp} \\ \hline r_{xy} & 1 & r_{11} & \cdots & r_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{yp} & r_{p1} & r_{p2} & \cdots & 1 \end{array} \right) = \begin{pmatrix} 1 & \mathbf{r}_{yx} \\ \mathbf{r}_{xy} & \mathbf{R}_{xx} \end{pmatrix}. \quad (18.36)$$

18.17 Subsets Selection

In practice, one often has more x 's than are needed for predicting y . Some of them may be redundant and could be discarded. In addition to logistical motivations for deleting variables, there are statistical incentives; for example, if an x is deleted from the fitted model, the variances of the $\hat{\beta}_j$'s and of the $\hat{y}$'s are reduced. Various aspects of model validation are reviewed by Box-Cox (2003, Section 7.3 and Chapter 9).

The two most popular approaches to subsets selection are to (1) examine all possible subsets and (2) use a stepwise technique. We discuss these in the next two sections.

18.17a All Possible Subsets

The optimal approach to subsets selection is to examine all possible subsets of the x 's. This may not be computationally feasible if the sample size and number of variables are large. Some programs take advantage of algorithms that find the optimum subset of each size without examining all of the subsets (see, for example, Fox and Williams (1974)).

We discuss three criteria for comparing subsets when searching for the best subset. To conform with established notation in the literature, the number of variables in a subset is denoted by $p - 1$, so that with the inclusion of an intercept, there are p parameters in the model. The corresponding total number of available variables from which a subset is to be selected is denoted by $k - 1$, with k parameters in the model.

1. R_p^2 . By its definition in (18.35) as the proportion of total (corrected) sum of squares accounted for by regression, R^2 is clearly a measure of model fit. The subscript p is an index of the subset size, since it includes the number of parameters in the model, including an intercept. However, R_p^2 does not reach a maximum for any value of p less than k because it cannot decrease when a variable is added to the model. The usual procedure is to find the subset with largest R_p^2 for each of $p = 1, 2, \dots, k$ and then choose a value of p beyond which the increases in R^2 appear to be unimportant. This judgment is, of course, subjective.

2. s_p^2 . Another useful criterion is the variance estimator for each subset as defined in (18.35)

$$s_p^2 = \frac{SSE_p}{n - p}. \quad (18.38)$$

For each of $p = 1, 2, \dots, k$, we find the subset with smallest c_p^2 . If k is fairly large, a typical pattern as p approaches k is for the minimal c_p^2 to decrease to an overall minimum less than c_1^2 and then increase. The minimum value of c_p^2 can be less than c_1^2 if the decrease in SSE _{p} with an additional variable does not offset the loss of a degree of freedom in the denominator. It is often suggested that the researcher choose the subset with absolute minimum c_p^2 . However, as Hosking (1976, p. 19) notes, this procedure may fit some noise unique to the sample and thereby include one or more extraneous predictor variables. An alternative suggestion is to choose p such that $\min_p c_p^2 = c_p^2$ or, more precisely, choose the smallest value of p such that $\min_p c_p^2 = c_p^2$, since there will not be a $p < k$ such that $\min_p c_p^2$ is exactly equal to c_p^2 .

3. C_p . The C_p criterion is due to Mallows (1964, 1973). In the following development, we follow Myers (1990, pp. 140–142). The expected squared error, $E(\hat{y}_i) = E(y_i | \hat{y}_i)$, is used in formulating the C_p criterion because it incorporates a variance component and a bias component. The goal is to find a model that achieves a good balance between the bias and variance of the fitted values $\hat{y}_i$. Bias arises when the $\hat{y}_i$ values are based on an incorrect model, in which $E(\hat{y}_i) \neq E(y_i)$. If $\hat{y}_i$ were based on the correct model, so that $E(\hat{y}_i) = E(y_i)$, then $E(\hat{y}_i) - E(y_i) | \hat{y}_i^2$ would be equal to $\text{var}(\hat{y}_i)$. In general, however, as we examine many competing models, for various values of p , $\hat{y}_i$ is not based on the correct model, and we have (see Problem 13.4)

$$\begin{aligned} E(\hat{y}_i) - E(y_i) | \hat{y}_i^2 &= E(\hat{y}_i) - E(y_i) + E(y_i) - E(y_i) | \hat{y}_i^2 \\ &= E(\hat{y}_i) - E(y_i) | \hat{y}_i^2 + [E(y_i) - E(y_i) | \hat{y}_i^2] \end{aligned} \quad (13.57)$$

$$= \text{var}(\hat{y}_i) + \text{bias in } \hat{y}_i | \hat{y}_i^2. \quad (13.58)$$

For a given value of p , the total expected squared error for the n observations in the sample, standardized by dividing by n^2 , becomes

$$\frac{1}{n^2} \sum_{i=1}^n E(\hat{y}_i) - E(y_i) | \hat{y}_i^2 = \frac{1}{n^2} \sum_{i=1}^n \text{var}(\hat{y}_i) + \frac{1}{n^2} \sum_{i=1}^n \text{bias in } \hat{y}_i | \hat{y}_i^2. \quad (13.59)$$

Before defining C_p as an estimate of (13.59), we can achieve some simplification. We first show that $\sum_{i=1}^n \text{var}(\hat{y}_i) / n^2$ is equal to p/n . Let the model for all n observations be designated by

$$\mathbf{y} = \mathbf{X}_p \mathbf{\beta}_p + \mathbf{e},$$

We assume that, in general, this prospective model is underspecified and that the true model (which produces $\mathbf{e}^2$) contains additional β 's and additional columns of the $\mathbf{X}$ matrix. If we designate the i th row of $\mathbf{X}_p$ by $\mathbf{x}'_{ip}$, then the first term on the right side of (13.59) becomes (see also Problem 13.5)

$$\begin{aligned}
\frac{1}{n^2} \sum_{j=1}^n \text{var}(x_j) &= \frac{1}{n^2} \sum_{j=1}^n \text{var}(b'_{j0} \hat{\beta}_0) \\
&= \frac{1}{n^2} \sum_{j=1}^n b'_{j0} b_0' (X_0' X_0)^{-1} b_{j0} \quad \text{[by (3.20)]} \\
&= \text{tr}(X_0' X_0 b_0 b_0' (X_0' X_0)^{-1} X_0) \quad \text{[by (3.22)]} \quad (10.40) \\
&= \text{tr}(X_0' X_0)^{-1} X_0' X_0 \quad \text{[by (2.17)]} \\
&= \text{tr} \hat{I}_p = p. \quad (10.41)
\end{aligned}$$

It can be shown (Miyata 1985, pp. 176–179) that

$$\sum_{i=1}^n (\text{bias in } \hat{y}_i)^2 = (n - p) \text{tr}(e_p^2 - \sigma^2). \quad (10.42)$$

Using (10.41) and (10.42), the final simplified form of the (standardized) total expected squared error in (10.39) is

$$\frac{1}{n^2} \sum_{i=1}^n E(x_i - E(x_i))^2 = p + \frac{n-p}{n^2} E(x_p^2 - \sigma^2). \quad (10.43)$$

In practice, σ^2 is usually estimated by s_p^2 , the MSE from the full model. We thus estimate (10.43) by

$$C_p = p + (n - p) \frac{s_p^2 - s_p^2}{s_p^2}. \quad (10.44)$$

An alternative form is

$$C_p = \frac{\text{MSE}_p}{s_p^2} = (n - 2p). \quad (10.45)$$

In (10.44), we see that if the bias is small for a particular model, C_p will be close to p . For this reason, the line $C_p = p$ is commonly plotted along with the C_p values of several candidate models. We look for small values of C_p that are near this line.

In a Monte Carlo study, Hinton (1985) compared several subset selection criteria based on MSE_p and C_p . The three best procedures were to choose (1) the subset with the smallest p such that $C_p = p$, (2) the subset with the smallest p such that $s_p^2 = s_p^2$, and (3) the subset with minimum s_p^2 . The first of these was found to give best results overall, with the second method close behind. The third method performed best in some cases where k was small.

10.1.7b Stepwise Selection

For many data sets, it may be impractical to examine all possible subsets, even with an efficient algorithm such as that of Furnival and Wilson (1974). In such cases, we

can use the familiar stepwise approach, which is widely available and has virtually no limit as to the number of variables or observations. A related stepwise technique was discussed in Sections 8.11.2 and 8.7 in connection with selection of dependent variables to separate groups in a MANOVA or discriminant analysis setting. In this section, we are concerned with selecting the independent variables (x 's) that best predict the dependent variable (y) in regression.

We first review the forward selection procedure, which typically uses an F -test at each step. At the first step, y is regressed on each x_j alone, and the x with the largest F -value is "added" into the model. In the second step, we search for the variable with the largest partial F -value for testing the significance of each variable in the presence of the variable first entered. Thus, if we denote the first variable to enter as x_1 , then at the second step we calculate the partial F -statistic

$$F = \frac{\text{MSR}(x_1, x_2)}{\text{MSE}(x_1, x_2)}$$

for each $j \neq 1$ and choose the variable that maximizes F , where $\text{MSR} = (\text{SSR}_j - \text{SSR}_1)/k$ and $\text{MSE} = \text{SSR}_j/(n - q - 1)$ are the mean squares for regression and error, respectively, as in (1.8.18). In this case, $\text{SSR}_j = \text{SSR}(x_1, x_j)$ and $\text{SSR}_1 = \text{SSR}(x_1)$. Note also that $k = 1$ because only one variable is being added, and MSE is calculated using only the variable already entered plus the candidate variable. This procedure continues at each step until the largest partial F for an entering variable falls below a preselected threshold F -value or until the corresponding p -value exceeds some predetermined level.

The stepwise selection procedure similarly adds the best variable in series at each step. Then after a variable has entered, each of the variables previously entered is examined by a partial F -test to see if it is no longer significant and can be dropped from the model.

The backward elimination procedure begins with all x 's in the model and deletes one at a time. The partial F -statistic for each variable in the presence of the others is calculated, and the variable with smallest F is eliminated. This continues until the smallest F at some step exceeds a preselected threshold value.

Since these sequential methods do not examine all subsets, they will often fail to find the optimum subset, especially if k is large. However, β_1^2 , s_1^2 , or C_p may not differ substantially between the optimum subset and the one found by stepwise selection. These sequential methods have been popular for at least a generation, and it is very likely they will continue to be used, even though increased computing power has put the optimal methods within reach for larger data sets.

There are some possible risks in the use of stepwise methods. The stepwise procedure may fail to detect a true predictor (an x_j for which $\beta_j \neq 0$) because s_j^2 is biased upward in an underspecified model, thus artificially reducing the partial F -value. On the other hand, a variable that is not a true predictor of y (an x_j for which $\beta_j = 0$) may enter because of chance correlations in a particular sample. In the presence of such "noise" variables, the partial F -statistic for the entering variable does not have an F -distribution because it is maximized at each step. The calculated p -values

become optimistic. This problem intensifies when the sample size is relatively small compared to the number of variables. Ranscher and Pao (1985) found that in such cases surprisingly large values of R^2 can occur, even when there is no relationship between y and the x 's in the population. In a related study, Fleiss and Chang (1987) included x 's that were authentic contributors as well as noise variables. They found that "for most samples, a large percentage of the selected variables is noise, particularly when the number of candidate variables is large relative to the number of observations. The adjusted R^2 of the selected variables is highly inflated" (p. 89).

10.3 MULTIPLE REGRESSION: RANDOM x 's

In Section 10.2, it was assumed that the x 's were fixed and would have the same values if another sample were taken; that is, the same $\mathbf{X}$ matrix would be used each time a y vector was observed. However, many regression applications involve x 's that are random variables.

Thus in the random- x case, the values of $x_1, x_2, \dots, x_p$ are not under the control of the experimenter. They occur randomly along with y . On each subject we observe $y, x_1, x_2, \dots, x_p$.

If we assume that $(y, x_1, x_2, \dots, x_p)$ has a multivariate normal distribution, then $\hat{\beta}$, $\hat{R}^2$, and the F -tests have the same formulation as in the fixed- x case [for details, see Ranscher (1978, Section 7.3, 300), Section 10.4]. Thus with the multivariate normal assumption, we can proceed with estimation and testing the same way as in the random- x case as with fixed x 's.

10.4 MULTIVARIATE MULTIPLE REGRESSION: ESTIMATION

In this section we extend the estimation results of Sections 10.2.2–10.2.4 to the multivariate y case. We assume the x 's are fixed.

10.4.1 The Multivariate Linear Model

We turn now to the multivariate multiple regression model, where multivariate refers to the dependent variable and multiple pertains to the independent variables. In this case, several y 's are measured corresponding to each set of x 's. Each of $y_1, y_2, \dots, y_r$ is to be predicted by all of $x_1, x_2, \dots, x_p$.

The r observed values of the vector of y 's can be listed as rows in the following matrix:

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{r1} & y_{r2} & \cdots & y_{rp} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \vdots \\ \mathbf{F}_r \end{pmatrix},$$

Thus each row of $\mathbf{Y}$ contains the values of the p dependent variables measured on a subject. Each column of $\mathbf{Y}$ consists of the n observations on one of the p variables and therefore corresponds to that vector in the (univariate) regression model (10.3).

The n values of $x_1, x_2, \dots, x_q$ can be placed in a matrix that turns out to be the same as the $\mathbf{X}$ matrix in the multiple regression formulation in Section 10.2.1:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix}.$$

We assume that $\mathbf{X}$ is fixed from sample to sample.

Since each of the p y 's will depend on the x 's in its own way, each column of $\mathbf{Y}$ will need different β 's. Thus we have a column of β 's for each column of $\mathbf{Y}$, and these columns form a matrix $\mathbf{B} = (\beta_1, \beta_2, \dots, \beta_p)$. Our multivariate model is therefore

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where $\mathbf{Y}$ is $n \times p$, $\mathbf{X}$ is $n \times (q+1)$, and $\mathbf{B}$ is $(q+1) \times p$. The notation $\mathbf{U}$ (the appropriate version of $\mathbf{E}$) is adopted here because of its resemblance to ϵ .

We illustrate the multivariate model with $p = 2$, $q = 3$:

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{pmatrix} + \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ \vdots & \vdots \\ u_{n1} & u_{n2} \end{pmatrix}.$$

The model for the first column of $\mathbf{Y}$ is

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{21} \\ \beta_{31} \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{pmatrix},$$

and for the second column, we have

$$\begin{pmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{02} \\ \beta_{12} \\ \beta_{22} \\ \beta_{32} \end{pmatrix} + \begin{pmatrix} u_{12} \\ u_{22} \\ \vdots \\ u_{n2} \end{pmatrix}.$$

If analogy with the univariate case (in Section 10.2.1), additional assumptions that lead to good estimates are as follows:

1. $E(Y_i) = X_i\beta$ or $E(\mathbf{Y}) = X\beta$.
2. $\text{cov}(Y_i) = \Sigma$ for all $i = 1, 2, \dots, n$, where Y_i is the i th row of $\mathbf{Y}$.
3. $\text{cov}(Y_i, Y_j) = 0$ for all $i \neq j$.

Assumption 1 states that the linear model is correct and that no additional x 's are needed to predict the y 's. Assumption 2 asserts that each of the n observation vectors (rows) in $\mathbf{Y}$ has the same covariance matrix. Assumption 3 declares that observation vectors (rows of $\mathbf{Y}$) are uncorrelated with each other. Thus we assume that the y 's within an observation vector (row of $\mathbf{Y}$) are correlated with each other but independent of the y 's in any other observation vector.

The covariance matrix Σ in assumption 2 contains the variances and covariances of $y_{i1}, y_{i2}, \dots, y_{ip}$ in any $\mathbf{Y}_i$:

$$\text{cov}(\mathbf{Y}_i) = \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}.$$

The covariance matrix $\text{cov}(Y_i, Y_j) = 0$ in assumption 3 contains the covariances of each of $y_{i1}, y_{i2}, \dots, y_{ip}$ with each of $y_{j1}, y_{j2}, \dots, y_{jp}$:

$$\begin{pmatrix} \text{cov}(y_{11}, y_{21}) & \text{cov}(y_{11}, y_{22}) & \cdots & \text{cov}(y_{11}, y_{2p}) \\ \text{cov}(y_{21}, y_{11}) & \text{cov}(y_{21}, y_{22}) & \cdots & \text{cov}(y_{21}, y_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(y_{p1}, y_{11}) & \text{cov}(y_{p1}, y_{22}) & \cdots & \text{cov}(y_{p1}, y_{2p}) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

10.4.2 Least Squares Estimation in the Multivariate Model

By analogy with the univariate case in (10.5), we estimate β with

$$\hat{\beta} = (X'X)^{-1}X'Y. \quad (10.46)$$

We call $\hat{\beta}$ the least squares estimator for β because it "minimizes" $\mathbf{E} = \mathbf{E}'\mathbf{E}$, a matrix analogous to SSR :

$$\mathbf{E} = \mathbf{E}'\mathbf{E} = (\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}).$$

The matrix $\hat{\beta}$ minimizes $\mathbf{E}$ in the following sense. If we let β_0 be an estimate that may possibly be better than $\hat{\beta}$ and add $X\hat{\beta} - X\beta_0$ to $\mathbf{Y} - X\hat{\beta}$, we find that this adds a positive definite matrix to $\mathbf{E} = (\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta})$ (Krechet 1998, Section 7.4.2). Thus we cannot improve on $\hat{\beta}$. The least squares estimate $\hat{\beta}$ also minimizes the scalar quantity $(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta})$ and $(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta})$. Note that by $(\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}) = \sum_{i=1}^n \sum_{j=1}^p y_{ij}^2 - \sum_{j=1}^p y_{.j}^2$.

We note further that in the model $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{Z}$, there is a column of $\mathbf{B}$ corresponding to each column of $\mathbf{Y}$; that is, each y_j , $j = 1, 2, \dots, p$, is predicted differently by $x_1, x_2, \dots, x_q$. (This is illustrated in Section 15.4.1 for $p = 2$.) In the estimate $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, we have a similar pattern. The matrix product $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is multiplied by each column of $\mathbf{Y}$ (see (2.48)). Thus the j th column of $\hat{\mathbf{B}}$ is the usual least squares estimate $\hat{\beta}_j$ for the j th dependent variable y_j . To give this a more precise expression, let us denote the p columns of $\mathbf{Y}$ by $\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \dots, \mathbf{y}_{(p)}$ to distinguish them from the n rows y_i , $i = 1, 2, \dots, n$. Then

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \dots, \mathbf{y}_{(p)}) \\ &= \{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{(1)}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{(2)}, \dots, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{(p)}\} \\ &= \{\hat{\beta}_{(1)}, \hat{\beta}_{(2)}, \dots, \hat{\beta}_{(p)}\}.\end{aligned}\quad (15.65)$$

Example 15.4.2. The results of a planned experiment involving a chemical reaction are given in Table 15.1 (Ben and Sule, 1959).

The input (independent) variables are

$$x_1 = \text{temperature}, \quad x_2 = \text{concentration}, \quad x_3 = \text{time}.$$

Table 15.1. Chemical Reaction Data

Experiment Number	Yield Variables			Input Variables		
	y_1	y_2	y_3	x_1	x_2	x_3
1	48.8	44.9	11.2	162	20	3
2	53.8	55.3	11.2	162	20	8
3	27.7	21.3	12.7	162	38	3
4	21.7	28.8	18.8	162	38	8
5	19.9	55.6	18.2	172	25	3
6	13.8	58.0	22.6	172	25	8
7	12.2	58.6	24.5	172	38	3
8	4.5	52.4	28.8	172	38	8
9	18.5	56.9	21.3	167	27.5	6.5
10	6.8	35.4	28.8	177	27.5	6.5
11	27.8	46.9	18.7	177	27.5	6.5
12	18.8	27.5	22.2	167	32.5	6.5
13	28.3	55.0	18.3	167	32.5	6.5
14	9.9	58.9	28.8	167	27.5	6.5
15	25.8	50.5	22.1	167	27.5	5.5
16	14.1	61.1	23.8	177	28	6.8
17	18.2	62.9	28.7	177	28	6.8
18	13.8	60.0	22.1	168	34	7.2
19	18.8	60.6	18.3	168	34	7.2

The (field independent) variables are

- y_1 = percentage of unalloyed starting material,
- y_2 = percentage converted to the desired product,
- y_3 = percentage of unwanted by-product.

Using (18.46), the least squares estimator $\hat{\mathbf{B}}$ for the regression of (y_1, y_2, y_3) on (x_1, x_2, x_3) is given by

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ = \begin{pmatrix} 352.11 & -25.04 & -164.08 \\ -1.35 & .40 & .54 \\ -1.42 & .29 & .59 \\ -2.24 & 1.06 & 1.15 \end{pmatrix}.$$

Note that the first column of $\hat{\mathbf{B}}$ gives $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ for regression of y_1 on x_1, x_2, x_3 ; the second column of $\hat{\mathbf{B}}$ gives $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ for regression of y_2 on x_1, x_2, x_3 ; and so on. (2)

18.4.3 Properties of Least Squares Estimators $\hat{\mathbf{B}}$

The least squares estimator $\hat{\mathbf{B}}$ can be obtained without imposing the assumptions $E(\mathbf{y}) = \mathbf{X}\mathbf{\beta}$, $\text{cov}(\mathbf{y}_0) = \mathbf{I}$, and $\text{cov}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{0}$. However, when these assumptions hold, $\hat{\mathbf{B}}$ has the following properties:

1. The estimator $\hat{\mathbf{B}}$ is unbiased, that is, $E(\hat{\mathbf{B}}) = \mathbf{B}$. This means that if we took repeated random samples from the same population, the average value of $\hat{\mathbf{B}}$ would be $\mathbf{B}$.
2. The least squares estimator $\hat{\beta}_{ji}$ in $\hat{\mathbf{B}}$ has minimum variance among all possible linear unbiased estimators. This result is known as the Gauss-Markov theorem. The restriction to unbiased estimators is necessary to exclude trivial estimators such as a constant, which has variance equal to zero, but is of no interest. This minimum-variance property of least squares estimators is remarkable for its distributional generality; normality of the y 's is not required.
3. All $\hat{\beta}_{ji}$'s in $\hat{\mathbf{B}}$ are correlated with each other. This is due to the correlations among the x 's and among the y 's. The $\hat{\beta}$'s within a given column of $\hat{\mathbf{B}}$ are correlated because $x_1, x_2, \dots, x_q$ are correlated. If $x_1, x_2, \dots, x_q$ were uncorrelated, so that only the $\hat{\beta}$'s within each column of $\hat{\mathbf{B}}$ would be uncorrelated. Thus the relationship of the x 's to each other affects the relationship of the $\hat{\beta}$'s within each column to each other. On the other hand, the $\hat{\beta}$'s in each column are correlated with $\hat{\beta}$'s in other columns because $y_1, y_2, \dots, y_p$ are correlated.

Because of the correlations among the columns of $\mathbf{X}$, we need multivariate tests for hypotheses about $\mathbf{B}$. We cannot use an F -test from Section 10.2.5 on each column of $\mathbf{B}$, because these F -tests would not take into account the correlations or preserve the α -level. Some appropriate multivariate tests are given in Section 10.5.

10.4.4 An Estimator for Σ

By analogy with (10.8) and (10.9), an unbiased estimator of $\text{cov}(y_i) = \Sigma$ is given by

$$S_y = \frac{\mathbf{E}}{n - q - 1} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})}{n - q - 1} \quad (10.48)$$

$$= \frac{\mathbf{Y}'\mathbf{I} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{I}}{n - q - 1}. \quad (10.49)$$

With the denominator $n - q - 1$, S_y is an unbiased estimator of Σ ; that is, $E(S_y) = \Sigma$.

10.4.5 Model-Corrected for Means

If the x 's are centered by subtracting their means, we have the centered $\mathbf{X}$ matrix, as in (10.11),

$$\mathbf{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1q} - \bar{x}_q \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2q} - \bar{x}_q \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nq} - \bar{x}_q \end{pmatrix}.$$

The $\mathbf{B}$ matrix can be partitioned as

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1' \\ \mathbf{B}_2' \end{pmatrix} = \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0q} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1q} \\ \vdots & \vdots & & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qq} \end{pmatrix}.$$

By analogy with (10.14) and (10.15), the estimates are

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y}, \quad (10.50)$$

$$\hat{\mathbf{B}}_2' = \mathbf{f}' - \mathbf{f}'\hat{\mathbf{B}}_1, \quad (10.51)$$

where $\mathbf{f}' = (\beta_{01}, \beta_{02}, \dots, \beta_{0q})$ and $\mathbf{f}' = (\beta_{11}, \beta_{12}, \dots, \beta_{1q})$. These estimates give the same results as $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in (10.46).

As in (10.28), the estimate $\hat{\mathbf{B}}_1$ in (10.33) can be expressed in terms of sample covariance matrices. We multiply and divide (10.30) by $n - 1$ to obtain

$$\begin{aligned}\hat{\mathbf{B}}_1 &= (n - 1)(\mathbf{X}_1' \mathbf{X}_1)^{-1} \frac{\mathbf{X}_1' \mathbf{Y}}{n - 1} = \left(\frac{\mathbf{X}_1' \mathbf{X}_1}{n - 1} \right)^{-1} \frac{\mathbf{X}_1' \mathbf{Y}}{n - 1} \\ &= \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy},\end{aligned}\quad (10.32)$$

where $\mathbf{S}_{xx}$ and $\mathbf{S}_{xy}$ are blocks from the overall sample covariance matrix of $y_1, y_2, \dots, y_p, x_1, x_2, \dots, x_q$:

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}, \quad (10.33)$$

10.6 MULTIVARIATE MULTIPLE REGRESSION: HYPOTHESIS TESTS

In this section we extend the two tests of Section 10.2.5 to the multivariate y case. We assume the x 's are fixed and the y 's are multivariate normal. For other tests and confidence intervals, see Renscher (1998, Chapter 7).

10.6.1 Test of Overall Regression

We first consider the hypothesis that none of the x 's predict any of the y 's, which can be expressed as $\mathbf{H}_0: \mathbf{B}_1 = \mathbf{0}$, where $\mathbf{B}_1$ includes all rows of $\mathbf{B}$ except the first.

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_0' \\ \mathbf{B}_1' \end{pmatrix} = \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{pmatrix}.$$

We do not wish to include $\mathbf{B}_0' = \mathbf{0}$ in the hypothesis, because this would restrict all y 's to have intercepts of zero. The alternative hypothesis is $\mathbf{H}_1: \mathbf{B}_1 \neq \mathbf{0}$, which implies that we want to know if even one $\beta_{ij} \neq 0, j = 1, 2, \dots, p; i = 1, 2, \dots, q$.

The numerator of (10.49) suggests a partitioning of the total sum of squares and products matrix $\mathbf{Y}'\mathbf{Y}$,

$$\mathbf{Y}'\mathbf{Y} = (\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}) + \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y},$$

by analogy to (10.29), we subtract ss_T from both sides to avoid inclusion of $\mathbf{B}_0 = \mathbf{0}$:

$$\begin{aligned}\mathbf{Y}'\mathbf{Y} - ss_T \mathbf{I} &= (\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}) + (\hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - ss_T \mathbf{I}) \\ &= \mathbf{E} + \mathbf{H},\end{aligned}\quad (10.50)$$

The overall regression sum of squares and products matrix $\mathbf{B} = \mathbf{B}'\mathbf{X}'\mathbf{Y} = \alpha\mathbf{f}'\mathbf{f}'$ can be used to test $H_0: \mathbf{B}_0: \mathbf{B}_1 = \mathbf{O}$. The notation $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y}$ and $\mathbf{H} = \mathbf{B}'\mathbf{X}'\mathbf{I} - \alpha\mathbf{f}'\mathbf{f}'$ conforms with usage of $\mathbf{E}$ and $\mathbf{H}$ in Chapter 6.

As in Chapter 6, we can test $H_0: \mathbf{B}_1 = \mathbf{O}$ by means of

$$A = \frac{|\mathbf{B}_1|}{|\mathbf{E} + \mathbf{H}|} = \frac{|\mathbf{Y}'\mathbf{I} - \mathbf{B}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - \alpha\mathbf{f}'\mathbf{f}'|}, \quad (10.55)$$

which is distributed as $A_{p,q,p-q-1}$ when $H_0: \mathbf{B}_1 = \mathbf{O}$ is true, where p is the number of y 's and q is the number of x 's. We reject H_0 if $A \leq A_{\alpha,p,q,p-q-1}$. The likelihood ratio approach leads to the same test statistic. If $\mathbf{H}$ is "large" due to large values of the β_{jk} 's, then $(\mathbf{E} + \mathbf{H})$ would be expected to be sufficiently greater than $(\mathbf{E})$ so that A would lead to rejection. If $\mathbf{H}$ large, we mean that the regression sums of squares on the diagonal are large. Critical values for A are available in Table A.9 using $v_1 = q$ and $v_2 = n - q - 1$. Note that these degrees of freedom are the same as in the interaction test for regressions of y on $x_1, x_2, \dots, x_q$ in (10.24). The F and g^T approximations for A in (6.12) and (6.36) can also be used.

There are two alternative expressions for Wilks' A in (10.55). We can express A in terms of the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_r$ of $\mathbf{B}^{-1}\mathbf{H}$:

$$A = \prod_{i=1}^r \frac{1}{1 + \lambda_i}, \quad (10.56)$$

where $r = \min(p, q)$. Wilks' A can also be written in the form

$$A = \frac{|\mathbf{B}|}{|\mathbf{B}_{xx}| |\mathbf{B}_{yy}|}, \quad (10.57)$$

where $\mathbf{B}$ is partitioned as in (10.53):

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{yy} & \mathbf{B}_{yx} \\ \mathbf{B}_{xy} & \mathbf{B}_{xx} \end{pmatrix}.$$

The form of A in (10.57) is the same as in the test for independence of $\mathbf{y}$ and $\mathbf{x}$ given in (7.33), where $\mathbf{y}$ and $\mathbf{x}$ are both random vectors. In the present section, the y 's are random variables and the x 's are fixed. Thus $\mathbf{B}_{yy}$ is the sample covariance matrix of the y 's in the usual sense, whereas $\mathbf{B}_{xx}$ consists of an analogous mathematical expression involving the constant x 's (see comments about $\mathbf{B}_{xx}$ in Section 10.2.4).

By the symmetry of (10.57) in $\mathbf{x}$ and $\mathbf{y}$, A is distributed as $A_{q,p,q-p-1}$ as well as $A_{p,q,p-q-1}$. This is equivalent to property 3 in Section 6.1.3. Hence, if we regressed the x 's on the y 's, we would get a different $\mathbf{B}$ but would have the same value of A for the test.

The union-intersection test of $H_0: \mathbf{B}_1 = \mathbf{O}$ uses Roy's test statistic analogous to (6.28).

$$\theta^2 = \frac{\lambda_1}{1 + \lambda_1}, \quad (10.26)$$

where λ_1 is the largest eigenvalue of $\mathbf{K}^{-1}\mathbf{H}$. Upper percentage points θ_0 are given in Table A.10. The accompanying parameters are

$$r = \min(p, q), \quad m = \frac{1}{2}(4q - p - 1), \quad N = \frac{1}{2}(n - q - p - 2).$$

The hypothesis is rejected if $\theta > \theta_0$.

As in Section 8.1.5, Pillai's test statistic is defined as

$$V^{(P)} = \sum_{i=1}^r \frac{\lambda_i}{1 + \lambda_i}, \quad (10.27)$$

and the Lawley-Hotelling test statistic is given by

$$U^{(LH)} = \sum_{i=1}^r \lambda_i, \quad (10.28)$$

where $\lambda_1, \lambda_2, \dots, \lambda_r$ are the eigenvalues of $\mathbf{K}^{-1}\mathbf{H}$. The $V^{(P)}$, upper percentage points are found in Table A.11, indexed by r , m , and n as defined earlier in connection with Roy's test. Upper percentage points for $v_0 V^{(P)}/v_0$ (see Section 8.1.5) are provided in Table A.12, where $v_0 = q$ and $v_1 = n - q - 1$. Alternatively, we can use the F -approximations for $V^{(P)}$ and $U^{(LH)}$ in Section 8.1.5.

When H_0 is true, all four test statistics have probability α of rejecting; that is, they all have the same probability of a Type I error. When H_0 is false, the power ranking of the tests depends on the configuration of the population eigenvalues, as was noted in Section 8.1. (The sample eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_r$ from $\mathbf{K}^{-1}\mathbf{H}$ are estimates of the population eigenvalues.) If the population eigenvalues are equal or nearly equal, the power ranking of the tests is $V^{(P)} \geq A \geq U^{(LH)} \geq \theta$. If only one population eigenvalue is nonzero, the powers are reversed: $\theta \geq V^{(P)} \geq A \geq U^{(LH)}$.

In the case of a single nonzero population eigenvalue, the rank of $\mathbf{H}_1$ is 1. There are various ways this could occur; for example, $\mathbf{H}_1$ could have only one nonzero row, which would indicate that only one of the x 's predicts the y 's. On the other hand, a single nonzero column implies that only one of the y 's is predicted by the x 's. Alternatively, $\mathbf{H}_1$ would have rank 1 if all rows were equal or linear combinations of each other, manifesting that all x 's act alike in predicting the y 's. Similarly, all columns equal to each other or linear functions of each other would signify only one dimension in the y 's as they relate to the x 's.

Example 18.5.1. For the chemical data of Table 10.1, we test the overall regression hypothesis $H_0: \mathbf{B}_1 = \mathbf{0}$. The $\mathbf{E}$ and $\mathbf{H}$ matrices are given by

$$\mathbf{E} = \begin{pmatrix} 86.174 & -21.764 & -65.929 \\ -21.764 & 249.461 & -179.496 \\ -65.929 & -179.496 & 250.197 \end{pmatrix},$$

$$\mathbf{H} = \begin{pmatrix} 1707.158 & -492.512 & -986.584 \\ -492.512 & 151.602 & 283.607 \\ -986.584 & 283.607 & 581.688 \end{pmatrix}.$$

The eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ are .263103, .1004, and .0003. The parameters for use in obtaining critical values of the three test statistics are

$$\begin{aligned} n_E = q = 3, \quad n_H = n - p - 1 = 18 - 3 - 1 = 14, \\ s = \min(3, 3) = 3, \quad m = \frac{1}{2}(q + p - 2) = -\frac{1}{2}, \\ B = \frac{1}{2}(n - q - p - 2) = 5.5. \end{aligned}$$

Using the eigenvalues, we obtain the test statistics

$$\begin{aligned} \Lambda &= \prod_{i=1}^3 \frac{1}{1 + \lambda_i} = \frac{1}{1 + .263103} \frac{1}{1 + .1004} \frac{1}{1 + .0003} \\ &= .0093 < \lambda_{.05, 3, 14, 5.5} = .509, \\ \bar{\rho} &= \frac{\lambda_1}{1 + \lambda_1} = .563 > \bar{\rho}_{.05, 3, 0, 5.5} = .568, \\ V^{(0)} &= \sum_{j=1}^3 \frac{\lambda_j}{1 + \lambda_j} = 1.058 > V_{.05, 3, 14, 5.5}^{(0)} = 1.040, \\ L^{(0)} &= \sum_{i=1}^3 \lambda_i = .264423, \quad \frac{r_E L^{(0)}}{n_E} = .15214, \end{aligned}$$

which exceeds the .05 critical values, 0.935 (interpolated), from Table A.21 (see Section 8.1.5). Thus all four tests reject H_0 . Note that the critical values given for $\bar{\rho}$ and $V^{(0)}$ are conservative, since 0 was used in place of $-.5$ for m .

In this case, the first eigenvalue, .263103, completely dominates the other two. In Example 10.4.2, we obtained

$$\hat{\mathbf{B}}_1 = \begin{pmatrix} -1.58 & .40 & .31 \\ -1.42 & .28 & .30 \\ -2.28 & 1.03 & 1.19 \end{pmatrix}.$$

The columns are approximately proportional to each other, indicating that there is essentially only one dimension in the y 's as they are predicted by the x 's. A similar

statement can be made about the rows and the dimensionality of the β 's as they predict the y 's. (1)

10.3.2 Test on a Subset of the β 's

We consider the hypothesis that the y 's do not depend on the last k of the x 's, $\beta_{q-k+1}, \beta_{q-k+2}, \dots, \beta_q$. By this we mean that none of the y 's is predicted by any of these k x 's. To express this hypothesis, write the $\mathbf{B}$ matrix in partitioned form,

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_r \\ \mathbf{B}_d \end{pmatrix},$$

where, as in Section 1.3.2.5b, the subscript r denotes the subset of β_{ij} 's to be retained in the reduced model and d represents the subset of β_{ij} 's to be deleted if they are not significant predictors of the y 's. Thus $\mathbf{B}_d$ has k rows. The hypothesis can be expressed as

$$\mathbf{B}_d : \mathbf{B}_d = \mathbf{0}.$$

If $\mathbf{X}_r$ contains the columns of $\mathbf{X}$ corresponding to $\mathbf{B}_r$, then the reduced model is

$$\mathbf{Y} = \mathbf{X}_r \mathbf{B}_r + \mathbf{E}. \quad (10.81)$$

To compare the fit of the full model and the reduced model, we use the difference between the regression sum of squares and products matrix for the full model, $\hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}$, and regression sum of squares and products matrix for the reduced model, $\hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}$. By analogy to (10.26), this difference becomes our $\mathbf{H}$ matrix:

$$\mathbf{H} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}. \quad (10.82)$$

Thus the test of $\mathbf{B}_d : \mathbf{B}_d = \mathbf{0}$ is a full and reduced model test of the significance of $\beta_{q-k+1}, \beta_{q-k+2}, \dots, \beta_q$ above and beyond $\beta_0, \beta_1, \dots, \beta_{q-k}$.

To make the test, we use the $\mathbf{E}$ matrix based on the full model, $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}$. Then

$$\begin{aligned} \mathbf{E} + \mathbf{H} &= (\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}) + (\hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}) \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}, \end{aligned}$$

and Wilks' Λ -statistic is given by

$$\begin{aligned} \Lambda(\beta_{q-k+1}, \dots, \beta_q | \beta_0, \dots, \beta_{q-k}) &= \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} \\ &= \frac{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}|}. \end{aligned} \quad (10.83)$$

which is distributed as $\lambda_{p, k, n-q-k}$ when $\mathbf{B}_0: \mathbf{B}_k = \mathbf{0}$ is true. Critical values are available in Table A.9 with $v_1 = k$ and $v_2 = n - q - 1$. Note that these degrees of freedom for the multivariate y case are the same as for the univariate y case (multiple regression) in Section 10.2.8b. The F - and χ^2 -approximations for λ in (5.1.5) and (5.1.6) can also be used.

As implied in the notation $\lambda(x_{p-q+1}, \dots, x_p/x_1, \dots, x_{p-q})$, Wilks' λ in (10.63) provides a full and reduced model test. We can express it in terms of λ for the full model and a similar λ for the reduced model. In the denominator of (10.63), we have $\mathbf{Y}'\mathbf{I} - \mathbf{R}_0'\mathbf{X}_0'\mathbf{Y}$, which is the error matrix for the reduced model $\mathbf{Y} = \mathbf{X}_0\mathbf{B}_0 + \mathbf{E}$ in (10.51). This error matrix could be used in a test for the significance of overall regression in the reduced model, as in (10.59),

$$\lambda_0 = \frac{(\mathbf{Y}'\mathbf{Y} - \mathbf{R}_0'\mathbf{X}_0'\mathbf{Y})}{(\mathbf{Y}'\mathbf{Y} - \sigma_0^2\mathbf{I})}. \quad (10.64)$$

Since λ_0 in (10.64) has the same denominator as λ in (10.63), we recognize (10.63) as the ratio of Wilks' λ for the overall regression test in the full model to Wilks' λ for the overall regression test in the reduced model,

$$\begin{aligned} \lambda(x_{p-q+1}, \dots, x_p/x_1, \dots, x_{p-q}) &= \frac{(\mathbf{Y}'\mathbf{Y} - \mathbf{R}'\mathbf{X}'\mathbf{Y})}{(\mathbf{Y}'\mathbf{Y} - \mathbf{R}_0'\mathbf{X}_0'\mathbf{Y})} \\ &= \frac{(\mathbf{Y}'\mathbf{Y} - \mathbf{R}'\mathbf{X}'\mathbf{Y})}{(\mathbf{Y}'\mathbf{Y} - \sigma_0^2\mathbf{I})} \\ &= \frac{(\mathbf{Y}'\mathbf{Y} - \mathbf{R}'\mathbf{X}'\mathbf{Y})}{(\mathbf{Y}'\mathbf{Y} - \sigma_0^2\mathbf{I})} \\ &= \frac{\lambda_1}{\lambda_0}, \end{aligned} \quad (10.65)$$

where λ_1 is given by (10.65). In (10.65), we have a convenient computational device. We run the overall regression test for the full model and again for the reduced model and take the ratio of the resulting λ values.

The Wilks' λ in (10.63), comparing the full and reduced models, is similar in appearance to (5.1.27). However, in (5.1.27), the full and reduced models involve the dependent variables $y_1, y_2, \dots, y_p$ in MANOVA, whereas in (10.63), the reduced model is obtained by deleting $p - q$ of the independent variables $x_1, x_2, \dots, x_p$ in regression. The parameters of the Wilks' λ distribution are different in the two cases. Note that, in (5.1.27), some of the dependent variables were denoted by $x_1, \dots, x_k$ for convenience.

Test statistics due to Roy, Pillai, and Lawley-Hotelling can be obtained from the eigenvalues of $\mathbf{R}^{-1}\mathbf{H} = (\mathbf{Y}'\mathbf{Y} - \mathbf{R}'\mathbf{X}'\mathbf{Y})^{-1}(\mathbf{R}'\mathbf{X}'\mathbf{Y} - \mathbf{R}_0'\mathbf{X}_0'\mathbf{Y})$. Critical values or approximate tests for these three test statistics are based on $v_1 = k$, $v_2 = n - q - 1$, and

$$J = \min\{p, k\}, \quad M = \frac{1}{2}(k - p - 1), \quad N = \frac{1}{2}(n - p - k - 2).$$

Example 10.5.2. The chemical data in Table 10.1 originated from an experiment seeking to locate optimum operating conditions. Therefore, a second-order model is of interest, and we add x_1^2 , x_2^2 , x_3^2 , x_1x_2 , x_1x_3 , x_2x_3 to the variables x_1 , x_2 , x_3 considered in Example 10.5.1. There are now $q = 9$ independent variables, and we obtain an overall Wilks' Λ of

$$\Lambda = .00449 = \Lambda_{(9, 1, 1, 1, 1)} = .024,$$

where $v_1 = q = 9$ and $v_2 = n - q - 1 = 19 - 9 - 1 = 9$. Hence whether the six second-order variables add a significant amount to x_1 , x_2 , x_3 for predicting the y 's, we calculate

$$\Lambda = \frac{\Lambda_1}{\Lambda_0} = \frac{.00449}{.0312} = .0438 = \Lambda_{(9, 1, 1, 1, 1)} = .049,$$

where $v_1 = h = 6$ and $v_2 = n - q - 1 = 19 - 9 - 1 = 9$. In this case, $\Lambda_0 = .0312$ is from Example 10.5.1, in which we considered the model with x_1 , x_2 , and x_3 . Thus we reject $H_0: B_2 = \mathbf{0}$ and conclude that the second-order terms add significant predictability to the model. $\square$

10.6 MEASURES OF ASSOCIATION BETWEEN THE y 's AND THE x 's

The most widely used measures of association between two sets of variables are the canonical correlations, which are treated in Chapter 11. In this section, we review other measures of association that have been proposed.

In (10.34), we have $R^2 = \mathbf{b}_{xy}'\mathbf{b}_{xy}^{-1}\mathbf{b}_{yx}/\mathbf{b}_{yy}$ for the univariate y case. By analogy, we define an R^2 -like measure of association between $y_1, y_2, \dots, y_p$ and $x_1, x_2, \dots, x_q$ as

$$R_M^2 = \frac{(\mathbf{b}_{y1}\mathbf{b}_{y2}^{-1}\mathbf{b}_{y3})}{\mathbf{b}_{y3}}, \quad (10.66)$$

where $\mathbf{b}_{y1}$, $\mathbf{b}_{y2}$, $\mathbf{b}_{y3}$, and $\mathbf{b}_{y3}$ are defined in (10.55) and the subscript M indicates multivariate y 's.

By analogy to $r_{xy} = a_{xy}/a_{yy}$ in (3.15), Rabin and Duvall (1970) suggested

$$R^2 = \frac{r(\mathbf{b}_{y1}\mathbf{b}_{y2}^{-1})}{\sqrt{r(\mathbf{b}_{y3})r(\mathbf{b}_{y3}^{-1})}}. \quad (10.67)$$

Kato (1982) discussed the generalized correlation-determinant

$$\text{GCD} = \frac{|\mathbf{L}'\mathbf{b}_{y1}\mathbf{M}\mathbf{M}'\mathbf{b}_{y2}\mathbf{L}|}{|\mathbf{L}'\mathbf{b}_{y3}\mathbf{L}||\mathbf{M}'\mathbf{b}_{y3}\mathbf{M}|}$$

for various choices of the transformation matrices $\mathbf{L}$ and $\mathbf{M}$.

In Section 6.1.8, we introduced several measures of association that quantify the amount of relationship between the y 's and the clustering variables in a MANOVA context. These are even more appropriate here in the multivariate regression setting, where both the x 's and the y 's are continuous variables. The R^2 -like indices given in (6.41), (6.43), (6.46), (6.48), and (6.51) range between 0 and 1 and will be briefly reviewed in the remainder of this section. For more complete commentary, see Section 6.1.8.

The two measures based on Wilks' Λ are

$$\begin{aligned} r_{\Lambda}^2 &= 1 - \Lambda, \\ \lambda_{\Lambda} &= 1 - \Lambda^{1/q}, \end{aligned}$$

where $q = \min(p, q)$. A measure based on Roy's θ is provided by θ itself,

$$\theta_1^2 = \frac{\lambda_1}{1 + \lambda_1} = \theta,$$

where λ_1 is the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{B}$. This was identified in Section 6.1.8 as the square of the first canonical correlation (see Chapter 11) between the y 's and the grouping variables in MANOVA. In the multivariate regression setting, θ is the square of the first canonical correlation, r_1^2 , between the y 's and the x 's.

Measures of association based on the Lawley–Hoffding and Pillai statistics are given by

$$\begin{aligned} \Lambda_{LH} &= \frac{\mathbf{D}^{(2)}/q}{1 + \mathbf{D}^{(2)}/q}, \\ \lambda_P &= \frac{\mathbf{P}^{(2)}}{n}. \end{aligned} \quad (6.60)$$

By (6.48) and (6.49), λ_P in (6.60) is the average of the r squared canonical correlations, $r_1^2, r_2^2, \dots, r_r^2$.

Example 18.6. We use the chemical data of Table 10.1 to illustrate some measures of association. For the three dependent variables y_1 , y_2 , and y_3 and the three independent variables x_1 , x_2 , and x_3 , the partitioned covariance matrix is

$$\begin{aligned} \Sigma &= \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \\ &= \begin{pmatrix} \begin{array}{ccc|ccc} 89.38 & -28.57 & -38.03 & -41.95 & -9.48 & -3.37 \\ -28.57 & 21.28 & 8.78 & 11.89 & 1.68 & 3.05 \\ -38.03 & 8.78 & 43.27 & 24.14 & 6.43 & 1.97 \\ \hline -41.95 & 11.89 & 24.14 & 88.67 & 12.17 & -1.77 \\ -9.48 & 1.68 & 6.43 & 12.17 & 17.98 & 1.22 \\ -3.37 & 3.05 & 1.97 & -1.77 & 1.22 & 1.67 \end{array} \end{pmatrix}. \end{aligned}$$

from which we obtain R_{DF}^2 and RF directly using (10.56) and (10.57),

$$R_{\text{DF}}^2 = \frac{(R_{YX}R_{Y1}^2R_{Y2})}{(R_{Y1})} = .70829,$$

$$RF = \frac{(n-2)R_{Y1}^2R_{Y2}}{\sqrt{(n-2)(1-R_{Y1}^2)}} = .403.$$

Using the results in Example 10.5.3, we obtain

$$\begin{aligned} r_{Y1}^2 &= 1 - A = 1 - .2917 = .7082, \\ A_{12} &= 1 - A^{1/2} = 1 - A^{.5} = .378, \\ r_{12}^2 &= \frac{A_4}{1 + A_4} = .593, \\ A_{123} &= \frac{(1^{.593})A_4}{1 + 2^{.593}/3} = \frac{26.4221/3}{1 + 26.4221/3} = .498, \\ A_F &= \frac{1^{.593}}{2} = \frac{1.498}{2} = .749. \end{aligned}$$

We have general agreement among r_{Y1}^2 , A_{12} , r_{12}^2 , and A_{123} . But R_{DF}^2 , RF , and A_F do not appear to be measuring the same level of association, especially R_{DF}^2 . It appears that more study is needed before one or more of these measures can be universally recommended. $\square$

10.7 SUBSET SELECTION

As in the univariate y -case in Section 10.2.7, there may be more potential predictor variables (x 's) than are useful in a given situation. Some of the x 's may be redundant in the presence of the other x 's.

We may also be interested in deleting some of the y 's if they are not well predicted by any of the x 's. This would lead to further simplification of the model.

We present two approaches to subset selection: stepwise procedures and methods involving all possible subsets.

10.7.1 Stepwise Procedures

Subset selection among the x 's is discussed in Section 10.7.1a, while subset selection among the y 's in Section 10.7.1b.

10.7.1a Finding a Subset of the x 's

We begin with the *forward* selection procedure based on Wilks' Λ . At the first step, we test the regression of the y x 's on each x_j . There will be two rows in the R matrix,

a row containing the intercepts and a row corresponding to x_{1j} :

$$\mathbf{B}_j = \begin{pmatrix} \beta_{0j} & \beta_{1j} & \cdots & \beta_{p_jj} \\ \beta_{1j} & \beta_{2j} & \cdots & \beta_{pj} \end{pmatrix}.$$

We use the overall regression test statistic (10.55),

$$A(x_j) = \frac{(\mathbf{Y}'\mathbf{Y} - \mathbf{B}_j'\mathbf{X}_j'\mathbf{Y})}{[\mathbf{Y}'\mathbf{Y} - n\bar{y}^2]},$$

which is distributed as $F_{p_j, 1, n-p_j}$, since $\mathbf{B}_j$ has two rows and $\mathbf{X}_j$ has two columns. After calculating $A(x_j)$ for each j , we choose the variable with minimum $A(x_j)$. Note that at the first step, we are not testing each variable in the presence of the others. We are searching for the variable x_j that best predicts the y 's by itself, not alone and beyond the other x 's.

At the second step, we seek the variable yielding the smallest partial A for each x adjusted for the variable first entered, where the partial A -statistic is given by (10.65). After one variable has entered, (10.65) becomes

$$A(x_j|x_1) = \frac{A(x_j, x_1)}{A(x_1)}, \quad (10.66)$$

where x_1 denotes the variable entered at the first step. We calculate (10.66) for each $x_j \neq x_1$ and choose the variable that minimizes $A(x_j|x_1)$.

If we denote the second variable to enter by x_2 , then at the third step we seek the x_j that minimizes

$$A(x_j|x_1, x_2) = \frac{A(x_j, x_1, x_2)}{A(x_1, x_2)}, \quad (10.67)$$

by property 7 in Section 6.1.3, the partial Wilks' A -statistic transforms to an exact F since $v_0 = k = 1$ at each step.

After m variables have been selected, the partial A would have the following form at the next step:

$$A(x_j|x_1, x_2, \dots, x_m) = \frac{A(x_j, x_2, \dots, x_m, x_1)}{A(x_1, x_2, \dots, x_m)}, \quad (10.71)$$

where the first m variables to enter are denoted $x_1, x_2, \dots, x_m$, and x_j is a candidate variable chosen from among the $q - m$ remaining variables. At this step, we would choose the x_j that minimizes (10.71). The partial Wilks' A in (10.71) is distributed as $F_{p_j, 1, n-m-p_j}$ and, by Table 6.1, transforms to a partial F -statistic distributed as $F_{p_j, n-m-p_j}$. [These distributions hold for a variable x_j chosen before seeing the data but not for the x_j that minimizes (10.71) or maximizes the corresponding partial F .]

The procedure continues, bringing in the "best" variable at each step, until a step is reached at which the minimum partial A exceeds a predetermined threshold value or

equivalently, the associated partial F falls below a predetermined value. Alternatively, the stopping rule can be cast in terms of the p -value of the partial A or F . If the smallest p -value at some step exceeds a predetermined value, the procedure stops.

For each x_j , there corresponds an entire row of the $\mathbf{B}$ matrix because x_j has a coefficient for each of the p y 's. Thus if it contains a significantly positive constant of the y 's, it should be retained.

The stepwise procedure is an extension of forward selection. Each time a variable enters, all the variables that have entered previously are checked by a partial A or F to see if the least "significant" one is now redundant and can be deleted.

The backward elimination procedure begins with all x 's (all rows of $\mathbf{B}$) included in the model and deletes one at a time using a partial A or F . At the first step, the partial A for each x_j is

$$A(x_1, \dots, x_p, x_{j+1}, \dots, x_q) = \frac{A(x_1, \dots, x_p)}{A(x_1, \dots, x_p, x_{j+1}, \dots, x_q)}, \quad (10.72)$$

which is distributed as $A_{p+1, q-p-1}$ and can be converted to $F_{p+1, q-p}$ by Table 6.1. The variable with largest A or smallest F is deleted. At the second step, a partial A or F is calculated for each of the $p-1$ remaining variables, and again the least important variable in the presence of the others is eliminated. This process continues until a step is reached at which the largest A or smallest F is significant, indicating that the corresponding variable is apparently not redundant in the presence of the others. Some predetermined p -value or threshold value of A or F is used to determine a stopping rule.

If no automated program is available for subset selection in the multivariate case, a forward-selection or backward elimination procedure could be carried out by means of a rather simple set of commands based on (10.71) or (10.72).

A sequential procedure such as stepwise selection will often fail to find the optimum subset, especially if a large pool of predictor variables is involved. However, the value of Wilks' A found by stepwise selection may not be far from that for the optimum subset.

The remarks in the final paragraph of Section 10.2.7b are pertinent in the multivariate context as well. Two predictors of the y 's in the population may be overlooked because of inflated error variances, or, on the other hand, x 's that are not true predictors may appear to be so in the sample. The latter problem can be severe for small sample sizes (Rendall and Pao 1985).

10.2.7b Finding a Subset of the y 's

After a subset of x 's has been found, the researcher may wish to know if those x 's predict all p of the y 's. If some of the y 's do not relate to any of the x 's, they could be deleted from the model to achieve a better simplification. The y 's can be checked for redundancy in a manner analogous to the stepwise discriminant approach presented in Sections 6.11.2 and 6.9, which finds subsets of dependent variables using a full and reduced model Wilks' A for the y 's. The partial A statistic for adding or deleting a y is similar to (10.69), (10.70), or (10.71), except that dependent variables are involved

rather than independent variables. Thus we add a y at the third step of a *forward* selection procedure, for example, where the first two variables already entered are denoted as y_1 and y_2 , we use (10.28) to obtain

$$R(y_3|y_1, y_2) = \frac{R(y_1, y_2, y_3)}{R(y_1, y_2)} \quad (10.28)$$

for each $y_3 \neq y_1$ or y_2 and we choose for y_3 that minimizes $R(y_3|y_1, y_2)$. [In (10.28) the dependent variable of interest was denoted by x instead of y_1 as here.] Similarly, if three y 's, designated as y_1, y_2 , and y_3 , were "in the model" and we were checking the feasibility of adding y_4 , the partial R -statistic would be

$$R(y_4|y_1, y_2, y_3) = \frac{R(y_1, y_2, y_3, y_4)}{R(y_1, y_2, y_3)}, \quad (10.29)$$

which is distributed as $R_{(q-p-1, q-p-4)}$, where q is the number of x 's presently in the model and 4 is the number of y 's presently in the model. The two Wilks' Λ values in the numerator and denominator of the right side of (10.29), $R(y_1, y_2, y_3, y_4)$ and $R(y_1, y_2, y_3)$, are obtained from (10.25). Since $p = 1$, $R_{(q-p-1, q-p-4)}$ in (10.29) transforms to $F_{(q-p-1, q-p-4)}$ (see Table 6.3).

In the first step of a *backward* elimination procedure, we would delete the y_1 that maximizes

$$R(y_1|y_2, \dots, y_{j-1}, y_{j+1}, \dots, y_p) = \frac{R(y_1, \dots, y_p)}{R(y_2, \dots, y_{j-1}, y_{j+1}, \dots, y_p)}, \quad (10.30)$$

which is distributed as $R_{(q-p, q-p-1)}$. In this case, $r_0 = q$ and $r_1 = n - q - 1$ so that the distribution of (10.30) is $R_{(q-p, q-p-1)}$, which can be transformed to an exact F . Note that q , the number of x 's, may have been reduced in a *subset* selection on the x 's, as in Section 10.7 (a). Similarly, p is the number of y 's and will decrease in subsequent steps.

Stopping rules for either the forward or backward approach could be defined in terms of p -values or threshold values of Λ or the equivalent F . A stepwise procedure could be defined as a modification of the forward approach.

If a software program is available that tests the significance of one x as in (10.72), it can be adapted to test one y as in (10.31) by use of property 3 of Section 6.1.3. The distribution of $R_{(q-p, q-p-1)}$ is the same as that of $R_{(n-p, n-p-1-p)}$, which can also be seen from the symmetry of Λ in (10.27),

$$\Lambda = \frac{S}{(S_{xx} - S_{xy}^2)},$$

which is distributed as $R_{(q-p, q-p-1)}$ or, equivalently, as $R_{(q-p, p-p-1)}$. Thus we can reverse the y 's and x 's, we list the x 's as dependent variables in the program, and the y 's as independent variables. Then the test of a single y in (10.31) or (10.32) can be carried out using (10.72) without any adjustment. The partial R -statistic in (10.72)

is distributed as $A_{p-1, p-q-1}$. If we interchange p and q , because the y 's and x 's are interchanged in dependent and independent variables, this becomes $A_{q-1, p-p-1}$. By property 3 of Section 6.1.3 (repeated above), this is equivalent to $A_{q-1, p-p-1+1+q} = A_{1, q, p-p+q}$, which is the distribution of (10.75).

10.7.2 All Possible Subsets

In Section 10.7.1a, we discussed the criteria R_p^2 , s_p^2 , and C_p for comparing all possible subsets of x 's to predict a variate y in multiple regression, where $p-1$ denotes the number of x 's in a subset selected from a pool of $k-1$ available independent variables. We now discuss some matrix analogues of these criteria for the multivariate y case, as suggested by Mallows (1973) and Späth, Contourinides, and Tashiro (1982).

In this section, in order to conform with notation in the literature, we will use p for the number of columns in $\mathbf{X}$ (and the number of rows in $\mathbf{Y}$), rather than for the number of y 's. The number of y 's will be indicated by m .

We now extend the three criteria R_p^2 , s_p^2 , and C_p to analogous matrix expressions $\mathbf{R}_p^2$, $\mathbf{S}_p$, and $\mathbf{C}_p$. These can be reduced to scalar form using trace or determinant.

1. $\mathbf{R}_p^2$. In the univariate y case, R_p^2 for a $(p-1)$ -variable subset of x 's is defined by (3.6.32) as

$$R_p^2 = \frac{\mathbf{R}_p \mathbf{X}_p' \mathbf{X}_p - n\beta^2}{n\beta - n\beta^2}.$$

A direct extension of R_p^2 for the multivariate y case is given by the matrix

$$\mathbf{R}_p^2 = (\mathbf{Y}'\mathbf{Y} - n\beta\beta')^{-1} (\mathbf{R}_p \mathbf{X}_p' \mathbf{Y} - n\beta\beta'), \quad (10.76)$$

where $p-1$ is the number of x 's selected from the $k-1$ available x 's. To convert $\mathbf{R}_p^2$ to scalar form, we can use $\text{tr}(\mathbf{R}_p^2)/m$, in which we divide by m , the number of y 's, so that $0 \leq \text{tr}(\mathbf{R}_p^2)/m \leq 1$. As in the univariate case, we identify the subset that maximizes $\text{tr}(\mathbf{R}_p^2)/m$ for each value of $p = 1, 2, \dots, k$. The criterion $\text{tr}(\mathbf{R}_p^2)/m$ does not attain its maximum until p reaches k , but we look for the value of p at which further increases are deemed unimportant. We could also use $|\mathbf{R}_p^2|$ in place of $\text{tr}(\mathbf{R}_p^2)/m$.

2. $\mathbf{S}_p$. A direct extension of the univariate criterion $s_p^2 = \text{MSR}_p = \text{SSR}_p/(m-p)$ is provided by

$$\mathbf{S}_p = \frac{\mathbf{E}_p}{n-p}, \quad (10.77)$$

where $\mathbf{E}_p = \mathbf{Y}'\mathbf{Y} - \mathbf{R}_p' \mathbf{X}_p' \mathbf{Y}$. To convert to scalar, we can use $\text{tr}(\mathbf{S}_p)$ or $|\mathbf{S}_p|$, either of which will behave in an analogous fashion to s_p^2 in the univariate case. The results in

Section 10.2. To apply here as well, one may wish to select the subset with minimum value of $\text{tr}(\hat{\Sigma}_{\beta_0})$ or perhaps the subset with smallest p such that $\text{tr}(\hat{\Sigma}_{\beta_0}) = \text{tr}(\hat{\Sigma}_{\beta_0})$. A similar application could be made with $\hat{\Sigma}_{\beta_0}$.

3. C_p . To extend the C_p criterion to the multivariate y case, we write the model under consideration as

$$\mathbf{Y} = \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon},$$

where $p - 1$ is the number of x 's in the subset and $k - 1$ is the number of x 's in the "full model." The predicted values of the y 's are given by

$$\hat{\mathbf{Y}} = \mathbf{X}_p \hat{\boldsymbol{\beta}}_p.$$

We are interested in predicted values of the observation vectors, namely $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_n$, which are given by the rows of $\hat{\mathbf{Y}}$:

$$\begin{pmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \\ \hat{\mathbf{y}}_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{p1}' \\ \mathbf{x}_{p2}' \\ \vdots \\ \mathbf{x}_{pn}' \end{pmatrix} \hat{\boldsymbol{\beta}}_p = \begin{pmatrix} \mathbf{x}_{p1}' \hat{\boldsymbol{\beta}}_p \\ \mathbf{x}_{p2}' \hat{\boldsymbol{\beta}}_p \\ \vdots \\ \mathbf{x}_{pn}' \hat{\boldsymbol{\beta}}_p \end{pmatrix}.$$

In general, the predicted vectors $\hat{\mathbf{y}}_i$ are biased estimators of $E(\mathbf{y}_i)$ in the correct model, because we are examining many competing models for which $E(\hat{\mathbf{y}}_i) \neq E(\mathbf{y}_i)$. In place of the univariate expected squared error $E[(\hat{y}_i - E(y_i))^2]$ in (10.32) and (10.38), we define a matrix of expected squares and products of errors, $E(\hat{\mathbf{y}}_i - E(\mathbf{y}_i))(E(\hat{\mathbf{y}}_i - E(\mathbf{y}_i))'$. We then add and subtract $E(\hat{\mathbf{y}}_i)$ to obtain (see Problem 10.8)

$$\begin{aligned} E(\hat{\mathbf{y}}_i - E(\mathbf{y}_i))(E(\hat{\mathbf{y}}_i - E(\mathbf{y}_i))' &= E(\hat{\mathbf{y}}_i - E(\hat{\mathbf{y}}_i) + E(\hat{\mathbf{y}}_i) - E(\mathbf{y}_i))(E(\hat{\mathbf{y}}_i - E(\hat{\mathbf{y}}_i) + E(\hat{\mathbf{y}}_i) - E(\mathbf{y}_i))' \\ &= E(\hat{\mathbf{y}}_i - E(\hat{\mathbf{y}}_i))(E(\hat{\mathbf{y}}_i - E(\hat{\mathbf{y}}_i))' + E(\hat{\mathbf{y}}_i - E(\hat{\mathbf{y}}_i))E(E(\hat{\mathbf{y}}_i) - E(\mathbf{y}_i))' \\ &\quad + E(E(\hat{\mathbf{y}}_i) - E(\mathbf{y}_i))E(E(\hat{\mathbf{y}}_i) - E(\mathbf{y}_i))'. \end{aligned} \quad (10.76)$$

By analogy to (10.38), our C_p matrix will be an estimate of the sum of (10.76), multiplied by $\mathbf{I}^{-1}$ for standardization.

We first find an expression for $\text{var}(\hat{\mathbf{y}}_i)$, which for convenience we write in row form,

$$\text{var}(\hat{\mathbf{y}}_i') = \text{var}(\mathbf{x}_{pi}' \hat{\boldsymbol{\beta}}_p) = \text{var}(\mathbf{x}_{pi}' \hat{\boldsymbol{\beta}}_{p(1)}, \mathbf{x}_{pi}' \hat{\boldsymbol{\beta}}_{p(2)}, \dots, \mathbf{x}_{pi}' \hat{\boldsymbol{\beta}}_{p(k)})$$

where $\mathbf{B} = (\hat{\boldsymbol{\beta}}_{p(1)}, \hat{\boldsymbol{\beta}}_{p(2)}, \dots, \hat{\boldsymbol{\beta}}_{p(k)})$, as in (10.47). This can be written as

$$\begin{aligned} \text{cov}(\hat{\beta}_i') &= \begin{pmatrix} \sigma_{11}x'_{ip}(\mathbf{X}_p\mathbf{X}_p)^{-1}x_{ip} & \cdots & \sigma_{1n}x'_{ip}(\mathbf{X}_p\mathbf{X}_p)^{-1}x_{ip} \\ \vdots & & \vdots \\ \sigma_{n1}x'_{ip}(\mathbf{X}_p\mathbf{X}_p)^{-1}x_{ip} & \cdots & \sigma_{nn}x'_{ip}(\mathbf{X}_p\mathbf{X}_p)^{-1}x_{ip} \end{pmatrix} \quad (10.78) \\ &= x'_{ip}(\mathbf{X}_p\mathbf{X}_p)^{-1}x_{ip}\Sigma, \end{aligned}$$

where n is the number of y 's and $\Sigma = \text{cov}(y_i)$ (see Problem 13.5). As in (10.40) (see also Problem 10.5), we can sum over the n observations and use (3.65) to obtain

$$\begin{aligned} \sum_{i=1}^n \text{cov}(\hat{\beta}_i') &= \sum_{i=1}^n x'_{ip}(\mathbf{X}_p\mathbf{X}_p)^{-1}x_{ip}\Sigma \\ &= \Sigma \sum_{i=1}^n x_{ip}(\mathbf{X}_p\mathbf{X}_p)^{-1}x_{ip} = p\Sigma. \end{aligned} \quad (10.80)$$

To sum the second term on the right of (10.78), we have, by analogy to (10.42),

$$\sum_{i=1}^n \text{bias in } \hat{\beta}_i \text{ bias in } \hat{\beta}_i' = (n-p)\text{tr}(\mathbf{B}_p - \Sigma), \quad (10.81)$$

where $\mathbf{B}_p$ is given by (13.77).

Now by (10.80) and (10.81), we can sum (13.76) and multiply by Σ^{-1} to obtain the matrix of total expected squares and products of y 's standardized by Σ^{-1} ,

$$\begin{aligned} \Sigma^{-1} \sum_{i=1}^n \mathcal{E}(\hat{\beta}_i - \mathcal{E}(\hat{\beta}_i)(\hat{\beta}_i - \mathcal{E}(\hat{\beta}_i))' &= \Sigma^{-1}[p\Sigma + (n-p)\mathcal{E}(\mathbf{B}_p - \Sigma)] \\ &= p\mathbf{I} + (n-p)(\Sigma^{-1}\mathcal{E}(\mathbf{B}_p - \Sigma)). \end{aligned} \quad (10.82)$$

Using $\mathbf{B}_p = \mathbf{B}_p/(n-p)$, the sample covariance matrix based on all $p+1$ variables, as an estimate of Σ , we can estimate (10.82) by

$$\mathbf{C}_p = p\mathbf{I} + (n-p)(\mathbf{B}_p^{-1}(\mathbf{B}_p - \mathbf{B}_p)) \quad (10.83)$$

$$= \mathbf{B}_p^{-1}\mathbf{B}_p + (2p-n)\mathbf{I} \quad \text{[by (10.77)],} \quad (10.84)$$

which is the form suggested by Mallows (1973). We can use $\text{tr}(\mathbf{C}_p)$ or $|\mathbf{C}_p|$ to induce this to a scalar. But if $2p-n$ is negative, $|\mathbf{C}_p|$ can be negative, and Spadix, Cook, Scruiven, and Troskie (1983) suggested a modification of $|\mathbf{C}_p|$,

$$|\mathbf{C}_p| = |\mathbf{B}_p^{-1}\mathbf{B}_p|, \quad (10.85)$$

which is always positive.

To find the optimal subset of x 's for each value of p , we could examine all possible subsets (or use a computational scheme such as that of Furnival and Wilson (1974)) and look for the "best" C_p matrix for each p . In (10.82), we see that when the bias is 0, the "population C_p " is equal to pI . Thus we seek a C_p that is "small" and near pI . In terms of trace, we seek for C_p close to pm , where m is the number of y 's in the vector of measurements that is, $\text{tr}(\mathbf{I}) = m$.

To find a "target" value for (10.82), we write $\mathbf{K}_p^{-1}\mathbf{K}_p$ in terms of C_p from (10.84),

$$\mathbf{K}_p^{-1}\mathbf{K}_p = \frac{C_p + (n - 2p)\mathbf{I}}{n - 1}. \quad (10.86)$$

When the bias is 0, we have $C_p = pI$, and (10.86) becomes

$$\mathbf{K}_p^{-1}\mathbf{K}_p = \frac{n - p}{n - 1}\mathbf{I}, \quad (10.87)$$

whence, by (2.87),

$$(\mathbf{K}_p^{-1}\mathbf{K}_p)^m = \left(\frac{n - p}{n - 1}\right)^m. \quad (10.88)$$

Thus we seek subsets such that

$$\text{tr}(C_p) \leq pm \quad \text{or} \quad (\mathbf{K}_p^{-1}\mathbf{K}_p) \leq \left(\frac{n - p}{n - 1}\right)^m.$$

In summary, when examining all possible subsets, any or all of the following criteria may be useful in finding the single best subset or the best subset for each p :

$$\text{tr}(\mathbf{K}_p^2), \quad |\mathbf{K}_p|, \quad \text{tr}(\mathbf{K}_p), \quad |\mathbf{K}_p|, \quad \text{tr}(C_p), \quad (\mathbf{K}_p^{-1}\mathbf{K}_p).$$

10.6 MULTIVARIATE REGRESSION: RANDOM x 's

In Sections 10.4 and 10.5, it was assumed that the x 's were fixed and would have the same values in repeated sampling. In many applications, the x 's are random variables. In such a case, the values of $x_1, x_2, \dots, x_k$ are not under the control of the experimenter but occur randomly along with $y_1, y_2, \dots, y_p$. On each subject, we observe $p + g$ values in the vector $(y_1, y_2, \dots, y_{p+g}, x_1, x_2, \dots, x_k)$.

If we assume that $Cy_1, y_2, \dots, y_p, x_1, x_2, \dots, x_k$ has a multivariate normal distribution, then all estimates and tests are the same (analogous to the fixed- x case) [for details, see Bancroft (1978, Section 7.7)]. Thus there is no essential difference in our procedures between the fixed- x case and the random- x case.

PROBLEMS

10.1 Show that $\sum_{i=1}^n (y_i - \mathbf{x}_i' \hat{\beta})^2 = (y - \mathbf{X}\hat{\beta})'(y - \mathbf{X}\hat{\beta})$ as in (10.6).

10.2 Show that $\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$ as in (10.10).

10.3 Show that $\sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 = 0$ as in (10.15).

10.4 Show that $E(y_i) - E(y_i)^2 = E(y_i) - E(y_i)^2 + (E(y_i) - E(y_i))^2$ as in (10.37).

10.5 Show that $\sum_{i=1}^n \text{var}(y_i)/n^2 = n[\mathbf{K}_p' \mathbf{X}_p' \mathbf{X}_p]^{-1} \mathbf{X}_p'$ as in (10.40).

10.6 Show that the alternative form of $\mathbf{C}_p$ in (10.42) is equal to the original form in (10.44).

10.7 Show that (10.46) is the same as (10.40), that is, $(Y - \mathbf{X}\hat{\beta})'(Y - \mathbf{X}\hat{\beta}) = \mathbf{Y}'\mathbf{I} - \hat{\beta}'\mathbf{X}'\mathbf{I}$.

10.8 Show that

$$\begin{aligned} E(y_i) - E(y_i | \hat{\beta}_i) &= E(y_i) \mathbf{I}' \\ E(y_i) - E(y_i | \hat{\beta}_i) &= E(y_i) \mathbf{I}' + (E(y_i) - E(y_i | \hat{\beta}_i)) E(y_i) - E(y_i) \mathbf{I}', \end{aligned}$$

thus verifying (10.78).

10.9 Show that $\text{cov}(y_i')$ has the form given in (10.79).

10.10 Show that the two forms of $\mathbf{K}_p$ in (10.37) and (10.44) are equal.

10.11 Explain why $(\mathbf{K}_p^{-1} \mathbf{K}_p) = \mathbf{I}$, as claimed following (10.55).

10.12 Show that $\mathbf{K}_p^{-1} \mathbf{K}_p = [\mathbf{K}_p]' + (n - 2p)(\mathbf{I})/(n - d)$, as in (10.56), where $\mathbf{K}_p$ is given in (10.45).

10.13 Show that if $\mathbf{C}_p = p\mathbf{I}$, then $\mathbf{K}_p^{-1} \mathbf{K}_p = (n - p)/(n - d)\mathbf{I}$ as in (10.57).

10.14 Use the diabetes data of Table 5.6.

- Find the least squares estimate $\hat{\mathbf{B}}$ for the regression of (y_0, y_1) on (x_0, x_1, x_2) .
- Test the significance of overall regression using all four test statistics.
- Determine what the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ reveal about the essential rank of $\mathbf{B}_1$ and the implications of this rank, such as the relative power of the four tests.
- Test the significance of each of x_1 , x_2 , and x_3 adjusted for the other two x 's.

10.15 Use the worm data of Table 5.7.

- Find the least squares estimate $\hat{\mathbf{B}}$ for the regression of (y_0, y_1) on (x_1, x_2) .
- Test the significance of overall regression using all four test statistics.

- (c) Determine what the eigenvalues of $\mathbf{B}^{-1}\mathbf{B}$ reveal about the essential rank of $\mathbf{B}_1$ and the implications of this rank, such as the relative power of the four tests.
- (d) Test the significance of x_1 adjusted for x_2 and of x_2 adjusted for x_1 .

18.16 Use the glucose data of Table 3.8.

- (a) Find the least squares estimate $\hat{\mathbf{B}}$ for the regression of (y_1, y_2, y_3) on (x_1, x_2, x_3) .
- (b) Test the significance of overall regression using all four test statistics.
- (c) Determine what the eigenvalues of $\mathbf{B}^{-1}\mathbf{B}$ reveal about the essential rank of $\mathbf{B}_1$ and the implications of this rank, such as the relative power of the four tests.
- (d) Test the significance of each of x_1, x_2 , and x_3 adjusted for the other two x 's.
- (e) Test the significance of each y adjusted for the other two by using (16.75).

18.17 Use the Seibin data of Table 7.1.

- (a) Find the least squares estimate $\hat{\mathbf{B}}$ for the regression of (y_1, y_2) on $(x_1, x_2, \dots, x_6)$ and test for significance.
- (b) Test the significance of (y_1, y_2) adjusted for the other x 's.
- (c) Test the significance of (x_4, x_5, x_6) adjusted for the other x 's.
- (d) Test the significance of (x_1, x_2, x_3) adjusted for the other x 's.

18.18 Use the Seibin data of Table 7.1.

- (a) Do a stepwise regression to select a subset of $x_1, x_2, \dots, x_6$ that adequately predicts (y_1, y_2) .
- (b) After selecting a subset of x 's, use the methods of Section 10.7.1b to check if either of the y 's can be deleted.

18.19 Use the temperature data of Table 7.1.

- (a) Find the least squares estimate $\hat{\mathbf{B}}$ for the regression of (y_1, y_2, y_3) on (x_1, x_2, x_3) and test for significance.
- (b) Find the least squares estimate $\hat{\mathbf{B}}$ for the regression of (y_1, y_2, y_3) on $(x_1, \dots, x_6)$ and test for significance.
- (c) Find the least squares estimate $\hat{\mathbf{B}}$ for the regression of (y_1, y_2, y_3) on $(x_1, \dots, x_6)$ and test for significance.

18.20 Using the temperature data of Table 7.1, carry out a stepwise regression to select a subset of $x_1, x_2, \dots, x_6$ that adequately predicts (y_1, y_2) .

Canonical Correlation

11.1 INTRODUCTION

Canonical correlation analysis is concerned with the amount of (linear) relationship between two sets of variables. We often measure two types of variables on each respondent—for example, a set of aptitude variables and a set of achievement variables, a set of personality variables and a set of ability measures, a set of price indices and a set of production indices, a set of student behaviors and a set of teacher behaviors, a set of psychological attributes and a set of physiological attributes, a set of ecological variables and a set of environmental variables, a set of academic achievement variables and a set of measures of job success, a set of closed-book exam scores and a set of open-book exam scores, and a set of personality variables of freshmen students and the same variables on the same students as seniors.

11.2 CANONICAL CORRELATIONS AND CANONICAL VARIABLES

We assume that two sets of variables $\mathbf{y}' = (y_1, y_2, \dots, y_p)$ and $\mathbf{x}' = (x_1, x_2, \dots, x_q)$ are measured on the same sampling unit. We denote the two sets of variables as $\mathbf{y}$ and $\mathbf{x}$ to conform to notation in Chapters 5, 7, and 18. In Section 7.6.2, we discussed the hypothesis that $\mathbf{y}$ and $\mathbf{x}$ were independent. In this chapter, we consider a measure of overall correlation between $\mathbf{y}$ and $\mathbf{x}$.

Canonical correlation is an extension of multiple correlation, which is the correlation between one y and several x 's (see Section 10.1.6). Canonical correlation analysis is often a useful complement to a multivariate regression analysis.

We first review multiple correlation. The sample variances and correlations among $y, x_1, x_2, \dots, x_q$ can be summarized in the matrices

$$\mathbf{S} = \begin{pmatrix} s_y^2 & \mathbf{s}_{yx}' \\ \mathbf{s}_{yx} & \mathbf{S}_{xx} \end{pmatrix}, \quad (11.1)$$

$$\mathbf{R} = \begin{pmatrix} 1 & \mathbf{r}_{yx}' \\ \mathbf{r}_{yx} & \mathbf{R}_{xx} \end{pmatrix}, \quad (11.2)$$

where $\mathbf{r}'_{y0} = (r_{y1}, r_{y2}, \dots, r_{yq})$ contains the sample correlations of y with $x_1, x_2, \dots, x_q$ and $\mathbf{R}_{xx}$ is the sample covariance matrix of the x 's [see (10.16)]. The partitioned matrix $\mathbf{R}$ is defined analogously: $\mathbf{r}'_{x0} = (r_{x1}, r_{x2}, \dots, r_{xq})$ contains the sample correlations of x with $x_1, x_2, \dots, x_q$, and $\mathbf{R}_{xx}$ is the sample covariance matrix of the x 's [see (10.18)].

By (10.54), the squared multiple correlation between y and the x 's can be computed from the partitioned covariance matrix (11.1) or correlation matrix (11.2) as follows:

$$R^2 = \frac{\mathbf{r}'_{y0} \mathbf{R}_{xx}^{-1} \mathbf{r}_{y0}}{s_y^2} = \mathbf{r}'_{y0} \mathbf{R}_{xx}^{-1} \mathbf{r}_{y0}. \quad (11.3)$$

In R^2 , the q correlations between y and the x 's in $\mathbf{r}_{y0}$ or the q correlations between y and the x 's in $\mathbf{r}_{y0}$ are clustered into a single measure of linear relationship between y and the x 's. The multiple correlation R can be defined alternatively as the maximum correlation between y and a linear combination of the x 's, that is, $R = \max_{\mathbf{c}} r_{y, \mathbf{c}x}$.

We now return to the case of several y 's and several x 's. The covariance structure associated with two subvectors y and x was first discussed in Section 5.8.1. By (5.42), the overall sample covariance matrix for $(y_1, \dots, y_p, x_1, \dots, x_q)$ can be partitioned as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix},$$

where $\mathbf{S}_{yy}$ is the $p \times p$ sample covariance matrix of the y 's, $\mathbf{S}_{yx}$ is the $p \times q$ matrix of sample covariances between the y 's and the x 's, and $\mathbf{S}_{xx}$ is the $q \times q$ sample covariance matrix of the x 's.

In Section 10.6, we discussed several measures of association between the y 's and the x 's. The first of these is defined in (10.60) as $R_{yx}^2 = (\mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}) / (\mathbf{S}_{yy})$, which is analogous to $R^2 = \mathbf{r}'_{y0} \mathbf{R}_{xx}^{-1} \mathbf{r}_{y0} / s_y^2$ in (11.3). By (2.30) and (2.31), R_{yx}^2 can be rewritten as $R_{yx}^2 = (\mathbf{S}_{yx}^T \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy})$. By (2.100), R_{yx}^2 can be expressed as

$$R_{yx}^2 = (\mathbf{S}_{yx}^T \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}) = \prod_{i=1}^s \lambda_i^2,$$

where $s = \min(p, q)$ and $\lambda_1^2, \lambda_2^2, \dots, \lambda_s^2$ are the eigenvalues of $\mathbf{S}_{yx}^T \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$. When written in this form, R_{yx}^2 seems to be a poor measure of association because $0 \leq \lambda_i^2 \leq 1$ for all i , and the product will usually be too small to meaningfully reflect the amount of association. (In Example 10.6, $R_{yx}^2 = .00028$ was a tiny fraction of the other measures of association.) The eigenvalues themselves, on the other hand, provide meaningful measures of association between the y 's and the x 's. The square roots of the eigenvalues, $\lambda_1, \lambda_2, \dots, \lambda_s$, are called *canonical correlations*.

The best overall measure of association is the largest squared canonical correlation (maximum eigenvalue) λ_1^2 of $\mathbf{S}_{yx}^T \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$, but the other eigenvalues (squared)

canonical correlations) of $\mathbf{S}_{xy}^{-1}\mathbf{S}_{yy}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ provide measures of supplemental dimensions of (linear) relationship between $\mathbf{y}$ and $\mathbf{x}$. As an alternative approach, it can be shown that r_1^2 is the maximum squared correlation between a linear combination of the $\mathbf{y}$'s, $u = \mathbf{a}'\mathbf{y}$, and a linear combination of the $\mathbf{x}$'s, $v = \mathbf{b}'\mathbf{x}$, that is,

$$r_1 = \max_{\mathbf{a}, \mathbf{b}} |\mathbf{a}'\mathbf{y}\mathbf{b}\mathbf{x}|. \quad (11.4)$$

We denote the coefficient vectors that yield the maximum correlation as $\mathbf{a}_1$ and $\mathbf{b}_1$. Thus r_1 (the positive square root of r_1^2) is the correlation between $u_1 = \mathbf{a}_1'\mathbf{y}$ and $v_1 = \mathbf{b}_1'\mathbf{x}$. The coefficient vectors $\mathbf{a}_1$ and $\mathbf{b}_1$ can be found as eigenvectors [see (11.7) and (11.8)]. The linear functions u_1 and v_1 are called the first canonical variables. There are additional canonical variables $u_2 = \mathbf{a}_2'\mathbf{y}$ and $v_2 = \mathbf{b}_2'\mathbf{x}$ corresponding to r_2 , $r_3, \dots, r_p$.

It was noted in Section 1.11.5 that the (nonzero) eigenvalues of $\mathbf{A}\mathbf{B}$ are the same as those of $\mathbf{B}\mathbf{A}$, as long as $\mathbf{A}\mathbf{B}$ and $\mathbf{B}\mathbf{A}$ are square but that the eigenvectors of $\mathbf{A}\mathbf{B}$ and $\mathbf{B}\mathbf{A}$ are not the same. If we let $\mathbf{A} = \mathbf{S}_{xy}^{-1}\mathbf{S}_{yy}$ and $\mathbf{B} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$, then $r_1^2, r_2^2, \dots, r_p^2$ can also be obtained from $\mathbf{B}\mathbf{A} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{xy}^{-1}\mathbf{S}_{yy}$ as well as from $\mathbf{A}\mathbf{B} = \mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$. Thus the eigenvalues can be obtained from either of the characteristic equations

$$|\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} - r^2\mathbf{I}| = 0, \quad (11.5)$$

$$|\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx} - r^2\mathbf{I}| = 0. \quad (11.6)$$

The coefficient vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ in the canonical variables $u_i = \mathbf{a}_i'\mathbf{y}$ and $v_i = \mathbf{b}_i'\mathbf{x}$ are the eigenvectors of these same two matrices:

$$(\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} - r^2\mathbf{I})\mathbf{a} = \mathbf{0}, \quad (11.7)$$

$$(\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx} - r^2\mathbf{I})\mathbf{b} = \mathbf{0}. \quad (11.8)$$

Thus the two matrices $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ and $\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$ have the same (nonzero) eigenvalues, as indicated in (11.5) and (11.6), but different eigenvectors, as in (11.7) and (11.8). Since $\mathbf{y}$ is $p \times 1$ and $\mathbf{x}$ is $q \times 1$, the $\mathbf{a}$'s are $p \times 1$ and the $\mathbf{b}$'s are $q \times 1$. This can also be seen in the sizes of the matrices in (11.7) and (11.8), that is, $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ is $p \times p$ and $\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$ is $q \times q$. Since p is typically not equal to q , the matrix that is larger in size will be singular, and the smaller one will be nonsingular. We illustrate for $p < q$. In this case $\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$ has the form

$$\begin{array}{c} p \\ \boxed{\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}} \end{array} \quad \begin{array}{c} p \\ \boxed{\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}} \end{array} \quad \begin{array}{c} p \\ \boxed{\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}} \end{array} \quad \begin{array}{c} p \\ \boxed{\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}} \end{array}$$

When $p < q$, the rank of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ is p , because $\mathbf{S}_{xx}^{-1}$ has rank q and $\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$ has rank p . In this case p eigenvalues are nonzero and the remaining $q - p$ eigenvalues are equal to zero. In general, there are $r = \min(p, q)$ values of the squared canonical correlation r_i^2 with r corresponding pairs of canonical variates $u_i = \mathbf{a}_i'y$ and $v_i = \mathbf{b}_i'x$. For example, if $p = 3$ and $q = 3$, there will be three canonical correlations, r_1, r_2 , and r_3 .

Thus we have r canonical correlations $r_1, r_2, \dots, r_r$ corresponding to the r pairs of canonical variates u_i and v_i :

$$\begin{array}{llll} r_1 & u_1 & = & \mathbf{a}_1'y & v_1 & = & \mathbf{b}_1'x \\ r_2 & u_2 & = & \mathbf{a}_2'y & v_2 & = & \mathbf{b}_2'x \\ \vdots & \vdots & & & \vdots & & \\ r_r & u_r & = & \mathbf{a}_r'y & v_r & = & \mathbf{b}_r'x. \end{array}$$

For each i , r_i is the sample correlation between u_i and v_i ; that is, $r_i = r_{u_i v_i}$. The pairs (u_i, v_i) , $i = 1, 2, \dots, r$, provide the r dimensions of relationship. For simplicity, we would prefer only one dimension of relationship, but this occurs only when $r = 1$, that is, when $p = 1$ or $q = 1$.

The r dimensions of relationship (u_i, v_i) , $i = 1, 2, \dots, r$, are uncorrelated. The information each pair provides is uncorrelated to the other pairs because $u_1, u_2, \dots, u_r$ are uncorrelated. They are not orthogonal because $u_1, u_2, \dots, u_r$ are eigenvalues of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$, which is nonsymmetric. Similarly, the v_i 's are uncorrelated, and each u_i is uncorrelated with all v_j , $j \neq i$, except, of course, v_i .

We examine the elements of the coefficient vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ for the information they provide about the contribution of the y 's and x 's to v_i . These coefficients can be standardized, as noted in the last paragraph in the present section and in Section 11.3.1.

As noted, the matrix $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ is not symmetric. Many algorithms for computation of eigenvalues and eigenvectors accept only symmetric matrices. Since $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ is the product of the two symmetric matrices $\mathbf{S}_{yy}^{-1}$ and $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$, we can proceed as in (8.13) and work with $(\mathbf{U}')^{-1}\mathbf{S}_{yy}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{U}^{-1}$, where $\mathbf{U}\mathbf{U}' = \mathbf{S}_{yy}$ is the Cholesky factorization of $\mathbf{S}_{yy}$ (see Section 2.7). The symmetric matrix $(\mathbf{U}')^{-1}\mathbf{S}_{yy}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{U}^{-1}$ has the same eigenvalues as $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ but has eigenvectors $\mathbf{U}\mathbf{a}_i$, where $\mathbf{a}_i$ is given in (11.7).

In effect, the pq correlations between the y 's and x 's in $\mathbf{S}_{xy}$ have been replaced by $r = \min(p, q)$ canonical correlations. These succinctly summarize the relationships between y and x . In fact, in a typical study, we do not need all r canonical correlations. The smallest eigenvalues can be disregarded to achieve even more simplifications. As in (8.13) for discriminant functions, we can judge the importance of each eigenvalue by its relative size:

$$\frac{r_i^2}{\sum_{j=1}^r r_j^2}. \quad (11.8)$$

The canonical correlations can also be obtained from the partitioned correlation matrix of the y 's and x 's,

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{pmatrix},$$

where $\mathbf{R}_{yy}$ is the $p \times p$ sample correlation matrix of the y 's, $\mathbf{R}_{yx}$ is the $p \times q$ matrix of sample correlations between the y 's and the x 's, and $\mathbf{R}_{xx}$ is the $q \times q$ sample correlation matrix of the x 's. The matrix $\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}$, $\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$ is analogous to $\mathbf{R}^2 = \mathbf{r}'_y\mathbf{R}_{yy}^{-1}\mathbf{r}_y$ in the univariate y -case. The characteristic equations corresponding to (11.5) and (11.6),

$$|\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} - r^2\mathbf{I}| = 0, \quad (11.7)$$

$$|\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} - r^2\mathbf{I}| = 0, \quad (11.8)$$

yield the same eigenvalues $r_1^2, r_2^2, \dots, r_p^2$ as (11.5) and (11.6) (the canonical correlations are scale invariant; see property 1 in Section 11.5).

If we use the partitioned correlation matrix in place of the covariance matrix in (11.7) and (11.8), we obtain the same eigenvalues (squared canonical correlations) but different eigenvectors:

$$(\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} - r^2\mathbf{I})\mathbf{c} = 0, \quad (11.9)$$

$$(\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} - r^2\mathbf{I})\mathbf{d} = 0. \quad (11.10)$$

The relationship between the eigenvectors $\mathbf{c}$ and $\mathbf{d}$ in (11.9) and (11.10) and the eigenvectors $\mathbf{a}$ and $\mathbf{b}$ in (11.7) and (11.8) is

$$\mathbf{c} = \mathbf{D}_y\mathbf{a} \quad \text{and} \quad \mathbf{d} = \mathbf{D}_x\mathbf{b}, \quad (11.11)$$

where $\mathbf{D}_y = \text{diag}(s_{y_1}, s_{y_2}, \dots, s_{y_p})$ and $\mathbf{D}_x = \text{diag}(s_{x_1}, s_{x_2}, \dots, s_{x_q})$.

The eigenvectors $\mathbf{c}$ and $\mathbf{d}$ in (11.9), (11.10), and (11.11) are *standardized coefficient vectors*. By analogy to (8.15), they would be applied to standardized variables. To show this, note that in terms of centered variables $y - \bar{y}$, we have

$$\begin{aligned} \mathbf{a} &= \mathbf{a}'(y - \bar{y}) = \mathbf{a}'\mathbf{D}_y\mathbf{D}_y^{-1}(y - \bar{y}) \\ &= \mathbf{c}'\mathbf{D}_y^{-1}(y - \bar{y}) \quad \text{[by (11.11)]} \\ &= c_1 \frac{y_1 - \bar{y}_1}{s_{y_1}} + c_2 \frac{y_2 - \bar{y}_2}{s_{y_2}} + \dots + c_p \frac{y_p - \bar{y}_p}{s_{y_p}}. \end{aligned} \quad (11.12)$$

Hence $\mathbf{c}$ and $\mathbf{d}$ are preferred to $\mathbf{a}$ and $\mathbf{b}$ for interpretation of the canonical variates u_i and v_i .

11.3. PROPERTIES OF CANONICAL CORRELATIONS

Two interesting properties of canonical correlations are the following (for other properties, see Rencher (1995), Section 8.3-8):

1. Canonical correlations are invariant to changes of scale on either the y 's or the x 's. For example, if the measurement scale is changed from inches to centimeters, the canonical correlations will not change; the corresponding eigenvectors will change. This property holds for simple and multiple correlations as well.
2. The first canonical correlation r_1 is the maximum correlation between linear functions of y and x . Therefore, r_1 exceeds the absolute value of the simple correlation between any y and any x or the multiple correlation between any y and all the x 's or between any x and all the y 's.

Example 11.3. For the chemical data of Table 10.1, we obtain the canonical correlations and illustrate property 2. We consider the extended set of nine x 's, as in Example 10.5.2. The matrix $\mathbf{R}_{yx}$ of correlations between the y 's and the x 's is

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
y_1	-.68	-.33	-.45	-.41	-.55	-.41	-.68	-.33	-.45
y_2	.45	.68	.36	.56	.44	.33	.45	.33	.36
y_3	.36	.33	.36	.45	.45	.36	.36	.33	.36

The three canonical correlations and their squares are

$$r_1 = .8959 \quad r_1^2 = .8026$$

$$r_2 = .5528 \quad r_2^2 = .3075$$

$$r_3 = .4625 \quad r_3^2 = .2139.$$

From the relative sizes of the squared canonical correlations, we would consider only the first two to be important. A hypothesis test for the significance of each is carried out in Example 11.4.2.

To confirm that property 2 holds in this case, we compare $r_1 = .8959$ to the individual correlations and the multiple correlations. We first note that .8959 is greater than individual correlations, since the absolute value of the largest correlation in $\mathbf{R}_{yx}$ is .68. The multiple correlation $R_{y_1|x}$ of each y_1 with the x 's is given by

$$R_{y_1|x} = .587, \quad R_{y_2|x} = .551, \quad R_{y_3|x} = .508,$$

and for the multiple correlation of each x with the y 's we have

$$\begin{array}{lll} R_{x_1|y} = .59, & R_{x_2|y} = .57, & R_{x_3|y} = .507, \\ R_{x_4|y} = .493, & R_{x_5|y} = .583, & R_{x_6|y} = .482, \\ R_{x_7|y} = .598, & R_{x_8|y} = .534, & R_{x_9|y} = .466. \end{array}$$

The first canonical correlation $\lambda_1 = .989$ exceeds all multiple correlations, and property 2 is satisfied. $\square$

11.4 TESTS OF SIGNIFICANCE

In the following two sections we discuss basic tests of significance associated with canonical correlations. For other aspects of model validation for canonical correlations and variates, see Rencher (1995, Section 9.5).

11.4.1 Tests of No Relationship between the y 's and the x 's

In Section 7.4.1, we considered the hypothesis of independence, $H_0: \Sigma_{yz} = \mathbf{O}$. If $\Sigma_{yz} = \mathbf{O}$, the covariance of every y_i with every x_j is zero, and all corresponding correlations are likewise zero. Hence, under H_0 , there is no (linear) relationship between the y 's and the x 's, and H_0 is equivalent to the statement that all canonical correlations $r_1, r_2, \dots, r_k$ are nonsignificant. Furthermore, H_0 is equivalent to the overall regression hypothesis in Section 10.5.1, $H_0: \mathbf{B}_1 = \mathbf{O}$, which also relates all the y 's to all the x 's. Thus by (7.30) or (10.37), the significance of $r_1, r_2, \dots, r_k$ can be tested by

$$\Lambda_1 = \frac{|\mathbf{B}|}{|\mathbf{B}_{yz}| |\mathbf{B}_{xz}|} = \frac{|\mathbf{B}|}{|\mathbf{B}_{yz}| |\mathbf{B}_{xx}|}, \quad (11.16)$$

which is distributed as $\Lambda_{k, p+q-1, q}$. We reject H_0 if $\Lambda_1 \geq \Lambda_{\alpha}$. Critical values Λ_{α} are available in Table A.8 using $v_1 = q$ and $v_2 = n - 1 - q$. The statistic Λ_1 in (11.16) is also distributed as $\Lambda_{k, p+q-1, p}$. As in (7.30), Λ_1 is expressible in terms of the squared canonical correlations:

$$\Lambda_1 = \prod_{i=1}^k (1 - r_i^2). \quad (11.17)$$

In this form, we can see that if one or more r_i^2 is large, Λ_1 will be small. We have used the notation Λ_1 in (11.16) and (11.17) because in Section 11.4.2 we will define Λ_2, Λ_3 , and so on to test the significance of r_k and succeeding x 's after the first.

If the parameters exceed the range of critical values for Wilks' Λ in Table A.8, we can use the χ^2 -approximation in (11.18),

$$\chi^2 = - \left[n - \frac{1}{2}(p+q+k) \right] \ln \Lambda_1, \quad (11.18)$$

which is approximately distributed as χ^2 with pq degrees of freedom. We reject H_0 if $\chi^2 \geq \chi_{\alpha}^2$. Alternatively, we can use the F -approximation given in (11.19):

$$F = \frac{1 - \Lambda_1^{1/q}}{\Lambda_1^{1/q}} \frac{q_2}{q_1}, \quad (11.19)$$

which has an approximate F -distribution with $\mathcal{R}_1$ and $\mathcal{R}_2$ degrees of freedom, where

$$\begin{aligned}\mathcal{R}_1 &= pq, & \mathcal{R}_2 &= nt - \frac{1}{2}pq + 1, \\ n &= n - \frac{1}{2}(p + q + 2), & t &= \sqrt{\frac{p^2q^2 + 4}{p^2 + q^2 - 2}}.\end{aligned}$$

We reject $\mathcal{H}_0$ if $F > F_\alpha$. When $pq = 1$, t is not equal to 1. If $\varepsilon = \min(p, q)$ is equal to either 1 or 2, then the F -approximation in (11.18) has an exact F -distribution. For example, if one of the two sets consists of only two variables, an exact test is afforded by the F -approximation in (11.18). In contrast, the χ^2 -approximation in (11.18) does not induce to an exact test for any parameter values.

The other three multivariate test statistics in Sections 6.1.4, 6.1.5, and 6.5.3 can also be used. Pillai's test statistic for the significance of canonical correlations is

$$V^{(R)} = \sum_{i=1}^R c_i^2. \quad (11.20)$$

Upper percentage points of $V^{(R)}$ are found in Table A.11, indexed by

$$x = \min(p, q), \quad m = \frac{1}{2}(q - p - 1), \quad R = \frac{1}{2}(n - p - p - 2).$$

For F -approximations for $V^{(R)}$, see Sections 6.1.5.

The Lawley–Hoffding statistic for canonical correlations is

$$U^{(R)} = \sum_{i=1}^R \frac{c_i^2}{1 - c_i^2}. \quad (11.21)$$

Upper percentage points for $u_0 U^{(R)} / v_0$ (see Section 6.1.5) are given in Table A.11, which is entered with p , $v_0 = q$, and $v_0 = n - q - 1$. For F -approximations, see Section 6.1.5.

Ray's largest root statistic is given by

$$P = c_1^2. \quad (11.22)$$

Upper percentage points are found in Table A.10, with x , m , and R defined as before for Pillai's test. An "upper bound" on P for Ray's test is given in (6.21). Even though this upper bound is routinely calculated in many software packages, it is not a valid approximation.

As noted at the beginning of this section, the following three tests are equivalent:

1. Test of $\mathcal{H}_0 : \Sigma_{yy} = \Omega$, independence of two sets of variables.
2. Test of $\mathcal{H}_0 : \mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}_3 = \mathcal{R}_4 = \mathcal{R}_5 = \mathcal{R}_6 = \mathcal{R}_7 = \mathcal{R}_8 = \mathcal{R}_9 = \mathcal{R}_{10} = 0$, significance of overall multivariate multiple regression.
3. Test of significance of the canonical correlations.

Even though these tests are equivalent, we have discussed them separately because each has an attention that is different from the others. The respective attentions are

1. Test of independence of three or more sets of variables (Section 7.4.2),
2. Test of full vs. reduced model in multivariate multiple regression (Section 11.3.2),
3. Test of significance of succeeding canonical correlations after the first (Section 11.4.2).

Example 11.4.1. For the chemical data of Table 10.1, with the extended set of nine x 's, we obtained canonical correlations .9899, .8528, and .4625 in Example 11.3. To test the significance of these, we calculate the following four test statistics and associated approximate F 's.

Statistic	Approximate F	df_1	df_2	p -Value for F
Wilks' $\Lambda = .00043$	6.143	27	23.09	< .0001
Pillai's $V^{(2)} = 2.18$	1.340	27	27	.0033
Lewis-Howell $U^{(2)} = 59.83$	11.888	27	17	< .0001
Roy's $J = .988$	48.938	9	9	< .0001

The F approximation for Roy's test is, of course, an "upper bound." Rejection of H_0 in these tests implies that at least r_1^2 is significantly different from zero. The question of how many r_j^2 's are significant is treated in the next section. $\square$

11.4.2 Test of Significance of Succeeding Canonical Correlations after the First

If the test in (11.17) based on all s canonical correlations rejects H_0 , we are not sure if the canonical correlations beyond the first are significant. To test the significance of $r_2, \dots, r_s$, we delete r_1^2 from Λ_0 in (11.17) to obtain

$$\Lambda_2 = \prod_{i=2}^s (1 - r_i^2). \quad (11.23)$$

If this test rejects the hypothesis, we conclude that at least r_2 is significantly different from zero. We can continue in this manner, testing each r_j in turn, until a test fails to reject the hypothesis. At the j th step, the test statistic is

$$\Lambda_j = \prod_{i=j+1}^s (1 - r_i^2), \quad (11.24)$$

which is distributed as $\Lambda_{p-(j+1),q-(j+1),p-j-Q}$ and tests the significance of $r_j, r_{j+1}, \dots, r_s$. (These test statistics are analogous to those for discriminant functions in

Section 11.4.1.) Note that each parameter is reduced by $k - 1$ from the parameter values p , q , and $n = 1 + p + q$ for A_0 in (11.16) or (11.17).

The usual χ^2 - and F -approximations can also be applied to A_k . The χ^2 -approximation analogous to (11.18) is given by

$$\chi^2 = - \left[n - \frac{1}{2}(p + q + 3) \right] \ln A_k, \quad (11.20)$$

which has $(p - k + 1)(q - k + 1)$ degrees of freedom. The F -approximation for A_k is a simple modification of (11.19) and the accompanying parameter definitions. In place of p , q , and n , we use $p - k + 1$, $q - k + 1$, and $n - k + 1$ to obtain

$$F = \frac{1 - A_k^{1/n}}{A_k^{1/2} - A_k^{1/n}},$$

where

$$\begin{aligned} df_1 &= (p - k + 1)(q - k + 1), \\ df_2 &= n - \frac{1}{2}[k(p - k + 1)(q - k + 1) + 1], \\ w &= n - \frac{1}{2}(p + q + 3), \\ t &= \frac{\left\{ (p - k + 1)^2(q - k + 1)^2 - d \right\}}{\left\{ (p - k + 1)^2 + (q - k + 1)^2 - 3 \right\}}. \end{aligned}$$

Example 11.4.2. We continue our analysis of the canonical correlations for the chemical data in Table 11.1 with these χ^2 and F tests. The tests are summarized in Table 11.3.

In the case of A_k , we have a discrepancy between the exact Wilks' λ -test and the approximate F -test. The test based on λ is not significant, whereas the F -test does reach significance. This illustrates the need to check critical values for exact tests whenever p -values for approximate tests are close to the nominal value of α . From the test using λ , we conclude that only $r_1 = .9959$ is significant. The relative sizes of the squared canonical correlations, .992, .968, and .714, would indicate two dimensions of relationship, but this is not confirmed by the Wilks' test, perhaps because of the small sample size relative to the number of variables ($p + q = 12$ and $n = 18$).

Table 11.3. Tests of Three Canonical Correlations of the Chemical Data

k	A_k	λ_k	Approximate F	df_1	df_2	p -Value for F
1	.9949	.994	6.937	27	28.1	α .0001
2	.9723	.968	2.718	16	16	.0009
3	.7084	.709	.650	7	9	.61

To illustrate the computations, we obtain the values in Table 11.1 for $k = 2$. Using (11.26), the computation for λ_2 is

$$\lambda_2 = \prod_{i=1}^2 (1 - r_i^2) = (1 - .806)(1 - .214) = .0725.$$

With $k = 2$, $p = 3$, $q = 3$, and $n = 35$, the critical value for λ_2 is obtained from Table A.9 as

$$\lambda_{.05, p-2+1, q-2+1, n-4} = \lambda_{.05, 2, 2, 31} = .089.$$

For the approximate F for λ_2 , we have

$$\begin{aligned} i &= \sqrt{\frac{(3-2+1)^2(3-2+1)^2-4}{(3-2+1)^2+(3-2+1)^2-3}} = 1, \\ w &= 35 - \frac{1}{2}(3+3+3) = 11.5, \\ d_1 &= 3 - 2 + 1 = 3 - 2 + 1 = 2, \\ d_2 &= (3-2)(2) - \frac{1}{2}(3-2+1)(3-2+1)(2+1) = 35, \\ F &= \frac{1 - (.0725)^{1/35}}{(.0725)^{1/35} - 1} = 2.114. \end{aligned} \quad (1)$$

11.5 INTERPRETATION

We now turn to an assessment of the information contained in the canonical correlations and canonical variates. As was done for discriminant functions in Section 8.7.4, distinction can be made between interpretation of the canonical variates and assessing the contribution of each variable. In the former, the signs of the coefficients are considered; in the latter, the signs are ignored and the coefficients are ranked in order of absolute value.

In Sections 11.5.1–11.5.3, we discuss three common tools for interpretation of canonical variates: (1) standardized coefficients, (2) the correlation between each variate and the canonical variate, and (3) rotation of the canonical variate coefficients. The second of these is the most widely recommended, but we note in Section 11.5.2 that it is the least useful. In fact, for reasons to be outlined, we recommend only the first, standardized coefficients. In Section 11.5.4, we describe redundancy analysis and discuss its shortcomings as a measure of association between two sets of variables.

11.5.1 Standardized Coefficients

The coefficients in the canonical variates $u_i = a_i'x$ and $v_i = b_i'y$ reflect differences in scaling of the variables as well as differences in contribution of the variables to

classical correlation. To remove the effect of scaling, $\mathbf{a}_i$ and $\mathbf{b}_i$ can be standardized by multiplying by the standard deviations of the corresponding variables as in (11.144)

$$\mathbf{c}_i = \mathbf{B}_x \mathbf{a}_i, \quad \mathbf{d}_i = \mathbf{B}_y \mathbf{b}_i,$$

where $\mathbf{B}_x = \text{diag}(s_{x_1}, s_{x_2}, \dots, s_{x_p})$ and $\mathbf{B}_y = \text{diag}(s_{y_1}, s_{y_2}, \dots, s_{y_q})$. Alternatively, $\mathbf{c}_i$ and $\mathbf{d}_i$ can be obtained directly from (11.12) and (11.13) as eigenvectors of $\mathbf{B}_{xy}^{-1} \mathbf{B}_x \mathbf{B}_{xy}^{-1} \mathbf{B}_y$ and $\mathbf{B}_{xy}^{-1} \mathbf{B}_y \mathbf{B}_{xy}^{-1} \mathbf{B}_x$, respectively. It was noted at the end of Section 11.1 that the coefficients in $\mathbf{c}_i$ are applied to standardized variables [see (11.13)]. Thus the effect of differences in scale or scaling of the variables is removed, and the coefficients $c_{1i}, c_{2i}, \dots, c_{pi}$ in $\mathbf{c}_i$ reflect the relative contribution of each of $x_1, x_2, \dots, x_p$ to $\mathbf{c}_i$. A similar statement can be made about $\mathbf{d}_i$.

The standardized coefficients show the contribution of the variables in the presence of each other. Thus if some of the variables are deleted and others added, the coefficients will change. This is precisely the behavior we desire from the coefficients in a multivariate setting.

Example 11A.1. For the chemical data in Table 18.1 with the extended set of nine x 's, we obtain the following standardized coefficients for the three classical variates

	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathbf{c}_3$
x_1	1.3380	4.4764	3.7541
x_2	20.08	2.8291	2.1280
x_3	.4879	2.1389	2.1442

	$\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_3$
y_1	3.8123	-38.8081	-12.5872
y_2	2.8550	-71.7180	-29.2298
y_3	1.6588	-7.6588	-32.7592
$x_4 x_5$	-3.9289	92.5017	11.6428
$x_4 x_6$	-2.2968	64.004	26.2549
$x_5 x_6$	.918	8066	1.2988
x_7^2	-2.8895	32.7911	4.8254
x_8^2	-1.2338	-3.5641	18.7879
x_9^2	.2783	.8711	.8798

Thus

$$\begin{aligned} \mathbf{v}_1 &= 1.24 \frac{x_1 - \bar{x}_1}{s_{x_1}} + .11 \frac{x_2 - \bar{x}_2}{s_{x_2}} + .47 \frac{x_3 - \bar{x}_3}{s_{x_3}}, \\ \mathbf{v}_2 &= 3.81 \frac{y_1 - \bar{y}_1}{s_{y_1}} + 3.85 \frac{y_2 - \bar{y}_2}{s_{y_2}} + \dots + .27 \frac{x_9^2 - \bar{x}_9^2}{s_{x_9^2}}. \end{aligned}$$

The variables that contribute most to the correlation between x_1 and x_2 are y_1 and $x_1, x_2, x_1x_2, x_1x_2^2, x_1^2$. The correlation between x_1 and x_2 is due largely to all three y 's and x_1, x_2, x_1x_2, x_1^2 . $\square$

11.5.2 Correlations between Variables and Canonical Variables

Many writers recommend the additional step of converting the standardized coefficients to correlations. Thus, for example, in $\hat{c}_1 = (c_{11}, c_{12}, \dots, c_{1p})$, instead of the second coefficient c_{12} we could examine $r_{x_2x_1}$, the correlation between x_2 and the first canonical variate u_1 . Such correlations are sometimes referred to as loadings or structure coefficients, and it is widely claimed that they provide a more valid interpretation of the canonical variates. Rencher (1995, PPVCh. 1995, Section 8.6.7) has shown, however, that a weighted sum of the correlations between x_1 and the canonical variates $u_1, u_2, \dots, u_p$ is equal to $R_{x_1}^2$, the squared multiple correlation between x_1 and the u 's. There is no information about how the y 's contribute jointly to canonical correlations with the x 's. Therefore, the correlations are useless in gauging the importance of a given variable in the context of the others. The researcher who uses these correlations for interpretation is unknowingly reducing the multivariate setting to a univariate one.

11.5.3 Rotation

In an attempt to improve interpretability, the canonical variate coefficients can be rotated (see Section 11.5.1) to increase the number of high- and low-coefficients and reduce the number of intermediate ones.

We do not recommend rotation of the canonical variate coefficients for two reasons (for proof and further discussion, see Rencher (1997a)).

1. Rotation destroys the optimality of the canonical correlations. For example, the first canonical correlation is reduced and is no longer equal to $\min_{\mathbf{a} \in \mathbf{A}} \|\mathbf{a}\|_{\mathbf{A}^*} / \|\mathbf{a}\|_{\mathbf{A}}$ (11.4).
2. Rotation introduces correlations among succeeding canonical variates. Thus, for example, u_1 and u_2 are correlated after rotation. Hence even though the loading coefficients may offer a subjectively more interpretable picture, this gain is offset by the increased complexity due to interrelationships among the canonical variates. For example, u_1 and u_2 no longer offer a new dimension of relationship uncorrelated with u_3 and u_4 . The dimensions now overlap, and some of the information in u_1 and u_2 is already available in u_3 and u_4 .

11.5.4 Redundancy Analysis

The redundancy is a measure of association between the y 's and the x 's based on the correlations between variables and canonical variates discussed in Section 11.5.2. Since these correlations provide only univariate information, the redundancy term

out to be a univariate rather than a multivariate measure of relationship. If the squared multiple correlation of y_j regressed on the x 's is denoted by $R_{j(x)}^2$, then the redundancy of the y 's given the x 's is the average squared multiple correlation:

$$\text{Red}(y) = \frac{\sum_{j=1}^p R_{j(x)}^2}{p}, \quad (11.26)$$

Similarly, the redundancy of the x 's given the y 's is the average

$$\text{Red}(x) = \frac{\sum_{j=1}^q R_{j(y)}^2}{q}, \quad (11.27)$$

where $R_{j(y)}^2$ is the squared multiple correlation of x_j regressed on the y 's. Since $\text{Red}(y)$ in (11.26) is the average squared multiple correlation of each y_j regressed on the x 's, it does not take into account the correlations among the y 's. It is thus an average univariate measure of relationship between the y 's and the x 's, not a multivariate measure at all. The two redundancy measures in (11.26) and (11.27) are not symmetric; that is, $\text{Red}(y) \neq \text{Red}(x)$.

Thus the so-called redundancy does not really quantify the redundancy arising from the y 's and x 's and is, therefore, not a useful measure of association between two sets of variables. For a measure of association we recommend r^2 itself.

11.6 RELATIONSHIPS OF CANONICAL CORRELATION ANALYSIS TO OTHER MULTIVARIATE TECHNIQUES

In Section 11.4.1, we noted the equivalence of the test for significance of the canonical correlations and the test for significance of overall regression, $H_0: \mathbf{B}_1 = \mathbf{0}$. Additional relationships between canonical correlation and multivariate regression are developed in Section 11.6.1. The relationship of canonical correlation analysis to MANOVA and discriminant analysis is discussed in Section 11.6.2.

11.6.1 Regression

There is a direct link between canonical variate coefficients and multivariate multiple regression coefficients. The matrix of regression coefficients of the y 's regressed on the x 's (corrected for their means) is given in (10.52) as $\mathbf{B}_1 = \mathbf{S}_{xy}^{-1}\mathbf{S}_{yx}$. This matrix can be used to relate $\mathbf{a}_1$ and $\mathbf{b}_1$:

$$\mathbf{b}_1 = \mathbf{B}_1 \mathbf{a}_1. \quad (11.28)$$

Since $\mathbf{a}_1$ and $\mathbf{b}_1$ are eigenvectors, (11.28) could also be written as $\mathbf{b}_1 = c\mathbf{B}_1 \mathbf{a}_1$, where c is an arbitrary scale factor. [Eq. (2.67) and (3.3.28), the canonical variate coefficient

vector $\mathbf{b}_0$ is expressible as a linear combination of the columns of $\mathbf{B}_0$. A similar expression for $\mathbf{a}_0$ can be obtained from the regression of $\mathbf{y}$ on $\mathbf{p}$: $\mathbf{a}_0 = \mathbf{S}_{yy}^{-1} \mathbf{S}_{y0} \mathbf{b}_0$.

In Section 11.2, canonical correlation was defined as an extension of multiple correlation. Correspondingly, canonical correlation reduces to multiple correlation when one of the two sets of variables has only one variable. When $p = 1$, for example, $\mathbf{B}_0$ becomes 1, and by (11.15), the single squared canonical correlation reduces to $r^2 = \mathbf{r}'_{yx} \mathbf{R}_{xx}^{-1} \mathbf{r}_{yx}$, which we recognise from (10.34) as R^2 .

The two 'Wilks' test statistics in multivariate regression in Sections 10.5.1 and 10.5.2, namely, the test for overall regression and the test on a subset of the x 's, can both be expressed in terms of canonical correlations. By (10.55) and (11.17), the test statistic for the overall regression hypothesis $H_0: \mathbf{B}_0 = \mathbf{0}$ can be written as

$$J_0' = \frac{(\mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y})}{(\mathbf{Y}'\mathbf{Y} - n\mathbf{I})} \quad (11.28)$$

$$= \prod_{i=1}^p (1 - r_i^2), \quad (11.29)$$

where r_i^2 is the i th squared canonical correlation.

A test statistic for $H_0: \mathbf{B}_0 = \mathbf{0}$, the hypothesis that the y 's do not depend on the last h of the x 's, is given by (10.62) as

$$J_0(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p | \mathbf{r}_1, \dots, \mathbf{r}_{p-h}) = \frac{J_0'}{A_h}, \quad (11.30)$$

where J_0' is given in (11.28) and A_h is given in (10.64) as

$$A_h = \frac{(\mathbf{Y}'\mathbf{Y} - \mathbf{B}'_h \mathbf{X}'_h \mathbf{Y})}{(\mathbf{Y}'\mathbf{Y} - n\mathbf{I})}, \quad (11.31)$$

by analogy with (11.30), A_h can be expressed in terms of the squared canonical correlations $r_1^2, r_2^2, \dots, r_l^2$ between $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_p$ and $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{p-h}$

$$A_h = \prod_{i=1}^l (1 - r_i^2), \quad (11.32)$$

where $l = \min(p, q - h)$. We have used the notation r_i^2 instead of r_{ij}^2 to emphasise that the canonical correlations in the reduced model differ from those in the full model. By (11.30) and (11.32), the full and reduced model test of $H_0: \mathbf{B}_0 = \mathbf{0}$ in (11.31) can now be expressed in terms of canonical correlations as

$$J_0(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p | \mathbf{r}_1, \dots, \mathbf{r}_{p-h}) = \frac{\prod_{i=1}^p (1 - r_i^2)}{\prod_{i=1}^l (1 - r_i^2)}, \quad (11.34)$$

If $p = 1$, as in multiple regression, then $r = r = 1$, and (11.34) reduces to

$$A = \frac{1 - R_1^2}{1 - R_0^2}, \quad (11.35)$$

where R_1^2 and R_0^2 are the squared multiple correlations for the full model and for the reduced model. The distribution of A in (11.35) is $F_{k, k-p-1}$ when R_0 is true. Since $p = 1$, there is an exact F -transformation from Table 6.1,

$$F = \frac{(1 - R_0)(n - q - 1)}{AR_0},$$

which is distributed as $F_{k, n-q-1}$ when R_0 is true. Substitution of $A = (1 - R_1^2)/(1 - R_0^2)$ from (11.35) yields the F -statistic expressed in terms of R^2 ,

$$F = \frac{R_0^2 - R_1^2(n - q - 1)}{(1 - R_1^2)A}, \quad (11.36)$$

as given in (10.57).

Subset selection in canonical correlation analysis can be handled by the methods for multivariate regression given in Section 10.7. A subset of x 's can be found by the procedure of Section 10.7.1a. After a subset of x 's is found, the approach in Section 10.7.1b can be used to select a subset of y 's.

Walleris (1982) discussed the relationship of canonical correlation analysis to multivariate regression and principal components. (Principal components are treated in Chapter 12.)

11.8.2 MANOVA and Discriminant Analysis

In Sections 6.1.8 and 8.4.1, it was noted that in a one-way MANOVA or discriminant analysis setting, $\lambda_k/(1 + \lambda_k)$ is equal to r_k^2 , where λ_k is the k th eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$ and r_k^2 is the k th squared canonical correlation between the p dependent variables and the $k = 1$ grouping variables. We now give a justification of this assertion.

Let the dependent variables be denoted by $y_1, y_2, \dots, y_p$, as usual. We represent the k groups by $k - 1$ dummy variables, $x_1, x_2, \dots, x_{k-1}$, defined for each observation of the k th group, $i \leq k - 1$, as $x_1 = 0, \dots, x_{i-1} = 0, x_i = 1, x_{i+1} = 0, \dots, x_{k-1} = 0$. For the k th group, all x 's are zero, thus Section 6.1.8 for an introduction to dummy variables.) To illustrate with $k = 4$, the x 's are defined as follows in each group:

Group	x_1	x_2	x_3
1	0	0	0
2	0	1	0
3	0	0	1
4	0	0	0

The MANOVA model is equivalent to multivariate regression of $y_1, y_2, \dots, y_p$ on the clustering grouping variables $x_1, x_2, \dots, x_{q-1}$. The MANOVA test of $H_0: \mu_1 = \mu_2 = \dots = \mu_q$ is equivalent to the multivariate regression test of $H_0: \mathbf{B}_1 = \mathbf{0}_k$ as given by (11.17).

$$\lambda = \prod_{j=1}^p (1 - r_j^2). \quad (11.37)$$

When we compute this form of λ , it is the MANOVA test statistic (5.14),

$$\lambda = \prod_{j=1}^p \frac{1}{1 + b_j}, \quad (11.38)$$

we obtain the relationships

$$r_j^2 = \frac{b_j}{1 + b_j}, \quad (11.39)$$

$$b_j = \frac{r_j^2}{1 - r_j^2}. \quad (11.40)$$

To establish this relationship more formally, we write (5.22) as

$$\mathbf{B}\mathbf{a} = \lambda\mathbf{E}\mathbf{a} \quad (11.41)$$

and (11.7) as

$$\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XZ}\mathbf{a} = r^2\mathbf{a}. \quad (11.42)$$

We multiply (11.42) on the left by $\mathbf{S}_{YZ}$ to obtain

$$\mathbf{S}_{YZ}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XZ}\mathbf{a} = r^2\mathbf{S}_{YZ}\mathbf{a}. \quad (11.43)$$

Using the centered matrix $\mathbf{X}_c$ in (10.14), with an analogous definition for $\mathbf{Y}_c$, we can write $\mathbf{B}_1$ in the form [see (10.27)]

$$\mathbf{B}_1 = \left(\frac{\mathbf{X}_c\mathbf{X}_c'}{n-1} \right)^{-1} \frac{\mathbf{X}_c\mathbf{Y}_c'}{n-1} = \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}.$$

In terms of centered matrices, $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{Y}$ in (10.49) can be written as

$$\begin{aligned} \frac{\mathbf{E}}{n-1} &= \frac{\mathbf{Y}_c'\mathbf{Y}_c}{n-1} - \frac{\mathbf{B}_1'\mathbf{S}_{XX}\mathbf{B}_1}{n-1} \\ &= \mathbf{S}_{YY} - \mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY} = \mathbf{S}_{YY} - \mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XZ}. \end{aligned} \quad (11.44)$$

since $\bar{Y}_{j0} = \bar{Y}_{j1}$. Similarly,

$$\frac{\bar{H}}{n-1} = \frac{\bar{H}_1' Y_1 Y_1}{n-1} = \bar{Y}_{j1} \bar{Y}_{j1}^{-1} \bar{Y}_{j1}. \quad (11.45)$$

Since MANOVA is equivalent to multivariate regression on dummy grouping variables, we can substitute these values of $\bar{K}$ and $\bar{H}$ into (11.41) to obtain

$$\bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j1} a = \lambda (\bar{Y}_{j0} - \bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j0}) a. \quad (11.46)$$

Subtracting $r^2 \bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j1} a$ from both sides of (11.46) gives

$$\bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j1} a = \frac{r^2}{1-r^2} (\bar{Y}_{j0} - \bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j0}) a. \quad (11.47)$$

A comparison of (11.46) and (11.47) shows that

$$\lambda = \frac{r^2}{1-r^2},$$

as in (11.40). Lindsay, Webster, and Huijsen (1985) discussed some advantages of using canonical correlation analysis in place of discriminant analysis in the several-group case.

PROBLEMS

- BL1 Show that the expression for canonical correlations in (11.31) can be obtained from the analogous expression in terms of variances and covariances in (11.7).
- BL2 Verify (11.28), $b_1 = \hat{b}_1 a_1$.
- BL3 Verify (11.33) for λ , when $p = d = t = 1$.
- BL4 Verify the expression in (11.35) for F in terms of R_1^2 and R_2^2 .
- BL5 Solve (11.36), $r_1^2 = \lambda_1/(1 + \lambda_1)$, for λ_1 to obtain (11.40).
- BL6 Verify (11.46), $\bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j1} a = \lambda (\bar{Y}_{j0} - \bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j0}) a$.
- BL7 Show that (11.47) can be obtained by subtracting $r^2 \bar{Y}_{j0} \bar{K}_{j0}^{-1} \bar{Y}_{j1} a$ from both sides of (11.46).
- BL8 Use the diabetes data of Table 1.4.
 - (a) Find the canonical correlations between (x_1, x_2) and (y_1, y_2, y_3) .
 - (b) Find the standardized coefficients for the canonical variates.
 - (c) Test the significance of each canonical correlation.

11.9 Use the wine data of Table 3.1.

- (a) Find the canonical correlations between (y_1, y_2) and (x_1, x_2) .
- (b) Find the standardized coefficients for the canonical variates.
- (c) Test the significance of each canonical correlation.

11.10 Use the glucose data of Table 3.2.

- (a) Find the canonical correlations between (y_1, y_2, y_3) and (x_1, x_2, x_3) .
- (b) Find the standardized coefficients for the canonical variates.
- (c) Test the significance of each canonical correlation.

11.11 Use the Seals data of Table 7.1.

- (a) Find the canonical correlations between (y_1, y_2) and $(x_1, x_2, \dots, x_6)$.
- (b) Find the standardized coefficients for the canonical variates.
- (c) Test the significance of each canonical correlation.

11.12 Use canonical correlation to carry out the tests in parts (b), (c), and (d) of Problem 11.11, using the Seals data. You will need to find the canonical correlations between (y_1, y_2) and the x 's in the indicated reduced models and use (11.34).

11.13 Using the temperature data of Table 7.2, find the canonical correlations and the standardized coefficients and carry out significance tests for the following:

- (a) (y_1, y_2, y_3) and (x_1, x_2, x_3)
- (b) $(y_1, y_2, \dots, y_3)$ and (x_1, x_2, x_3)
- (c) $(y_1, y_2, \dots, y_3)$ and (x_1, x_2)
- (d) $(y_1, y_2, \dots, y_3)$ and $(x_1, x_2, \dots, x_3)$

Principal Component Analysis

12.1 INTRODUCTION

In principal component analysis, we seek to maximise the variance of a linear combination of the variables. For example, we might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking.

Essentially, principal component analysis is a one-sample technique applied to data with no groupings among the observations as in Chapters 8 and 9 and no partitioning of the variables into subsets y and x , as in Chapters 10 and 11. All the linear combinations that we have considered previously were related to either variables or to the data structure. In regression, we have linear combinations of the independent variables that best predict the dependent variable(s); in canonical correlation, we have linear combinations of a subset of variables that maximally correlate with linear combinations of another subset of variables; and discriminant analysis involves linear combinations that maximally separate groups of observations. Principal components, on the other hand, are concerned only with the core structure of a single sample of observations on p variables. None of the variables is designated as dependent, and no grouping of observations is assumed. For a discussion of the use of principal components with data consisting of several samples or groups, see Rencher (1995, Section 9.5).

The first principal component is the linear combination with maximal variance; we are essentially searching for a dimension along which the observations are maximally separated or spread out. The second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component, and so on. In general, the principal components define different dimensions from those defined by discriminant functions or canonical variates.

In some applications, the principal components are an end in themselves and may be amenable to interpretation. More often they are obtained for use as input to another analysis. For example, two situations in regression where principal components may be useful are (1) if the number of independent variables is large relative to

the number of observations, a test may be ineffective or even impossible, and (2) if the independent variables are highly correlated, the estimates of regression coefficients may be unstable. In such cases, the independent variables can be reduced to a smaller number of principal components that will yield a better test or more stable estimates of the regression coefficients. For details of this application, see Rencher (1995, Section 5.6).

As another illustration, suppose that in a MANOVA application p is close to n , so that a test has low power, or that $p > n$, in which case we have too many dependent variables that a test cannot be made. In such cases, we can replace the dependent variables with a smaller set of principal components and then carry out the test.

In these illustrations, principal components are used to reduce the number of dimensions. Another useful dimension reduction device is to evaluate the first two principal components for each observation vector and construct a scatter plot to check for multivariate normality, outliers, and so on.

Finally, we note that in the term *principal components*, we use the adjective *principal*, denoting what kind of components—main, primary, fundamental, major, and so on. We do not use the noun *principle* as a modifier for components.

12.1 GEOMETRIC AND ALGEBRAIC BASES OF PRINCIPAL COMPONENTS

12.1.1 Geometric Approach

As noted in Section 12.1, principal component analysis deals with a single sample of n observation vectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ that form a swarm of points in a p -dimensional space. Principal component analysis can be applied to any distribution of $\mathbf{y}$, but it will be easier to visualize geometrically if the swarm of points is ellipsoidal.

If the variables $y_1, y_2, \dots, y_p$ in $\mathbf{y}$ are correlated, the ellipsoidal swarm of points is not oriented parallel to any of the axes represented by $y_1, y_2, \dots, y_p$. We wish to find the natural axes of the swarm of points (the axes of the ellipsoid) with origin at $\bar{\mathbf{y}}$, the mean vector of $y_1, y_2, \dots, y_p$. This is done by translating the origin to $\bar{\mathbf{y}}$ and then rotating the axes. After rotation so that the axes become the natural axes of the ellipsoid, the new variables (principal components) will be uncorrelated.

We could indicate the translation of the origin to $\bar{\mathbf{y}}$ by writing $\mathbf{y}_i - \bar{\mathbf{y}}$, but we will not usually do so for economy of notation. We will write $\mathbf{y}_i - \bar{\mathbf{y}}$ when there is an explicit need; otherwise we assume that $\mathbf{y}_i$ has been centered.

The axes can be rotated by multiplying each $\mathbf{y}_i$ by an orthogonal matrix $\mathbf{A}$ (see (12.13)), where the orthogonal matrix was denoted by $\mathbf{C}$:

$$\mathbf{z}_i = \mathbf{A}\mathbf{y}_i. \quad (12.1)$$

Since $\mathbf{A}$ is orthogonal, $\mathbf{A}^t\mathbf{A} = \mathbf{I}$, and the distance to the origin is unchanged:

$$\|\mathbf{z}_i\|^2 = (\mathbf{A}\mathbf{y}_i)^t(\mathbf{A}\mathbf{y}_i) = \mathbf{y}_i^t\mathbf{A}^t\mathbf{A}\mathbf{y}_i = \mathbf{y}_i^t\mathbf{y}_i.$$

[see (2.103)]. Thus an orthogonal matrix transforms y_0 to a point y_1 that is the same distance from the origin, and the axes are effectively rotated.

Finding the axes of the ellipsoid is equivalent to finding the orthogonal matrix A that rotates the axes to line-up with the natural extensions of the means of points so that the new variables (principal components) $z_1, z_2, \dots, z_p$ in $\mathbf{z} = A\mathbf{y}$ are uncorrelated. Thus we want the sample covariance matrix of $\mathbf{z}$, $S_z = A\mathbf{S}\mathbf{A}'$ [see (3.64)], to be diagonal:

$$S_z = A\mathbf{S}A' = \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p^2 \end{pmatrix}, \quad (12.2)$$

where $\mathbf{S}$ is the sample-covariance matrix of $y_1, y_2, \dots, y_p$. By (2.113), $C\mathbf{S}C' = \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, where the λ_i 's are eigenvalues of $\mathbf{S}$ and C is an orthogonal matrix whose columns are normalized eigenvectors of $\mathbf{S}$. Thus the orthogonal matrix A that diagonalizes $\mathbf{S}$ is the transpose of the matrix C :

$$A = C' = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix}, \quad (12.3)$$

where $\mathbf{a}_i$ is the i th normalized ($\|\mathbf{a}_i\| = 1$) eigenvector of $\mathbf{S}$. The principal components are the transformed variables $z_1 = \mathbf{a}_1'\mathbf{y}$, $z_2 = \mathbf{a}_2'\mathbf{y}$, $\dots$, $z_p = \mathbf{a}_p'\mathbf{y}$ in $\mathbf{z} = A\mathbf{y}$. For example, $z_1 = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1p}y_p$.

By (2.113), the diagonal elements of $A\mathbf{S}A'$ on the right side of (12.2) are eigenvalues of $\mathbf{S}$. Hence the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$ of $\mathbf{S}$ are the (sample) variances of the principal components: $\lambda_i = \mathbf{a}_i'\mathbf{S}$.

$$\lambda_i^2 = \lambda_i. \quad (12.4)$$

Since the rotation lines up with the natural extensions of the means of points, $z_1 = \mathbf{a}_1'\mathbf{y}$ has the largest (sample) variance and $z_p = \mathbf{a}_p'\mathbf{y}$ has the smallest variance. This also follows from (12.4), because the variance of z_1 is λ_1 , the largest eigenvalue, and the variance of z_p is λ_p , the smallest eigenvalue. If some of the eigenvalues are small, we can neglect them and represent the points fairly well with fewer than p dimensions. For example, if $p = 3$ and λ_3 is small, then the means of points is an "elliptical pancake," and a two-dimensional representation will adequately portray the configuration of points.

Because the eigenvalues are variances of the principal components, we can speak of "the proportion of variance explained" by the first k components:

$$\begin{aligned}\text{Proportion of variance} &= \frac{\lambda_1 + \lambda_1 + \cdots + \lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \\ &= \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_j}{\sum_{j=1}^p \lambda_j},\end{aligned}\quad (12.5)$$

since $\sum_{j=1}^p \lambda_j = n(S)$ by (1.13). Thus we try to represent the p -dimensional points $(x_1, x_2, \dots, x_p)$ with a few principal components $(z_1, z_2, \dots, z_k)$ that account for a large proportion of the total variance. If a few variables have relatively large variances, they will figure disproportionately in $\sum_{j=1}^p \lambda_j$ and in the principal components. For example, if x_{22} is strikingly larger than the other variances, then in $z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$, the coefficient a_{12} will be large and all other a_{1j} will be small.

When a table analogous to (12.5) is used for discriminant functions and canonical variates [see (8.13) and (11.16)], it is frequently referred to as *percent of variance*. However, in the case of discriminant functions and canonical variates, the eigenvalues are not variances, so they are *not* principal components.

If the variables are highly correlated, the maximal dimensionality is much smaller than p . In this case, the first few eigenvalues will be large, and (12.5) will be close to 1 for a small value of j . On the other hand, if the correlations among the variables are all small, the dimensionality is close to p and the eigenvalues will be nearly equal. In this case, the principal components essentially duplicate the variables, and no useful reduction in dimension is achieved.

Any two principal components $z_i = a_i'x$ and $z_j = a_j'x$ are orthogonal for $i \neq j$, that is, $a_i'a_j = 0$, because a_i and a_j are eigenvectors of the symmetric matrix S (see Section 2.11.6). Principal components also have the secondary property of being uncorrelated in the sample [see (1.2.2) and (3.63)]: that is, the covariance of z_i and z_j is zero:

$$c_{z_i z_j} = a_i'Sa_j = 0 \quad \text{for } i \neq j. \quad (12.6)$$

Discriminant functions and canonical variates, on the other hand, have the weaker property of being uncorrelated but not the stronger property of orthogonality. Thus what we get the first two discriminant functions or canonical variates on perpendicular coordinate axes, there is some distortion of their true relationship because the actual angle between their axes is not 90°.

If we change the scale on one or more of the y 's, the shape of the cloud of points will change, and we will need different components to represent the new points. Hence the principal components are not scale-invariant. We therefore need to be concerned with the units in which the variables are measured. If possible, all variables should be expressed in the same units. If the variables have widely disparate variances, we could standardize them before extracting eigenvalues and eigenvectors. This is equivalent to finding principal components of the correlation matrix R and is treated in Section 12.5.

If one variable has a much greater variance than the other variables, the cloud of points will be elongated and will be nearly parallel to the axis corresponding to the

variable with large variance. The first principal component will largely represent that variable, and the other principal components will have negligibly small variances. Such principal components based on $\mathbf{B}$ do not involve the other $p - 1$ variables, and we may prefer to analyze the correlation matrix $\mathbf{R}$.

Example 12.3.1. To illustrate principal components as a rotation when $p = 2$, we use two variables from the snout data of Table 2.2: y_1 is head length and y_2 is head width for the first snout. The mean vector and covariance matrix are

$$\bar{\mathbf{y}} = \begin{pmatrix} 183.7 \\ 151.1 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 85.28 & 51.87 \\ 51.87 & 54.56 \end{pmatrix}.$$

The eigenvalues and eigenvectors of $\mathbf{S}$ are

$$\lambda_1 = 131.51, \quad \lambda_2 = 18.14, \\ \mathbf{u}_1' = (u_{11}, u_{12}) = (.825, .565), \quad \mathbf{u}_2' = (u_{21}, u_{22}) = (-.565, .825).$$

The symmetric pattern in the eigenvectors is due to their collinearity: $\mathbf{u}_1'\mathbf{u}_2 = u_{11}u_{21} + u_{12}u_{22} = 0$.

The observations are plotted in Figure 12.1, along with the translated and rotated axes. The major axis is the line passing through $\bar{\mathbf{y}}' = (183.7, 151.1)$ in the direction determined by $\mathbf{u}_1' = (.825, .565)$; the slope is $u_{12}/u_{11} = .565/.825$. Alternatively, the equation of the major axis can be obtained by setting $z_2 = 0$:

$$\begin{aligned} z_2 &= 0 = a_{21}(y_1 - \bar{y}_1) + a_{22}(y_2 - \bar{y}_2) \\ &= -.565(y_1 - 183.7) + .825(y_2 - 151.1). \end{aligned}$$

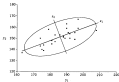


Figure 12.1. Principal-component transformation for the snout data.

The lengths of the semimajor and semiminor axes are proportional to $\sqrt{\lambda_1} = 11.5$ and $\sqrt{\lambda_2} = 4.3$, respectively.

Note that the line formed by the major axis can be considered to be a regression line. It is fit to the points so that the perpendicular distance of the points to the line is minimized, rather than the usual vertical distance (see Section 12.3). $\square$

12.1.2 Algebraic Approach

An algebraic approach to principal components can be briefly described as follows. As noted in Section 12.1, we seek a linear combination with maximal variance. By (8.58), the sample variance of $\mathbf{z} = \mathbf{a}'\mathbf{y}$ is $\mathbf{a}'\mathbf{S}\mathbf{a}$. Since $\mathbf{a}'\mathbf{S}\mathbf{a}$ has no maximum if $\mathbf{a}$ is unrestricted, we seek the maximum of

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}}. \quad (12.7)$$

By an argument similar to that used in (8.94)–(8.10), the maximum value of λ is given by the largest eigenvalue in the expression

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0} \quad (12.8)$$

(see Problem 11.1). The eigenvector $\mathbf{a}_1$ corresponding to the largest eigenvalue λ_1 is the coefficient vector in $\mathbf{z}_1 = \mathbf{a}_1'\mathbf{y}$, the linear combination with maximum variance.

Unlike discriminant analysis or canonical correlation, there is no interest involved before obtaining eigenvectors for principal components. Therefore, $\mathbf{S}$ can be singular, in which case some of the eigenvalues are zero and can be ignored. A singular $\mathbf{S}$ would arise, for example, when $n = p$, that is, when the sample size is less than the number of variables.

This tolerance of principal component analysis for a singular $\mathbf{S}$ is important in certain research situations. For example, suppose that one has a one-way MANOVA with 30 observations in each of three groups and that $p = 30$, so that there are 30 variables in each of these 30 observation vectors. A MANOVA test involving $\mathbf{E}^{-1}\mathbf{H}$ cannot be carried out directly in this case because $\mathbf{E}$ is singular, but we could induce the 30 variables to a small number of principal components and then do a MANOVA test on the components. The principal components would be based on $\mathbf{S}$ obtained from the 30 observations ignoring groups. For entry into the MANOVA program, we would evaluate the principal components for each observation vector. If we are retaining k components, we calculate

$$\begin{aligned} \mathbf{z}_{11} &= \mathbf{a}_1'\mathbf{y}_1 \\ \mathbf{z}_{12} &= \mathbf{a}_2'\mathbf{y}_1 \\ &\vdots \\ \mathbf{z}_{k1} &= \mathbf{a}_k'\mathbf{y}_1 \end{aligned} \quad (12.9)$$

for $i = 1, 2, \dots, 30$. These are sometimes referred to as *component scores*. In vector form, (12.14) can be written as

$$\mathbf{z}_i = \mathbf{A}_2 \mathbf{y}_i, \quad (12.15)$$

where

$$\mathbf{z}_i = \begin{pmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{i6} \end{pmatrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_6' \end{pmatrix}.$$

We then use $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{30}$ as input to the BLANCOVA program.

Note that in this case with $p = 6$, the k components would not likely be stable, that is, they would be different in a new sample. However, this is of no concern here because we are using the components only to extract information from the sample at hand in order to compare the three groups.

Example 12.2.2. Consider the football data of Table 8.3. In Example 8.3, we saw that high school football players (group 1) differed from the other two groups, college football players and college-age nonfootball players. Therefore, to obtain a homogeneous group of observations, we delete group 1 and use groups 2 and 3 combined. The covariance matrix is as follows:

$$\mathbf{S} = \begin{pmatrix} .370 & .602 & .149 & .044 & .107 & .308 \\ .892 & 2.628 & .804 & .886 & .163 & .373 \\ .149 & .804 & .458 & .031 & -.013 & .128 \\ .044 & .602 & .031 & 1.474 & .252 & -.054 \\ .107 & .163 & -.013 & .252 & .488 & -.086 \\ .308 & .373 & .128 & -.054 & -.086 & .324 \end{pmatrix}.$$

The total variance is

$$\sum_{j=1}^6 s_{jj} = \sum_{j=1}^6 \lambda_j = 5.743.$$

The eigenvalues of $\mathbf{S}$ are as follows:

Eigenvalue	Proportion of Variance	Cumulative Proportion
3.323	.578	.578
1.274	.222	.800
.479	.083	.883
.029	.005	.888
.107	.019	.907
.080	.014	1.000

The first two principal components account for 81.6% of the total variance. The corresponding eigenvalues are as follows:

	λ_1	λ_2
WORM	.267	-.142
CYCLUM	.493	-.219
FRUITE	.264	-.224
HYPERO	.356	.491
BARRED	.366	.202
LOP	.128	-.187

Thus the first two principal components are

$$\begin{aligned} z_1 = \mathbf{z}'_1 \mathbf{z} &= .267z_1 + .493z_2 + .264z_3 + .356z_4 + .366z_5 + .128z_6, \\ z_2 = \mathbf{z}'_2 \mathbf{z} &= -.142z_1 - .219z_2 - .224z_3 + .491z_4 + .202z_5 - .187z_6. \end{aligned}$$

Notice that the large coefficient in z_1 and the large coefficient in z_2 , .493 and .491, respectively, correspond to the two largest variances on the diagonal of $\mathbf{S}$. The two variables with large variances, y_2 and y_4 , have a notable influence on the first two principal components. However, z_1 and z_2 are still uncorrelated linear functions. If the six variances were closer in size, the six variables would enter more evenly into the first two principal components. On the other hand, if the variances of y_2 and y_4 were substantially larger, z_1 and z_2 would be essentially equal to y_2 and y_4 , respectively.

Note that y_3 and y_6 did not contribute at all when this data set was used to separate groups in Examples 8.3, 8.4, 8.5.1, and 9.5.1. However, these two variables are very useful here in the first two dimensions showing the spread of individual observations. $\square$

12.3 PRINCIPAL COMPONENTS AND PERPENDICULAR REGRESSION

It was noted in Section 12.2.1 that principal components constitutes a rotation of axes. Another geometric property of the line formed by the first principal component is that it minimizes the total sum of squared perpendicular distances from the points to the line. This is easily demonstrated in the bivariate case. The first principal component line is plotted in Figure 12.3 for the first two variables of the worm data, as in Example 12.2.1. The perpendicular distance from each point to the line is simply z_2 , the second coordinate in the transformed coordinates (z_1, z_2) . Hence the sum of squares of perpendicular distances is

$$\sum_{i=1}^n z_i^2 = \sum_{i=1}^n [a_i y_i - \bar{y}]^2, \quad (12.13)$$

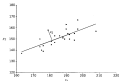


Figure 12.2. The first principal component is a perpendicular regression line.

where $\mathbf{a}_2$ is the second eigenvector of $\mathbf{B}_2$ and we use $p_1 = \bar{y}$ because the axes have been translated to the new origin $\bar{\mathbf{y}}$. Since $\mathbf{a}_2^T(\mathbf{y}_i - \bar{\mathbf{y}}) = (\mathbf{y}_i - \bar{\mathbf{y}})^T \mathbf{a}_2$, we can write (12.12) in the form

$$\begin{aligned} \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n \mathbf{a}_2^T (\mathbf{y}_i - \bar{\mathbf{y}}) (p_1 - \bar{y}) / \mathbf{a}_2 \\ &= \mathbf{a}_2^T \left[\sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}}) (p_1 - \bar{y}) \right] \mathbf{a}_2 \quad \text{[by (2.44)]} \\ &= (n - 1) \mathbf{a}_2^T \mathbf{B}_2 \mathbf{a}_2 = (n - 1) \lambda_2 \quad \text{[by (3.27)],} \end{aligned} \quad (12.13)$$

which is a minimum by a remark following (12.4).

For the two variables x_1 and x_2 , as plotted in Figure 12.2, the ordinary regression line of x_2 on x_1 minimizes the sum of squares of vertical distances from the points to the line. Similarly, the regression of x_1 on x_2 minimizes the sum of squares of horizontal distances from the points to the line. The first principal component line represents a "perpendicular" regression line that lies between the other two. The three lines are compared in Figure 12.3 for the partial area data. The equation of the first principal component line is easily obtained by setting $\lambda_2 = 0$:

$$\begin{aligned} \mathbf{a}_2^T (\mathbf{p} - \bar{\mathbf{p}}) &= 0, \\ \mathbf{a}_{21} (p_1 - \bar{p}_1) + \mathbf{a}_{22} (p_2 - \bar{p}_2) &= 0, \\ -.855(p_1 - \bar{p}_1) + .528(p_2 - \bar{p}_2) &= 0, \end{aligned}$$

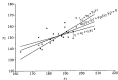


Figure 12.3. Regression lines compared with first principal component line.

12.4 PLOTTING OF PRINCIPAL COMPONENTS

The plots in Figures 12.1 and 12.2 were illustrations of principal components as a rotation of axes when $p = 2$. When $p > 2$, we can plot the first two components (z_1, z_2) for each observation vector and plot these n points. The plot is equivalent to a projection of the p -dimensional data vectors onto the plane that shows the greatest spread of the points.

The plot of the first two components may reveal some important features of the data set. In Example 12.4(a), we show a principal component plot that exhibits a pattern typical of a sample from a multivariate normal distribution. One of the objectives of plotting is to check for departures from normality, such as outliers or nonlinearity. In Examples 12.4(b) and 12.4(c), we illustrate principal component plots showing a nonnormal pattern characterized by the presence of outliers. Jackson (1990) provided a test for adequacy of representation of observation vectors in terms of principal components.

Granados-Ruiz (1997, p. 308) pointed out that, in general, the first few principal components are sensitive to outliers that inflate variances or distort correlations, and the last few are sensitive to outliers that introduce artificial dimensions or mask singularities. We could examine the bivariate plots of at least the first two and the last two principal components in a search for outliers that may exert undue influence.

Devlin, Granados-Ruiz, and Kottnering (1991) recommended the extraction of principal components from robust estimates of Σ or $\mathbf{B}$ that reduce the influence of outliers. Campbell (1989) and Ruysschaert (1991) discussed direct robust estimation of principal components. Critchley (1992) developed methods for detection of influential observations in principal component analysis.

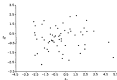


Figure 12.4. Plot of first four components for the modified football data.

Another feature of the data that a plot of the first two components may reveal is a tendency of the points to cluster. The plot may reveal groupings of points; this is illustrated in Example 12.4(d).

Example 12.4(a). For the modified football data in Example 12.1.2, the first two principal components were given as follows:

$$\begin{aligned} z_1 = \hat{\mathbf{e}}_1^T \mathbf{y} &= .297y_1 + .873y_2 + .264y_3 + .326y_4 + .666y_5 + .128y_6, \\ z_2 = \hat{\mathbf{e}}_2^T \mathbf{y} &= -.342y_1 - .317y_2 - .231y_3 + .494y_4 + .332y_5 - .487y_6. \end{aligned}$$

These are evaluated for each observation vector and plotted in Figure 12.4. (For convenience in making $\mathbf{y} - \bar{\mathbf{y}}$ was used in the computations.) The pattern is typical of that from a multivariate normal distribution. Note that the variance along the z_1 axis is greater than the variance in the z_2 direction, as expected. $\square$

Example 12.4(b). In Figures 4.3 and 6.10, the $(\hat{\mu} - \hat{\mu})$ plot and bivariate scatter plots for the same horse data of Table 3.6 exhibit a nonnormal pattern. A principal component analysis using the covariance matrix is given in Table 12.1, and the first two

Table 12.1. Principal Components for the Horse Race Data of Table 3.6

Number	Eigenvalue	First Two Eigenvectors		
		Variable	$\mathbf{a}_1$	$\mathbf{a}_2$
1	21.01	AGE (S)	.474	.592
2	3.74	AGE (S.S)	.493	.409
3	.31	AGE (P)	.319	-.508
4	.11	AGE (S.S)	.317	-.627

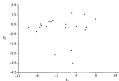


Figure 12.3. First two principal components for the same/boat data in Table 3.8.

principal components are plotted in Figure 12.5. The presence of these outliers that cause a nonnormal pattern is evident. These outliers do not appear when the four variables are examined individually. $\square$

Example 12.4(c). A rather extreme example of the effect of an outlier is given by Dorlin, Grossmanstein, and Kensoring (1981). The data set involved $p = 34$ environmental variables for $n = 29$ chemical compounds. The first two principal components are plotted in Figure 12.6. The sample correlation $\rho_{PC1,PC2}$ is indeed zero for all 29 points, as it must be (see 12.6f), but if the apparent outlier is excluded from the computation, then $\rho_{PC1,PC2} = .89$ for the remaining 28 points. If the outlier were deleted from the data set, the axes of the principal components would pass through the natural extensions of the data clouds. $\square$

Example 12.4(d). Jellies (1987) applied principal component analysis to a sample of 40 slate sculpins (winged aptodes) on which the following 19 variables had been measured:

LENGTH	body length
WIDTH	body width
POWERTO2	forewing length
HEADWTO2	head-wing length
SPINAC	number of spinules
ANTSEG 1	length of antennal segment I
ANTSEG 2	length of antennal segment II
ANTSEG 3	length of antennal segment III
ANTSEG 4	length of antennal segment IV
ANTSEG 5	length of antennal segment V

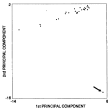


Figure 11.4. First two principal components for economic data.

ANTSPIN	number of antenae spines
TARSUS.1	leg length, tarsus III
TIBIA.1	leg length, tibia III
PEMUS.1	leg length, femur III
ROSTRUM	rostrum
OVIPUS	ovipositor
OVSPIN	number of ovipositor spines
FOLD	wing fold
HADKS	number of hindwing hooks

An objective in the study was to determine the number of distinct taxa present in the habitat where the sample was taken. Since subgenus are difficult to identify by the usual taxonomic methods, principal component analysis was used to search for groupings among the 40 individuals in the sample.

The correlation matrix is given in Table 11.2, and the eigenvalues and first four eigenvectors are in Tables 11.3 and 11.4, respectively. The eigenvectors are scaled so that the largest value in each is 1. The first principal component is largely an index of size. The second component is associated with *SPINAC*, *OVIPUS*, *OVSPIN*, and *FOLD*.

The first two components were computed for each of the 40 individuals and plotted in Figure 11.7. Since the first two components account for 85% of the total variance, the plot represents the data with very little distortion. There are four major

Table 12.1. Correlation Matrix for Winged Apterid Variables (Lower Triangle)

	F_1												
F_1	1.000												
F_2	.834	1.000											
F_3	.927	.541	1.000										
F_4	.989	.544	.979	1.000									
F_5	.934	.487	.949	.499	1.000								
F_6	.769	.321	.896	.313	.758	1.000							
F_7	.864	.368	.886	.889	.719	.823	1.000						
F_8	.789	.334	.846	.883	.556	.698	.731	1.000					
F_9	.833	.369	.882	.836	.462	.752	.793	.795	1.000				
F_{10}	.843	.378	.883	.881	.567	.836	.913	.787	.883	1.000			
F_{11}	-.458	-.495	-.522	-.488	-.174	-.307	-.383	-.497	-.356	-.381			
F_{12}	.917	.541	.949	.943	.545	.846	.928	.884	.848	.902			
F_{13}	.939	.563	.978	.933	.438	.849	.934	.876	.877	.924			
F_{14}	.933	.534	.948	.949	.432	.823	.886	.878	.883	.884			
F_{15}	.893	.399	.882	.868	.351	.833	.891	.796	.818	.833			
F_{16}	.893	.382	.886	.823	.313	.812	.833	.819	.826	.712			
F_{17}	.327	.308	.356	.272	.346	.253	.367	.367	.388	.384			
F_{18}	.376	-.312	-.347	-.336	-.236	-.384	-.302	-.378	-.386	-.329			
F_{19}	.761	.316	.746	.771	.385	.469	.362	.793	.671	.658			
	X_1												
F_{20}	1.000												
F_{21}	-.467	1.000											
F_{22}	-.447	.981	1.000										
F_{23}	-.438	.971	.991	1.000									
F_{24}	-.485	.928	.928	.921	1.000								
F_{25}	-.338	.723	.714	.676	.725	1.000							
F_{26}	-.332	.595	.589	.598	.578	.786	1.000						
F_{27}	.493	-.357	-.375	-.378	-.355	-.386	.349	1.000					
F_{28}	-.423	.695	.726	.731	.694	.287	.695	-.275	1.000				
		F_{29}								F_{30}			
F_{29}		1.000											
F_{30}		-.467	1.000										
F_{31}		-.447	.981	1.000									
F_{32}		-.438	.971	.991	1.000								
F_{33}		-.485	.928	.928	.921	1.000							
F_{34}		-.338	.723	.714	.676	.725	1.000						
F_{35}		-.332	.595	.589	.598	.578	.786	1.000					
F_{36}		.493	-.357	-.375	-.378	-.355	-.386	.349	1.000				
F_{37}		-.423	.695	.726	.731	.694	.287	.695	-.275	1.000			

groups, apparently corresponding to species. The groupings form an interesting S-shape. □

12.5. PRINCIPAL COMPONENTS FROM THE CORRELATION MATRIX

Generally, extracting components from $\mathbf{K}$ rather than $\mathbf{R}$ remains closer to the spirit and intent of principal component analysis, especially if the components are to be used in further computations. However, in some cases, the principal components will be more interpretable if $\mathbf{R}$ is used. For example, if the variances differ widely or if the measurement units are not commensurate, the components of $\mathbf{R}$ will be dominated by the variables with large variances. The other variables will contribute very little. For a more balanced representation in such cases, components of $\mathbf{K}$ may be used (see, for example, Problem 12.9).

Table K1.3. Eigenvalues of the Covariance Matrix of the Winged Aged Data

Component	Eigenvalue	Percent of Variance	Cumulative Percent
1	11.861	71.6	71.6
2	1.770	11.1	82.7
3	.748	4.7	87.4
4	.502	3.1	90.5
5	.378	2.4	92.9
6	.268	1.7	94.6
7	.193	1.2	95.8
8	.177	.8	96.7
9	.140	.7	97.4
10	.113	.6	98.1
11	.091	.4	98.6
12	.074	.4	99.0
13	.060	.3	99.3
14	.042	.2	99.5
15	.036	.2	99.7
16	.024	.1	99.8
17	.020	.1	99.9
18	.011	.1	100.0
19	.004	.0	100.0
11.000			

Table K1.4. Eigenvalues for the First Four Components of the Winged Aged Data

Variable	Eigenvalues			
	1	2	3	4
LENGTH	.96	-.06	.03	-.12
WIDTH	.96	-.12	.01	-.16
POWERS	.93	-.06	-.06	-.11
EDWARDS	.98	-.10	.03	-.09
SPRAC	.61	.34	-.28	1.00
ANTSEG-1	.91	.56	.04	.05
ANTSEG-2	.96	.50	.06	-.04
ANTSEG-3	.98	-.43	.06	-.34
ANTSEG-4	.96	-.06	.36	-.04
ANTSEG-5	.94	.06	.31	.06
ANTSEG	-.46	.37	1.00	.27
TARGET-1	.99	-.06	.03	-.26
TEMP-1	1.00	-.06	.06	-.34
TEMP-2	.99	-.12	.02	-.34
BOSTER-90	.96	.06	.06	-.06
ENTPOG	.76	.73	-.03	-.06
ENTPOG	.41	1.00	-.36	-.06
PCOLD	-.75	.66	.04	-.80
BOPOG	.76	-.50	.06	.12

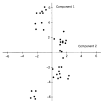


Figure 12.3. Plotted values of the first two components for individual insects.

As with any change of scale, when the variables are standardized in transforming from $\mathbf{X}$ to $\mathbf{Z}$, the shape of the clouds of points will change. Note, however, that after transforming to $\mathbf{Z}$, any further changes of scale on the variables would not affect the components because changes of scale do not change $\mathbf{R}$. Thus the principal components from $\mathbf{R}$ are scale invariant.

To illustrate how the eigenvalues and eigenvectors change when converting from $\mathbf{S}$ to $\mathbf{R}$, we use a simple bivariate example in which one variance is substantially larger than the other. Suppose that $\mathbf{S}$ and the corresponding $\mathbf{R}$ have the values

$$\mathbf{S} = \begin{pmatrix} 1 & .4 \\ .4 & .25 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} 1 & .8 \\ .8 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors from $\mathbf{R}$ are

$$\begin{aligned} \lambda_1 &= .2569, & \mathbf{a}_1' &= (.180, .987), \\ \lambda_2 &= .7431, & \mathbf{a}_2' &= (.987, -.180). \end{aligned}$$

The patterns we see in λ_1 , λ_2 , $\mathbf{a}_1$, and $\mathbf{a}_2$ are quite predictable. The symmetry in $\mathbf{a}_1$ and $\mathbf{a}_2$ is due to their orthogonality, $\mathbf{a}_1'\mathbf{a}_2 = 0$. The large variance of y_1 in $\mathbf{S}$ is reflected in the first principal component $\tau_1 = .160y_1 + .987y_2$, where y_2 is weighted heavily. Thus the first principal component τ_1 essentially duplicates y_2 and does not show the mutual effect of y_1 and y_2 . As expected, τ_1 accounts for virtually all of the

total variance:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{23.65}{28} = .845.$$

The eigenvalues and eigenvectors of $\mathbf{B}$ are

$$\begin{aligned}\lambda_1 &= 1.8, & \mathbf{b}_1' &= (.703, .707), \\ \lambda_2 &= .2, & \mathbf{b}_2' &= (.703, -.707).\end{aligned}$$

The first principal component of $\mathbf{B}$,

$$z_1 = .707 \frac{y_1 - \bar{y}_1}{1} + .707 \frac{y_2 - \bar{y}_2}{2},$$

accounts for a high proportion of variance,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.8}{2} = .9,$$

because the variables are fairly highly correlated ($r = .8$). But the standardized variables $(y_1 - \bar{y}_1)/1$ and $(y_2 - \bar{y}_2)/2$ are equally weighted in z_1 , due to the equality of the diagonal elements ("variances") of $\mathbf{B}$.

We now list some general comparisons of principal components from $\mathbf{B}$ with those from $\mathbf{S}$.

1. The percent of variance in (1.2.5) accounted for by the components of $\mathbf{B}$ will differ from the percent for $\mathbf{S}$, as illustrated above.
2. The coefficients of the principal components from $\mathbf{B}$ differ from those obtained from $\mathbf{S}$, as illustrated above.
3. If we express the components from $\mathbf{B}$ in terms of the original variables, they still will not agree with the components from $\mathbf{S}$. By transforming the standardized variables back to the original variables in the above illustration, the components of $\mathbf{B}$ become

$$\begin{aligned}z_1 &= .707 \frac{y_1 - \bar{y}_1}{1} + .707 \frac{y_2 - \bar{y}_2}{2} \\ &= .707 y_1 + .141 y_2 + \text{const.} \\ z_2 &= .707 \frac{y_1 - \bar{y}_1}{1} - .707 \frac{y_2 - \bar{y}_2}{2} \\ &= .707 y_1 - .141 y_2 + \text{const.}\end{aligned}$$

As expected, these are very different from the components extracted directly from $\mathbf{S}$. This problem arises, of course, because of the lack of scale invariance of the components of $\mathbf{S}$.

4. The principal components from $\mathbf{R}$ are scale invariant, because $\mathbf{R}$ itself is scale invariant.
5. The components from a given matrix $\mathbf{R}$ are not unique to that $\mathbf{R}$. For example, in the bivariate case, the eigenvalues of

$$\mathbf{R} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

are given by

$$\lambda_1 = 1 + r, \quad \lambda_2 = 1 - r, \quad (12.13)$$

and the eigenvectors are $\mathbf{a}_1' = (.707, .707)$ and $\mathbf{a}_2' = (.707, -.707)$, which give principal components

$$\begin{aligned} z_1 &= .707 \frac{Y_1 - \bar{Y}_1}{s_1} + .707 \frac{Y_2 - \bar{Y}_2}{s_2}, \\ z_2 &= .707 \frac{Y_1 - \bar{Y}_1}{s_1} - .707 \frac{Y_2 - \bar{Y}_2}{s_2}. \end{aligned} \quad (12.14)$$

The components in (12.14) do not depend on r . For example, they serve equally well for $r = .50$ and for $r = .99$. For $r = .50$, the proportion of variance explained by z_1 is $\lambda_1/(\lambda_1 + \lambda_2) = (1 + .50)/(1 + .50 + 1 - .50) = 1.50/2 = .75$. For $r = .99$, the ratio is $1.99/2 = .995$. Thus the statement that the first component from a correlation matrix accounts for, say, 90% of the variance is not very meaningful. In general, for $p > 2$, the components from $\mathbf{R}$ depend only on the ratios (relative values) of the correlations, not on their actual values, and components of a given $\mathbf{R}$ matrix will serve for other $\mathbf{R}$ matrices [see Rencher (1995, Section 9.4)].

12.6 DECIDING HOW MANY COMPONENTS TO RETAIN

In every applications, a decision must be made on how many principal components should be retained in order to effectively summarize the data. The following guidelines have been proposed:

1. Retain sufficient components to account for a specified percentage of the total variance, say, 80%.
2. Retain the components whose eigenvalues are greater than the average of the eigenvalues, $\sum_{i=1}^p \lambda_i/p$. For a correlation matrix, this average is 1.
3. Use the scree graph, a plot of λ_i versus i , and look for a natural break between the "large" eigenvalues and the "small" eigenvalues.
4. Test the significance of the "large" components, that is, the components corresponding to the larger eigenvalues.

We now discuss these four criteria for choosing the components to keep. Note, however, that the smallest components may carry valuable information that should not be routinely ignored (see Section 12.7).

In method 1, the challenge lies in selecting an appropriate threshold percentage. If we aim too high, we run the risk of including components that are either sample specific or variable specific. By sample specific we mean that a component may not generalize to the population or to other samples. A variable-specific component is dominated by a single variable and does not represent a composite summary of several variables.

Method 2 is widely used and is the default in many software packages. By (11.17), $\sum_{i=1}^p \lambda_i = 100$, and the average eigenvalue is also the average variance of the individual variables. Thus method 2 retains those components that account for more variance than the average variance of the variables. In cases where the data can be successfully summarized in a relatively small number of dimensions, there is often a wide gap between the two eigenvalues that fall on both sides of the average. In Example 12.2.1, the average eigenvalue (of 8) for the football data is .977, which is amply bracketed by $\lambda_4 = 1.57$ and $\lambda_5 = .48$. In the winged sprint data in Example 12.4(4), the second and third eigenvalues (of 8) are 2.030 and .346, leaving a comfortable margin on both sides of 1. In some cases, one may wish to move the cutoff point slightly to accommodate a visible gap in eigenvalues.

The scree graph in method 2 is named for its similarity in appearance to a cliff with rocky debris at its bottom. The scree graph for the modified football data of Example 12.2.2 exhibits an ideal pattern, as shown in Figure 11.8. The first two eigenvalues form a steep curve followed by a bend and then a straight line (and with shallower slope). The recommendation is to retain those eigenvalues in the steep curve before the first one on the straight line. Thus in Figure 11.8, two components would be retained. In practice, the turning point between the steep curve and the straight line may not be as distinct as this or there may be more than one discernible bend. In such cases, this approach is not as conclusive. The scree graph for the winged sprint

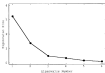


Figure 11.8. Scree graph for eigenvalues of modified football data.

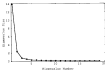


Figure 12.8. Scree plot for eigenvalues of winged-spliid data.

data in Example 12.4(d) is plotted in Figure 12.8. The plot would suggest that two components be retained (possibly three).

The remainder of this section is devoted to method 4, tests of significance. The tests assume multivariate normality, which is not required for estimation of principal components.

It may be useful to make a preliminary test of complete independence of the variables, as in Section 7.4.3: $H_0: \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$, or equivalently, $H_0: \rho_{ij} = 0$. The test statistic is given in (7.35) and (7.36). If the results indicate that the variables are independent, there is no point in extracting principal components, since (except for sampling fluctuation) the variables themselves already form the principal components.

To test the significance of the “large” components, we test the hypothesis that the last k population eigenvalues are small and equal, $H_{0k}: \nu_{p-k+1} = \nu_{p-k+2} = \dots = \nu_p$, where $\nu_1, \nu_2, \dots, \nu_p$ denote the population eigenvalues, namely, the eigenvalues of Σ . The implication is that the first sample components capture all the essential dimensions, whereas the last components reflect noise. If H_0 is true, the last k sample eigenvalues will tend to have the pattern shown by the straight line with small slope in the ideal scree graph, such as in Figure 12.8 or 12.9.

To test $H_{0k}: \nu_{p-k+1} = \dots = \nu_p$ using a likelihood ratio approach, we calculate the average of the last k eigenvalues of S ,

$$\bar{\Sigma} = \sum_{i=p-k+1}^p \frac{\lambda_i}{k},$$

and use the test statistic

$$u = \left(n - \frac{2p+11}{6} \right) \left(2 \ln \bar{\Sigma} - \sum_{i=p-k+1}^p \ln \lambda_i \right), \quad (12.10)$$

which has an approximate χ^2 -distribution. We reject H_0 if $s \geq \chi_{\alpha, r}^2$, where $r = \frac{1}{2}(k-1)(d+2)$.

To carry out this procedure, we could begin by testing $H_{01}: \gamma_{p-1} = \gamma_p$. If this is accepted, we could then test $H_{02}: \gamma_{p-2} = \gamma_{p-1} = \gamma_p$ and continue testing in this fashion until H_{0k} is rejected for some value of k .

In practice, when the variables are fairly highly correlated and the data can be successfully represented by a small number of principal components, the first three methods will typically agree on the number of components to retain, and the test in method 4 will often indicate a large number of components.

Example 12.4. We apply the preceding four criteria to the modified football data of Example 12.1.2.

For method 1, we simply examine the eigenvalues and their proportion of variance explained, as obtained in Example 12.1.2:

Eigenvalue	Proportion of Variance	Cumulative Proportion
3.523	.714	.579
1.074	.214	.814
.479	.093	.908
.328	.062	.967
.137	.027	.993
.088	.015	1.000

To account for 92% of the variance, we would keep two components. This percent of variance is high enough for most descriptive purposes. For certain other applications, such as input to another analysis, we might wish to retain three components, which would account for 96% of the variance.

To apply method 2, we find the average eigenvalue to be

$$\bar{\lambda} = \sum_{j=1}^6 \frac{\lambda_j}{6} = \frac{5.74324}{6} = .957.$$

Since only λ_1 and λ_2 exceed .957, we would retain two components.

For method 3, the scree graph in Figure 12.8 indicates conclusively that two components should be retained.

To implement method 4, we carry out the significance tests in (12.15). The values of the test statistic s for $k = 2, 3, \dots, 6$ are as follows:

Eigenvalue	k	s	df	χ_{α}^2
3.52348	6	245.57	20	34.41
1.07408	5	125.80	14	25.68
.47987	4	44.18	9	16.92
.32488	3	25.84	5	11.07
.13888	2	4.62	2	5.99
.08793	1			

The tests indicate that only the last two (population) eigenvalues are equal, and we should retain the first four. This differs from the results of the other three criteria, which are in close agreement that two components should be retained. $\square$

12.7 INFORMATION IN THE LAST FEW PRINCIPAL COMPONENTS

Up to this point, we have focused on using the first few principal components to summarize and simplify the data. However, the last few components may carry useful information in some applications.

Since the eigenvalues serve as variances of the principal components, the last few principal components have smaller variances. If the variance of a component is zero or close to zero, the component represents a linear relationship among the variables that is essentially constant; that is, the relationship holds for all $\mathbf{x}_i$'s in the sample. Thus if the last eigenvalue is near zero, it signifies the presence of a collinearity that may provide new information for the researcher. Suppose, for example, that there are five variables and $\mathbf{x}_1 = \sum_{j=2}^5 \gamma_j \mathbf{x}_j$ ¹⁴. Then $\mathbf{K}$ is singular, and having round-off error, λ_2 will be zero. Thus $\mathbf{z}_2^T = \hat{0}$, and $\mathbf{z}_2$ is constant. As noted early in Section 12.2, the $\mathbf{z}_i$'s are centered, because the origin of the principal components is translated to $\bar{\mathbf{y}}$. Hence the constant value of $\mathbf{z}_2$ is its mean, which is zero:

$$\mathbf{z}_2 = \mathbf{a}_2^T \mathbf{y} = a_{21}y_1 + a_{22}y_2 + \cdots + a_{25}y_5 = \bar{0}.$$

Since this must reflect the dependency of y_1 on y_2, y_3, y_4 , and y_5 , the eigenvector $\mathbf{a}_2^T$ will be proportional to $(1, 1, 1, 1, -4)$.

12.8 INTERPRETATION OF PRINCIPAL COMPONENTS

In Section 12.5, we noted that principal components obtained from $\mathbf{K}$ are not compatible with those obtained from $\mathbf{S}$. Because of this lack of scale invariance of principal components from $\mathbf{K}$, the coefficients cannot be converted to standardized form, as can be done with coefficients in discriminant functions in Chapter 9 and canonical variates in Chapter 11. Hence interpretation of principal components is not as clear-cut as with previous linear functions that we have discussed. We must choose between components of $\mathbf{S}$ or $\mathbf{K}$, knowing they will have a different interpretation. If the variables have widely disparate variances, we may use $\mathbf{K}$ instead of $\mathbf{S}$ to improve interpretation.

For certain patterns of elements in $\mathbf{S}$ or $\mathbf{K}$, the form of the principal components can be predicted. This and its interpretation is discussed in Section 12.8.1. As with discriminant functions and canonical variates, some writers have advocated rotation and the use of correlations between the variables and the principal components. We again applied the test of these two approaches to interpretation in Sections 12.8.2 and 12.8.3.

12.8.1 Special Patterns in $\mathbf{S}$ or $\mathbf{R}$

In the covariance or correlation matrix, we may recognize a distinguishing pattern from which the structure of the principal components can be deduced. For example, we noted in Section 12.1 that if one variable has a much larger variance than the other variables, this variable will dominate the first component, which will account for most of the variance. Another case in which a component will duplicate a variable occurs when the variable is uncorrelated with the other variables. We now demonstrate this by showing that if all p variables are uncorrelated, the variables themselves are the principal components. If the variables were uncorrelated (orthogonal), $\mathbf{S}$ would have the form

$$\mathbf{S} = \begin{pmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{pp} \end{pmatrix}, \quad (12.16)$$

and the characteristic equation would be

$$0 = |\mathbf{S} - \lambda \mathbf{I}| = \prod_{i=1}^p (s_{ii} - \lambda) \quad \text{by (2.83A),}$$

which has solutions

$$\lambda_i = s_{ii}, \quad i = 1, 2, \dots, p. \quad (12.17)$$

The corresponding normalized eigenvectors have a 1 in the i th position and 0's elsewhere:

$$\mathbf{a}_i^* = (0, \dots, 0, 1, 0, \dots, 0). \quad (12.18)$$

Thus the i th component is

$$z_i = \mathbf{a}_i^* \mathbf{S} = y_i.$$

In practice, the sample correlations (or covariances) among variables will not be zero, but if the correlations are all small, the principal components will largely duplicate the variables.

By the Borel-Fischer theorem, in Section 2.11.4, if all correlations or covariances are positive, all elements of the first eigenvector $\mathbf{a}_1$ are positive. Since the remaining eigenvectors $\mathbf{a}_2, \mathbf{a}_3, \dots, \mathbf{a}_p$ are orthogonal to $\mathbf{a}_1$, they must have both positive and negative elements. When all elements of $\mathbf{a}_1$ are positive, the first component is a weighted average of the variables and is sometimes referred to as a measure of size. Likewise, the positive and negative coefficients in subsequent components

may be regarded as defining shape. This pattern is often seen when the variables are random measurements of an organism.

Example 12.8.1. In the modified football data of Example 12.2.2, there are a few negative covariances in $\mathbf{S}$, but they are small, and all elements of the first eigenvector remain positive. The second eigenvector therefore has positive and negative elements:

First Two Eigenvectors		
	$\mathbf{a}_1$	$\mathbf{a}_2$
WEIGH	.207	-.142
CIRCUM	.203	-.218
FRONT	.261	-.238
HYPER	.125	.891
BARREL	.005	.222
LOF	.128	-.167

With all positive coefficients, the first component τ_1 is an overall measure of head size (τ_1 increases if all six variables increase). The second component τ_2 is a shape component that contrasts the vertical measurements (HYPER and BARREL) with the three lateral measurements and CIRCUM (τ_2 increases if HYPER and BARREL increase and the other four variables decrease). $\square$

12.8.2 Rotation

The principal components are initially obtained by rotating axes in order to line up with the natural extensions of the system, whereupon the new variables become uncorrelated and reflect the directions of maximum variance. If the resulting components do not have a satisfactory interpretation, they can be further rotated, seeking directions in which many of the coefficients of the linear combinations are near zero to simplify interpretation.

However, the new rotated components are correlated, and they do not necessarily account for maximum variance. They are, therefore, no longer principal components in the usual sense, and their routine use is questionable. For improved interpretation, you may wish to try factor analysis (Chapter 13), in which rotation does not destroy any properties. (In factor analysis, the statistic does not involve the space of the variables $x_1, x_2, \dots, x_p$, but another space, that of the factor loadings.)

12.8.3 Correlations between Variables and Principal Components

The use of correlations between variables and principal components is widely recommended as an aid to interpretation. It was noted in Sections 8.7.3 and 11.3.2 that analogous correlations for discriminant functions and canonical variates are not useful in a multivariate context because they provide only univariate information about how each variable operates by itself, ignoring the other variables. Rencher (1995b) obtained a similar result for principal components.

We denote the correlation between the i th variable y_i and the j th principal component z_j by $r_{y_i z_j}$. Because of the orthogonality of the z_j 's, we have the simple relationship

$$r_{y_i z_1}^2 + r_{y_i z_2}^2 + \cdots + r_{y_i z_d}^2 = R_{y_i(z_1, \dots, z_d)}^2 \quad (12.18)$$

where d is the number of components retained and $R_{y_i(z_1, \dots, z_d)}^2$ is the squared multiple correlation of y_i with the z_j 's. Thus $r_{y_i z_1}^2$ forms part of $R_{y_i(z_1, \dots, z_d)}^2$, which shows how y_i relates to the z_j 's by itself, not what it contributes in the presence of the other y 's. The correlations are, therefore, not informative about the joint contribution of the y 's in a principal component.

Note that the simple partitioning of R^2 into the sum of squares of correlations in (12.18) does not happen in practice when the independent variables (y 's) are correlated. However, here the z 's are principal components and are, therefore, orthogonal.

Since we do not recommend rotation or correlations for interpretation, we are left with the coefficients themselves, obtained from the eigenvectors of either $\mathbf{S}$ or $\mathbf{R}$.

Example 12.8.3. In Example 12.8.1, the eigenvectors of $\mathbf{S}$ from the modified football data give a satisfactory interpretation of the first two principal components as *lean size* and *shape*. We give these in Table 12.5, along with the correlations between each of the variables $y_1, y_2, \dots, y_6$ and the first two principal components z_1 and z_2 . For comparison we also give $R_{y_i(z_1, z_2)}^2$ for each variable.

The correlations rank the variables somewhat differently in their contribution to the components, since they form part of the composite information provided by R^2 for each variable by itself. For example, for the first component, the correlations rank the variables in the order 2, 3, 1, 4, 6, 5, whereas the coefficients (first eigenvector) from $\mathbf{S}$ rank them in the order 2, 4, 3, 1, 6, 5.

12.9 SELECTION OF VARIABLES

We have previously discussed subset selection in connection with Wilks' Λ (Section 9.11.2), discriminant analysis (Section 8.9), classifi- cation analysis (Section 9.8),

Table 12.5. Eigenvectors Obtained from $\mathbf{S}$, Correlations between Variables and Principal Components, and R^2 for the First Two Principal Components

Variable	Eigenvectors from $\mathbf{S}$		Correlations		$R_{y_i(z_1, z_2)}^2$
	z_1	z_2	$r_{y_i z_1}$	$r_{y_i z_2}$	
1	.21	-.14	.62	-.27	.46
2	.87	-.11	.98	-.16	.99
3	.36	-.13	.76	-.40	.86
4	.33	.69	.46	.86	.88
5	.67	.22	.37	.37	.17
6	.13	-.39	.46	-.39	.32

and regression Sections 10.2.7 and 10.3). In each case the criterion for selection of variables is the relationship of the variables to some external factor, such as dependent variable(s), separation of groups, or correct classification rates. In the context of principal components, we have no dependent variable, as in regression, and no groupings among the observations, as in discriminant analysis. Without external influences, we simply wish to find the subset that best captures the internal variation (and covariation) of the variables.

Jolliffe (1972, 1973) discussed eight selection methods and referred to the process as discarding variables. The eight methods were based on three basic approaches: multiple correlation, clustering of variables, and principal components. One of the correlation methods, for example, proceeds in a stepwise fashion, deleting at each step the variable that has the largest multiple correlation with the other variables. The clustering methods partition the variables into groups or clusters (see Chapter 14) and select a variable from each cluster.

We describe Jolliffe's principal component methods in the context of selecting a subset of 10 variables out of 50 variables. One of his techniques associates a variable with each of the first 10 principal components and retains these 10 variables. Another approach is to associate a variable with each of the last 40 principal components and delete the 40 variables. To associate a variable with a principal component, we choose the variable corresponding to the largest coefficient (in absolute value) in the component, providing the variable has not previously been selected. We can use components extracted from either $\mathbf{R}$ or $\mathbf{B}$. For example, in the two principal components for the football data in Example 12.2.1, we would choose variable 2 and 4, which clearly have the largest coefficients in the two components. Jolliffe's methods could also be applied iteratively, with the principal components being recomputed after a variable is retained or deleted.

Jolliffe (1972) compared the eight methods using both real and simulated data and found that the methods based on principal components performed well in comparison to the regression and cluster-based methods. But he concluded that no single method was uniformly best.

McCabe (1984) suggested several criteria for selection, most of which are based on the conditional covariance matrix of the variables not selected, given those selected. He denoted the selected variables as *principal variables*. Let $\mathbf{y}$ be partitioned as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

where $\mathbf{y}_1$ contains the selected variables and $\mathbf{y}_2$ consists of the variables not selected. The corresponding covariance matrix is

$$\text{cov}(\mathbf{y}) = \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

By (4.8), the conditional covariance matrix is given by (assuming normality)

$$\text{cov}(\mathbf{y}_2|\mathbf{y}_1) = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12},$$

which is estimated by $\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$. To find a subset $\mathbf{y}_1$ of size m , two of McCabe's criteria are to choose the subset $\mathbf{y}_1$ that

1. minimizes $|\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}|$ and
2. maximizes $\sum_{i=1}^p r_i^2$, where $r_i, i = 1, 2, \dots, m^* = \min(m, p-m)$ are the canonical correlations between the m selected variables in $\mathbf{y}_1$ and the $p-m$ deleted variables in $\mathbf{y}_2$.

Clearly, these criteria would be evaluated for all possible subsets of size m to obtain the best subset of such size. McCabe suggested a regression approach for obtaining a percent of variance explained by a subset of variables to be compared with the percent of variance accounted for by the same number of principal components.

PROBLEMS

(11.1) Show that the solutions to $\lambda = \alpha(\mathbf{S}\mathbf{a})$'s in (11.7) are given by the eigenvalues and eigenvectors in (11.8), so that λ in (11.7) is maximized by the largest eigenvalue of $\mathbf{B}$.

(11.2) Show that the eigenvalues of

$$\mathbf{B} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

are $1 \pm r$, as in (11.17), and that the eigenvectors are as given in (11.14).

(11.3) (a) Give a justification based on the likelihood ratio for the test statistic v in (11.13).

(b) Give a justification for the degrees of freedom $v = \frac{1}{2}(k-1)(d+1)$ for the test statistic in (11.15).

(11.4) Show that when $\mathbf{B}$ is diagonal as in (11.18), the eigenvectors have the form $\mathbf{a}_i = (0, \dots, 0, 1, 0, \dots, 0)$, as given in (11.18).

(11.5) Show that $r_{1k_1}^2 + r_{2k_1}^2 + \dots + r_{mk_1}^2 = \mathbf{R}_{1k_1, \dots, mk_1}^2$ in (11.23).

(11.6) Carry out a principal component analysis of the diabetes data of Table 3.4. Use all five variables, including y 's and x 's. Use both $\mathbf{B}$ and $\mathbf{R}$. Which do you think is more appropriate here? Show the percent of variance explained.

Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either **S** or **R**?

- 12.7** Do a principal component analysis of the probe level data of Table 3.5. Use both **S** and **R**. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either **S** or **R**?
- 12.8** Carry out a principal component analysis on all six variables of the glucose data of Table 3.8. Use both **S** and **R**. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either **S** or **R**?
- 12.9** Carry out a principal component analysis on the hematology data of Table 4.3. Use both **S** and **R**. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either **S** or **R**? Does the large variance of y_6 affect the pattern of the components of **R**?
- 12.10** Carry out a principal component analysis separately for males and females in the psychological data of Table 5.1. Compare the results for the two groups. Use **S**.
- 12.11** Carry out a principal component analysis separately for the two species in the beetle data of Table 5.5. Compare the results for the two groups. Use **S**.
- 12.12** Carry out a principal component analysis on the engine data of Table 5.6 as follows:
- Use the pooled covariance matrix.
 - Ignore groups and use a covariance matrix based on all 40 observations.
 - Which of the approaches in (a) or (b) appears to be more successful?
- 12.13** Repeat the previous problem for the dystrophy data of Table 5.7.
- 12.14** Carry out a principal component analysis on all 10 variables of the fishes data of Table 7.1. Use both **S** and **R**. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either **S** or **R**?
- 12.15** Carry out a principal component analysis on the temperature data of Table 7.2. Use both **S** and **R**. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either **S** or **R**?

Factor Analysis

13.1 INTRODUCTION

In factor analysis we represent the variables $y_1, y_2, \dots, y_p$ as linear combinations of a few random variables $f_1, f_2, \dots, f_m$ ($m < p$) called factors. The factors are underlying (unobservable) variables that “generate” the y 's. Like the original variables, the factors vary from individual to individual, but unlike the variables, the factors cannot be measured or observed. The existence of these hypothetical variables is therefore open to question.

If the original variables $y_1, y_2, \dots, y_p$ are at least moderately correlated, the basic dimensionality of the system is less than p . The goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors.

Suppose the pattern of the high and low correlations in the correlation matrix is such that the variables in a particular subset have high correlations among themselves, but low correlations with all the other variables. Then there may be a single underlying factor that gave rise to the variables in the subset. If the other variables can be similarly grouped into subsets with a like pattern of correlations, then a few factors can represent these groups of variables. In this case the pattern in the correlation matrix corresponds directly to the factors. For example, suppose the correlation matrix has the form

$$\begin{pmatrix} 1.00 & .80 & .05 & .03 & .05 \\ .80 & 1.00 & .08 & .02 & .05 \\ .05 & .05 & 1.00 & .80 & .90 \\ .03 & .03 & .80 & 1.00 & .80 \\ .05 & .05 & .08 & .80 & 1.00 \end{pmatrix}.$$

Then variables 1 and 2 correspond to a factor, and variables 3, 4, and 5 correspond to another factor. In some cases where the correlation matrix does not have such a simple pattern, factor analysis will still partition the variables into clusters.

Factor analysis is related to principal component analysis in that both seek a simplification in a set of variables but they differ in many respects (see Section 13.8). For example, two differences in basic approach are as follows:

1. Principal components are defined as linear combinations of the original variables. In factor analysis, the original variables are expressed as linear combinations of the factors.
2. In principal component analysis, we explain a large part of the total variance of the variables, $\sum_1^p \sigma_i^2$. In factor analysis, we seek to account for the variations or covariations among the variables.

In practice, there are some data sets for which the factor analysis model does not provide a satisfactory fit. Thus, factor analysis remains somewhat subjective in many applications, and it is considered controversial by some statisticians. Sometimes a few easily interpretable factors emerge, but for other data sets, neither the number of factors nor the interpretation is clear. Some possible reasons for these failures are discussed in Section 13.7.

13.2 ORTHOGONAL FACTOR MODEL

13.2.1 Model Definition and Assumptions

Factor analysis is basically a one-sample procedure (for possible applications to data with groups, see Kuehler (1994, Section 13.8)). We assume a random sample $y_1, y_2, \dots, y_p$ from a homogeneous population with mean vector μ and covariance matrix Σ .

The factor analysis model expresses each variable as a linear combination of underlying common factors $f_1, f_2, \dots, f_m$ with an accompanying error term to account for that part of the variable that is unique (not in common with the other variables). For $y_1, y_2, \dots, y_p$ in any observation vector y , the model is as follows:

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1m}f_m + \varepsilon_1 \\ y_2 - \mu_2 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \cdots + \lambda_{2m}f_m + \varepsilon_2 \\ &\vdots \\ y_p - \mu_p &= \lambda_{p1}f_1 + \lambda_{p2}f_2 + \cdots + \lambda_{pm}f_m + \varepsilon_p \end{aligned} \quad (13.1)$$

Ideally, m should be substantially smaller than p ; otherwise we have not achieved a parsimonious description of the variables as functions of a few underlying factors. We might regard the f 's in (13.1) as random variables that engender the y 's. The coefficients λ_{ij} are called *loadings* and serve as weights, showing how each y_i individually depends on the f 's. (In this chapter we defer to some more usage in the factor analysis literature and use the notation λ_{ij} for loadings rather than eigenvalues.) With appropriate assumptions, λ_{ij} indicates the importance of the j th factor f_j to the i th variable y_i and can be used in interpretation of f_j . We describe or interpret f_2 , for example, by examining its coefficients, $\lambda_{21}, \lambda_{22}, \dots, \lambda_{2p}$. The larger loadings show f_2 to be corresponding y 's. From these y 's, we infer a meaning or description of f_2 . After estimating the λ_{ij} 's (and rotating them; see Sections 13.2.2 and 13.2), it is hoped they will partition the variables into groups corresponding to factors.

The system of equations (13.1) bears a superficial resemblance to the multiple regression model (10.1), but there are fundamental differences. For example, (1) the y 's are unobserved and (2) the model in (13.1) represents only one observation vector, whereas (10.1) depicts all n observations.

It is assumed that for $j = 1, 2, \dots, m$, $E(f_{ij}) = 0$, $\text{var}(f_{ij}) = 1$, and $\text{cov}(f_{ij}, f_{ik}) = 0$, $j \neq k$. The assumptions for u_i , $i = 1, 2, \dots, p$, are similar, except that we must allow each u_i to have a different variance, since it shows the residual part of y_i that is not in common with the other variables. Thus we assume that $E(u_i) = 0$, $\text{var}(u_i) = \phi_i$, and $\text{cov}(u_i, u_j) = 0$, $i \neq j$. In addition, we assume that $\text{cov}(u_i, f_{ij}) = 0$ for all i and j . We refer to ϕ_i as the specific variance.

These assumptions are natural consequences of the basic model (13.1) and the goals of factor analysis. Since $E(y_i - \mu_i) = 0$, we need $E(f_{ij}) = 0$, $j = 1, 2, \dots, m$. The assumption $\text{cov}(f_{ij}, f_{ik}) = 0$ is made for parsimony in expressing the y 's as functions of as few factors as possible. The assumptions $\text{var}(f_{ij}) = 1$, $\text{var}(u_i) = \phi_i$, $\text{cov}(f_{ij}, f_{ik}) = 0$, and $\text{cov}(u_i, f_{ij}) = 0$ yield a simple expression for the variance of y_i :

$$\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \phi_i, \quad (13.2)$$

which plays an important role in our development. Note that the assumption $\text{cov}(u_i, u_j) = 0$ implies that the factors account for all the correlations among the y 's, that is, all that the y 's have in common. Thus the emphasis in factor analysis is on modeling the covariances or correlations among the y 's.

Model (13.1) can be written in matrix notation as

$$\mathbf{y} - \boldsymbol{\mu} = \mathbf{A}\mathbf{f} + \mathbf{u}, \quad (13.3)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)'$, $\mathbf{f} = (f_1, f_2, \dots, f_m)'$, $\mathbf{u} = (u_1, u_2, \dots, u_p)'$, and

$$\mathbf{A} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & & \vdots \\ \lambda_{pn} & \lambda_{pn} & \cdots & \lambda_{pm} \end{pmatrix}. \quad (13.4)$$

We illustrate the model in (13.1) and (13.3) with $p = 2$ and $m = 3$. The model for each variable in (13.1) becomes

$$\begin{aligned}
y_1 - \mu_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \varepsilon_1 \\
y_2 - \mu_2 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \varepsilon_2 \\
y_3 - \mu_3 &= \lambda_{31}f_1 + \lambda_{32}f_2 + \varepsilon_3 \\
y_4 - \mu_4 &= \lambda_{41}f_1 + \lambda_{42}f_2 + \varepsilon_4 \\
y_5 - \mu_5 &= \lambda_{51}f_1 + \lambda_{52}f_2 + \varepsilon_5.
\end{aligned}$$

In matrix notation as in (13.1), this becomes

$$\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ y_3 - \mu_3 \\ y_4 - \mu_4 \\ y_5 - \mu_5 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}. \quad (13.2)$$

$$y\mathbf{F} - \boldsymbol{\mu} = \mathbf{LF} + \mathbf{v}.$$

The assumptions listed between (13.1) and (13.2) can be expressed concisely using vector and matrix notation: $E(f_j) = 0$, $j = 1, 2, \dots, m$, becomes

$$E(\mathbf{f}) = \mathbf{0}, \quad (13.3)$$

$\text{var}(f_j) = 1$, $j = 1, 2, \dots, m$, and $\text{cov}(f_j, f_k) = 0$, $j \neq k$, becomes

$$\text{cov}(\mathbf{f}) = \mathbf{I}, \quad (13.4)$$

$E(\varepsilon_i) = 0$, $i = 1, 2, \dots, p$, becomes

$$E(\mathbf{v}) = \mathbf{0}, \quad (13.5)$$

$\text{var}(\varepsilon_i) = \psi_i$, $i = 1, 2, \dots, p$, and $\text{cov}(\varepsilon_i, \varepsilon_k) = 0$, $i \neq k$, becomes

$$\text{var}(\mathbf{v}) = \boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}, \quad (13.6)$$

and $\text{cov}(f_i, f_j) = 0$ for all i and j becomes

$$\text{cov}(\mathbf{f}, \mathbf{v}) = \mathbf{0}. \quad (13.7)$$

The notation $\text{cov}(\mathbf{f}, \mathbf{v})$ indicates a rectangular matrix containing the covariances of the f 's with the v 's.

$$\text{cov}(\mathbf{z}, \mathbf{z}) = \begin{pmatrix} \sigma_{z_1 z_1}^2 & \sigma_{z_1 z_2}^2 & \cdots & \sigma_{z_1 z_p}^2 \\ \sigma_{z_2 z_1}^2 & \sigma_{z_2 z_2}^2 & \cdots & \sigma_{z_2 z_p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{z_p z_1}^2 & \sigma_{z_p z_2}^2 & \cdots & \sigma_{z_p z_p}^2 \end{pmatrix}.$$

It was noted following (13.7) that the emphasis in factor analysis is on modeling the covariances among the y 's. We wish to express the $\frac{1}{2}p(p-1)$ covariances (and the p variances) of the variables $y_1, y_2, \dots, y_p$ in terms of a simplified structure involving the p m loadings λ_{ij} and the p specific variances ψ_i ; that is, we wish to express Σ in terms of $\mathbf{A}$ and Ψ . We can do this using the model (13.5) and the assumptions (13.7a), (13.7b), and (13.10). Since μ does not affect variances and covariances of y , we have, from (13.8),

$$\Sigma = \text{cov}(\mathbf{y}) = \text{cov}(\mathbf{A}\mathbf{f} + \mathbf{z}).$$

By (13.10), $\mathbf{A}\mathbf{f}$ and $\mathbf{z}$ are uncorrelated; therefore, the covariance matrix of their sum is the sum of their covariance matrices:

$$\begin{aligned} \Sigma &= \text{cov}(\mathbf{A}\mathbf{f}) + \text{cov}(\mathbf{z}) \\ &= \mathbf{A} \text{cov}(\mathbf{f}) \mathbf{A}' + \Psi && \text{by (3.34) and (13.7)} \\ &= \mathbf{A}\mathbf{A}' + \Psi && \text{by (13.7)} \\ &= \mathbf{A}\mathbf{A}' + \Psi. \end{aligned} \tag{13.11}$$

If $\mathbf{A}$ has only a few columns, say five or three, then $\Sigma = \mathbf{A}\mathbf{A}' + \Psi$ in (13.11) represents a simplified structure for Σ , in which the covariances are modeled by the λ_{ij} 's alone since Ψ is diagonal. For example, in the illustration in (13.5) with $m = 2$ factors, $\sigma_{y_1 y_2}$ would be the product of the first two rows of $\mathbf{A}$, that is,

$$\sigma_{y_1 y_2} = \text{cov}(y_1, y_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22}$$

where $(\lambda_{11}, \lambda_{21})$ is the first row of $\mathbf{A}$ and $(\lambda_{12}, \lambda_{22})$ is the second row of $\mathbf{A}$. If y_1 and y_2 have a great deal in common, they will have similar loadings on the common factors f_1 and f_2 . That is, $(\lambda_{11}, \lambda_{21})$ will be similar to $(\lambda_{12}, \lambda_{22})$. In this case, either $\lambda_{11}\lambda_{21}$ or $\lambda_{12}\lambda_{22}$ is likely to be high. On the other hand, if y_1 and y_2 have little in common, then their loadings λ_{11} and λ_{21} on f_1 will be different and their loadings λ_{12} and λ_{22} on f_2 will likewise differ. In this case, the products $\lambda_{11}\lambda_{21}$ and $\lambda_{12}\lambda_{22}$ will tend to be small.

We can also find the covariances of the y 's with the f 's in terms of the λ 's. Consider, for example, $\text{cov}(y_1, f_1)$. By (13.5), $y_1 = \mu_1 = \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1m}f_m + \psi_1$. From (13.7), f_1 is uncorrelated with all other f_j 's, and by (13.10), f_1 is uncorrelated with ψ_1 . Thus

$$\begin{aligned}
\text{cov}(y_i, f_j) &= E[(y_i - \mu_i)(f_j - \mu_{f_j})] \\
&= E[\lambda_{i1}f_1 + \lambda_{i2}f_2 + \cdots + \lambda_{im}f_m](f_j) \\
&= E[\lambda_{i1}f_1f_j + \lambda_{i2}f_2^2 + \cdots + \lambda_{im}f_mf_j] \\
&= \lambda_{i1}\text{cov}(f_1, f_j) + \lambda_{i2}\text{var}(f_2) + \cdots + \lambda_{im}\text{cov}(f_m, f_j) \\
&= \lambda_{ij}
\end{aligned}$$

since $\text{var}(f_j^2) = 1$. Hence the loadings themselves represent covariances of the variables with the factors. In general,

$$\text{cov}(y_i, f_j) = \lambda_{ij}, \quad i = 1, 2, \dots, p, \quad j = 1, 2, \dots, m. \quad (11.12)$$

Since λ_{ij} is the ij th element of $\mathbf{A}$, we can write (11.12) in the form

$$\text{cov}(\mathbf{y}, \mathbf{f}) = \mathbf{A}. \quad (11.13)$$

If standardized variables are used, (11.11) is replaced by $\mathbf{P}_p = \mathbf{A}\mathbf{K}' + \mathbf{\Psi}$, and the loadings become correlations:

$$\text{corr}(y_i, f_j) = \lambda_{ij}. \quad (11.14)$$

In (11.2), we have a partitioning of the variance of y_i into a component due to the common factors, called the *communality*, and a component unique to y_i , called the *specific variance*:

$$\begin{aligned}
\sigma_{y_i}^2 = \text{var}(y_i) &= (\lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2) + \phi_i \\
&= h_i^2 + \phi_i \\
&= \text{communality} + \text{specific variance},
\end{aligned}$$

where

$$\text{Communality} = h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2, \quad (11.15)$$

$$\text{Specific variance} = \phi_i.$$

The communality h_i^2 is also referred to as *common variance*, and the specific variance ϕ_i has been called *specificity unique variance*, or *residual variance*.

Assumptions (11.6)–(11.10) lead to the simple covariance structure of (11.11), $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}$, which is an essential part of the factor analysis model. In schematic form, $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}$ has the following appearance:

$$\Sigma = \begin{array}{|c|} \hline \\ \hline \end{array} + \begin{array}{|c|} \hline \\ \hline \end{array}$$

The diagonal elements of Σ can be easily modeled by adjusting the diagonal elements of Ψ , but AA' is a simplified configuration for the off-diagonal elements. Hence the critical aspect of the model involves the covariances, and this is the major emphasis of factor analysis, as noted in Section 13.1 and in comments following (13.2) and (13.10).

If is a true population covariance matrix Σ that can be expressed exactly as $\Sigma = AA' + \Psi$, where Ψ is diagonal and A is $p \times m$, with m relatively small. In practice, many sample covariance matrices do not come satisfactorily close to this ideal pattern. However, we do not relax the assumption because the structure $\Sigma = AA' + \Psi$ is essential for estimation of A .

One advantage of the factor analysis model is that when it does not fit the data, the estimate of A clearly reflects this failure. In such cases, there are two problems in the estimates: (1) it is unclear how many factors there should be, and (2) it is unclear what the factors are. In other statistical procedures, failure of assumptions may not lead to such obvious consequences in the estimates or tests. In factor analysis, the assumptions are essentially self-checking, whereas in other procedures, we typically have to check the assumptions with residual plots, tests, and so on.

13.1.2 Nonuniqueness of Factor Loadings

The loadings in the model (13.5) can be multiplied by an orthogonal matrix without impacting their ability to reproduce the covariance matrix, i.e. $\Sigma = AA' + \Psi$. To see this, let T be an arbitrary orthogonal matrix. Then by (13.12), $TT' = I$, and we can insert TT' into the basic model (13.5) to obtain

$$y - \mu = A(TT')T'x + e.$$

We then associate T with A and associate T' with Γ so that the model becomes

$$y - \mu = A^*\Gamma'x, \quad (13.16)$$

where

$$A^* = AT, \quad (13.17)$$

$$\Gamma' = T'I. \quad (13.18)$$

If $\mathbf{A}$ in $\Sigma = \mathbf{A}\mathbf{A}' + \Psi$ is replaced by $\mathbf{A}' = \mathbf{A}\mathbf{T}$, we have

$$\begin{aligned}\Sigma &= \mathbf{A}'\mathbf{A}' + \Psi = \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})' + \Psi \\ &= \mathbf{A}\mathbf{T}\mathbf{T}'\mathbf{A}' + \Psi = \mathbf{A}\mathbf{A}' + \Psi,\end{aligned}$$

since $\mathbf{T}\mathbf{T}' = \mathbf{I}$. Thus the new loadings $\mathbf{A}' = \mathbf{A}\mathbf{T}$ in (13.17) reproduce the covariance matrix, just as $\mathbf{A}$ does in (13.11).

$$\Sigma = \mathbf{A}'\mathbf{A}' + \Psi = \mathbf{A}\mathbf{A}' + \Psi. \quad (13.18)$$

The new factors $\mathbf{F}' = \mathbf{T}\mathbf{F}$ in (13.18) satisfy the assumptions (13.6), (13.7), and (13.10), that is, $E(\mathbf{F}') = \mathbf{0}$, $\text{cov}(\mathbf{F}') = \mathbf{I}$, and $\text{cov}(\mathbf{F}', \mathbf{e}) = \mathbf{0}$.

The communalities $k_i^2 = k_{i1}^2 + k_{i2}^2 + \cdots + k_{iM}^2$, $i = 1, 2, \dots, p$, as defined in (13.12), are also unaffected by the transformation $\mathbf{A}' = \mathbf{A}\mathbf{T}$. This can be seen as follows. The communality k_i^2 is the sum of squares of the i th row of $\mathbf{A}$. If we denote the i th row of $\mathbf{A}$ by $\mathbf{A}_i'$, then the sum of squares in vector notation is $k_i^2 = \mathbf{A}_i'\mathbf{A}_i$. The i th row of $\mathbf{A}' = \mathbf{A}\mathbf{T}$ is $\mathbf{A}_i'\mathbf{T} = \mathbf{A}_i\mathbf{T}$, and the corresponding communality is

$$k_i'^2 = \mathbf{A}_i'\mathbf{T}\mathbf{T}'\mathbf{A}_i = \mathbf{A}_i'\mathbf{T}\mathbf{T}\mathbf{A}_i = \mathbf{A}_i'\mathbf{A}_i = k_i^2.$$

Thus the communalities remain the same for the new loadings. Note that $k_i^2 = k_{i1}^2 + k_{i2}^2 + \cdots + k_{iM}^2 = \mathbf{A}_i'\mathbf{A}_i$ is the distance from the origin to the point $\mathbf{A}_i' = (a_{i1}, a_{i2}, \dots, a_{iM})$ in the M -dimensional space of the factor loadings. (Since the distance $\mathbf{A}_i'\mathbf{A}_i$ is the same as $\mathbf{A}_i'\mathbf{T}\mathbf{T}'\mathbf{A}_i$, the points $\mathbf{A}_i'$ are rotated from the points $\mathbf{A}_i$. [This also follows, because $\mathbf{A}_i' = \mathbf{A}_i\mathbf{T}$, where $\mathbf{T}$ is orthogonal. Multiplication of a vector by an orthogonal matrix is equivalent to a rotation of axes; see (2.305).])

The inherent potential to rotate the loadings to a new frame of reference without affecting any assumptions or properties is very useful in interpretation of the factors and will be exploited in Section 13.5.

Note that the coefficients (loadings) in (13.1) are applied to the factors, not to the variables, as they are in discriminant functions and principal components. Thus in factor analysis, the observed variables are not involved in the rotation, as they are in discriminant functions and principal components.

13.3 ESTIMATION OF LOADINGS AND COMMUNALITIES

In the Sections 13.3.1–13.3.4, we discuss four approaches to estimation of the loadings and communalities.

13.3.1 Principal Component Method

The first technique we consider is commonly called the principal component method. This name is perhaps unfortunate in that it adds to the confusion between factor

analysis and principal component analysis. In the principal component method for estimation of loadings, we do not actually calculate any principal components. The reason for the name is given following (13.20).

From a random sample $y_1, y_2, \dots, y_n$, we obtain the sample covariance matrix $\mathbf{S}$ and then attempt to find an estimator $\hat{\mathbf{A}}$ that will approximate the fundamental expression (13.11) with $\mathbf{S}$ in place of Σ :

$$\mathbf{S} \approx \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Psi}, \quad (13.20)$$

In the principal component approach, we neglect $\hat{\Psi}$ and factor $\mathbf{S}$ into $\mathbf{S} = \hat{\mathbf{A}}\hat{\mathbf{A}}'$.

In order to factor $\mathbf{S}$, we use the spectral decomposition in (2.139),

$$\mathbf{S} = \mathbf{C}\mathbf{B}\mathbf{C}', \quad (13.21)$$

where $\mathbf{C}$ is an orthogonal matrix constructed with normalized eigenvectors ($\lambda_i/\alpha_i = 1$) of $\mathbf{S}$ as columns and $\mathbf{B}$ is a diagonal matrix with the eigenvalues $\theta_1, \theta_2, \dots, \theta_p$ of $\mathbf{S}$ on the diagonal:

$$\mathbf{B} = \begin{pmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \theta_p \end{pmatrix}. \quad (13.22)$$

We use the notation θ_i for eigenvalues instead of the usual λ_i in order to avoid confusion with the notation λ_{ij} used for the loadings.

To finish factoring $\mathbf{C}\mathbf{B}\mathbf{C}'$ in (13.21) into the form $\hat{\mathbf{A}}\hat{\mathbf{A}}'$, we observe that since the eigenvalues θ_i of the positive semidefinite matrix $\mathbf{S}$ are all positive or zero, we can factor $\mathbf{B}$ into

$$\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2},$$

where

$$\mathbf{B}^{1/2} = \begin{pmatrix} \sqrt{\theta_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\theta_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{\theta_p} \end{pmatrix}.$$

With this factoring of $\mathbf{B}$, (13.21) becomes

$$\begin{aligned} \mathbf{S} &= \mathbf{C}\mathbf{B}\mathbf{C}' = \mathbf{C}\mathbf{B}^{1/2}\mathbf{B}^{1/2}\mathbf{C}' \\ &= \mathbf{A}\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{1/2}\mathbf{A}', \end{aligned} \quad (13.23)$$

This is of the form $\mathbf{S} = \hat{\mathbf{A}}\mathbf{Y}$, but we do not define $\hat{\mathbf{A}}$ to be $\mathbf{C}\mathbf{D}^{1/2}$ because $\mathbf{C}\mathbf{D}^{1/2}$ is $p \times p$, and we are seeking a $\hat{\mathbf{A}}$ that is $p \times m$ with $m < p$. We therefore define $\mathbf{D}_m = \text{diag}(\theta_1, \theta_2, \dots, \theta_m)$ with the m largest eigenvalues $\theta_1 > \theta_2 > \dots > \theta_m$ and $\mathbf{C}_m = (\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_m)$ containing the corresponding eigenvectors. We then estimate $\hat{\mathbf{A}}$ by the first m columns of $\mathbf{C}\mathbf{D}^{1/2}$,

$$\hat{\mathbf{A}} = \mathbf{C}_m \mathbf{D}_m^{1/2} = (\sqrt{\theta_1} \mathbf{c}_1, \sqrt{\theta_2} \mathbf{c}_2, \dots, \sqrt{\theta_m} \mathbf{c}_m) \quad (13.24)$$

[see (2.56)], where $\hat{\mathbf{A}}$ is $p \times m$, $\mathbf{C}_m$ is $p \times m$, and $\mathbf{D}_m^{1/2}$ is $m \times m$.

We illustrate the structure of the $\hat{\mathbf{A}}_j$'s in (13.24) for $p = 5$ and $m = 2$:

$$\begin{aligned} \begin{pmatrix} \hat{a}_{11} & \hat{a}_{12} \\ \hat{a}_{21} & \hat{a}_{22} \\ \hat{a}_{31} & \hat{a}_{32} \\ \hat{a}_{41} & \hat{a}_{42} \\ \hat{a}_{51} & \hat{a}_{52} \end{pmatrix} &= \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \\ c_{51} & c_{52} \end{pmatrix} \begin{pmatrix} \sqrt{\theta_1} & 0 \\ 0 & \sqrt{\theta_2} \end{pmatrix} \\ &= \begin{pmatrix} \sqrt{\theta_1} c_{11} & \sqrt{\theta_2} c_{12} \\ \sqrt{\theta_1} c_{21} & \sqrt{\theta_2} c_{22} \\ \sqrt{\theta_1} c_{31} & \sqrt{\theta_2} c_{32} \\ \sqrt{\theta_1} c_{41} & \sqrt{\theta_2} c_{42} \\ \sqrt{\theta_1} c_{51} & \sqrt{\theta_2} c_{52} \end{pmatrix} \quad [\text{by (2.56)}] \quad (13.25) \end{aligned}$$

We can use in (13.25) the source of the term 'principal component' solution. The columns of $\hat{\mathbf{A}}$ are proportional to the eigenvectors of $\mathbf{S}$, so that the loadings in the j th factor are proportional to coefficients in the j th principal component. The factors are thus related to the first m principal components, and it would seem that interpretations would be the same as for principal components. But after rotation of the loadings, the interpretations of the factors is usually different. The researcher will ordinarily prefer the rotated factors for reasons to be treated in Section 13.3.

By (2.57), the i th diagonal element of $\hat{\mathbf{A}}\mathbf{Y}'$ is the sum of squares of the i th row of $\hat{\mathbf{A}}$, is $\hat{\mathbf{A}}_i' \hat{\mathbf{A}}_i = \sum_{j=1}^m \hat{a}_{ij}^2$. Hence to complete the approximation of $\mathbf{S}$ in (13.23), we define

$$\hat{\phi}_i = \lambda_i = \sum_{j=1}^m \hat{a}_{ij}^2 \quad (13.26)$$

and write

$$\mathbf{S} \approx \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Phi}, \quad (13.27)$$

where $\hat{\Phi} = \text{diag}(\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p)$. Thus in (13.27) the variances on the diagonal of $\mathbf{S}$ are modelled exactly, but the off-diagonal covariances are only approximate. Again, this is the challenge of factor analysis.

In this method of estimation, the sums of squares of the rows and columns of $\hat{\mathbf{A}}$ are equal to communalities and eigenvalues, respectively. This is easily shown. By (13.20) and by analogy with (13.15), the i th communality is estimated by

$$\hat{\lambda}_i^2 = \sum_{j=1}^p \hat{\lambda}_{ij}^2, \quad (13.26)$$

which is the sum of squares of the i th row of $\hat{\mathbf{A}}$. The sum of squares of the j th column of $\hat{\mathbf{A}}$ is the j th eigenvalue of $\hat{\mathbf{B}}$

$$\begin{aligned} \sum_{i=1}^n \hat{\lambda}_{ij}^2 &= \sum_{i=1}^n \lambda_{ij}^2 \hat{c}_{ij}^2 \quad \text{by (13.23)} \\ &= \theta_j \sum_{i=1}^n \hat{c}_{ij}^2 \\ &= \theta_j, \end{aligned} \quad (13.28)$$

since the normalized eigenvectors (columns of $\hat{\mathbf{C}}$) have length 1.

By (13.26) and (13.28), the variance of the i th variable is partitioned into a part due to the factors and a part due uniquely to the variable:

$$\begin{aligned} \sigma_{ii} &= \hat{\lambda}_i^2 + \theta_i \\ &= \hat{\lambda}_{1i}^2 + \hat{\lambda}_{2i}^2 + \cdots + \hat{\lambda}_{pi}^2 + \theta_i. \end{aligned} \quad (13.29)$$

Thus the j th factor contributes $\hat{\lambda}_{ji}^2$ to σ_{ii} . The contribution of the j th factor to the total sample variance, $\text{tr}(\hat{\mathbf{B}}) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{nn}$, is, therefore,

$$\text{Variance due to } j\text{th factor} = \sum_{i=1}^n \hat{\lambda}_{ji}^2 = \hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + \cdots + \hat{\lambda}_{nj}^2, \quad (13.30)$$

which is the sum of squares of loadings in the j th column of $\hat{\mathbf{A}}$. By (13.28), this is equal to the j th eigenvalue, θ_j . The proportion of total sample variance due to the j th factor is, therefore,

$$\frac{\sum_{i=1}^n \hat{\lambda}_{ji}^2}{\text{tr}(\hat{\mathbf{B}})} = \frac{\theta_j}{\text{tr}(\hat{\mathbf{B}})}. \quad (13.31)$$

If the variables are not communalities, we use use standardized variables and work with the correlation matrix $\hat{\mathbf{R}}$. The eigenvalues and eigenvectors of $\hat{\mathbf{R}}$ are then used in place of those of $\hat{\mathbf{B}}$ in (13.24) to obtain estimates of the loadings. In practice, $\hat{\mathbf{R}}$ is used more often than $\hat{\mathbf{B}}$ and is the default in most software packages. Since the emphasis in factor analysis is on reproducing the variances or correlations rather

than the variances, use of $\mathbf{R}$ is more appropriate in factor analysis than in principal components. In applications, $\mathbf{R}$ often gives better results than $\mathbf{S}$.

If we are factoring $\mathbf{R}$, the properties corresponding to (13.32) is

$$\frac{\sum_{i=1}^p \lambda_i^2}{h(\mathbf{R})} = \frac{g}{p}, \quad (13.33)$$

where g is the number of variables.

We can assess the fit of the factor analysis model by computing the left and right sides of (13.33). The error matrix

$$\mathbf{E} = \mathbf{R} - (\hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Phi})$$

has zeros on the diagonal but nonzero off-diagonal elements. The following inequality gives a bound on the size of the elements in $\mathbf{E}$:

$$\sum_{j=1}^p e_{ij}^2 \leq e_{ii+1}^2 + e_{ii+2}^2 + \cdots + e_{ij}^2, \quad (13.34)$$

that is, the sum of squared entries in the matrix $\mathbf{E} = \mathbf{R} - (\hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Phi})$ is at most equal to the sum of squares of the deleted eigenvalues of $\mathbf{R}$. If the eigenvalues are small, the residuals in the error matrix $\mathbf{R} - (\hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Phi})$ are small and the fit is good.

Example 13.1. To illustrate the principal component method of estimation, we use a simple data set collected by Brown, Williams, and Baulow (1984). A 12-year-old girl made five ratings on a 5-point semantic differential scale for each of seven of her acquaintances. The ratings were based on the five adjectives kind, intelligent, happy, likable, and just. Her ratings are given in Table 13.1.

Table 13.1. Perception Scale Ratings on Five Adjectives for Brown People

People	Kind	Intelligent	Happy	Likable	Just
POOR ^a	1	3	3	1	1
MOTHER	4	4	5	4	4
POOR2	4	4	4	4	4
FATHER	4	4	4	4	4
TEACHER	1	4	1	1	4
MEAN ^b	4	1	3	4	4
POOR3	4	1	4	4	1

^aFemale acquaintance.

^bMale acquaintance.

The correlation matrix for the five variables (adjectives) is as follows, with the larger values bolded:

$$\mathbf{R} = \begin{pmatrix} 1.000 & .298 & \mathbf{.881} & \mathbf{.997} & .545 \\ .298 & 1.000 & -.822 & .326 & \mathbf{.837} \\ \mathbf{.881} & -.822 & 1.000 & \mathbf{.987} & .180 \\ \mathbf{.997} & .326 & \mathbf{.987} & 1.000 & .844 \\ .545 & \mathbf{.837} & .180 & .844 & 1.000 \end{pmatrix}. \quad (13.39)$$

The boldface values indicate two groups of variables: (1,3,4) and (2,5). We would therefore expect that the correlations among the variables can be explained fairly well by two factors.

The eigenvalues of $\mathbf{R}$ are 3.263, 1.536, .368, .031, and 0. Thus $\mathbf{R}$ is singular, which is possible in a situation such as this with only seven observations on five variables recorded in a single-digit scale. The multicollinearity among the variables induced by the fifth eigenvalue, 0, would be manifested from the corresponding eigenvector, as noted in Section 12.7 (see Problem 13.6).

By (13.34), the first two factors account for $(3.263 + 1.536)/5 = .56$ of the total sample variance. We therefore extract two factors. The first two eigenvectors are

$$\mathbf{e}_0 = \begin{pmatrix} .517 \\ .288 \\ .434 \\ .517 \\ .286 \end{pmatrix} \quad \text{and} \quad \mathbf{e}_1 = \begin{pmatrix} -.186 \\ .651 \\ -.473 \\ -.168 \\ .508 \end{pmatrix}.$$

Table 13.3. Factor Loadings by the Principal Component Method for the Reception Data of Table 13.1

Variable	Loadings		Communality, $\hat{\rho}_i^2$	Specific Variance, $\hat{\psi}_i$
	$\hat{\lambda}_{0i}$	$\hat{\lambda}_{1i}$		
Kind	.660	-.208	.498	.021
Intelligent	.319	.807	.624	.076
Riggy	.783	-.387	.600	.040
Likable	.071	-.218	.007	.013
Just	.356	.667	.340	.060
Variances accounted for	3.263	1.536	4.800	
Proportion of total variance	.653	.308	.960	
Cronbach's proportion	.653	.308	.960	

When these are multiplied by the square roots of the respective eigenvalues 3.265 and 1.508 as in (13.25), we obtain the loadings in Table 13.2.

The communalities in Table 13.2 are obtained from the sum of squares of the rows of the loadings, as in (13.28). The first row, for example, is $(.365)^2 + (-.251)^2 = .395$. The specific variances are obtained from (13.26) as $\hat{\sigma}_i^2 = 1 - R_i^2$ using 1 in place of σ_{ii} because we are factoring $\mathbf{R}$ rather than $\mathbf{S}$. The variance accounted for by each factor is the sum of squares of the corresponding columns of the loadings, as in (13.30). By (13.29), the variance accounted for is also equal to the eigenvalue in each case. Notice that the variance accounted for by the two factors adds to the sum of the communalities, since the latter is the sum of all squared loadings. By (13.33), the proportion of total variance for each factor is the variance accounted for divided by 3.

The two factors account for 95% of the total variance and therefore represent the five variables very well. To see how well the two-factor model reproduces the correlation matrix, we obtain

$$\begin{aligned} \hat{\mathbf{R}}\hat{\mathbf{R}}' + \Psi &= \begin{pmatrix} .365 & -.251 \\ .508 & .603 \\ .783 & -.587 \\ .371 & -.213 \\ .764 & .667 \end{pmatrix} \begin{pmatrix} .365 & .519 & .765 & .576 & .764 \\ -.251 & .607 & -.587 & -.213 & .667 \end{pmatrix} \\ &+ \begin{pmatrix} .605 & 0 & 0 & 0 & 0 \\ 0 & .379 & 0 & 0 & 0 \\ 0 & 0 & .200 & 0 & 0 \\ 0 & 0 & 0 & .613 & 0 \\ 0 & 0 & 0 & 0 & .233 \end{pmatrix} \\ &= \begin{pmatrix} 1.000 & .343 & .606 & .606 & .329 \\ .317 & 1.000 & -.099 & .779 & .604 \\ .606 & -.099 & 1.000 & .683 & .351 \\ .606 & .305 & .683 & 1.000 & .348 \\ .328 & .604 & .351 & .343 & 1.000 \end{pmatrix}, \end{aligned}$$

which is very close to the original $\mathbf{R}$. We will not attempt to interpret the factors at this point but will wait until they have been rotated in Section 13.4.2. $\square$

13.3.2 Principal Factor Method

In the principal component approach to estimation of the loadings, we exploited Ψ and factored $\mathbf{S}$ or $\mathbf{R}$. The principal factor method (also called the principal axis method) uses an initial estimate $\hat{\Psi}$ and factors $\mathbf{S} - \hat{\Psi}$ or $\mathbf{R} - \hat{\Psi}$ to obtain

$$\mathbf{S} - \hat{\Psi} \approx \hat{\mathbf{L}}\hat{\mathbf{L}}', \quad (13.36)$$

$$\mathbf{R} - \hat{\Psi} \approx \hat{\mathbf{L}}\hat{\mathbf{L}}', \quad (13.37)$$

where $\hat{\mathbf{A}}$ is $p \times m$ and is calculated as in (13.24) using eigenvalues and eigenvectors of $\mathbf{S} - \hat{\Phi}$ or $\mathbf{R} - \hat{\Psi}$.

The i th diagonal element of $\mathbf{S} - \hat{\Phi}$ is given by $s_{ii} - \hat{\phi}_{ii}$, which is the i th communality $\hat{h}_i^2 = s_{ii} - \hat{\phi}_{ii}$ [see (13.38)]. Likewise, the diagonal elements of $\mathbf{R} - \hat{\Psi}$ are the communalities $\hat{h}_i^2 = 1 - \hat{\psi}_{ii}$. Clearly, $\hat{\phi}_{ii}$ and $\hat{\psi}_{ii}$ have different values for $\mathbf{S}$ than for $\mathbf{R}$. With these diagonal values, $\mathbf{S} - \hat{\Phi}$ and $\mathbf{R} - \hat{\Psi}$ have the form

$$\mathbf{S} - \hat{\Phi} = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{pmatrix}, \quad (13.39)$$

$$\mathbf{R} - \hat{\Psi} = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{pmatrix}. \quad (13.40)$$

A popular initial estimate for a communality in $\mathbf{R} - \hat{\Psi}$ is $\hat{h}_i^2 = R_{ii}^2$, the squared multiple-correlation between y_i and the other $p - 1$ variables. This can be found as

$$\hat{h}_i^2 = R_{ii}^2 = 1 - \frac{1}{x^{(i)}}, \quad (13.41)$$

where $x^{(i)}$ is the i th diagonal element of $\mathbf{R}^{-1}$.

For $\mathbf{S} - \hat{\Phi}$, an initial estimate of communality analogous to (13.40) is

$$\hat{h}_i^2 = s_{ii} - \frac{1}{x^{(i)}}, \quad (13.42)$$

where s_{ii} is the i th diagonal element of $\mathbf{S}$ and $x^{(i)}$ is the i th diagonal element of $\mathbf{S}^{-1}$. It can be shown that (13.41) is equivalent to

$$\hat{h}_i^2 = s_{ii} - \frac{1}{x^{(i)}} = r_{ii} \hat{h}_i^2, \quad (13.43)$$

which is a reasonable estimate of the amount of variance that y_i has in common with the other y 's.

To use (13.40) or (13.41), $\mathbf{R}$ or $\mathbf{S}$ must be nonsingular. If $\mathbf{R}$ is singular, we can use the absolute value of the square of the largest correlation in the i th row of $\mathbf{R}$ as an estimate of communality.

After obtaining communality estimates, we calculate eigenvalues and eigenvectors of $\mathbf{S} - \hat{\Phi}$ or $\mathbf{R} - \hat{\Psi}$ and use (13.24) to obtain estimates of factor loadings, $\hat{\mathbf{A}}$. Then the columns and rows of $\hat{\mathbf{A}}$ can be used to obtain new eigenvalues (variance

explained) and communalities, respectively. The sum of squares of the j th column of $\hat{\mathbf{A}}$ is the j th eigenvalue of $\mathbf{S} - \hat{\Psi}$ or $\mathbf{R} - \hat{\Psi}$, and the sum of squares of the i th row of $\hat{\mathbf{A}}$ is the communality of y_i . The proportion of variance explained by the j th factor is

$$\frac{h_j^2}{\text{tr}(\mathbf{S} - \hat{\Psi})} = \frac{h_j^2}{\sum_{i=1}^p h_i^2}$$

or

$$\frac{h_j^2}{\text{tr}(\mathbf{R} - \hat{\Psi})} = \frac{h_j^2}{\sum_{i=1}^p h_i^2},$$

where h_j^2 is the j th eigenvalue of $\mathbf{S} - \hat{\Psi}$ or $\mathbf{R} - \hat{\Psi}$. The matrices $\mathbf{S} - \hat{\Psi}$ and $\mathbf{R} - \hat{\Psi}$ are not necessarily positive semidefinite and will often have some small negative eigenvalues. In such a case, the cumulative proportion of variance will exceed 1 and then decline to 1 as the negative eigenvalues are added. [Note that loadings cannot be obtained by (13.34) for the negative eigenvalues.]

Example 13.5.2. To illustrate the principal factor method, we use the perception data from Table 13.1. The correlation matrix as given in Example 13.5.1 is singular. Hence in place of multiple correlations as communality estimates, we use (the absolute value of) the largest correlation in each row of $\mathbf{R}$. [The multiple correlation of y with several variables is greater than the simple correlation of y with any of the individual variables; see, for example, Kendall (2003, p. 240).] The diagonal elements of $\mathbf{R} - \hat{\Psi}$ as given by (13.38) are, therefore, .893, .837, .805, .893, and .877, which are obtained from $\mathbf{R}$ in (13.15). The eigenvalues of $\mathbf{R} - \hat{\Psi}$ are 3.202, 1.395, .080, -.8002, and -.088, whose sum is 4.546. The first five eigenvalues of $\mathbf{R} - \hat{\Psi}$ are

$$\mathbf{a}_1 = \begin{pmatrix} .548 \\ .372 \\ .431 \\ .549 \\ .375 \end{pmatrix} \quad \text{and} \quad \mathbf{a}_2 = \begin{pmatrix} -.178 \\ .636 \\ -.660 \\ -.179 \\ .549 \end{pmatrix}.$$

When these are multiplied by the square roots of the respective eigenvalues, we obtain the principal factor loadings. In Table 13.3, these are compared with the loadings obtained by the principal component method in Example 13.5.1. The two sets of loadings are very similar, as we would have expected because of the large size of the communalities. The communalities in Table 13.3 are for the principal factor loadings, as noted above. The proportions of variance in each case for the principal factor loadings is obtained by dividing the variance accounted for (eigenvalue) by the sum of the eigenvalues, 4.546; for example, $3.202/4.546 = .704$. $\square$

Table 13.3 Loadings Obtained by Two Different Methods for Perception Data of Table 13.1

Variables	Principal-Component Loadings		Principal Factor Loadings		Communalities
	$\hat{A}_1$	$\hat{A}_2$	$\hat{F}_1$	$\hat{F}_2$	
Kind	.908	-.254	.880	-.238	.895
Intelligence	.514	.807	.487	.774	.657
Happy	.788	-.887	.776	-.844	.881
Likability	.871	-.219	.862	-.208	.868
Sex	.764	.687	.687	.648	.807
Variance accounted for	1.363	1.538	1.282	1.565	
Proportion of total variance	.681	.769	.641	.782	
Cumulative proportion	.681	.850	.764	1.00	

13.3.3 Iterated Principal Factor Method

The principal factor method can easily be iterated to improve the estimates of communality. After obtaining $\hat{\mathbf{A}}$ from $\mathbf{S} - \hat{\Phi}$ or $\mathbf{R} - \hat{\Phi}$ in (13.30) or (13.37) using initial communality estimates, we can obtain new communality estimates from the loadings in $\hat{\mathbf{A}}$ using (13.28).

$$\hat{h}_i^2 = \sum_{j=1}^p \hat{A}_{ij}^2.$$

These values of $\hat{h}_i^2$ are substituted into the diagonal of $\mathbf{S} - \hat{\Phi}$ or $\mathbf{R} - \hat{\Phi}$, from which we obtain a new value of $\hat{\mathbf{A}}$ using (13.24). This process is continued until the communality estimates converge. (For some data sets, the iterative procedure does not converge.) Then the eigenvalues and eigenvectors of the final version of $\mathbf{S} - \hat{\Phi}$ or $\mathbf{R} - \hat{\Phi}$ are used in (13.24) to obtain the loadings.

The principal factor method and iterated principal factor method will typically yield results very close to those from the principal component method when either of the following is true.

1. The correlations are fairly large, with a resulting small value of m .
2. The number of variables, p , is large.

A shortcoming of the iterative approach is that sometimes it leads to a communality estimate $\hat{h}_i^2$ exceeding 1 (rather than being ≤ 1). Such a result is known as a *flay-*

worst-case (Haywood 1971). If $\hat{\alpha}_i^2 > 1$, then $\hat{\alpha}_i < 0$ by (13.26) and (13.28), which is clearly improper, since we cannot have a negative specific variance. Thus when a communality exceeds 1, the iterative process should stop, with the program reporting that a solution cannot be reached. Some software programs have an option of continuing the iterations by setting the communality equal to 1 in all subsequent iterations. The resulting solution with $\hat{\alpha}_i = 0$ is somewhat questionable because it implies exact dependence of a variable on the factors, a possible but unlikely outcome.

Example 13.3.3. We illustrate the limited principal factor method using the Sales data in Table 7.1. The correlation matrix is as follows:

$$\mathbf{R} = \begin{pmatrix} 1.00 & .36 & .22 & .13 & .20 & -.04 & .13 & .03 & -.07 & .09 \\ .36 & 1.00 & -.09 & .15 & .20 & -.17 & .17 & .24 & .18 & .06 \\ .22 & -.09 & 1.00 & .16 & .70 & -.34 & -.45 & -.34 & -.11 & .68 \\ .13 & .15 & .16 & 1.00 & .49 & -.09 & -.18 & .01 & .02 & .37 \\ .20 & .20 & .70 & .49 & 1.00 & -.32 & -.34 & -.19 & .28 & .87 \\ -.04 & -.17 & -.34 & -.09 & -.32 & 1.00 & -.42 & -.57 & -.11 & -.26 \\ .13 & .17 & -.45 & -.18 & -.34 & -.42 & 1.00 & .02 & .23 & -.36 \\ .03 & .24 & -.34 & .01 & -.19 & -.57 & .02 & 1.00 & .49 & -.17 \\ -.07 & .18 & -.11 & .02 & .28 & -.11 & .23 & .49 & 1.00 & .28 \\ .09 & .06 & .68 & .37 & .87 & -.26 & -.36 & -.17 & .28 & 1.00 \end{pmatrix}.$$

The eigenvalues of $\mathbf{R}$ are 3.17, 2.36, 1.43, 1.28, .54, .47, .23, .12, .10, and .04. There is a notable gap between 1.28 and .54, and we therefore extract four factors (see Section 13.4). The first four eigenvalues account for a proportion

$$\frac{3.17 + 2.36 + 1.43 + 1.28}{10} = .84$$

of $\text{tr}(\mathbf{R})$.

For initial communality estimates, we use the squared multiple correlations between each variable and the other nine variables. These are given in Table 13.4, along with the final communalities after iteration. We multiply the first four eigenvectors of the final limited version of $\mathbf{R} - \hat{\Phi}$ by the square roots of the respective eigenvalues, as in (13.54), to obtain the factor loadings given in Table 13.4. We will not attempt to interpret the factors until after they have been rotated in Example 13.5.2b). $\square$

13.3.4 Maximum Likelihood Method

If we assume that the observations $y_1, y_2, \dots, y_n$ constitute a random sample from $N_p(\mu, \Sigma)$, then $\mathbf{A}$ and $\hat{\Phi}$ can be estimated by the method of maximum likelihood. It can be shown that the estimates $\hat{\mathbf{A}}$ and $\hat{\Phi}$ satisfy the following:

$$\hat{\Phi}\hat{\mathbf{A}} = \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\mathbf{A}}'\hat{\Phi}^{-1}\hat{\mathbf{A}}, \quad (13.60)$$

Table 13.4. Iterated Principal Factor Loadings and Communalities for the Soda Data

Variable	Loadings				Initial Communalities	Final Communalities
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$		
Taste	.22	.38	.82	.12	.37	1.00
Color	.37	.60	.43	-.20	.34	.98
pH	.80	.08	.83	-.40	.78	.79
Acidity 1	.41	.22	-.37	.34	.40	.56
Acidity 2	.34	.26	-.37	.05	.36	.56
Sales-share	-.13	-.67	.08	.56	.77	.79
Reducing sugar	-.33	.60	.83	-.11	.78	.79
Total sugar	-.43	.60	-.34	-.05	.87	.69
Alcohol	.12	.36	-.37	.34	.66	.79
Proxyl-nitrogen	.34	.21	-.17	-.02	.80	.79
Variance accounted for	1.00	2.37	1.24	.96	7.66	7.97

$$\hat{\Psi} = \text{diag}(\hat{\sigma}^2 - \hat{\lambda}'\hat{\lambda}), \quad (13.42)$$

$$\hat{\lambda}'\hat{\Psi}^{-1}\hat{\lambda} \quad \text{is diagonal.} \quad (13.43)$$

These equations must be solved iteratively, and in practice the procedure may fail to converge or may yield a Heywood case (Section 13.5.3).

We note that the proportion of variance accounted for by the factors, as given by (13.37) or (13.39), will not necessarily be as desirable index for maximum likelihood factors, as it is for factors obtained from the principal component or principal factor method.

Example 13.3.6. We illustrate the maximum likelihood method with the Soda data of Table 7.1. The correlation matrix and its eigenvalues were given in Example 13.3.5. We extract four factors, as in Example 13.3.5. The iterative solution of (13.41), (13.42), and (13.43) yielded the loadings and communalities given in Table 13.4.

The pattern of the loadings is different from that obtained using the iterated principal factor method in Example 13.3.4, but we will not compare them until after solution in Example 13.5.2(b). Note that the four values of variance accounted for are not in descending order. $\square$

13.4 CHOOSING THE NUMBER OF FACTORS, m

Several criteria have been proposed for choosing m , the number of factors. We consider four criteria, which are similar to those given in Section 12.6 for choosing the number of principal components retain.

Table 13.8. Maximum Likelihood Loadings and Communalities for the *Depression Data*

Variables	Loadings				Communalities
	F_1	F_2	F_3	F_4	
Ten	1.00	0	0	0	1.00
Out	.45	-.35	.11	.19	.29
pH	.25	.58	-.30	-.40	.76
Activity 1	.10	.47	.10	.57	.58
Activity 2	.20	.58	.05	.68	1.00
Intermarker	-.58	-.35	-.48	.39	.86
Reducing sugar	.11	-.38	.76	-.32	.79
Total sugar	.28	-.22	.86	.32	.88
Alcohol	-.57	.35	.11	.68	.72
Fungal-nitrogen	.05	.79	-.05	-.18	.87
Variance accounted for	1.33	1.60	2.36	1.00	7.32

1. Choose m equal to the number of factors necessary for the variance accounted for to achieve a predetermined percentage, say 80%, of the total variance $\text{tr}(\mathbf{B})$ or $\text{tr}(\mathbf{R})$.
2. Choose m equal to the number of eigenvalues greater than the average eigenvalue. For $\mathbf{R}$ the average is 1; for $\mathbf{B}$ it is $\sum_{i=1}^p \theta_i / p$.
3. Use the scree test based on a plot of the eigenvalues of $\mathbf{B}$ or $\mathbf{R}$. If the graph drops sharply, followed by a straight line with much smaller steps, choose m equal to the number of eigenvalues before the straight line begins.
4. Test the hypothesis that m is the correct number of factors, $\mathbf{R}_0: \mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}$, where $\mathbf{A}$ is $p \times m$.

Method 1 applies particularly to the principal component method. By (13.32), the proportion of total sample variance (variance accounted for) due to the j th factor from $\mathbf{B}$ is $\sum_{i=1}^n \lambda_{ij}^2 / \text{tr}(\mathbf{B})$. The corresponding proportion from $\mathbf{R}$ is $\sum_{i=1}^n \lambda_{ij}^2 / p$, as in (13.33). The contribution of all m factors to $\text{tr}(\mathbf{B})$ or p is therefore $\sum_{i=1}^n \sum_{j=1}^m \lambda_{ij}^2$, which is the sum of squares of all elements of $\mathbf{A}$. For the principal component method, we see by (13.36) and (13.35) that this sum is also equal to the sum of the first m eigenvalues or to the sum of all p communalities:

$$\sum_{i=1}^n \sum_{j=1}^m \lambda_{ij}^2 = \sum_{j=1}^m \theta_j^2 = \sum_{j=1}^m \phi_j. \quad (13.40)$$

Thus we choose m sufficiently large so that the sum of the communalities or the sum of the eigenvalues (variance accounted for) constitutes a relatively large portion of $\text{tr}(\mathbf{B})$ or p .

Method 3 can be extended to the principal factor method, where prior estimates of communalities are used to form $\mathbf{S} = \hat{\Phi}$ or $\mathbf{R} = \hat{\Phi}$. However, $\mathbf{S} = \hat{\Phi}$ or $\mathbf{R} = \hat{\Phi}$ will often have some negative eigenvalues. Therefore, as values of m range from 1 to p , the cumulative proportion of eigenvalues, $\sum_{i=1}^m R_i / \sum_{i=1}^p R_i$, will exceed 1.0 and then reduce to 1.0 as the negative eigenvalues are added. Hence a percentage such as 80% will be reached for a lower value of m than would be the case for $\mathbf{S}$ or $\mathbf{R}$, and a better strategy might be to choose m equal to the value for which the percentage first exceeds 100%.

In the limited principal factor method, m is specified before iteration, and $\sum_{i=1}^m \hat{R}_i^2$ is obtained after iteration as $\sum_{i=1}^m R_i^2 = \text{tr}(\mathbf{R} - \hat{\Phi})$. To choose m before iterating, one could use *a priori* considerations of the eigenvalues of $\mathbf{S}$ or $\mathbf{R}$, as in the principal component method.

Method 3 is a popular criterion of long standing and is the default in many software packages. Although heuristically based, it often works well in practice. A variation to method 3 that has been suggested for use with $\mathbf{R} = \hat{\Phi}$ is to let m equal the number of positive eigenvalues. (There will typically be some negative eigenvalues of $\mathbf{R} = \hat{\Phi}$.) However, this criterion will often result in too many factors, since the sum of the positive eigenvalues will exceed the sum of the communalities.

The scree test in method 3 was named after the geological term *scree*, referring to the debris at the bottom of a rocky cliff. It also performs well in practice.

In method 4 we wish to test

$$H_0: \Sigma = \Lambda\Lambda' + \Phi \quad \text{vs.} \quad H_1: \Sigma \neq \Lambda\Lambda' + \Phi,$$

where Λ is $p \times m$. The test statistic, a function of the likelihood ratio, is

$$\left(n - \frac{2p + 4m + 11}{6} \right) \ln \left(\frac{(\Lambda\Lambda' + \Phi)}{|\Sigma|} \right), \quad (13.47)$$

which is approximately χ^2_{2m} when H_0 is true, where $r = \frac{1}{2}(p - m)^2 - p - m + 1$ and Λ and Φ are the maximum likelihood estimators. Rejection of H_0 implies that m is too small and more factors are needed.

In practice, when n is large, the test in method 4 often shows more factors to be significant than do the other three methods. We may therefore consider the value of m indicated by the test to be an upper bound on the number of factors with practical importance.

For many data sets, the choice of m will not be obvious. This indeterminacy leaves many statisticians skeptical as to the validity of factor analysis. A researcher may begin with one of the methods (say, method 3) for an initial choice of m , will inspect the resulting percent of $\text{tr}(\mathbf{R})$ or $\text{tr}(\hat{\Sigma})$, and will then examine the rotated loadings for interpretability. If the percent of variance or the interpretation does not seem satisfactory, the experimenter will try other values of m in a search for an acceptable compromise between percent of $\text{tr}(\mathbf{R})$ and interpretability of the factors. Admittedly, this is a subjective procedure, and for such data sets one could well question the outcome (see Section 13.7).

When a data set is successfully fitted by a factor analysis model, the first three methods will almost always give the same value of m , and there will be little question as to what this value should be. Thus for a "good" data set, the entire procedure becomes much more objective.

Example 13.4(a). We compare the four methods of choosing m for the perception data used in Examples 13.3.1 and 13.3.2.

Method 1 gives $m = 2$, because one eigenvalue accounts for 85% of $\text{tr}(\mathbf{R})$, and two eigenvalues account for 96%.

Method 2 gives $m = 2$, since $\lambda_2 = 1.36$ and $\lambda_3 = .17$.

For method 3, we examine the scree plot in Figure 13.1. It is clear that $m = 2$ is indicated.

Method 4 is not available for the perception data because $\mathbf{R}$ is singular (first eigenvalue is zero), and the test involves $\|\mathbf{R}\|$.

Hence for the perception data, all three available methods agree on $m = 2$. $\square$

Example 13.4(b). We compare the four methods of choosing m for the Seidman data used in Examples 13.3.3 and 13.3.4.

Method 1 gives $m = 4$ for the principal component method, because four eigenvalues of $\mathbf{R}$ account for 82% of $\text{tr}(\mathbf{R})$. For the principal factor method with initial communality estimates R_{ii}^2 , the eigenvalues of $\mathbf{R} - \hat{\Phi}$ and corresponding proportions are as follows.

Eigenvalues	2.68	1.07	.94	.88	.12	.08	.08	-.08	-.13	-.22
Proportions	.65	.23	.14	.16	.02	.01	.00	-.01	-.02	-.03
Cumulative proportions	.65	.88	.92	1.00	1.02	1.03	1.03	1.02	1.00	0.97

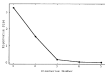


Figure 13.1. Scree graph for the perception data.

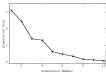


Figure 13.2. Scree plot for the Soester data.

The proportions are obtained by dividing the eigenvalues by their sums, 8.63. Thus the cumulative proportions first exceeds 1.00 for $m = 4$.

Method 2 gives $m = 4$, since $\lambda_4 = 1.30$ and $\lambda_5 = .81$, where λ_4 and λ_5 are eigenvalues of $\mathbf{B}$.

For method 3, we examine the scree plot in Figure 13.2. There is a discernible bend in slope at the fifth eigenvalue.

For method 4, we use $m = 4$ in the approximate chi squared statistic in (13.67) and obtain $\chi^2 = 5.079$, with degrees of freedom

$$r = \frac{1}{2}(p - m)^2 - p - m = \frac{1}{2}(10 - 4)^2 - 10 - 4 = 11.$$

Since $5.079 < \chi^2_{0.01,11} = 19.68$, we do not reject the hypothesis that four factors are adequate.

Thus for the Soester data, all four methods agree on $m = 4$. □

13.2 ROTATION

13.2.1 Introduction

As noted in Section 13.2.2, the factor loadings (rows of $\mathbf{A}$) in the population model are unique only up to multiplication by an orthogonal matrix that rotates the loadings. The rotated loadings preserve the essential properties of the original loadings; they reproduce the covariance matrix and satisfy all basic assumptions. The estimated loading matrix $\hat{\mathbf{A}}$ can likewise be rotated to obtain $\hat{\mathbf{A}}^* = \hat{\mathbf{A}}\mathbf{T}$, where $\mathbf{T}$ is orthogonal. Since $\mathbf{T}\mathbf{T}' = \mathbf{I}$ by (2.102), the rotated loadings provide the same estimate of the covariance matrix as before:

$$\hat{\Sigma} \hat{\mathbf{A}}' \hat{\mathbf{A}}' + \hat{\Phi} = \hat{\Sigma} \mathbf{T}' \mathbf{T} \hat{\mathbf{A}}' + \hat{\Phi} = \hat{\Sigma} \hat{\mathbf{A}}' + \hat{\Phi}. \quad (13.68)$$

Geometrically, the loadings in the i th row of $\hat{\mathbf{A}}$ constitute the coordinates of a point in the loading space corresponding to $\mathbf{x}_i$. Rotation of the p points gives their coordinates with respect to new axes (clusters) but otherwise leaves their basic geometric configuration intact. We hope to find a new frame of reference in which the factors are more interpretable. To this end, the goal of rotation is to place the axes close to as many points as possible. If there are clusters of points (corresponding to groupings of $\mathbf{y}$'s), we seek to move the axes in order to pass through or near these clusters. This would associate each group of variables with a factor (axis) and make interpretations more objective. The resulting axes then represent the natural factors.

If we can achieve a rotation in which every point is close to an axis, then each variable loads highly on the factor corresponding to the axis and has small loadings on the remaining factors. In this case, there is no ambiguity. Such a happy state of affairs is called *simple structure*, and interpretation is greatly simplified. We readily observe which variables are associated with each factor, and the factor is defined or named accordingly.

In order to identify the natural groupings of variables, we seek a rotation to an interpretable pattern for the loadings, in which the variables load highly on only one factor. The number of factors on which a variable has moderate or high loadings is called the *complexity* of the variable. In the ideal situation referred to previously as *simple structure*, the variables all have a complexity of 1. In this case, the variables have been clearly clustered into groups corresponding to the factors.

We consider two basic types of rotations: *orthogonal* and *oblique*. The rotation in (13.46) involving an orthogonal matrix is an *orthogonal rotation*; the original perpendicular axes are rotated rigidly and remain perpendicular. In an orthogonal rotation, angles and distances are preserved, communalities are unchanged, and the basic configuration of the points remains the same. Only the reference axes differ. In an *oblique "rotation"* (transformation), the axes are not required to remain perpendicular and are thus free to pass closer to clusters of points.

In Sections 13.5.2 and 13.5.3, we discuss orthogonal and oblique rotations, followed by some guidelines for interpretation in Section 13.5.4.

13.5.2 Orthogonal Rotation

It was noted above in Section 13.5.1 that orthogonal rotations preserve communalities. This is because the rows of $\hat{\mathbf{A}}$ are rotated, and the distance to the origin is unchanged, which, by (13.28), is the communality. However, the variances accounted for by each factor as given in (13.34) will change, as will the corresponding proportions in (13.32) or (13.33). The proportions due to the rotated loadings will not necessarily be in descending order.

In Sections 13.5.2a and 13.5.2b, we consider two approaches to orthogonal rotations.

13.5.2a Graphical Approach

If there are only two factors ($m = 2$), we can use a graphical rotation based on a visual inspection of a plot of factor loadings. In this case, the rows of $\hat{\mathbf{A}}$ are pairs of

loadings, $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2})$, $i = 1, 2, \dots, p$, corresponding to $\gamma_1, \gamma_2, \dots, \gamma_p$. We choose an angle ϕ through which the axes can be rotated to move them closer to groupings of points. The new rotated loadings $(\hat{\lambda}_{i1}^*, \hat{\lambda}_{i2}^*)$ can be measured directly on the graph as coordinates of the axes or calculated from $\hat{A}^* = \hat{A}T$ using

$$T = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}. \quad (13.49)$$

Example 13.3.2a. In Example 13.3.1, the initial factor loadings for the perception data did not provide an interpretation consistent with the two groupings of variables apparent in the pattern of correlations in $\mathbf{R}$. The two pairs of loadings $(\hat{\lambda}_{11}, \hat{\lambda}_{12})$ corresponding to the five variables are plotted in Figure 13.3. An orthogonal rotation through -59° would bring the axes (factorial) closer to the two clusters of points (variables) identified in Example 13.3.1. With the rotation, each cluster of variables corresponds much more closely to a factor. Using $\hat{A}$ from Example 13.3.1 and -59° in T as given in (13.49), we obtain the following rotated loadings:



Figure 13.3. Plot of the two loadings for each of the five variables in the perception data of Table 13.1.

$$\begin{aligned}\hat{\mathbf{K}}^* &= \hat{\mathbf{K}}\mathbf{B}^* = \begin{pmatrix} .968 & -.211 \\ .518 & .807 \\ .795 & -.387 \\ .971 & -.210 \\ .794 & .807 \end{pmatrix} \begin{pmatrix} .809 & .574 \\ -.874 & .819 \end{pmatrix} \\ &= \begin{pmatrix} .927 & .367 \\ -.007 & .959 \\ .988 & -.001 \\ .908 & .389 \\ .194 & .959 \end{pmatrix}.\end{aligned}$$

In Table 13.6, we compare the rotated loadings in $\hat{\mathbf{K}}^*$ with the original loadings in $\hat{\mathbf{K}}$.

The interpretation of the rotated loadings is clear. As indicated by the boldface loadings in Table 13.6, the first factor is associated with variables 1, 3, and 4 (kind, happy, and likable). The second factor is associated with the other two variables: intelligent and just. This same grouping of variables is indicated by the pattern in the correlation matrix in (13.35) and can also be seen in the two clusters of points in Figure 13.3. The first factor might be described as representing a person's perceived humanity or amiability, while the second involves more logical or rational practices.

Note that if the angle between the rotated axis is allowed to be less than 90° (an oblique rotation), the lower axis representing f_2^* could come closer to the points corresponding to variables 1 and 4 so that the coordinates on f_2^* , .367 and -.389, could be reduced. However, the basic interpretation would not change; variables 1 and 4 would still be associated with f_1^* . $\square$

Table 13.6. Graphically Rotated Loadings for the Perception Data of Table 13.1

Variable	Principal Component Loadings		Graphically Rotated Loadings		Communality
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1^*$	$\hat{f}_2^*$	
Kind	.968	-.211	.927	.367	.993
Intelligent	.518	.807	-.007	.959	.921
Happy	.795	-.387	.988	-.001	.990
Likable	.971	-.210	.908	.389	.987
Just	.794	.807	.194	.959	.940
Values accounted for	2.260	1.508	2.696	2.106	4.802
Proportion of total variance	.853	.588	.539	.421	.660
Correlation proportion	.853	.588	.539	.421	.660

13.5.2b Varimax Rotation

The graphical approach to rotation is generally limited to $m = 2$. For $m > 2$, various analytical methods have been proposed. The most popular of these is the varimax technique, which seeks rotated loadings that maximize the variance of the squared loadings in each column of $\hat{A}^*$. If the loadings in a column were evenly equal, the variance would be least (0). As the squared loadings approach 0 and 1 (the factor loading K), the variance will approach a maximum. Thus the varimax method attempts to make the loadings either large or small to facilitate interpretation.

The varimax procedure cannot guarantee that all variables will load highly on only one factor. In fact, no procedure could do this for all possible data sets. The configuration of the points in the loading space remains fixed; we merely rotate the axes to be as close to as many points as possible. In many cases, the points are not well clustered, and the axes simply cannot be rotated so as to be near all of them. This problem is compounded by having to choose an m to be changed; the coordinates $(x_{1j}, x_{2j}, \dots, x_{mj})$ change, and the relative position of the points is altered.

The varimax rotation is available in virtually all factor analysis software programs. The output typically includes the rotated loading matrix $\hat{A}^*$, the variance accounted for (sum of squares of each column of $\hat{A}^*$), the communalities (sum of squares of each row of $\hat{A}^*$), and the orthogonal matrix T used to obtain $\hat{A}^* = \hat{A}T$.

Example 13.5.2b(iii). In Example 13.5.2a, a graphical rotation was devised visually to achieve interpretable loadings for the perception data of Table 13.1. As we would expect, the varimax method yields a similar result. The varimax rotated loadings are given in Table 13.7. For comparison, we have included the original unrotated loadings from Table 13.3 and the graphically rotated loadings from Table 13.6.

Table 13.7. Varimax Rotated Factor Loadings for the Perception Data of Table 13.1

Variable	Principal Component Loadings		Graphically Rotated Loadings		Varimax Rotated Loadings		Communalities
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$	
Kind	.608	-.211	.621	.267	.621	.268	.693
Intelligent	.319	.007	-.027	.609	.023	.609	.623
Happy	.783	-.287	.698	-.551	.878	-.303	.660
Literate	.971	-.230	.836	.283	.941	.217	.987
Just	.764	.007	.754	.008	.753	.010	.640
Variances accounted for	1.263	1.338	1.696	2.108	1.831	1.691	4.892
Proportion of total variance	.631	.669	.739	.821	.862	.798	.660
Correlation proportion	.631	.660	.739	.660	.862	.660	.660

The orthogonal matrix **T** for the varimax rotation is

$$\mathbf{T} = \begin{pmatrix} .858 & .512 \\ -.512 & .858 \end{pmatrix}.$$

By (13.46), $-\sin \phi = -.512$, and the angle of rotation is given by $\phi = -\sin^{-1}(.512) = -30.8^\circ$. Thus the varimax rotation shows an angle of rotation of -30.8° as compared to the -35° we selected visually, but the results are very close and the interpretation is exactly the same. $\square$

Example 13.5.2(a,b). In Examples 13.5.3 and 13.5.4, we obtained the desired principal factor loadings and maximum likelihood loadings for the Sweden data. In Table 13.8, we show the varimax rotation of these two sets of loadings. The similarities in the two sets of rotated loadings are striking. The interpretation in each case is the same. The variances accounted for are virtually identical.

The rotation in each case has achieved a satisfactory simple structure and most variables show a complexity of 1. The boldface loadings indicate the variables associated with each factor for interpretation purposes. These may be meaningful to the researcher. For example, factor 1 is associated with cake-meat, rotating angle, and total sugar, whereas factor 2 is aligned with taste and color. $\square$

13.5.3 Oblique Rotation

The term *oblique rotation* refers to a transformation in which the axes do not remain perpendicular. Technically, the term *oblique rotation* is a misnomer, since rotation implies an orthogonal transformation that preserves distances. A more accurate term

Table 13.8. Varimax Rotated Loadings for the Sweden Data

Variables	Desired Principal Factor Rotated Loadings				Maximum Likelihood Rotated Loadings			
	F_1	F_2	F_3	F_4	F_1	F_2	F_3	F_4
Taste	.18	-.08	.88	-.08	.18	-.08	.88	-.10
Color	-.11	.14	.85	.14	-.07	.14	.89	.17
pH	.88	-.12	.05	-.13	.82	-.10	.08	-.15
Acidity 1	.28	-.08	.09	.84	.28	-.08	.11	.83
Acidity 2	.88	-.08	.10	.03	.86	-.08	.10	.10
Cake-meat	-.03	-.78	.04	.07	-.08	-.80	.08	.10
Rotating angle	-.07	.78	.08	.03	-.07	.78	.10	.08
Total sugar	-.28	.82	.10	.13	-.27	.78	.11	.16
Alcohol	-.08	.25	.05	.88	-.08	.25	.08	.84
Pumpkin-seedgen	.14	-.07	-.08	.28	.16	-.07	-.08	.12
Variances accounted for	2.82	2.32	1.27	1.27	2.81	2.34	1.28	1.28

rotation, would be oblique transformation, but the term oblique rotation is well established in the literature.

Instead of the orthogonal transformation matrix $\mathbf{T}$ used in (13.16), (13.17), and (13.18), an oblique rotation uses a general nonsingular transformation matrix $\mathbf{Q}$ to obtain $\mathbf{F}^* = \mathbf{Q}^* \mathbf{F}$, and by (3.74),

$$\text{cov}(\mathbf{F}^*) = \mathbf{Q}^* \mathbf{\Sigma} \mathbf{Q} = \mathbf{Q}^* \mathbf{Q} \neq \mathbf{I}. \quad (13.20)$$

Thus the new factors are correlated. Since distances and angles are not preserved, the communalities for $\mathbf{F}^*$ are different from those for $\mathbf{F}$. Some program packages report communalities obtained from the original loadings, rather than the oblique loadings.

When the axes are not required to be perpendicular, they can more easily pass through the major clusters of points in the loading space (assuming there are such clusters). For example, in Figure 13.4, we have plotted the varimax rotated loadings for two factors extracted from the same data of Table 3.7 (see Example 13.5.3 at the end of this section). Oblique axes with an angle of 35° would pass much closer to the points, and the resulting loadings would be very close to 0 and 1. However, the interpretation would not change, since the same points (variables) would be associated with the oblique axes as with the orthogonal ones.

Various analytical methods for achieving oblique rotations have been proposed and are available in program packages. Typically, the output of one of these procedures includes a pattern matrix, a structure matrix, and a matrix of correlations among the oblique factors. For interpretation, we would usually prefer the pattern

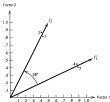


Figure 13.4. Orthogonal and oblique rotations for the same data.

matrix rather than the structure matrix. The loadings in a row of the pattern matrix are the natural coordinates of the point (variable) on the oblique axis and serve as coefficients in the model relating the variable to the factors.

One test for an oblique rotation is to check on the orthogonality of the factors. The orthogonality in the original factors is imposed by the model and maintained by an orthogonal rotation. If an oblique rotation produces a correlation matrix that is nearly diagonal, we can be more confident that the factors are indeed orthogonal.

Example 13.3. The correlation matrix for the same data of Table 13.1 is

$$\mathbf{R} = \begin{pmatrix} 1.000 & .585 & .711 & .504 \\ .709 & 1.000 & .699 & .509 \\ .711 & .693 & 1.000 & .839 \\ .504 & .509 & .839 & 1.000 \end{pmatrix}.$$

The varimax rotated loadings for two factors obtained by the principal component method are given in Table 13.5 and plotted in Figure 13.4. An analytical oblique rotation (Hartman–Kaiser orthogonal method in SAS) produced oblique axes with an angle of 38° , the same as obtained by a graphical approach. The correlation between the two factors is .79 (obtained from $\mathbf{Q}^*\mathbf{Q}$ in (13.90)), which is related to the angle by $\cos(15), .79 = \cos(38^\circ)$. The pattern loadings are given in Table 13.5.

The oblique loadings give a much cleaner simple structure than the varimax loadings, but the interpretation is essentially the same if we neglect loadings below .45 on the varimax rotation.

In Figure 13.4, it is evident that a single factor would be adequate since the angle between axes is less than 45° . The suggestion to let $m = 1$ is also supported by the first three criteria in Section 13.4: the eigenvalues of $\mathbf{R}$ are 3.30, .58, .27, and .45. The first accounts for 80%, the second for an additional 9%. The large correlations, .76, between the two oblique factors constitutes additional evidence that a single-factor model would suffice here. In fact, the pattern in $\mathbf{R}$ itself indicates the presence of only one factor. The four variables form only one cluster, since all are highly correlated. There are no small correlations between groupings of variables. (3)

Table 13.5. Varimax and Orthoblique Loadings for the Same Data

Variable	Varimax Loadings		Orthoblique Pattern matrix	
	F_1	F_2	F_1	F_2
1	.42	.42	.58	.89
2	.49	.48	-.58	.86
3	.87	.11	.87	-.10
4	.86	.13	.83	.10

13.5.4 Interpretation

In Sections 13.5.1, 13.5.2, and 13.5.3, we have discussed the usefulness of rotation as an aid to interpretation. Our goal is to achieve a simple structure in which each variable loads highly on only one factor, with small loadings on all other factors. In practice, we often fail to achieve this goal, but rotation usually produces loadings that are closer to the desired simple structure.

We now suggest general guidelines for interpreting the factors by examination of the matrix of rotated factor loadings. Moving horizontally from left to right across the m loadings in each row, identify the highest loading (in absolute value). If the highest loading is of a significant size (a subjective determination, see the next paragraph), circle or underline it. This is done for each of the p variables. There may be other significant loadings in a row besides the one circled. If these are considered, the interpretation is less simple. On the other hand, there may be variables with such small communalities that no significant loading appears on any factor. In this case, the researcher may wish to increase the number of factors and run the program again so that these variables might associate with a new factor.

To assess significance of factor loadings, $\hat{L}_{ij}$, obtained from **FA**, a threshold value of β has been advocated by many writers. For most successful applications, however, a critical value of β is too low and will result in variables of complexity greater than 1. A larger value of β or β is typically more useful. The β criterion is loosely based on the critical value for significance of an ordinary correlation coefficient, r . However, the distribution of the sample loadings is not the same as that of r arising from the bivariate normal. In addition, the critical value should be increased because $m\beta$ values of $\hat{L}_{ij}$ are being tested. On the other hand, if m is large, the critical value might possibly need to be reduced somewhat. Since $\hat{L}_i^2 = \sum_{j=1}^p \hat{L}_{ij}^2$ is bounded by 1, an increase in m reduces the average squared loading in a row.

After identifying potentially significant loadings, the experimenter then attempts to discover some meaning in the factors and, ideally, to label or name them. This can usually be done if the group of variables associated with each factor makes sense to the researcher. But in many situations, the groupings are not so logical, and revisions can be tried, such as adjusting the size of loading deemed to be important, changing m , using a different method of estimating the loadings, or employing another type of rotation.

13.6 FACTOR SCORES

In many applications, the researcher wishes only to ascertain whether a factor analysis model fits the data and to identify the factors. In other applications, the experimenter wishes to obtain factor scores, $\hat{\mathbf{F}} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_m)'$, $i = 1, 2, \dots, n$, which are defined as estimates of the underlying factor values for each observation. There are two potential uses for such scores: (1) the behavior of the observations in terms of the factors may be of interest, and (2) we may wish to use the factor scores

as input to another analysis, such as MANOVA. The latter usage resembles a similar application of principal components.

Since the f 's are not observed, we must estimate them as functions of the observed y 's. The most popular approach to estimating the factors is based on regression (Thomson, 1950). We will discuss this method and also briefly describe an informal technique that can be used when $\mathbf{B}$ (or $\mathbf{S}$) is singular. For other approaches see Hansen (1976, Chapter 16).

Since $E(f_i) = 0$, we relate the f 's to the y 's by a centered regression model

$$\begin{aligned} f_1 &= \beta_{11}(y_1 - \bar{y}_1) + \beta_{12}(y_2 - \bar{y}_2) + \cdots + \beta_{1p}(y_p - \bar{y}_p) + \alpha_1, \\ f_2 &= \beta_{21}(y_1 - \bar{y}_1) + \beta_{22}(y_2 - \bar{y}_2) + \cdots + \beta_{2p}(y_p - \bar{y}_p) + \alpha_2, \\ &\vdots \\ f_n &= \beta_{n1}(y_1 - \bar{y}_1) + \beta_{n2}(y_2 - \bar{y}_2) + \cdots + \beta_{np}(y_p - \bar{y}_p) + \alpha_n, \end{aligned} \quad (13.51)$$

which can be written in matrix form as

$$\mathbf{f} = \mathbf{B}_f(\mathbf{y} - \bar{\mathbf{y}}) + \boldsymbol{\alpha}. \quad (13.52)$$

We have used the notation $\mathbf{q}$ to distinguish this error term α in the original factor model $\mathbf{y} - \boldsymbol{\mu} = \mathbf{A}\mathbf{f} + \boldsymbol{\alpha}$ given in (13.3). Our approach is to estimate $\mathbf{B}_f$ and use the predicted value $\hat{\mathbf{f}} = \mathbf{B}_f(\mathbf{y} - \bar{\mathbf{y}})$ to estimate $\mathbf{f}$.

The model (13.52) holds for each observation:

$$f_i = \mathbf{B}_f'(y_i - \bar{\mathbf{y}}) + \alpha_i, \quad i = 1, 2, \dots, n.$$

In transposed form, the model becomes

$$\bar{\mathbf{f}}_i' = (y_i - \bar{\mathbf{y}})' \mathbf{B}_f + \alpha_i', \quad i = 1, 2, \dots, n,$$

and these n equations can be combined into a single model,

$$\begin{aligned} \mathbf{F}' &= \begin{pmatrix} \bar{\mathbf{f}}_1' \\ \bar{\mathbf{f}}_2' \\ \vdots \\ \bar{\mathbf{f}}_n' \end{pmatrix} = \begin{pmatrix} (y_1 - \bar{y}_1)' \mathbf{B}_f \\ (y_2 - \bar{y}_2)' \mathbf{B}_f \\ \vdots \\ (y_n - \bar{y}_n)' \mathbf{B}_f \end{pmatrix} + \begin{pmatrix} \alpha_1' \\ \alpha_2' \\ \vdots \\ \alpha_n' \end{pmatrix} \\ &= \begin{pmatrix} y_1 - \bar{y}_1' \\ y_2 - \bar{y}_2' \\ \vdots \\ y_n - \bar{y}_n' \end{pmatrix} \mathbf{B}_f + \boldsymbol{\Xi}' \\ &= \mathbf{Y}' \mathbf{B}_f + \boldsymbol{\Xi}' \quad \text{Eq. (13.11) (I)} \end{aligned} \quad (13.53)$$

The model (13.54) has the appearance of a centered multivariate multiple regression model as in Section 10.4.5, with $\mathbf{Y}_j$ in place of $\mathbf{X}_j$. By (10.50), the estimate for $\mathbf{B}_1$ would be

$$\hat{\mathbf{B}}_1 = (\mathbf{Y}_1' \mathbf{Y}_1)^{-1} \mathbf{Y}_1' \mathbf{F}. \quad (13.56)$$

However, $\mathbf{F}$ is unobserved. To evaluate $\hat{\mathbf{B}}_1$ in spite of this, we first use (10.52) to rewrite (13.54) in terms of covariance matrices,

$$\hat{\mathbf{B}}_1 = \mathbf{S}_{YF}^{-1} \mathbf{S}_{Y1}. \quad (13.57)$$

In derivation of the present chapter, $\mathbf{S}_{YF}$ is represented by $\mathbf{S}$; for $\mathbf{S}_{Y1}$ we use $\hat{\mathbf{A}}$ since $\hat{\mathbf{A}}$ estimates $\text{cov}(\mathbf{y}, \mathbf{F}) = \mathbf{A}$ in (13.13). Thus, based on the assumptions in Section 13.2.1, we can write (13.57) as

$$\hat{\mathbf{B}}_1 = \mathbf{S}^{-1} \hat{\mathbf{A}}. \quad (13.58)$$

Then from model (13.53), the estimated (predicted) $\mathbf{F}$ values are given by

$$\begin{aligned} \mathbf{F} &= \begin{pmatrix} \hat{F}_1 \\ \hat{F}_2 \\ \vdots \\ \hat{F}_p \end{pmatrix} = \mathbf{Y}_1 \mathbf{B}_1 \\ &= \mathbf{Y}_1 \mathbf{S}^{-1} \hat{\mathbf{A}}. \end{aligned} \quad (13.59)$$

If $\mathbf{B}$ is factored instead of $\mathbf{S}$, (13.58) and (13.59) become

$$\hat{\mathbf{B}}_1 = \mathbf{B}^{-1} \hat{\mathbf{A}}, \quad (13.60)$$

$$\hat{\mathbf{F}} = \mathbf{Y}_1 \mathbf{B}^{-1} \hat{\mathbf{A}}, \quad (13.61)$$

respectively, where $\mathbf{Y}_1$ is the observed matrix of standardized variables, $(y_{1j} = \mathcal{T}_j(x_{1j}))$.

We would optimally obtain factor scores for the related factors rather than the original factors. Thus $\hat{\mathbf{A}}$ in (13.59) or (13.61) would be replaced by $\hat{\mathbf{A}}^*$.

In order to obtain factor scores by (13.59) or (13.61), $\mathbf{S}$ or $\mathbf{B}$ must be nonsingular. When $\mathbf{B}$ (or $\mathbf{S}$) is singular, we can obtain factor scores by a simple method based directly on the related loadings. We cluster the variables into groups (factors) according to the loadings and find a score for each factor by averaging the variables associated with the factor. If the variables are not commensurate, the variables should be standardized before averaging. An alternative approach would be to weight the variables by their loadings when averaging.

Example 13.4. The speaking rate of four voices was artificially manipulated by means of a rate changer without altering the pitch (Jacobs, Strong, and Rescher 1973). There were five rates for each voice:

FF = 45% faster,

F = 15% faster,

N = normal rate,

S = 25% slower,

SS = 45% slower.

The resulting 20 voices were played to 30 judges, who rated them on 15 paired-opposite adjectives (validated with a 4-point scale between poles). The following adjectives were used: intelligent, ambitious, polite, active, confident, happy, just, flexible, kind, sincere, dependable, religious, good-looking, variable, and strong. The results were averaged over the 30 judges to produce 30 observation vectors of 15 variables each. The averaging produced very reliable data, so that even though there were only 30 observations on 15 variables, the factor analysis model fit very well. The correlation matrix is as follows:

$$R = \begin{pmatrix} 1.00 & .06 & -.17 & .00 & .02 & .00 & .15 & .10 & -.02 & -.16 & .12 & -.13 & -.15 & -.15 & .13 \\ .06 & 1.00 & -.06 & .03 & .07 & .76 & -.16 & .10 & -.10 & -.42 & .11 & -.10 & -.08 & -.10 & -.42 \\ -.17 & -.06 & 1.00 & -.16 & -.15 & .07 & .05 & .75 & .00 & .11 & .08 & .00 & .11 & .14 & .15 \\ .00 & .03 & -.16 & 1.00 & .05 & .73 & -.23 & -.02 & -.43 & -.17 & .16 & -.13 &td="15" style="border: none;">.08 & .04 & .30 \\ .02 & .07 & -.15 & .05 & 1.00 & .70 & .10 & .10 & -.05 & -.16 & .05 & -.10 & .07 & .00 & .11 \\ .00 & .76 & .07 & .73 & .04 & 1.00 & .27 & .13 & .12 & .08 & .08 & .08 & .00 & .03 & .10 \\ .15 & -.16 & .05 & -.20 & .20 & .27 & 1.00 & .00 & .01 & .76 & .76 & .11 & .13 & .14 & .13 \\ .10 & .10 & .75 & -.02 & .10 & .13 & .05 & 1.00 & .04 & .10 & .03 & .11 & .11 & .08 & .10 \\ -.02 & -.10 & .08 & -.05 & -.05 & -.12 & .01 & .04 & 1.00 & .11 & .16 & .10 & .10 & .16 & .13 \\ -.16 & -.02 & .10 & -.27 & -.16 & .06 & .76 & .76 & .01 & 1.00 & .12 & .08 & .10 & .10 & .10 \\ .02 & .10 & .07 & .10 & .08 & .08 & .76 & .03 & .16 & .12 & 1.00 & .12 & .13 & .11 & .10 \\ -.15 & -.08 & .08 & -.15 & -.10 & .08 & .01 & .11 & .05 & .06 & .12 & 1.00 & .11 & .12 & .14 \\ .76 & .08 & .11 & .10 & .05 & .08 & .13 & .10 & .10 & .13 & .13 & 1.00 & .08 & .02 \\ .76 & .08 & .11 & .10 & .08 & .01 & .13 & .08 & .10 & .10 & .11 &td="15" style="border: none;">.13 & .06 & 1.00 & .03 \\ .76 & .07 & .10 & .10 & .10 & .10 & .13 & .10 & .10 & .10 & .14 & .12 &td="15" style="border: none;">.12 & 1.00 \end{pmatrix}$$

The eigenvalues of R are 7.91, 5.83, .31, .26, ..., .002, with the scree plot in Figure 13.5. Clearly, by any criterion for choosing m , there are two factors.

All four major methods of factor extraction discussed in Section 13.3 produced nearly identical results (after rotation). We give the initial and rotated loadings obtained from the principal component method in Table 13.6.

The two rotated factors were labeled *competence* and *homeliness*. The same two factors emerged consistently in similar studies with different voices and different judges.

The first groupings of variables can also be seen in the correlation matrix. For example, in the first row, the large correlations correspond to the boldface rotated

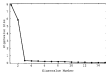


Figure 13.8. Scree graph for voice data.

Table 13.16. Initial and Varimax Rotated Loadings for the Voice Data

Variable	Initial Loadings		Rotated Loadings		Communalities
	F_1	F_2	F_1	F_2	
Intelligent	.70	-.65	.86	-.68	.95
Ambitious	.48	-.58	.66	-.59	.66
Polite	.80	.80	-.32	.88	.83
Active	.37	-.64	.86	-.43	.87
Childlike	.73	-.64	.87	-.68	.95
Happy	.83	-.47	.86	.33	.81
Just	.70	.56	.55	.89	.86
Liable	.89	.39	.45	.87	.95
Kind	.56	.79	-.63	.96	.89
Secure	.83	.83	-.33	.87	.95
Dependable	.83	.37	.86	.76	.86
Religious	.88	.76	-.67	.86	.83
Goodlooking	.80	-.39	.89	.33	.81
Healthy	.80	-.32	.86	.40	.87
Strong	.80	-.33	.86	.41	.86
Variances accounted for	3.40	3.40	7.11	6.63	13.76
Proportion of total variance	.33	.33	.67	.64	.92
Cumulative proportion	.33	.67	.67	.92	.92

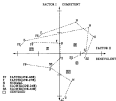


Figure 13.6. Factor scores of adjective ratings of voices with four levels of manipulation.

loadings for F_2 , whereas in the final run, the large correlations correspond to the backward rotated loadings for F_2 .

The factor scores were of primary interest in this study. The goal was to ascertain the effect of the rate manipulations on the two factors, that is, to determine the perceived change in competence and benevolence when the speaking rate is increased or decreased.

The two factor scores were obtained for each of the 20 voices; these are plotted in Figure 13.6, where a consistent effect of the manipulation of speaking rate on all four voices can clearly be seen. Decreasing the speaking rate causes the speaker to be rated less competent; increasing the rate causes the speaker to be rated less benevolent. The mean vectors (contour) are also given in Figure 13.6 for the four systems. □

13.7 VALIDITY OF THE FACTOR ANALYSIS MODEL

For many statisticians, factor analysis is controversial and does not belong in a toolkit of legitimate multivariate techniques. The reasons for this mistrust include the following: the difficulty in choosing *as* the many methods of extracting factors, the many rotation techniques, and the subjectivity in interpretation. Some statisticians also criticize factor analysis because of the indeterminacy of the factor loading matrix Λ or Λ_r first noted in Section 13.2.2. However, it is the ability to rotate that gives factor analysis its utility, if not its charm.

The basic question is whether the factors really exist. The model (3.11) for the covariance matrix is $\Sigma = \Lambda\Lambda' + \Psi$ or $\Sigma - \Psi = \Lambda\Lambda'$, where $\Lambda\Lambda'$ is of rank m . Many populations have covariance matrices that do not approach this pattern unless m is large. Thus the model will not fit data from such a population when we try to impose a small value of m . On the other hand, for a population in which Σ is reasonably close to $\Lambda\Lambda' + \Psi$ for small m , the sampling procedure leading to $\hat{\Sigma}$ may obscure this pattern. The researcher may believe there are underlying factors but has difficulty collecting data that will reveal them. In many cases, the basic problem is that $\hat{\Sigma}$ (or $\mathbf{R}$) contains both structure and error, and the methods of factor analysis cannot separate the two.

A statistical consultant in a university setting or elsewhere all too often sees the following scenario. A researcher designs a long questionnaire, with answers to be given in, say, a five-point semantic differential scale or Likert scale. The respondents, who vary in attitude from uninterested to zealous, hurriedly mark answers that in many cases are not even good subjective responses to the questions. Then the researcher submits the results to a handy factor analysis program. Being disappointed in the results, he or she appeals to a statistician for help. They attempt to improve the results by trying different methods of extraction, different rotations, different values of m , and so on. But it is all to no avail. The scree plot looks more like the foothills than a steep cliff with gently sloping debris at the bottom. There is no clear value of m . They have to extract 10 or 12 factors to account for, say, 60% of the variance, and interpretation of this large number of factors is hopeless. If a few underlying dimensions exist, they are totally obscured by both systematic and random errors in marking the questionnaire. A factor analysis model simply does not fit such a data set, unless a large value of m is used, which gives useless results.

It is not necessarily the "discreteness" of the data that causes the problem, but the "noisiness" of the data. The specified variables are not measured accurately. In some cases, discrete variables yield satisfactory results, such as in Examples 15.3.1, 15.3.2, 15.5.1a, and 15.5.2b(i), where a 12-year-old girl, responding carefully to a semantic differential scale, produced data leading to an unambiguous factor analysis. On the other hand, continuous variables do not guarantee good results (see Example 15.7a(i)).

In cases in which some factors are found that provide a satisfactory fit to the data, we should still be tentative in interpretation until we can independently establish the existence of the factors. If the same factors emerge in repeated sampling from the same population or a similar one, then we can have confidence that application of the model has uncovered some real factors. Thus it is good practice to repeat the experiment to check the stability of the factors. If the data set is large enough, it could be split in half and a factor analysis performed on each half. The two solutions could be compared with each other and with the solution for the complete set.

If there is replication in the data set, it may be helpful to average over the replications. This was done to great advantage in Example 15.6, where several judges rated the same reviews. Averaging over the judges produced variables that apparently possessed very low noise. Similar experimentation with different judges always pro-

load the same factors. Unfortunately, replication of this type is unavailable in most situations.

As with other techniques in this book, factor analysis assumes that the variables are at least approximately linearly related to each other. We could make bivariate scatter plots to check this assumption.

A basic prerequisite for a factor analysis application is that the variables not be independent. To check this requirement, we could test $H_0: \mathbf{P}_p = \mathbf{I}$ by using the test in Section 7.4.3.

Some writers have suggested that $\mathbf{R}^{-1}$ should be a near-diagonal matrix in order to successfully fit a factor analysis model. To assess how close $\mathbf{R}^{-1}$ is to a diagonal matrix, Kaiser (1970) proposed a measure of sampling adequacy,

$$\text{MSA} = \frac{\sum_{i,j} r_{ij}^2}{\sum_{i,j} r_{ij}^2 + \sum_{i,j} r_{ij}^2}, \quad (11.58)$$

where r_{ij}^2 is the square of an element from $\mathbf{R}$ and q_{ij}^2 is the square of an element from $\mathbf{Q} = \mathbf{I} - \mathbf{R}\mathbf{R}^{-1}\mathbf{R}$, with $\mathbf{I} = (\text{diag}[\mathbf{R}^{-1}]^{1/2})^{-1}$. As $\mathbf{R}^{-1}$ approaches a diagonal matrix, MSA approaches 1. Kaiser and Rice (1974) suggest that MSA should exceed .5 for satisfactory results to be expected. We show some results for MSA in Example 11.5(b).

In summary, there are many data sets to which factor analysis should not be applied. One indication that $\mathbf{R}$ is inappropriate for factoring is the failure of the methods in Section 11.4 to clearly and rather objectively choose a value for m . If the scree plot does not have a pronounced bend or the eigenvalues do not show a large gap around 1, then $\mathbf{R}$ is likely to be unsuitable for factoring. In addition, the communality estimates after factoring should be fairly large.

To balance the "good" examples in this chapter, we now give an example involving a data set that cannot be successfully modeled by factor analysis. Likewise, the problems at the end of the chapter include both "good" and "bad" data sets.

Example 11.5(a). As an illustration of an application of factor analysis that is less successful than previous examples, in this chapter, we consider the diabetes data of Table 5.8. The correlation matrix for the five variables is as follows:

$$\mathbf{R} = \begin{pmatrix} 1.00 & .09 & -.19 & .07 & .21 \\ .09 & 1.00 & -.01 & .01 & -.10 \\ -.19 & -.01 & 1.00 & .28 & .08 \\ .07 & .01 & .28 & 1.00 & .21 \\ .21 & -.10 & .08 & .21 & 1.00 \end{pmatrix}.$$

The correlations are all small and the variables do not appear to have much in common. The MSA value is .49. The eigenvalues are 1.40, 1.21, 1.04, .71, and .05. These factors would be required to account for 79% of the variance and four factors to reach 87%. This is not a useful reduction in dimensionality. The eigenvalues are plotted in a scree graph in Figure 11.3. The lack of a clear value of m is apparent.

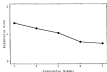


Figure 15.7. Scree graph for diabetes data.

It is evident from the small correlations in **R** that the communalities of the variables will be large. The principal component method, which essentially estimates the initial communalities as 1, gave very different final communality estimates than did the iterated principal factor method:

Communalities					
Principal component method	.76	.81	.71	.67	.64
Iterated principal factor method	.91	.96	.88	.97	.93

The communalities obtained by the iterated approach reflect more accurately the small correlations among the variables.

The various rotated factor loadings for three factors extracted by the iterated principal factor method are given in Table 15.11. The first factor is associated with variables 3 and 4, the second factor with variables 1 and 3, and the third with variable

Table 15.11. Various Rotated Factor Loadings for Iterated Principal Factors from the Diabetes Data

Variable	Rotated Loadings			Communalities
	F_1	F_2	F_3	
1	-.08	.88	.12	.81
2	.88	.38	.49	.86
3	.87	-.17	-.06	.88
4	.87	.22	.82	.97
5	.19	.47	-.27	.59
Variance accounted for	.88	.59	.24	1.00

2. This clustering of variables can be seen in $\mathbf{R}$, where variables 1 and 5 have a correlation of .31, variables 3 and 4 have a correlation of .29, and variable 2 has very low correlations with all other variables. However, these correlations (.31 and .29) are small, and in this case the collapsing of five variables to three factors is not a useful reduction in dimensionality, especially since the first three eigenvalues account for only 77% of $\text{tr}(\mathbf{R})$. The TSS is not convincingly greater than 60%, which we would expect from three original variables picked at random. This conclusion is borne out by a test of $H_0: \mathbf{P}_p = \mathbf{I}$. Using (3.17) and (3.18), we obtain

$$\begin{aligned}a &= |\mathbf{R}| = .3821\%, & r &= 20 - 1 = 19, & p &= 5, \\a' &= -\left[p - \frac{1}{2}(2p + 5)\right] \log a = -(18 - \frac{19}{2})(-.3117) = 3.625.\end{aligned}$$

With $\frac{1}{2}p(p-1) = 10$ degrees of freedom, the .05 critical value for this approximate χ^2 -test is 18.31, and we have no basis to question the independence of the five variables. Thus the three factors we obtained are very likely an artifact of the present sample and would not reappear in another sample from the same population. $\square$

Example 13.7(K). For datasets used in previous examples in this chapter, the values of MSA from (3.60) are calculated as follows:

$$\text{Seizure data: } \text{MSA} = .31,$$

$$\text{Sore data: } \text{MSA} = .81,$$

$$\text{Voice data: } \text{MSA} = .71,$$

$$\text{Diabetes data: } \text{MSA} = .46.$$

The MSA value cannot be computed for the perception data, because $\mathbf{R}$ is singular.

These results do not support great confidence in the MSA index as a rule guide in the suitability of $\mathbf{R}$ for factoring. We see a wide disparity in the MSA values for the first three data sets. Yet all three yielded successful factor analyses. These three MSA values seem to be inversely related to the number of factors: In the sore data, there were indications that one factor would suffice; the voice data clearly had two factors; and for the Seizure data, there were four factors.

The MSA for the diabetes data is close to that of the Seizure data. Yet the diabetes data are totally unsuitable for factor analysis, whereas the factor analysis of the Seizure data is very convincing. $\square$

13.8 THE RELATIONSHIP OF FACTOR ANALYSIS TO PRINCIPAL COMPONENT ANALYSIS

Both factor analysis and principal component analysis have the goal of reducing dimensionality. Because the objectives are similar, many authors discuss principal

-component analysis as another type of factor analysis. This can be confusing, and we wish to underscore the distinguishing characteristics of the two techniques.

Two of the differences between factor analysis and principal component analysis were mentioned in Section 13.1: (1) In factor analysis, the variables are expressed as linear combinations of the factors, whereas the principal components are linear functions of the variables, and (2) in principal component analysis, the emphasis is on explaining the total variance $\sum \lambda_{ii}$, as contrasted with the attempt to explain the covariances in factor analysis.

Additional differences are that (3) principal component analysis requires essentially no assumptions, whereas factor analysis makes several key assumptions; (4) the principal components are unique (assuming distinct eigenvalues of $\mathbf{S}$), whereas the factors are subject to an arbitrary rotation; and (5) if we change the number of factors, the estimated factors change. This does not happen in principal components.

The ability to rotate to improve interpretability is one of the advantages of factor analysis over principal components. If finding and describing some underlying factors is the goal, factor analysis may prove more useful than principal components; we would prefer factor analysis if the factor model fits the data well and we like the interpretation of the rotated factors. On the other hand, if we wish to define a smaller number of variables for input into another analysis, we would ordinarily prefer principal components, although this can sometimes be accomplished with factor scores. Occasionally, principal components are interpretable, as in the size and shape components in Example 13.3.1.

PROBLEMS

13.1 Show that the assumptions lead to (13.1), $\text{var}(y_i) = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_m^2 + \sigma_{\epsilon_i}^2$.

13.2 Verify directly that $\text{covr}(\mathbf{Y}) = \mathbf{A}$ as in (13.13).

13.3 Show that $\mathbf{F} = \mathbf{T}\mathbf{f}$ in (13.16) satisfies the assumptions (13.6) and (13.7), $\mathbf{A}(\mathbf{F}) = \mathbf{B}$ and $\text{var}(\mathbf{F}) = \mathbf{I}$.

13.4 Show that $\sum_{i=1}^p \sigma_{ij}^2 \leq \sigma_{i+1,i+1}^2 + \sigma_{i+2,i+2}^2 + \cdots + \sigma_p^2$ as in (13.30), where the σ_{ij} 's are the elements of $\mathbf{B} = \mathbf{B} = (\mathbf{A}\mathbf{A}') + \hat{\Phi}$ and the σ_j^2 's are eigenvalues of $\mathbf{B}$.

13.5 Show that $\sum_{i=1}^p \sum_{j=1}^p \sigma_{ij}^2$ is equal to the sum of the first p eigenvalues and also equal to the sum of all p -crossloadings, as in (13.46).

13.6 In Example 13.3.2, the correlation matrix for the perception data was shown to have an eigenvalue equal to 0. Find the multicollinearity among the five variables that this implies.

13.7 Use the words data of Table 5.9.

- Obtain principal component loadings for two factors.
- Do a graphical rotation of the two factors.
- Do a varimax rotation and compare the results with those in part (b).

- EX.8** Use the nurses' bone data of Table 3.5.
- Estimate loadings by the principal component method and do a varimax rotation. Use two factors.
 - Do all variables have a complexity of 1? Carry out an oblique rotation to improve the loadings.
 - What is the angle between the oblique axes? Would a single factor ($m = 1$) be more appropriate here?
- EX.9** Carry out a factor analysis of the meerkat data of Table 4.2. Combine the six groups into a single sample.
- Estimate the loadings for two factors by the principal component method and do a varimax rotation.
 - Did the rotation improve the loadings?
- EX.10** Use the fish data of Table 4.17. Combine the three groups into a single sample.
- Obtain loadings on two factors by the principal component method and do a varimax rotation.
 - Notice the similarity of loadings for y_1 and y_2 . Is there any indication in the correlation matrix as to why this is so?
 - Compute factor scores.
 - Using the factor scores, carry out a MANOVA comparing the three groups.
- EX.11** Carry out a factor analysis of the fish data in Table 5.5. Combine the two groups into a single sample.
- From an examination of the eigenvalues greater than 1, the scree plot, and the percentages, is there a clear choice of m ?
 - Estimate two factors by the principal component method and carry out a varimax rotation.
 - Is the rotation an improvement? Try an oblique rotation.
- EX.12** Use the engineer data of Table 5.6. Combine the two groups into a single sample.
- Using a scree plot, the number of eigenvalues greater than 1, and the percentages, is there a clear choice of m ?
 - Estimate three factors by the principal component method and carry out a varimax rotation.
 - Estimate three factors by the principal factor method and carry out a varimax rotation.
 - Compare the results of parts (b) and (c).

EX.15 Use the probe-word data of Table 3.3.

- Obtain loadings for two factors by the principal component method and carry out a varimax rotation.
- Notice the near-duplication of loadings for x_1 and x_2 . Is there any indication in the correlation matrix as to why this is so?
- Is the rotation satisfactory? Try an oblique rotation.

Cluster Analysis

14.1 INTRODUCTION

In cluster analysis we search for patterns in a data set by grouping the (individual) observations into clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other. We hope to find the natural groupings in the data, groupings that make sense to the researcher.

Cluster analysis differs fundamentally from classification analysis (Chapter 9). In classification analysis, we allocate the observations to a known number of predefined groups or populations. In cluster analysis, neither the number of groups nor the groups themselves are known in advance.

To group the observations into clusters, many techniques begin with similarities between all pairs of observations. In many cases the similarities are based on some measure of distance. Other cluster methods use a preliminary choice for cluster centers or a comparison of within- and between-cluster variability. It is also possible to cluster the variables, in which case the similarity could be a correlation; see Section 14.7.

We can search for clusters graphically by plotting the observations. If there are only two variables ($p = 2$), we can do this in a scatter plot (see Section 5.3). For $p = 2$, we can plot the data in two dimensions using principal components (see Section 12.4) or biplots (see Section 15.3). For an example of a principal component plot, see Figure 12.7 in Section 12.4, in which four clear groupings of points can be observed. Another approach to plotting is provided by principal pursuit, which seeks two-dimensional projections that reveal clusters (see Friedman and Tibshirani (1994); Huber (1985); Sibson (1984); Jones and Sibson (1987); Tompkins (1998); Pomeroy (1995); Nassim (1995); Ripley (1996, pp. 266–303)).

Cluster analysis has also been referred to as classification, pattern recognition (specifically, unsupervised learning), and numerical taxonomy. The techniques of cluster analysis have been extensively applied to data in many fields, such as medicine, population, sociology, criminology, anthropology, archaeology, geology, geography, remote sensing, market research, economics, and engineering.

We shall concentrate largely on quantitative variables (for categorical variables, see Gordon (1989) or Everitt (1993)). The data matrix (see (3.17)) can be written as

$$\mathbf{T} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = (\mathbf{x}_{11}, \mathbf{x}_{12}, \dots, \mathbf{x}_{1p}) \quad (14.1)$$

where $\mathbf{x}_i'$ is a row (observation vector) and $\mathbf{x}_{ij}$ is a column (corresponding to a variable). We generally wish to group the n $\mathbf{x}_i'$'s below into p clusters. We may also wish to cluster the columns $\mathbf{x}_{ij}$, $j = 1, 2, \dots, p$ (see Section 14.7).

Two common approaches to clustering the observation vectors are hierarchical clustering and partitioning. In hierarchical clustering we typically start with n clusters, one for each observation, and end with a single cluster containing all n observations. At each step, an observation or a cluster of observations is absorbed into another cluster. We can also reverse this process, that is, start with a single cluster containing all n observations and end with a cluster of a single item each (see Section 14.3.1b). In partitioning, we simply divide the observations into p clusters. This can be done by starting with an initial partitioning, or with cluster centers and then reassigning the observations according to some optimality criterion. Other clustering methods that we will discuss are based on fitting mixtures of multivariate normal distributions or searching for regions of high density sometimes called modes.

There is an abundant literature on cluster analysis. Useful monographs and reviews have been given by Gordon (1989), Everitt (1993), Khuriel and Nall (2006, Chapter 6), Kaufman and Rousseeuw (1990), Xie (1998, Chapter 7), Anderberg (1973), and Hartigan (1979a).

14.2 MEASURES OF SIMILARITY OR DISSIMILARITY

Since cluster analysis attempts to identify the observation vectors that are similar and group them into clusters, many techniques use an index of similarity or proximity between each pair of observations. A convenient measure of proximity is the distance between two observations. Since a distance increases as two units become further apart, distance is actually a measure of dissimilarity.

A common distance function is the Euclidean distance between two vectors $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ and $\mathbf{y} = (y_1, y_2, \dots, y_p)'$, defined as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} = \sqrt{\sum_{j=1}^p (x_j - y_j)^2}. \quad (14.2)$$

To adjust for differing variances and covariances among the p variables, we could use the statistical distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \mathbf{W}^{-1} (\mathbf{x} - \mathbf{y})} \quad (14.3)$$

[see (3.76)], where $\mathbf{S}$ is the sample covariance matrix. After the clusters are formed, $\mathbf{S}$ could be computed as the pooled within-cluster covariance matrix, but we do not know beforehand what the clusters will be. If we compute $\mathbf{S}$ on the unpartitioned sample, there will be distortion of the variances and covariances because of the groups in the data (assuming there really are some natural clusters). We therefore usually use the Euclidean distance given by (14.2). In some clustering procedures, it is not necessary to take the square root in (14.2) or (14.3).

Other distance measures have been suggested, for example, the Minkowski metric

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^p |x_j - y_j|^r \right]^{1/r}, \quad (14.4)$$

For $r = 2$, $d(\mathbf{x}, \mathbf{y})$ in (14.4) becomes the Euclidean distance given in (14.2). For $p = 2$ and $r = 1$, (14.4) measures the "city block" distance between two observations. There are distance measures for categorical data; see Gordon (1999, Chapter 2).

For the n observation vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, we can compute an $n \times n$ matrix $\mathbf{D} = (d_{ij})$ of distances (or dissimilarities), where $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ is usually given by (14.2), $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}$. We sometimes use $\mathbf{D} = (d_{ij}^2)$, where $d_{ij}^2 = d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$ is the square of (14.2). The matrix $\mathbf{D}$ typically is symmetric with diagonal elements equal to zero.

The scale of measurement of the variables is an important consideration when using the Euclidean distance measure in (14.2). Changing the scale can affect the relative distances among the items. For example, suppose three items have the following bivariate measurements (x_1, x_2) : (2, 5), (4, 3), (1, 9). Using d_{ij} as given by (14.2), the matrix $\mathbf{D} = (d_{ij})$ for these items is

$$\mathbf{D}_1 = \begin{pmatrix} 0.0 & 1.6 & 6.4 \\ 1.6 & 0.0 & 1.6 \\ 6.4 & 1.6 & 0.0 \end{pmatrix}.$$

However, if we multiply x_1 by 100 as, for example, in changing from meters to centimeters, the matrix becomes

$$\mathbf{D}_2 = \begin{pmatrix} 0 & 200 & 800 \\ 200 & 0 & 200 \\ 800 & 200 & 0 \end{pmatrix},$$

and the largest distance is now d_{13} instead of d_{23} . The distance rankings have been altered by scaling.

To remove this problem, each variable could be standardized in the usual way by subtracting the mean and dividing by the standard deviation of the variable. However, such scaling would ordinarily be based on the entire dataset, that is, on all n values in

each column of $\mathbf{X}$ in (14.1). In this case, the variables that best separate clusters might no longer do so after division by standard deviations that include between-cluster variation. If we use standardized variables, the clusters could be less well separated. The question of scaling is, therefore, not an easy one. However, standardization of this type is recommended by many authors.

If (14.2), the squared Euclidean distance between two observations $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ and $\mathbf{y} = (y_1, y_2, \dots, y_p)'$ is $d^2(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^p (x_j - y_j)^2$. This can be expanded as

$$d^2(\mathbf{x}, \mathbf{y}) = (x_1 - y_1)^2 + p\bar{x}^2 - 2\bar{x}\bar{y} + 2x_1y_1(1 - r_{xy}), \quad (14.5)$$

where $\bar{x}^2 = \sum_{j=1}^p x_j^2/p - \bar{x}^2$ and $\bar{x} = \sum_{j=1}^p x_j/p$, with similar expressions for $\bar{y}^2$ and $\bar{y}$. The correlation r_{xy} in (14.5) is given by

$$r_{xy} = \frac{\sum_{j=1}^p x_j y_j - \bar{x}\bar{y}}{\sqrt{\sum_{j=1}^p x_j^2 - p\bar{x}^2} \sqrt{\sum_{j=1}^p y_j^2 - p\bar{y}^2}}. \quad (14.6)$$

In Figure 14.1, we illustrate the profile (see Sections 5.9 and 6.8) for each of two observation vectors $\mathbf{x}$ and $\mathbf{y}$. The squared Euclidean distance in (14.5) may be used to compare the profiles of $\mathbf{x}$ and $\mathbf{y}$ in terms of levels, variation, and shape, where $\bar{x}$ and $\bar{y}$ are the levels of the two profiles, v_x and v_y are the variations of the profiles, and the correlation r_{xy} is a measure of the closeness of the shapes of the two profiles. The closer r_{xy} is to 1, the greater is the similarity in shape of the two profiles. Note that $\bar{x}$ and v_x are the mean and variation of the p variables within the observation vector $\mathbf{x}$, not over the n observations in the data set. A similar comment can be made about $\bar{y}$ and v_y . Likewise, the correlation r_{xy} is between the two observation vectors $\mathbf{x}$ and $\mathbf{y}$, not between two variables. The use of r_{xy} has been questioned by Jardine and Sibson (1971) and Wishart (1971), but Shamos, Hoffman, and Carpenter (1973)

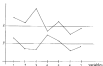


Figure 14.1. Profile for two observation vectors $\mathbf{x}$ and $\mathbf{y}$.

found the correlation to be superior to the Euclidean distance for finding the clusters in a particular data set.

14.3 HIERARCHICAL CLUSTERING

14.3.1 Introduction

Hierarchical methods and other clustering algorithms represent an attempt to find "good" clusters in the data using a computationally efficient technique. It is not generally feasible to examine all possible clustering possibilities for a data set, especially a large one. The number of ways of partitioning a set of n items into g clusters is given by

$$N(n, g) = \frac{1}{g!} \sum_{j=0}^{g-1} \binom{g}{j} (-1)^{g-j} j^n \quad (14.7)$$

(see Geman and Gelfand (1974, Chapter 4), Jensen (1981), and Sahai (1984, p. 377)). This can be approximated by $g^n/g!$, which is large even for moderate values of n and g . For example, if (n, g) is $(25, 10)$ is 2.6×10^{15} . The total possible number of clusters for a set of n items is $\sum_{g=1}^n N(n, g)$, which, for $n = 25$, is greater than (10^{15}) . Hence, hierarchical methods and other approaches permit us to search for a reasonable solution without having to look at all possible arrangements.

As noted in Section 14.1, hierarchical clustering algorithms involve a sequential process. In each step of the agglomerative hierarchical approach, an observation or a cluster of observations is merged into another cluster. In this process, the number of clusters shrinks and the clusters themselves grow larger. We start with n clusters (individual items) and end with one single cluster containing the entire data set. An alternative approach, called the divisive method, starts with a single cluster containing all n items and partitions a cluster into two clusters at each step (see Section 14.3.1b). The end result of the divisive approach is n clusters of one item each. Agglomerative methods are more commonly used than divisive methods. In either type of hierarchical clustering, a decision must be made as to the optimal number of clusters (see Section 14.3c).

At each step of an agglomerative hierarchical approach, the two closest clusters are merged into a single new cluster. The process is thus irreversible in the sense that any two items that are ever lumped together in a cluster cannot be separated later in the procedure; any early mistake cannot be corrected. Similarly, in a divisive hierarchical method, items cannot be moved to other clusters. An optional approach is to carry out a hierarchical procedure followed by a partitioning procedure in which items can be moved from one cluster to another (see Section 14.4.1).

Since an agglomerative hierarchical procedure combines the two closest clusters at each step, we must consider the question of measuring the similarity or dissimilarity of two clusters. Different approaches to measuring distance between clusters

give rise to different hierarchical methods. Agglomerative techniques are discussed in Sections 14.3.2–14.3.5, and the divisive approach is considered in Section 14.3.6.

14.3.2 Single Linkage (Nearest Neighbor)

In the single linkage method, the distance between two clusters A and B is defined as the minimum distance between a point in A and a point in B :

$$d(A, B) = \min\{d(\mathbf{y}_i, \mathbf{y}_j) : \mathbf{y}_i \text{ in } A \text{ and } \mathbf{y}_j \text{ in } B\}, \quad (14.8)$$

where $d(\mathbf{y}_i, \mathbf{y}_j)$ is the Euclidean distance in (14.2) or some other distance between the vectors $\mathbf{y}_i$ and $\mathbf{y}_j$. This approach is also called the *nearest neighbor method*.

At each step in the single linkage method, the distance (14.8) is found for every pair of clusters, and the two clusters with smallest distance are merged. The number of clusters is therefore reduced by 1. After two clusters are merged, the procedure is repeated for the next step: the distances between all pairs of clusters are calculated again, and the pair with minimum distance is merged into a single cluster.

The results of a hierarchical clustering procedure can be displayed graphically using a *tree diagram*, also known as a *dendrogram*, which shows all the steps in the hierarchical procedure, including the distances at which clusters are merged. Dendrograms are shown in Figures 14.2 and 14.3 in Examples 14.3.2(a) and 14.3.2(b).

Example 14.3.2(a). Hartigan (1975a, p. 28) compared the crime rates per 100,000 population for various cities. The data are in Table 14.1 (taken from the PCN 1.8).

Table 14.1. City Crime Rates per 100,000 Population

City	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto Theft
Atlanta	16.3	29.8	100	147	1102	993	499
Boston	4.2	13.3	122	79	982	669	354
Chicago	11.6	29.7	140	162	898	699	545
Dallas	18.1	34.2	164	209	1608	991	632
Denver	6.9	41.3	175	138	1134	1398	789
Detroit	13.6	39.7	277	208	1666	1083	798
San Diego	2.3	8.8	68	100	1017	728	468
Minneapolis	3.6	12.7	42	28	1497	1082	497
Houston	10.8	26.6	269	186	1389	787	697
San Jose	10.6	40.2	155	226	1494	953	565
Los Angeles	9.7	51.8	260	355	1982	1386	862
New Orleans	10.3	39.7	269	289	1636	1036	779
New York	9.4	19.4	833	267	1674	1092	588
Portland	8.0	26.0	187	144	1186	1281	488
Tucson	3.1	22.9	83	148	1266	756	483
Washington	12.3	27.6	324	217	1496	1083	799

Statistical Abstract. In order to illustrate the use of the distance matrix in single linkage clustering, we use the first six observations in Table 14.1 (Atlanta through Detroit).

The distance matrix D is given by

City	Distance between Cities					
Atlanta	0	300.0	316.4	390.2	491.6	719.2
Boston	300.0	0	447.4	833.1	913.0	891.1
Chicago	316.4	447.4	0	924.8	1023.4	971.5
Dallas	390.2	833.1	924.8	0	521.7	464.5
Detroit	491.6	913.0	1023.4	521.7	0	358.7
Fortuit	719.2	891.1	971.5	464.5	358.7	0

The smallest distance is 358.7 between Detroit and Detroit, and therefore these two cities are joined at the first step to form $C_1 = \{\text{Detroit, Detroit}\}$. In the next step, the distance matrix is calculated for Atlanta, Boston, Chicago, Dallas, and C_1 :

Atlanta	0	300.0	316.4	390.2	491.6
Boston	300.0	0	447.4	833.1	891.1
Chicago	316.4	447.4	0	924.8	971.5
Dallas	390.2	833.1	924.8	0	464.5
C_1	491.6	891.1	971.5	464.5	0

Note that all elements of this distance matrix are contained in the original distance matrix. This same pattern will hold in subsequent distance matrices and is also characteristic of the complete linkage method (see Example 14.3.3a). The smallest distance is 447.4 between Boston and Chicago. Therefore $C_2 = \{\text{Boston, Chicago}\}$. At the next step, the distance matrix is calculated for Atlanta, Dallas, C_1 , and C_2 :

Atlanta	0	316.4	390.2	491.6
C_1	316.4	0	833.1	891.1
Dallas	390.2	833.1	0	464.5
C_2	491.6	891.1	464.5	0

The smallest distance is 464.5 between Dallas and C_1 , so that $C_3 = \{\text{Dallas, } C_1\}$. The distance matrix for Atlanta, C_2 , and C_3 is given by

Atlanta	0	316.4	390.2
C_1	316.4	0	833.1
C_3	390.2	833.1	0

The smallest distance is 316.4, which defines $C_4 = \{\text{Atlanta, } C_1\}$. The distance matrix for C_2 and C_4 is

C_2	0	890.2
C_4	890.2	0

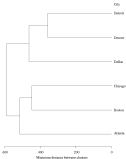


Figure 14.2. Dendrogram for single linkage of the first six observations in the city crime data in Table 14.1 (See Example 14.3.2(a)).

The last cluster is given by $\hat{C}_3 = \{\hat{C}_2, \hat{C}_1\}$. The dendrogram for the steps in this example is given in Figure 14.2. The order in which the clusters were formed and the relative distances at which they formed can all be seen. (Note that the distance scale runs from right to left.) $\square$

Example 14.3.2(b). To further illustrate the single linkage method of clustering, we use the complete city crime data from Table 14.1. The dendrogram in Figure 14.3 shows the cluster groupings obtained by the single linkage method. $\square$

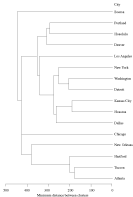


Figure 14.3. Dendrogram for single linkage of the complete city value data from Table 14.1 [see Example 14.3.2(a)].

14.3.3 Complete Linkage (Farthest Neighbor)

In the complete linkage approach, also called the *farthest neighbor method*, the distance between two clusters A and B is defined as the maximum distance between a point in A and a point in B :

$$D(A, B) = \max\{d(y_i, y_j) : y_i \text{ in } A \text{ and } y_j \text{ in } B\}. \quad (14.5)$$

At each step, the distance (14.5) is found for every pair of clusters, and the two clusters with the smallest distance are merged.

Example 14.3.3(a). As in Example 14.3.2(a) for single linkage clustering, we illustrate the use of the distance matrix in complete linkage clustering with the first six observations of the city crime data in Table 14.1. The initial distance matrix is exactly the same as in Example 14.3.2(a):

City	Distance					
Atlanta	0	506.6	58.4	586.2	681.6	716.2
Boston	506.6	0	447.4	833.0	915.0	881.1
Chicago	58.4	447.4	0	934.8	1053.4	971.8
Dallas	586.2	833.0	934.8	0	527.7	494.9
Denver	681.6	915.0	1053.4	527.7	0	582.7
Detroit	716.2	881.1	971.8	494.9	582.7	0

The smallest distance is 58.2 between Denver and Detroit, and these two therefore form the first cluster, $C_1 = \{\text{Denver, Detroit}\}$. Note that since the first cluster is based on the initial distance matrix, it will be the same regardless of which hierarchical clustering method is used.

In the next step, the distance matrix is calculated for Atlanta, Boston, Chicago, Dallas, and C_1 :

Atlanta	0	506.6	58.4	586.2	716.2
Boston	506.6	0	447.4	833.0	915.0
Chicago	58.4	447.4	0	934.8	1053.4
Dallas	586.2	833.0	934.8	0	527.7
C_1	716.2	915.0	1053.4	527.7	0

Note that this distance matrix differs from its analog for the second step in Example 14.3.2(a) only in the distances between C_1 and the other cities. All elements of this matrix and subsequent distance matrices below are contained in the original distance matrix for the six cities. The smallest distance is 447.4 between Boston and Chicago. Therefore, $C_2 = \{\text{Boston, Chicago}\}$. At the next step, distances are calculated for Atlanta, Dallas, C_1 , and C_2 :

Atlanta	0	586.2	586.2	716.2
C_1	586.2	0	934.8	881.1
Dallas	586.2	934.8	0	527.7
C_2	881.1	881.1	527.7	0

The smallest distance, 527.7, defines $C_3 = \{\text{Dallas, } C_1\}$. The distance matrix for Atlanta, C_3 , and C_2 is given by

Atlanta	0	586.6	715.2
C_1	586.6	0	1033.4
C_2	586.2	1033.4	0

The smallest distance is 586.2 between Atlanta and C_2 , so that $C_3 = \{\text{Atlanta}, C_2\}$. The distance matrix for C_3 and C_1 is

C_3	0	1073.4
C_1	1073.4	0

The last cluster is given by $C_4 = \{C_3, C_1\}$. The dendrogram in Figure 14.4 shows the steps in this example. $\square$



Figure 14.4. Dendrogram for complete linkage of the first six observations in the city crime data in Table 14.1 (see Example 14.3.3(a)).

Example 14.3.5b4. To further illustrate the complete linkage method, we use the complete crime data in Table 14.1. The dendrogram in Figure 14.5 shows the clusters found for this data set by the complete linkage approach. There are some differences between these groupings and the groupings from single linkage in Figure 14.3. $\square$

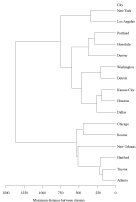


Figure 14.5. Dendrogram for complete linkage of the complete city crime data of Table 14.1 [see Example 14.3.5(b)].

14.3.4 Average Linkage

In the *average linkage* approach, the distance between two clusters A and B is defined as the average of the $n_A n_B$ distances between the n_A points in A and the n_B points in B :

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{x}_i, \mathbf{y}_j), \quad (14.10)$$

where the sum is over all $\mathbf{x}_i$ in A and all $\mathbf{y}_j$ in B . At each step, we join the two clusters with the smallest distance, as measured by (14.10).

Example 14.3.4. Figure 14.6 shows the dendrogram resulting from the average linkage method applied to the city crime data in Table 14.1. The solution is the same as the complete linkage solution for this data set given in Example 14.3.3(b) and Figure 14.5. $\square$

14.3.5 Centroid

In the *centroid* method, the distance between two clusters A and B is defined as the Euclidean distance between the mean vectors (often called *centroids*) of the two clusters:

$$D(A, B) = d(\bar{\mathbf{y}}_A, \bar{\mathbf{y}}_B), \quad (14.11)$$

where $\bar{\mathbf{y}}_A$ and $\bar{\mathbf{y}}_B$ are the mean vectors for the observation vectors in A and the observation vectors in B , respectively, and $d(\bar{\mathbf{y}}_A, \bar{\mathbf{y}}_B)$ is defined as (14.2). We define $\bar{\mathbf{y}}_A$ and $\bar{\mathbf{y}}_B$ in the usual way, that is, $\bar{\mathbf{y}}_A = \sum_{i=1}^{n_A} \mathbf{x}_i / n_A$. The two clusters with the smallest distance between centroids are merged at each step.

After two clusters A and B are joined, the centroid of the new cluster AB is given by the weighted average

$$\bar{\mathbf{y}}_{AB} = \frac{n_A \bar{\mathbf{y}}_A + n_B \bar{\mathbf{y}}_B}{n_A + n_B}. \quad (14.12)$$

Example 14.3.5. Figure 14.7 shows the dendrogram resulting from using the centroid clustering method on the complete-city crime data in Table 14.1.

Note the two crossovers in the dendrogram in Figure 14.7. Boston and Chicago join at a distance of 441.4. Then that cluster joins with (Atlanta, Tucson, Hartford) at a distance of 441.1. Finally, all five join with New Orleans at a distance of 593.8. Crossovers are discussed in Section 14.3.8a. $\square$

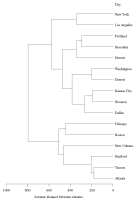


Figure 14.6. Dendrogram for average linkage clustering of the data in Table 14.1 (see Example 14.3-4).

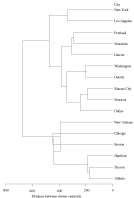


Figure 14.3. Dendrogram for the sequential clustering of the complete city crime data in Table 14.1 (see Example 14.3.2).

14.3.6 Median

If two clusters A and B are combined using the centroid method, and if A contains a larger number of items than B , then the new centroid $\bar{x}_{AB} = (n_A\bar{x}_A + n_B\bar{x}_B)/(n_A + n_B)$ may be much closer to $\bar{x}_A$ than to $\bar{x}_B$. To avoid weighting the mean vectors according to cluster size, we can use the median (midpoint) of the line joining A and B as the point for computing new distances to other clusters:

$$m_{AB} = \frac{1}{2}(\bar{x}_A + \bar{x}_B). \quad (14.13)$$

These clusters with the smallest distance between medians are merged at each step.

Note that the median in (14.13) is not the ordinary median in the statistical sense. The terminology arises from a median of a triangle, namely, the line from a vertex to the midpoint of the opposite side.

Example 14.3.6. Figure 14.8 shows the dendrogram resulting from using the median distance-clustering method on the complete city crime data in Table 14.1. In Figure 14.8, we use the same two measures as in Figure 14.7. $\square$

14.3.7 Ward's Method

Ward's method, also called the incremental sum-of-squares method, uses the within-cluster (squared) distances and the between-cluster (squared) distances (Ward 1963, Ward and 1963a). If AB is the cluster obtained by combining clusters A and B , then the sum of within-cluster distances (of the items from the cluster mean vectors) are

$$SS E_A = \sum_{i=1}^{n_A} (p_i - \bar{x}_A)^T (p_i - \bar{x}_A), \quad (14.14)$$

$$SS E_B = \sum_{i=1}^{n_B} (p_i - \bar{x}_B)^T (p_i - \bar{x}_B), \quad (14.15)$$

$$SS E_{AB} = \sum_{i=1}^{n_{AB}} (p_i - \bar{x}_{AB})^T (p_i - \bar{x}_{AB}), \quad (14.16)$$

where $\bar{x}_{AB} = (n_A\bar{x}_A + n_B\bar{x}_B)/(n_A + n_B)$, as in (14.12), and n_A , n_B , and $n_{AB} = n_A + n_B$ are the numbers of points in A , B , and AB , respectively. Since these sums of distances are equivalent to within-cluster sums of squares, they are denoted by $SS E_A$, $SS E_B$, and $SS E_{AB}$.

Ward's method joins the two clusters A and B that minimize the increase in $SS E$, defined as

$$J_{AB} = SS E_{AB} - (SS E_A + SS E_B). \quad (14.17)$$

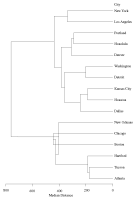


Figure 14.8. Dendrogram for the median clustering method applied to the complete city-state data in Table 14.1 (see Example 14.3.6).

It can be shown that the increase I_{AB} in (14.17) has the following two equivalent forms:

$$I_{AB} = n_A(\bar{Q}_A - \bar{Q}_{AB})^2(\bar{Q}_A - \bar{Q}_{AB}) + n_B(\bar{Q}_B - \bar{Q}_{AB})^2(\bar{Q}_B - \bar{Q}_{AB}) \quad (14.18)$$

$$= \frac{n_A n_B}{n_A + n_B} (\bar{Q}_A - \bar{Q}_B)^2 (\bar{Q}_A - \bar{Q}_B). \quad (14.19)$$

Thus by (14.19), minimizing the increase in SSI is equivalent to minimizing the between-cluster distances. If A consists only of q_1 and B consists only of q_2 , then SSK_A and SSK_B are zero, and (14.17) and (14.19) reduce to

$$I_0 = SSK_{AB} = \frac{1}{2}(q_1 - q_2)^2(q_1 - q_2) = \frac{1}{2}d^2(q_1, q_2).$$

Ward's method is related to the centroid method in Section 14.3.5. If the distance $d(\bar{Q}_A, \bar{Q}_B)$ in (14.11) is squared and compared to (14.19), the only difference is the coefficient $n_A n_B / (n_A + n_B)$ for Ward's method. Thus the cluster sizes have an impact on Ward's method, but not on the centroid method. Writing $n_A n_B / (n_A + n_B)$ in (14.19) as

$$\frac{n_A n_B}{n_A + n_B} = \frac{1}{1/n_A + 1/n_B},$$

we see that as n_A and n_B increase, $n_A n_B / (n_A + n_B)$ increases. Writing the coefficient as

$$\frac{n_A n_B}{n_A + n_B} = \frac{n_A}{1 + n_A/n_B},$$

we see that as n_B increases with n_A fixed, $n_A n_B / (n_A + n_B)$ increases. Therefore, compared to the centroid method, Ward's method is more likely to join smaller clusters or clusters of equal size.

Example 14.3.7. Figure 14.8 shows the dendrogram resulting from using Ward's clustering method on the complete city crime data in Table 14.1. The vertical axis is $I_{\text{tot}} / \sum_{i=1}^n (Q_i - \bar{Q})(Q_i - \bar{Q})$, where $\bar{Q}$ is the overall mean vector for the data. $\square$

14.3.8 Flexible Beta Method

Suppose clusters A and B have just been merged to form cluster AB . A general formula for the distance between AB and any other cluster C was given by Lance and Williams (1967):

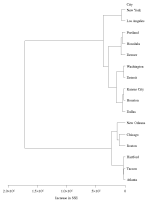


Figure 14.8. Dendrogram for Ward's method applied to the complete city crime data in Table 14.1 (see Example 14.3.7).

$$D(C, AB) = \alpha_1 D(C, A) + \alpha_2 D(C, B) + \beta D(A, B) + \gamma (D(C, A) - D(C, B)). \quad (14.28)$$

The distances $D(C, A)$, $D(C, B)$, and $D(A, B)$ are from the distance matrix before joining A and B . The distances from AB to other clusters are given by (14.28) would be used (along with distances between other pairs of clusters) to form the next distance matrix for choosing the pair of clusters with smallest distance. This pair would then be joined at the next step.

To simplify (14.28), Lance and Williams (1967) suggested the following constraints on the parameter values:

$$\begin{aligned} \alpha_1 + \alpha_2 + \beta &= 1, \\ \alpha_1 &= \alpha_2, \\ \gamma &= 0, \\ \beta &\leq 1. \end{aligned}$$

With $\alpha_1 = \alpha_2$ and $\gamma = 0$, we have $(\alpha_1 = 1 - \beta$ or $\alpha_1 = \alpha_2 = (1 - \beta)/2$), and we need only choose a value of β . The resulting hierarchical clustering procedure is called the *flexible beta method*.

The choice of β determines the characteristics of the flexible beta clustering procedure. Lance and Williams (1967) suggested the use of a small negative value of β , such as $\beta = -.25$. If there are (or might be) outliers in the data, the use of a smaller value of β , such as $\beta = -.5$, may be more likely to include these outliers into single clusters.

The distances defined for the agglomerative hierarchical methods in Sections 14.3.2–14.3.7 can all be expressed as special cases of (14.28). The specific parameter values are given in Table 14.2. For the centroid, median, and Ward's methods, the

Table 14.2. Parameter Values for (14.28)

Cluster Method	α_1	α_2	β	γ
Single linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Average linkage	$\frac{\alpha_1}{\alpha_1 + \alpha_2}$	$\frac{\alpha_2}{\alpha_1 + \alpha_2}$	0	0
Centroid	$\frac{\alpha_1}{\alpha_1 + \alpha_2}$	$\frac{\alpha_2}{\alpha_1 + \alpha_2}$	$\frac{-\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2}$	0
Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0
Ward's method	$\frac{\alpha_1 + \alpha_2}{\alpha_1 + \alpha_2 + \alpha_3}$	$\frac{\alpha_1 + \alpha_2}{\alpha_1 + \alpha_2 + \alpha_3}$	$\frac{-\alpha_3}{\alpha_1 + \alpha_2 + \alpha_3}$	0
Flexible beta	$(1 - \beta)/2$	$(1 - \beta)/2$	β (≤ 1)	0

distances in (14.20) must be squared distances (assuming Euclidean distances). For the other methods in Table 14.2, the distances may be either squared or unsquared.

We illustrate the choice of parameter values in Table 14.2 for the single linkage method. Using $\alpha_1 = \alpha_2 = \frac{1}{2}$, $\beta = 0$, and $\gamma = -\frac{1}{2}$ as in the first row of Table 14.2, (14.20) becomes

$$D(C, AB) = \frac{1}{2}D(C, A) + \frac{1}{2}D(C, B) - \frac{1}{2}D(C, A) - D(C, B), \quad (14.21)$$

if $D(C, A) > D(C, B)$, then $|D(C, A) - D(C, B)| = D(C, A) - D(C, B)$, and (14.21) reduces to

$$D(C, AB) = D(C, B). \quad (14.22)$$

On the other hand, if $D(C, A) < D(C, B)$, then $|D(C, A) - D(C, B)| = D(C, B) - D(C, A)$, and (14.21) reduces to

$$D(C, AB) = D(C, A). \quad (14.23)$$

Thus, (14.21) can be written as

$$D(C, AB) = \min\{D(C, A), D(C, B)\}, \quad (14.24)$$

which is equivalent to (14.6), the definition of distance for the single linkage method.

Example 14.3.8. Figures 14.9 and 14.11 show dendrograms produced when using the flexible beta clustering method on the complete city crime data in Table 14.1, with $\beta = -.25$ and $\beta = -.75$. The two results are similar. $\square$

14.3.9 Properties of Hierarchical Methods

14.3.9a Monotonicity

If an item or a cluster joins another cluster at a distance that is less than the distance for the previous merger of two clusters, we say that an *inversion* or a *reversal* has occurred. The reversal is represented by a crossover in the dendrogram. Examples of reversals can be found in Figures 14.3 and 14.5.

A hierarchical method in which reversals cannot occur is said to be *monotonic*, because the distance at each step is greater than the distance at the previous step. A distance measure or clustering method that is monotonic is also called *ultrametric*.

We now show that the single linkage and complete linkage methods are monotonic. Let d_k be the distance at which two clusters are joined at the k th step. We can describe steps k and $k+1$ in terms of four clusters A , B , C , and D . Suppose $D(A, B) < d_k$ so that the distance between any other pair among these four clusters, so that A and B are joined at step k to form AB . Then

$$d_{k+1} = D(A, B) = \min\{D(A, C), D(B, C), D(C, D)\}. \quad (14.25)$$

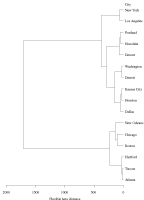


Figure 14.18. Dendrogram for the flexible beta method with $\beta = -0.25$ applied to the complete city distance data in Table 14.1. (see Example 14.3.8).

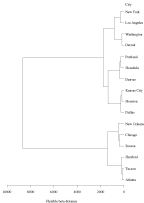


Figure 14.11: Dendrogram for the flexible beta method with $\beta = -0.75$ applied to the complete-city-crime-data in Table 14.1 (see Example 14.3.6).

[If $D(A, B)$ is less than these three distances, it is less than the other two possible distances, $D(A, C)$ and $D(B, C)$.] Suppose at step $k+1$ we join A and C or we join C and B . If we merge C and B , then by (14.23), $a_k = D(A, B) = D(C, B) = a_{k+1}$. If we join A and C , then for single linkage (14.24) gives

$$a_{k+1} = D(C, AB) = \min\{D(A, C), D(B, C)\} = a_k = D(A, B)$$

by (14.25), both of $D(A, C)$ and $D(B, C)$ exceed $D(A, B)$, and this also holds for complete linkage. Thus, the single linkage and complete linkage methods are monotonic.

For the methods in Table 14.2 other than single linkage and complete linkage, we have $\gamma = 0$, then by (14.28) and (14.29),

$$D(C, AB) = (\alpha_1 + \alpha_2 + \beta)D(A, B). \quad (14.26)$$

Thus we need $\alpha_1 + \alpha_2 + \beta \leq 1$ for monotonicity. Using this criterion, we see that all methods in Table 14.1 (beyond the first two) are monotonic except the centroid and median methods. (These two methods showed cross-over in the dendrograms in Figures 14.7 and 14.8.) Because of lack of monotonicity, some authors do not recommend the centroid and median methods.

14.3.6 Conservation or Dilution

We now consider the characteristics of the distances or proximities between the original points. As clusters form, the properties of this space of distances may be altered somewhat. A clustering method that does not alter the spatial properties is referred to by Lance and Williams (1967) as *space-conserving*. A method that is not space-conserving may either *contract* or *dilate* the space.

A method is *space-contracting* if newly formed clusters appear to move closer to individual observations, so that an individual item tends to join an existing cluster rather than join with another individual item to form a new cluster. This tendency is also called *clumping*.

A method is *space-dilating* if newly formed clusters appear to move away from individual observations, so that individual item tends to form new clusters rather than join existing clusters. In this case, clusters appear to be more distant than they are.

Debus and Wierle (1979) described the spatial properties as follows. Suppose that the distances among three clusters satisfy

$$D(A, B) = D(A, C) = D(B, C).$$

Then a cluster method is space-conserving if

$$D(A, C) \leq D(AB, C) \leq D(B, C). \quad (14.27)$$

A method is *space-contracting* if the first inequality in (14.27) does not hold and *space-dilating* if the second inequality does not hold.

The single linkage method is very space-contracting, with marked chaining tendencies. For this reason, single linkage is not recommended by some authors. Complete linkage on the other hand, is very space-dilating, with a tendency to artificially improve cluster boundaries.

Other hierarchical methods fall in between the extremes represented by single linkage and complete linkage. The centroid and average linkage methods are largely space-conserving, whereas Ward's method is space-contracting. Whenever a method produces *severals* for a particular data set, it can be considered to be space-contracting. Thus, for example, the centroid method is space-conserving unless it has *severals*, whereupon it becomes space-contracting.

The flexible beta method is space-contracting for $\beta > 0$, space-conserving for $\beta = 0$, and space-dilating for $\beta < 0$. A small degree of dilation may help define cluster boundaries, but too much dilation may lead to too many clusters in the early stages. Thus the recommended value of $\beta = -.15$ may represent a good compromise.

Example 14.3.9b. To illustrate chaining in the single linkage method, consider the data plotted in Figure 14.12 (similar to Overin 1953, p. 68). There are two distinct clusters, *A* and *C*, with intervening points labeled *B* that do not belong to *A* or *C*.

In Figure 14.13, the three-cluster solution for single linkage clustering places C_1 and C_{11} into one cluster and all other points into another cluster. The three-cluster solution has two clusters with *C*'s and a cluster with *A*'s and *B*'s.

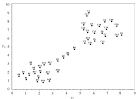


Figure 14.12. Two distinct clusters with intervening individuals.

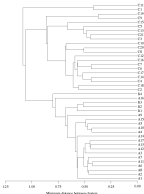


Figure 14.13. Single-linkage clustering of the data in Figure 14.12.

A dendrogram for average linkage clustering of the data in Figure 14.12 is given in Figure 14.14. For this data set, the average linkage method is more robust to chaining. The hierarchical solution separates the C's from the A's and B's. The same cluster solution completely separates the three groups, A, B, and C. $\square$

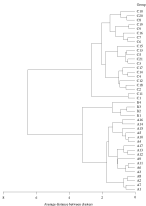


Figure 14.14. Average linkage clustering of the data in Figure 14.12.

14.1.3c Other Properties

The single linkage method has been criticized by many authors because of its chaining tendency and its sensitivity to errors in distances between observations. On the other hand, the single linkage approach is better than the other methods at identifying clusters that have curly shapes instead of spherical or elliptical shapes, and it is somewhat robust to outliers in the data.

Ward's method and the average linkage method are also relatively insensitive to outliers. For example, in the average linkage method, outliers tend to remain isolated in the early stages and to join with other outliers rather than to join with large clusters or with less compact clusters. This is due to two properties of the average linkage method: (1) the average distance between two groups (squared Euclidean distance) increases as the points in the groups are more spread out, and (2) the average distance increases as the size of the groups increases.

These two properties of the average linkage method are illustrated in one dimension in Figure 14.13 (similar to Johnson 1992, pp. 524–525), where cluster A has one point at z_1 and cluster B has two points, z_2 and z_3 , located at $z_1 - h$ and $z_1 + h$. The average squared distance between A and B is

$$\begin{aligned}\bar{D}^2 &= \frac{1}{2}[(z_1 - z_2 + h)^2 + (z_1 - z_3 - h)^2] \\ &= \frac{1}{2}[(z_1 - z_2)^2 + h^2 + 2h(z_1 - z_2) + (z_1 - z_2)^2 + h^2 - 2h(z_1 - z_2)] \\ &= (z_1 - z_2)^2 + h^2.\end{aligned}$$

Thus the average distance between A and B increases as the spread of z_2 and z_3 increases (that is, as h increases).

To illustrate the second property of the average linkage method, suppose cluster B in Figure 14.13 consists of a single point located at z_2 . Then, the distance between A and B is $(z_1 - z_2)^2$, and A is closer to B than it is if B consists of two points.



Figure 14.13. Clusters in one dimension.

The centroid method is fairly robust to outliers. Complete linkage is somewhat sensitive to outliers and tends to produce clusters of the same size and shape. Ward's method tends to yield spherical clusters of the same size.

Many studies conclude that the best overall performers are Ward's method and the average linkage method. However, there seems to be an interaction between methods and data sets; that is, some methods work better for certain data sets, and other methods work better for other data sets.

A good strategy is to try several methods. If the results agree to some extent, you may have found some natural clusters in the data.

14.3.18 Divisive Methods

In the agglomerative hierarchical methods covered in Sections 14.3.1–14.3.9, we begin with n items and end with a single cluster containing all n items. As noted in the several paragraphs of Section 14.3.1, a divisive hierarchical method starts with a single cluster of n items and divides it into two groups. At each step thereafter, one of the groups is divided into two subgroups. The ultimate result of a divisive algorithm is a cluster of one item each. The results can be shown in a dendrogram.

Divisive methods suffer from the same potential drawback as the agglomerative methods—namely, once a partition is made, an item cannot be moved into another group if does not belong to it at the time of the partitioning. However, if larger clusters are of interest, then the divisive approach may sometimes be preferred over the agglomerative approach, in which the larger clusters are reached only after a large number of joinings of smaller groups.

Divisive algorithms are generally of two classes: monothetic and polythetic. In a monothetic approach, the division of a group into two subgroups is based on a single variable, whereas, the polythetic approach uses all p variables to make the split.

If the variables are binary (quantitative variables can be converted to binary variables), the monothetic approach can easily be applied. Division into two groups is based on presence or absence of an attribute. The variable (attribute) is chosen that maximizes a chi-square statistic or an information statistic; see Drezlil (1993, pp. 87–90) or Gordon (1999, pp. 126–134).

For a monothetic approach using a quantitative variable y , we seek to maximize the between-group sum of squares,

$$SSB = n_1\bar{Y}_1 - \bar{Y}^2 + n_2\bar{Y}_2 - \bar{Y}^2,$$

where n_1 and n_2 are the two group sizes (with $n_1 + n_2 = n$), $\bar{Y}_1$ and $\bar{Y}_2$ are the group means, and $\bar{Y}$ is the overall mean based on all n observations. The sum of squares SSB would be calculated for all possible splits into two groups of sizes n_1 and n_2 and for each of the p variables. The final division would be based on the variable that maximizes SSB ($\sum_{i=1}^n (y_i - \bar{Y})^2$).

For a polythetic approach, we consider a technique proposed by MacNaughton-Smith et al. (1988). To divide a group, we work with a splitter group and the remainder. We seek the item in the remainder whose average distance (dissimilarity) from

other items in the remainder, minus its average distance from items in the splinter group, is largest. If the largest difference is positive, the item is shifted to the splinter group. If the largest difference is negative, the procedure stops, and the division is complete. We can start the splinter group with the item that has the largest average distance from the other items in the group.

Example 14.3.18. In Table 14.3 we have the track records of eight countries (Division 1997). Based on the distance matrix for these eight observations, the average distance from each observation to the other seven observations is given in Table 14.4. Since USA has the greatest average distance to the other countries, USA becomes the first observation in the splinter group. Now, the average distance between each observation in the remainder to the other six observations in the remainder is calculated. Then the (average)-distance between USA and each item in the remainder is calculated. (This may be found using the distance matrix since there is only one observation in the splinter group.) Finally, the difference between the average distance to the remainder and the average distance to the splinter group is calculated. The results are in Table 14.5. Because Australia has a positive difference in Table 14.5, it is added to the splinter group with USA. This process is repeated for the six countries in the remainder; the results are given in Table 14.6. Since no difference in Table 14.6 is positive, the process stops, giving the following clusters:

Table 14.3. Athlete Records for Eight Countries

Country	1	2	3	4	5	6	7	8
Australia	30.71	30.06	44.84	1.74	1.71	15.38	25.65	136.50
Belgium	30.54	30.08	45.04	1.73	1.60	15.22	25.43	150.04
Canada	30.77	30.22	45.68	1.76	1.63	15.33	26.09	182.13
COR	30.12	30.33	44.87	1.73	1.36	15.17	25.42	120.02
GB	30.11	30.21	44.93	1.70	1.31	15.04	25.31	120.13
Kenya	30.40	30.00	44.91	1.73	1.23	15.30	25.80	120.15
USA	30.93	31.73	45.80	1.73	1.53	15.20	25.43	120.22
USSR	30.07	30.00	44.60	1.73	1.39	15.30	25.53	130.15

Event: 50 100m (S), 200m (S), 400m (S), 800m (S), 1500m (S), 5000m (S), 10000m (S), 20000m (S), 50000m (S), 100000m (S), 200000m (S), 400000m (S), 800000m (S), 1600000m (S), 3200000m (S), 6400000m (S), 12800000m (S), 25600000m (S), 51200000m (S), 102400000m (S), 204800000m (S), 409600000m (S), 819200000m (S), 1638400000m (S), 3276800000m (S), 6553600000m (S), 13107200000m (S), 26214400000m (S), 52428800000m (S), 104857600000m (S), 209715200000m (S), 419430400000m (S), 838860800000m (S), 1677721600000m (S), 3355443200000m (S), 6710886400000m (S), 13421772800000m (S), 26843545600000m (S), 53687091200000m (S), 107374182400000m (S), 214748364800000m (S), 429496729600000m (S), 858993459200000m (S), 1717986918400000m (S), 3435973836800000m (S), 6871947673600000m (S), 13743895347200000m (S), 27487790694400000m (S), 54975581388800000m (S), 109951162777600000m (S), 219902325555200000m (S), 439804651110400000m (S), 879609302220800000m (S), 1759218604441600000m (S), 3518437208883200000m (S), 7036874417766400000m (S), 14073748835532800000m (S), 28147497671065600000m (S), 56294995342131200000m (S), 112589990684262400000m (S), 225179981368524800000m (S), 450359962737049600000m (S), 900719925474099200000m (S), 1801439850948198400000m (S), 3602879701896396800000m (S), 7205759403792793600000m (S), 14411518807585587200000m (S), 28823037615171174400000m (S), 57646075230342348800000m (S), 115292150460684697600000m (S), 230584300921369395200000m (S), 461168601842738790400000m (S), 922337203685477580800000m (S), 1844674407370955161600000m (S), 3689348814741910323200000m (S), 7378697629483820646400000m (S), 14757395258967641292800000m (S), 29514790517935282585600000m (S), 59029581035870565171200000m (S), 118059162071741130342400000m (S), 236118324143482260684800000m (S), 472236648286964521369600000m (S), 944473296573929042739200000m (S), 1888946593147858085478400000m (S), 3777893186295716170956800000m (S), 7555786372591432341913600000m (S), 15111572745182864683827200000m (S), 30223145490365729367654400000m (S), 60446290980731458735308800000m (S), 120892581961462917470617600000m (S), 241785163922925834941235200000m (S), 483570327845851669882470400000m (S), 967140655691703339764940800000m (S), 1934281311383406679529881600000m (S), 3868562622766813359059763200000m (S), 7737125245533626718119526400000m (S), 15474250491067253436239052800000m (S), 30948500982134506872478105600000m (S), 61897001964269013744956211200000m (S), 123794003928538027489912422400000m (S), 247588007857076054979824844800000m (S), 495176015714152109959649689600000m (S), 990352031428304219919299379200000m (S), 1980704062856608439838598758400000m (S), 3961408125713216879677197516800000m (S), 7922816251426433759354395033600000m (S), 15845632502852867518708790067200000m (S), 31691265005705735037417580134400000m (S), 63382530011411470074835160268800000m (S), 126765060022822940149670320537600000m (S), 253530120045645880299340641075200000m (S), 507060240091291760598681282150400000m (S), 1014120480182583521197362564300800000m (S), 2028240960365167042394725128601600000m (S), 4056481920730334084789450257203200000m (S), 8112963841460668169578900514406400000m (S), 16225927682921336339157801028812800000m (S), 32451855365842672678315602057625600000m (S), 64903710731685345356631204115251200000m (S), 129807421463370690713262408230502400000m (S), 259614842926741381426524816461004800000m (S), 519229685853482762853049632922009600000m (S), 1038459371706965525706099265844019200000m (S), 2076918743413931051412198531688038400000m (S), 4153837486827862102824397063376076800000m (S), 8307674973655724205648794126752153600000m (S), 16615349947311448411297588253504307200000m (S), 33230699894622896822595176507008614400000m (S), 66461399789245793645190353014017228800000m (S), 132922799578491587290380706028034457600000m (S), 265845599156983174580761412056068915200000m (S), 531691198313966349161522824112137830400000m (S), 1063382396627932698323045648224275660800000m (S), 2126764793255865396646091296448551321600000m (S), 4253529586511730793292182592897102643200000m (S), 8507059173023461586584365185794205286400000m (S), 17014118346046923173168730371588410572800000m (S), 34028236692093846346337460743176821145600000m (S), 68056473384187692692674921486353642291200000m (S), 136112946768375385385349842972707284582400000m (S), 272225893536750770770699685945414569164800000m (S), 544451787073501541541399371890829138329600000m (S), 1088903574147003083082798743781658276659200000m (S), 2177807148294006166165597487563316553318400000m (S), 4355614296588012332331194975126633106636800000m (S), 8711228593176024664662389950253266213273600000m (S), 17422457186352049329324779900506532426547200000m (S), 34844914372704098658649559801013064853094400000m (S), 69689828745408197317299119602026129706188800000m (S), 139379657490816394634598239204052259412377600000m (S), 278759314981632789269196478408104518824755200000m (S), 557518629963265578538392956816209037649510400000m (S), 1115037259926531157076785913632418075299020800000m (S), 2230074519853062314153571827264836150598041600000m (S), 4460149039706124628307143654529672301196083200000m (S), 8920298079412249256614287309059344602392166400000m (S), 17840596158824498513228574618118689204784332800000m (S), 35681192317648997026457149236237378409568665600000m (S), 71362384635297994052914298472474756819137331200000m (S), 142724769270595988105828596944949513638274662400000m (S), 285449538541191976211657193889899027276549324800000m (S), 570899077082383952423314387779798054553098649600000m (S), 1141798154164767904846628775559596109106197299200000m (S), 2283596308329535809693257551119192218212394598400000m (S), 4567192616659071619386515102238384436424789196800000m (S), 9134385233318143238773030204476768872849578393600000m (S), 18268770466636286477546060408953537745699156787200000m (S), 36537540933272572955092120817907075491398313574400000m (S), 73075081866545145910184241635814150982796627148800000m (S), 146150163733090291820368483271628301965593254297600000m (S), 292300327466180583640736966543256603931186508595200000m (S), 584600654932361167281473933086513207862373017190400000m (S), 1169201309864722334562947866173026415724746034380800000m (S), 2338402619729444669125895732346052831449492068761600000m (S), 4676805239458889338251791464692105662898984137523200000m (S), 9353610478917778676503582929384211325797968275046400000m (S), 18707220957835557353007165858768422651595936550092800000m (S), 37414441915671114706014331717536845303191873100185600000m (S), 74828883831342229412028663435073690606383746200371200000m (S), 149657767662684458824057326870147381212767492400742400000m (S), 299315535325368917648114653740294762425534984801484800000m (S), 598631070650737835296229307480589524851069969602969600000m (S), 1197262141301475670592458614961179049702139939205939200000m (S), 2394524282602951341184917229922358099404279878411878400000m (S), 4789048565205902682369834459844716198808559756823756800000m (S), 9578097130411805364739668919689432397617119513647513600000m (S), 19156194260823610729479337839378864795234239027295027200000m (S), 38312388521647221458958675678757729590468478054590054400000m (S), 76624777043294442917917351357515459180936956109180108800000m (S), 153249554086588885835834702715030918361873912218360217600000m (S), 306499108173177771671669405430061836723747824436720435200000m (S), 612998216346355543343338810860123673447495648873440870400000m (S), 1225996432692711086686677621720247346894991297746881740800000m (S), 2451992865385422173373355243440494693789982595493763481600000m (S), 4903985730770844346746710486880989387579965190987526963200000m (S), 9807971461541688693493420973761978775159930381975053926400000m (S), 19615942923083377386986841947523957550319860763950107852800000m (S), 39231885846166754773973683895047915100639721527900215705600000m (S), 78463771692333509547947367790095830201279443055800431411200000m (S), 156927543384667019095894735580191660402558886111600862822400000m (S), 313855086769334038191789471160383320805117772223201725644800000m (S), 627710173538668076383578942320766641610235544446403451289600000m (S), 1255420347077336152767157884641533283220471088892806902579200000m (S), 2510840694154672305534315769283066566440942177785613805158400000m (S), 5021681388309344611068631538566133132881884355571227610316800000m (S), 10043362776618689222137263077132266265763768711142455220633600000m (S), 20086725553237378444274526154264532531527537422284910441267200000m (S), 40173451106474756888549052308529065063055074844569820882534400000m (S), 80346902212949513777098104617058130126110149689139641765068800000m (S), 160693804425899027554196209234116260252220299378279283530137600000m (S), 321387608851798055108392418468232520504440598756558567060275200000m (S), 642775217703596110216784836936465041008881197513117134120550400000m (S), 1285550435407192220433569673872930082017762395026234268241100800000m (S), 2571100870814384440867139347745860164035524790052468536482201600000m (S), 5142201741628768881734278695491720328071049580104937072964403200000m (S), 10284403483257537763468557390983440656142099160209874145928806400000m (S), 20568806966515075526937114781966881312284198320419748291857612800000m (S), 41137613933030151053874229563933762624568396640839496583715225600000m (S), 82275227866060302107748459127867525249136793281678993167430451200000m (S), 164550455732120604215496918255735050498273586563357986334860902400000m (S), 329100911464241208430993836511470100996547173126715972669721804800000m (S), 658201822928482416861987673022940201993094346253431945339443609600000m (S), 1316403645856964833723975346045880403986188692506863890678887219200000m (S), 2632807291713929667447950692091760807972377385013727781357774438400000m (S), 5265614583427859334895901384183521615944754770027455562715548876800000m (S), 10531229166855718669791802768367043231889509540054911125431097753600000m (S), 21062458333711437339583605536734086463779019080109822250862195507200000m (S), 42124916667422874679167211073468172927558038160219644501724391014400000m (S), 84249833334845749358334422146936345855116076320439289003448782028800000m (S), 168499666669691498716668844293872691710232152640878578006897564057600000m (S), 336999333339382997433337688587745383420464305281757156013795128115200000m (S), 673998666678765994866675377175490766840928610563514312027590256230400000m (S), 1347997333357531989733350754350981533681857221127028624055180512460800000m (S), 2695994666715063979466701508701963067363714442254057248110361024921600000m (S), 5391989333430127958933403017403926134727428884508114496220722049843200000m (S), 10783978666860255917866806034807852269454857769016228992441444099686400000m (S), 21567957333720511835733612069615704538909715538032457984882888199372800000m (S), 43135914667441023671467224139231409077819431076064915969765776398745600000m (S), 86271829334882047342934448278462818155638862152129831939531552797491200000m (S), 172543658669764094685868896556925636311277724304259663879063105594982400000m (S), 345087317339528189371737793113851272622555448608519327758126211189964800000m (S), 690174634679056378743475586227702545245110897217038655516252422379929600000m (S), 1380349269358112757486951172455405090490221794434077311032504844759859200000m (S), 2760698538716225514973902344910810180980443588868154622065009689519718400000m (S), 5521397077432451029947804689821620361960887177736309244130019379039436800000m (S), 11042794154864902059895609379643240723921774355472618488260038758078873600000m (S), 2208558

Table 14.5. Average Distances to Remainder and Splitter Group for Seven-Country

Country	Average Distance to Remainder (1)	Average Distance to Splitter Group (2)	Difference (1)–(2)
Australia	1.720	1.120	.600
GB	1.628	1.060	.568
GER	.948	2.879	–1.931
USSR	1.159	2.469	–1.310
Canada	1.492	2.898	–1.406
Kenya	.595	2.173	–1.578
Belgium	.595	2.329	–1.734

Table 14.6. Average Distances to Remainder and Splitter Group for Six Countries

Country	Average Distance to Remainder (1)	Average Distance to Splitter Group (2)	Difference (1)–(2)
GB	1.146	1.266	–.120
GER	.767	1.872	–1.105
USSR	1.149	2.073	–1.224
Canada	1.249	2.457	–1.208
Kenya	.895	1.994	–1.099
Belgium	.883	2.059	–1.176

$C_1 = \{\text{USA, Australia}\}$, $C_2 = \{\text{GB, GER, USSR, Canada, Kenya, Belgium}\}$. We could continue and divide C_2 into two groups in the same way. $\square$

14.4 NONHIERARCHICAL METHODS

In this section, we discuss three nonhierarchical techniques: partitioning, mixtures of distributions, and density estimation. Among these three methods, partitioning is the most commonly used.

14.4.1 Partitioning

In the partitioning approach, the observations are separated into g clusters without using a hierarchical approach based on a matrix of distances or similarities between all pairs of points. The methods described in this section are sometimes called optimization methods rather than partitioning.

An attractive strategy would be to examine all possible ways to partition n items into g clusters and find the optimal clustering according to some criterion. However, the number of possible partitions as given by (14.7) is prohibitively large for even moderate values of n and g . Thus we seek simpler techniques.

14.4.1a *k*-Means

We now consider an approach to partitioning that is usually called the *k*-means method. (We will continue to use the notation g rather than k for the number of clusters.) The method allows the items to be moved from one cluster to another, a reallocation that is not available in the hierarchical methods.

We first select g items to serve as seeds. These are later replaced by the centroids (mean vectors) of the clusters. There are various ways we can choose the seeds: select g items at random (perhaps separated by a specified minimum distance), choose the first g points in the dataset (again subject to a minimum distance requirement), select the g points that are mutually farthest apart, find the g points of maximum density, or specify g regularly spaced points in a gridlike pattern (these would not be actual data points).

For these methods of selecting seeds, the number of clusters, g , must be specified. Alternatively, a minimum distance between seeds may be specified, and then all items that satisfy this criterion are chosen as seeds.

After the seeds are chosen, each remaining point in the data set is assigned to the cluster with the nearest seed (based on Euclidean distance). As soon as a cluster has more than one member, the cluster seed is replaced by the centroid.

After all items are assigned to clusters, each item is examined to see if it is closer to the centroid of another cluster than to the centroid of its own cluster. If so, the item is moved to the new cluster and the new cluster centroids are updated. This process is continued until no further improvement is possible.

The *k*-means procedure is somewhat sensitive to the initial choice of seeds. It might be advisable to try the procedure again with another choice of seeds. If different initial choices of seeds produce widely different final clusters, or if convergence is extremely slow, there may be no natural clusters in the data.

The *k*-means partitioning method can also be used as a possible improvement on hierarchical techniques. We first cluster the items using a hierarchical method and then use the centroids of these clusters as seeds for a *k*-means approach, which will allow points to be reallocated from one cluster to another.

Example 14.4.1a. Protein consumption in 25 European countries for nine food groups is given in Table 14.7 (Hend et al. 1994, p. 295). In order to evaluate the sensitivity of the *k*-means clustering method to the initial choice of seeds, we use the following four methods of choosing seeds:

1. Select at random g observations that are at least a distance r apart.
2. Select the first g observations that are at least a distance r apart.
3. Select the g observations that are mutually farthest apart.
4. Use the g centroids from the g -cluster solution from the average linkage (hierarchical) clustering method.

To help choose g , the number of clusters, we plot the first two principal components in Figure 14.15. It appears that there may be at least five clusters. For the

Table 14.5. Chicken Data

Country	Bird Meat	White Meat	Eggs	BBB	Hals	Cheeks	Starchy Food	Fats	Protein/egg
Austria	10.1	1.4	5	8.8	3	42.5	8	5.5	3.7
Austria	8.9	14.8	4.8	18.8	2.1	28.8	3.8	1.8	4.3
Belgium	13.8	8.3	4.1	17.8	4.8	28.8	8.7	2.1	4.6
Belgium	7.8	8.8	1.8	8.3	3.2	58.7	1.1	3.7	4.2
Czech	8.7	11.4	2.8	12.5	2.8	88.3	5.8	1.1	4.0
Denmark	10.8	10.8	3.7	25.8	8.8	21.3	4.8	3	2.4
E. Germany	8.4	11.8	3.7	11.1	5.4	28.8	8.5	8	3.6
Finland	9.5	4.8	2.7	59.7	5.8	28.5	5.1	1.8	3.4
France	18.8	9.8	5.5	18.5	5.7	28.1	4.8	2.4	6.5
France	18.2	5.8	2.8	17.8	8.8	48.7	2.5	7.8	6.8
Hungary	5.8	12.8	2.8	8.7	3	48.1	4.8	5.8	4.2
Ireland	13.8	18.8	8.7	28.8	2.2	28.8	8.2	1.8	2.8
Italy	9.8	5.1	2.8	18.7	8.4	38.8	2.1	4.3	8.7
Netherlands	9.5	13.8	3.8	28.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.5	8.7	23.8	4.8	1.8	2.7
Poland	8.8	18.2	2.7	18.5	3.8	38.1	5.8	2.8	6.6
Portugal	8.2	5.7	1.1	4.8	34.2	27.8	5.8	4.7	7.8
Romania	8.2	8.3	1.8	11.1	1.8	88.8	3.1	8.8	2.8
Spain	7.1	8.4	3.1	8.8	3.8	28.2	8.7	8.8	7.2
Sweden	8.8	7.8	3.5	28.7	7.5	18.5	3.7	1.8	2.8
Switzerland	18.1	18.1	3.1	28.8	2.3	25.8	2.8	2.8	4.8
UK	17.4	5.7	4.7	28.8	4.3	38.5	4.7	3.4	3.3
USSR	9.5	4.8	2.1	18.8	3.8	48.8	8.4	3.4	2.8
W. Germany	11.4	12.5	4.1	18.8	3.4	18.8	5.2	1.5	3.8
Yugoslavia	4.4	5.8	1.2	8.5	8	55.8	5.8	5.7	3.2

first method, we select five observations at random that are at least a distance $r = 1$ from each other. The five chosen seeds are Ireland, UK, Poland, Greece, and East Germany. Using these seeds, the k -means method produced the clusters identified in Table 14.8 along with the distance of each observation from its cluster centroid.

To view the clusters, we plot the first two discriminant functions (see Section 8.4.1) in Figure 14.13. The first two discriminant functions show good separation for clusters 2, 3, and 4 but poor separation for clusters 1 and 5.

We now select the first five observations as cluster seeds. With these seeds, the k -means clustering method produced the clusters in Table 14.9. The first two discriminant functions are plotted in Figure 14.18. Good separation of clusters is seen except for clusters 2 and 3.

We next choose as cluster seeds the five observations that are mutually farthest apart. These seeds give rise to the clusters in Table 14.10. The first two discriminant functions are plotted in Figure 14.19. Clusters 1, 3, and 4 seem very well separated, but clusters 2 and 5 show considerable overlap.

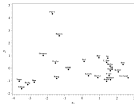


Figure 14.16. First two principal components z_1 and z_2 for the protein data in Table 14.7.

Table 14.8. k-Means-Cluster Solution for Seeds Chosen at Random

Country	Cluster	Distance from Centroid	Country	Cluster	Distance from Centroid
Portugal	1	1.488	Sweden	4	1.594
Spain	1	1.488	E. Germany	4	1.666
Netherlands	2	1.123	Norway	4	2.081
Austria	2	1.217	France	4	2.603
Czech	2	1.385	Romania	3	1.060
Switzerland	2	1.657	Yugoslavia	3	1.358
Poland	2	1.714	Bulgaria	3	1.341
Ireland	3	1.594	Italy	3	2.060
UK	3	1.828	Hungary	3	2.443
Finland	3	2.261	USSR	3	2.613
Belgium	4	1.281	Albania	3	2.728
W. Germany	4	1.485	Greece	3	2.741

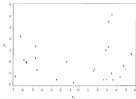


Figure 14.17. First two discriminant functions x_1 and x_2 for the countries in Table 14.9.

Table 14.9. K-Means Cluster Solution Using the First Five Observations as Seeds

Country	Cluster	Distance from Cluster	Country	Cluster	Distance from Cluster
Austria	1	.000	Romania	4	1.415
Netherlands	2	.648	Belgium	4	1.587
Austria	3	1.000	Yugoslavia	4	1.790
N. Germany	3	1.007	Italy	4	1.898
Switzerland	2	1.489	Greece	4	2.050
Belgium	3	1.568	Poland	3	1.709
Sweden	3	1.462	Czech	3	1.956
Denmark	3	1.666	USSR	3	2.288
Ireland	3	1.832	E. Germany	3	2.285
Norway	3	1.927	Spain	3	2.588
UK	3	2.076	Hungary	3	2.598
Finland	3	2.340	Portugal	3	2.829
France	3	2.829			

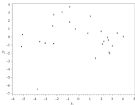


Figure 14.15. The two-discriminant functions D_1 and D_2 for the clusters in Table 14.15.

Table 14.15. K-Means Cluster Solution Using as Seeds the First Observations Plotted Appear

Country	Cluster	Distance from Centroid	Country	Cluster	Distance from Centroid
Scandinavia	1	.701	France	1	2.598
Yugoslavia	1	1.189	Poland	1	2.405
Belgium	1	1.433	UK	2	2.037
Austria	1	1.621	Greece	3	1.073
Finland	1	1.940	Italy	3	1.073
Belgium	2	.686	Portugal	4	1.400
R. Germany	2	1.031	Spain	4	1.400
Netherlands	2	1.435	Norway	5	1.054
Austria	2	1.693	Sweden	5	1.191
Denmark	2	1.706	Switzerland	6	1.543
Switzerland	2	1.713	Denmark	6	1.708
Ireland	2	1.839	USSR	6	2.780
R. Germany	2	2.042			

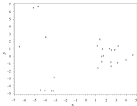


Figure 14.15. First two discriminant functions α_1 and α_2 for the cluster in Table 14.10

Finally, we obtain a five-cluster solution from average linkage and use the centroids of these clusters as the new seeds. The clusters in Table 14.11 result. The first two discriminant functions are plotted in Figure 14.16. All five clusters are well separated in the first two discriminant functions. These clusters show some resemblance to those in the principal component plot given in Figure 14.10. (2)

Table 14.11. 5-Means Cluster Solution Using Seeds from Average Linkage

Country	Cluster	Distance		Country	Cluster	Distance	
			from Centroid				from Centroid
Romania	1		.079	Norway	1		2.187
Yugoslavia	1		1.182	UK	1		2.554
Belgium	1		1.559	France	2		2.690
Albania	1		1.679	Poland	2		2.680
Belgium	2		1.132	Greece	3		1.079
W. Germany	2		1.265	Italy	3		1.079
Netherlands	2		1.267	Portugal	4		1.490
Sweden	2		1.608	Spain	4		1.490
Ireland	2		1.744	Czech	5		1.557
Denmark	2		1.795	Poland	5		1.579
Switzerland	2		1.821	USSR	5		1.764
Austria	2		2.037	Hungary	5		2.029
E. Germany	2		2.284				

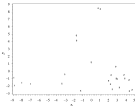


Figure 14.28. Two two-discriminant functions z_1 and z_2 for the clusters in Table 14.11.

14.4.13 Other Partitioning Criteria

We now consider three partitioning methods that are not based directly on the distance from a point to the centroid of a cluster. These methods are based on the between-cluster and within-cluster sum of squares and products matrices $\mathbf{B}$ and $\mathbf{E}$ defined in (3.8) and (3.10) for one-way MANOVA. For well defined clusters, we would like $\mathbf{E}$ to be “small” and $\mathbf{B}$ to be “large.”

The three criteria are as follows:

1. Minimize $\text{tr}(\mathbf{E})$.
2. Minimize $|\mathbf{E}|$.
3. Maximize $\text{tr}(\mathbf{E}^{-1}\mathbf{B})$.

Using criterion 1, for example, we would move an item with observation vector $\mathbf{y}$ to the cluster for which $\text{tr}(\mathbf{E})$ is minimized after the move.

We can express the first criterion in two alternate forms. By (3.10), we have

$$\begin{aligned} \text{tr}(\mathbf{E}) &= \text{tr} \left[\sum_{i=1}^J \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mathbf{F}_i)(\mathbf{x}_{ij} - \mathbf{F}_i)' \right] \\ &= \sum_i \text{tr} \left[\sum_j (\mathbf{x}_{ij} - \mathbf{F}_i)(\mathbf{x}_{ij} - \mathbf{F}_i)' \right] \quad \text{By (2.96)} \end{aligned} \quad (14.28)$$

$$= \sum_i n(\mathbf{K}_i), \quad (14.28)$$

where $\mathbf{K}_i = \sum_{j=1}^n (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)'$ is the sum of squares and products matrix of deviations of observations from the mean vector for the i th cluster. In (14.28) we use the notation of Section 8.1.1 for a balanced design, in which n is the number of observations in each cluster.

We can write $n(\mathbf{K}_i)$ in (14.28) in the form

$$\begin{aligned} n(\mathbf{K}_i) &= n \sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \\ &= \sum_j n(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \quad \text{(by (2.65))} \\ &= \sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) \quad \text{(by (2.67))} \end{aligned} \quad (14.29)$$

Thus $n(\mathbf{K}_i)$ is the sum of the (squared) Euclidean distances from the individual points in the centroid of the i th cluster.

A second form of (14.28) was given by Seber (1984, p. 277) as

$$n(\mathbf{K}) = \frac{1}{n} \sum_i \sum_{j=1}^n (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)'(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i). \quad (14.30)$$

Hence minimizing $n(\mathbf{K})$ is equivalent to minimizing the sum of squared Euclidean distances between all pairs of points in a cluster.

The second criterion, minimizing $\|\mathbf{K}\|$, is related to $\Delta = \|\mathbf{B}\|/(\mathbf{K} + \mathbf{H})$ in (6.13). Minimizing $\|\mathbf{K}\|$ is equivalent to minimizing Wilks' Λ for the clusters.

Another way to look at minimizing $\|\mathbf{K}\|$ is to consider the effect of adding a point $\mathbf{y}$ to a cluster with centroid $\bar{\mathbf{y}}$. Let $\mathbf{u} = \mathbf{y} - \bar{\mathbf{y}}$. By (14.28), $\mathbf{K}$ is a sum of terms of the form $\mathbf{u}\mathbf{u}' = (\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})'$. Thus (ignoring the change in centroid with the added observation $\mathbf{y}$), the increase in $\|\mathbf{K}\|$ is

$$\begin{aligned} \|\mathbf{K} + \mathbf{u}\mathbf{u}'\| - \|\mathbf{K}\| &= \|\mathbf{K}\|(1 + \mathbf{u}\mathbf{K}^{-1}\mathbf{u}) - \|\mathbf{K}\| \quad \text{(by (2.65))} \\ &= \|\mathbf{K}\|\mathbf{u}'\mathbf{K}^{-1}\mathbf{u}. \end{aligned}$$

Hence, the minimum increase in $\|\mathbf{K}\|$ is obtained by adding $\mathbf{y}$ to the cluster for which the standardized distance $\mathbf{u}'\mathbf{K}^{-1}\mathbf{u}$ of $\mathbf{y}$ from $\bar{\mathbf{y}}$ is the smallest. By comparison, the $n(\mathbf{K})$ criterion would add $\mathbf{y}$ to the cluster for which $\mathbf{u}'\mathbf{u}$ is minimum [see (14.30)].

The third criterion, maximizing $n(\mathbf{K}^{-1}\mathbf{H})$, is related to the Lawley–Hoeffding statistic $L^{\text{LH}} = n(\mathbf{K}^{-1}\mathbf{H}) = \sum_{i=1}^p \lambda_i$ in (6.22), where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigenvalues of $\mathbf{K}^{-1}\mathbf{H}$ and $p = \min(p, g - 1)$. Associated with each λ_i is the eigenvector $\mathbf{u}_i$ and the discriminant function $\Delta_i = \mathbf{u}_i'\mathbf{y}$ (see Section 8.4). The largest eigenvalue, λ_{i_1} , and the accompanying first discriminant function, $\Delta_1 = \mathbf{u}_{i_1}'\mathbf{y}$, have the greatest

influence on $\text{tr}(\mathbf{E}^{-1}\mathbf{E})$. Maximizing $\text{tr}(\mathbf{E}^{-1}\mathbf{E})$ has the inclination to produce elliptical clusters of the same size. These clusters would tend to follow a straight-line trend, especially if the first eigenvalue dominates the others. If the initial clusters or seeds are lined up in a different direction than the "true clusters," maximizing $\text{tr}(\mathbf{E}^{-1}\mathbf{E})$ may not correct the error in subsequent iterations.

Since $\text{tr}(\mathbf{E})$ involves only the diagonal elements, the first criterion ignores the correlations and tends to yield spherical clusters. The second criterion, minimizing $|\mathbf{E}|$, takes correlations into account and tends to produce elliptical clusters. These clusters have a tendency to be of the same shape because $\mathbf{E}_i/v_i$ is a pooled estimator of the covariance matrix. A modification that may be useful is $|\sum_{i=1}^n \mathbf{E}_i|$, where $\mathbf{E}_i$ is the error matrix for the i th cluster [see (14.28)].

Finally, we compare the three criteria in terms of invariance to nonsingular linear transformations $\mathbf{v}_i = \mathbf{A}\mathbf{p}_i + \mathbf{b}$, where $\mathbf{A}$ is a constant nonsingular matrix and $\mathbf{b}$ is a vector of constants. The first criterion, minimizing $\text{tr}(\mathbf{E})$, is not invariant to such linear transformations, whereas the other two criteria are invariant to these transformations. Therefore, minimizing $\text{tr}(\mathbf{E})$ will likely give different partitions for the raw data and standardized data.

14.4.2 Other Methods

We discuss mixtures of distributions in Section 14.4.2a and density estimation in Section 14.4.2b.

14.4.2a Mixtures of Distributions

In this method, we assume the existence of g distributions (usually multivariate normal), and we wish to assign each of the n items in the sample to the distribution it most likely belongs to. Such an approach imitates the classification analysis in Chapter 9. Along with partitioning in Section 14.4.1, this method has the property that points can be transferred from one cluster to another, but it requires more assumptions than partitioning.

We define the density of a mixture of g distributions as the weighted average

$$f(\mathbf{y}) = \sum_{i=1}^g \alpha_i f(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (14.32)$$

where $0 \leq \alpha_i \leq 1$, $\sum_{i=1}^g \alpha_i = 1$, and $f(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the multivariate normal distribution $N(\mathbf{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ given in (4.2).

Clusters could be formed in two ways. The first approach is to assign an item with observation vector $\mathbf{y}$ to the cluster C_i with largest value of the estimated posterior probability

$$\hat{f}(\mathbf{y}; \boldsymbol{\mu}_i) = \frac{\alpha_i f(\mathbf{y}; \boldsymbol{\mu}_i, \hat{\boldsymbol{\Sigma}}_i)}{\hat{f}(\mathbf{y})} \quad (14.33)$$

[see Raudenbush (1988, Sections 6.2.4 and 6.3.1)], where $\hat{\alpha}_i$, $\hat{\mu}_i$, and $\hat{\Sigma}_i$ are maximum likelihood estimates and $\hat{h}(y)$ is given by (14.32) with estimates inserted for parameters. The posterior probability (14.33) is an estimate of the probability that an item with observation vector y belongs to the i th cluster, C_i .

The second approach is to assign an item with observation vector y to the cluster with largest value of

$$\ln \hat{h}_i = \frac{1}{2} \ln |\hat{\Sigma}_i| - \frac{1}{2} (y - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} (y - \hat{\mu}_i) \quad (14.34)$$

[see (5.14)]. For either of these approaches [based on (14.33) or (14.34)], we need the estimates $\hat{\alpha}_i$, $\hat{\mu}_i$, and $\hat{\Sigma}_i$. These estimates are obtained by maximizing the likelihood function $\hat{L} = \prod_{j=1}^g h(y_j)$, where $h(y_j)$ is given by (14.32). The results are

$$\hat{\alpha}_i = \frac{1}{n} \sum_{j=1}^n \hat{P}(C_i | y_j), \quad i = 1, 2, \dots, g-1,$$

$$\hat{\mu}_i = \frac{1}{n \hat{\alpha}_i} \sum_{j=1}^n x_j \hat{P}(C_i | y_j), \quad i = 1, 2, \dots, g,$$

$$\hat{\Sigma}_i = \frac{1}{n \hat{\alpha}_i} \sum_{j=1}^n (x_j - \hat{\mu}_i)(x_j - \hat{\mu}_i)' \hat{P}(C_i | y_j), \quad i = 1, 2, \dots, g$$

[Silver 1985, p. 111], where $\hat{P}(C_i | y_j)$ is given by (14.33). These three equations must be solved iteratively. For a given value of g , we can begin with initial estimates or guesses for the parameters and adjust them by iteration (this approach is related to the EM algorithm mentioned in Section 3.11). If g is not known, we can begin with $g = 1$ and then successively try $g = 2$, $g = 3$, and so on, until the results are satisfactory.

The total number of parameters to be estimated is $g\alpha p$. There are p parameters in each μ_i , $\frac{1}{2}p(p+1)$ unique parameters in each Σ_i , and $g-1$ values of α_i (the remaining α_g is found by $\sum_{i=1}^g \alpha_i = 1$), for a total of

$$\frac{1}{2}p(p+1) + p + 2g - 1 \quad (14.35)$$

parameters. If the sample size is not sufficiently large to estimate all of these parameters, we could assume a common covariance matrix Σ , which reduces the number of parameters by $\frac{1}{2}(g-1)p(p+1)$.

The method of mixtures is insensitive to full-rank linear transformations and is somewhat robust to the assumption of normality. The technique works better if the g densities are well-separated or the sample sizes are large.

Example 14.4.2a. To illustrate the clustering method based on mixtures of distributions, we use the protein consumption data of Table 14.7. Because of the small number of countries in the data set, there are not enough degrees of freedom to estimate

a different covariance matrix for each cluster. Hence we assume equal covariance matrices and estimate a pooled covariance matrix $\hat{\Sigma}$. For illustration purposes, we choose $g = 5$, as in Example 14.6.3a.

We use the five clusters found by Ward's method to obtain initial estimates of α_i , μ_i , and Σ . Then the maximum likelihood equations are solved iteratively to find the following estimates:

$$\hat{\alpha}_1 = .2804, \quad \hat{\alpha}_2 = .3200, \quad \hat{\alpha}_3 = .3199,$$

$$\hat{\alpha}_4 = .1899, \quad \hat{\alpha}_5 = .3200,$$

$$\hat{\mu}_1 = \begin{pmatrix} 9.64 \\ 6.87 \\ 2.56 \\ 14.08 \\ 2.58 \\ 26.27 \\ 3.79 \\ 4.71 \\ 4.86 \end{pmatrix}, \quad \hat{\mu}_2 = \begin{pmatrix} 13.26 \\ 10.69 \\ 3.88 \\ 21.36 \\ 3.38 \\ 24.78 \\ 4.69 \\ 3.66 \\ 4.38 \end{pmatrix}, \quad \hat{\mu}_3 = \begin{pmatrix} 6.11 \\ 5.71 \\ 1.43 \\ 9.43 \\ .64 \\ 26.09 \\ 5.80 \\ 4.90 \\ 5.40 \end{pmatrix},$$

$$\hat{\mu}_4 = \begin{pmatrix} 9.86 \\ 7.88 \\ 3.43 \\ 26.88 \\ 6.24 \\ 22.88 \\ 4.88 \\ 1.88 \\ 3.12 \end{pmatrix}, \quad \hat{\mu}_5 = \begin{pmatrix} 7.05 \\ 4.23 \\ 1.63 \\ 9.28 \\ 6.47 \\ 26.50 \\ 4.80 \\ 3.88 \\ 4.23 \end{pmatrix},$$

$$\hat{\Sigma} = \begin{pmatrix} 4.158 & -0.182 & .081 & -.363 & 1.008 & .689 & -.197 & .087 & .000 \\ -0.182 & 9.411 & .863 & .264 & -1.936 & -0.238 & 1.240 & -0.893 & -.309 \\ .081 & .863 & 2.711 & .582 & -.296 & -.089 & .388 & -.136 & -.088 \\ -.363 & .264 & .582 & 9.798 & -1.236 & .041 & 1.815 & -.804 & .802 \\ 1.008 & -0.238 & -.089 & -1.236 & 3.608 & .367 & .114 & .339 & 1.000 \\ .689 & -0.238 & -.089 & -.363 & .367 & 9.432 & -.777 & 1.708 & .357 \\ -.197 & .087 & .388 & -.804 & .114 & -.777 & 1.608 & -.849 & -.288 \\ .087 & -.893 & -.136 & -.804 & .339 & 1.708 & -.849 & 5.051 & .585 \\ .000 & -.309 & -.088 & .802 & 1.000 & .357 & -.288 & .585 & 1.000 \end{pmatrix}.$$

Then assigning each country to the cluster for which it has the highest posterior probability of membership as in (14.54) yields the following clusters:

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Austria, Czech, Greece,	Arctic,	Belgium,	Denmark,	E. Germany,
Finland, Italy,	Belgium, France,	Komoro,	Finland,	Portugal, Spain,
Poland, USSR,	Iceland,	Yugoslavia	Norway,	
	Netherlands,		Sweden	
	Switzerland, UK,			
	W. Germany			

□

14.4.2b Density Estimation

In the method of *density estimation*, or *density searching*, we seek regions of high density sometimes called *modes*. No assumption is made about the form of the density, as was done in Section 14.4.2a. We could estimate the density using a kernel function as in Section 9.3.1. Alternatively, we simply attempt to separate regions with a high concentration of points from regions with a low density.

To find regions of high density, we first choose a radius r and a value of k , the number of points in a k -nearest neighbor scheme. For each of the n points in the data, the number of points within a sphere of radius r is found. A point is called a *dense point* if at least k other points are contained in its sphere.

If a dense point is more than a distance r from all other dense points, it becomes the nucleus of a new cluster. If a dense point is within a distance r from at least one dense point that belongs to a cluster, it is added to the cluster. If the dense point is within a distance r of two or more clusters, those clusters are combined. Two clusters are also combined if the smallest distance between their dense points is less than the average of the 24 smallest distances between the original n points. The value of r can be gradually increased so that more points become dense. Another option is to begin with the specified value of r for each point and then gradually increase r until k observations are contained in its sphere.

Example 14.4.2b. To illustrate the density estimation method, we use the protein data. For each pair of values of k and r , the value of r was allowed to increase if needed, as described above. For the following values of k and r , the number of clusters obtained are given.

$k \backslash r$	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8
2	3	3	3	4	4	4	4	4	3	3	3	3	3
3	3	3	3	3	3	3	3	3	2	2	2	2	2
4	3	3	3	3	3	3	3	3	2	2	2	2	2

The five-cluster solution found by setting $r = 1.8$ and $k = 2$ is

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Austria, Belgium, France, Ireland, Netherlands, Switzerland, U.K., W. Germany	Denmark, Holland, Norway, Sweden	Albania, Belgium, Hungary, Romania, Yugoslavia	Czechs, E. Germany, Poland, USSR	Greece, Italy, Portugal, Spain

This partitioning into five clusters is perhaps more reasonable than that found in Example 14.4.2a. The first five discriminant functions for these five clusters are plotted in Figure 14.21. $\square$

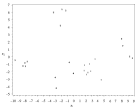


Figure 14.11. First two discriminant functions for the clusters from Example 14.4.2b.

14.2 CHOOSING THE NUMBER OF CLUSTERS

In hierarchical clustering, we can select g clusters from the dendrogram by cutting across the branches at a given level of the distance measure used by one of the tests. This is illustrated in Figure 14.12, which is the dendrogram for the average linkage method (Section 14.3.4) applied to the city-crime data in Table 14.1 (see Figure 14.10). Cutting the dendrogram at a level of 700 yields two clusters. Cutting it at 545 gives three clusters.

We wish to determine the value of g that provides the best fit to the data. One approach is to look for large changes in distances at which clusters are formed. For example, in Figure 14.12, the largest change in levels occurs in going from two clusters to a single cluster. The change in distance between the two-cluster solution and the three-cluster solution is 82 units squared. The difference between the three-cluster solution and the four-cluster solution is 79 units squared, and the change between the four- and five-cluster solutions is only 38 units squared. In this case we would choose two clusters.

A formalization of this procedure was proposed by Mojena (1977): choose the number of groups given by the first stage in the dendrogram at which

$$a_j > \bar{a} + k\hat{\sigma}_a, \quad j = 1, 2, \dots, n, \quad (14.26)$$

where $a_1, a_2, \dots, a_n$ are the distance values for stages with $n, n-1, \dots, 1$ clusters, $\bar{a}$ and $\hat{\sigma}_a$ are the mean and standard deviation of the a 's, and k is a constant. Mojena

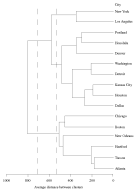


Figure 14.11. Cutting the dendrogram to choose the number of clusters.

(1977) suggested using a value of k in the range 2.75 to 3.5, but Milligan and Cooper (1985) recommended $k = 1.25$, based on a simulation study.

An index that can be used with either hierarchical or partitioning methods is

$$c = \frac{\ln(B_1/g) - D}{\ln(B_2/g) - g}, \quad (14.37)$$

The value of g that maximizes c is chosen. A related approach is to choose the value of g that minimizes

$$d = g^2 \bar{R}_g. \quad (14.36)$$

To compare two cluster solutions with g_1 and g_2 clusters, where $g_2 > g_1$, we can use the test statistic

$$F = \frac{\text{tr}(\bar{R}_{g_1}) - \text{tr}(\bar{R}_{g_2})}{\text{tr}(\bar{R}_{g_2}) \left[\left(\frac{n - g_1}{n - g_2} \right) \left(\frac{g_2}{g_1} \right)^{1/p} - 1 \right]}, \quad (14.37)$$

which has an approximate F -distribution with $p(g_2 - g_1)$ and $p(n - g_2)$ degrees of freedom (Hosmer (1989)). The matrices $\bar{R}_{g_1}$ and $\bar{R}_{g_2}$ are within-cluster sums of squares and products matrices corresponding to g_1 and g_2 . The hypothesis is that the cluster solutions with g_1 and g_2 clusters are equally valid, and rejection implies that the cluster solution with g_2 clusters is better than the solution with g_1 clusters ($g_2 > g_1$). The F -approximation in (14.37) may not be sufficiently accurate to justify the use of p -values.

14.6 CLUSTER VALIDITY

To check the validity of a cluster solution, it may be possible to test the hypothesis that there are no clusters or groups in the population from which the sample at hand was taken. For example, the hypothesis could be that the population represents a single unimodal distribution such as the multivariate normal, or that the observations arise from a uniform distribution. Formal tests of hypotheses of this type concerning cluster validity are reviewed by Gordon (1999, Section 7.1).

A cross-validation approach can also be used to check the validity or stability of a clustering result. The data are randomly divided into two subsets, say A and B , and the cluster analysis is carried out separately on each of A and B . The results should be similar if the clusters are valid. An alternative approach is the following (Gordon (1999, Section 7.1); McQueen (1996)):

1. Use some clustering method to partition subset A into g clusters.
2. Partition subset B into g clusters in two ways:
 - (a) Assign each item in B to the cluster in A that it is closest to by using, for example, the distance to cluster centroids.
 - (b) Use the same clustering method on B that was used on A .
3. Compare the results of (a) and (b) in step 2.

14.7 CLUSTERING VARIABLES

In some cases, it may be of interest to cluster the p variables rather than the n observations. For a similarity measure between each pair of variables, we would usually use the correlation. Since most clustering methods use dissimilarities (such as distances), we need to convert the correlation matrix $\mathbf{R} = (r_{ij})$ to a dissimilarity matrix. This can conveniently be done by replacing each r_{ij} by $1 - |r_{ij}|$ or $1 - r_{ij}^2$. Using the resulting dissimilarity matrix, we can apply a clustering method such as a hierarchical technique to cluster the variables.

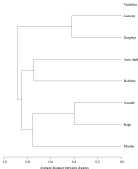


Figure 14.13. Dendrogram for clustering the variables of Table 14.1 using average linkage (see Example 14.7).



Figure 14.24. Dendrogram for clustering the variables of Table 14.1 using Ward's method (see Example 14.7).

Clustering of variables can sometimes be done successfully with factor analysis, which groups the variables corresponding to each factor, see Sections 13.1 and 13.5.

Example 14.7. We illustrate clustering of variables using the city crime data in Table 14.1. We first calculate the correlation matrix $\mathbf{R} = (r_{ij})$ and then transform $\mathbf{R}$ to a dissimilarity matrix $\mathbf{D} = (1 - r_{ij}^2)$. The variables are then clustered using both average linkage and Ward's clustering methods, and the dendrograms are given in Figures 14.23 and 14.24, respectively. Both clustering methods yield the same solution.

Table 14.12. Rotated Factor Loadings for City Crime Data

Variables	Factor 1	Factor 2	Factor 3
Murder	-.363	.734	.142
Rape	.564	.659	.168
Robbery	.177	.553	.726
Assault	.398	.746	.098
Burglary	.764	.221	.181
Larceny	.847	-.044	.244
Auto-theft	.248	.897	.684

We next carry out a factor analysis of the data and compare the resulting groups of variables with the clusters obtained with the average linkage and Ward's methods. The factor loadings are estimated using the principal factor method (Section 13.3.2) with squared multiple correlations as initial communality estimates, and the loadings are then rotated with a varimax rotation (Section 13.3.2b). The rotated factor patterns is given in Table 14.12. The highest loading in each row is **bolded**. The first factor deals with crimes associated with the home. The second factor involves crimes that are violent in nature. The third factor consists of crimes of theft outside the home. Note that the three-cluster solution found by both average linkage and Ward's methods are identical to the grouping of variables in the factor analysis solution, namely, (1) murder, rape, and assault; (2) robbery and auto theft, and (3) burglary and larceny. Since all three methods agree, we have some confidence in the validity of the solution. $\square$

PROBLEMS

14.1 Show that $s^2(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^p \alpha_j^2 - \gamma_j^2$ from (14.2) is equal to (14.5), $s^2(\mathbf{x}, \mathbf{y}) = \alpha_0^2 - \alpha_0^2 \beta^2 + \beta^2 \mathcal{T} - \mathcal{T}^2 + 2\alpha_0 \alpha_1 (1 - r_{10})$, where $\alpha_j^2 = \sum_{i=1}^n \alpha_i^2 \alpha_j^2 \alpha_i^2$, $\mathcal{T} = \sum_{i=1}^n \alpha_i^2 \beta_i^2$, and r_{10} is defined in (14.6).

14.2 (a) Show that $\mathcal{L}_{AB} = n\alpha(\mathcal{T}_A - \mathcal{T}_{AB})^2(\mathcal{T}_B - \mathcal{T}_{AB}) + n\alpha(\mathcal{T}_B - \mathcal{T}_{AB})^2(\mathcal{T}_A - \mathcal{T}_{AB})$ as in (14.17).

(b) Show that (14.18) is equal to (14.19); that is,

$$\begin{aligned} n\alpha(\mathcal{T}_A - \mathcal{T}_{AB})^2(\mathcal{T}_A - \mathcal{T}_{AB}) + n\alpha(\mathcal{T}_B - \mathcal{T}_{AB})^2(\mathcal{T}_B - \mathcal{T}_{AB}) \\ = \frac{n\alpha^2\alpha}{\alpha_A + \alpha_B}(\mathcal{T}_A - \mathcal{T}_B)^2(\mathcal{T}_A - \mathcal{T}_B). \end{aligned}$$

14.3 Using the hints provided in each case, show that the parameter values for (14.20) in Table 14.7 produce appropriate distances for the following cluster methods.

- (a) Complete linkage. Use an approach analogous to that in Section 14.3.4 for the single linkage method.
- (b) Average linkage. Write (14.25) in terms of parameter values for average linkage in Table 14.2. Then use (14.9).
- (c) Centroid method. Show that

$$\begin{aligned} (Q_C - T_{AB})^2 (Q_C - T_{AB}) &= \frac{n_A}{n_A + n_B} (Q_C - T_A)^2 (Q_C - T_B)^2 \\ &\quad + \frac{n_B}{n_A + n_B} (Q_C - T_B)^2 (Q_C - T_A)^2 \\ &\quad - \frac{n_A n_B}{(n_A + n_B)^2} (T_A - T_B)^2 (Q_A - T_B)^2, \end{aligned} \quad (14.40)$$

where $T_{AB} = (n_A T_A + n_B T_B) / (n_A + n_B)$.

- (d) Median method. Use $n_A = n_B$ in (14.12) and (14.40) [see part (c)].
- (e) Ward's method. Show that

$$L_{C(A,B)} = \frac{n_A + n_C}{n_A + n_B + n_C} L_{BC} + \frac{n_B + n_C}{n_A + n_B + n_C} L_{AC} - \frac{n_C}{n_A + n_B + n_C} L_{AB},$$

where L_{AB} is defined in (14.17).

14.4 Show that for all methods in Table 14.2 for which $\gamma = 0$, we have $\text{DOC}(AB) > (n_A + n_B) / \text{FID}(A, B)$ as in (14.26).

14.5 Verify the statement in the last paragraph of Section 14.4.1, namely, that the first criterion in Section 14.4 is not invariant to nonsingular linear transformations $x_{ij} = Ax_{ij} + b_i$, where A is a $p \times p$ nonsingular matrix, and that the other two criteria are invariant to such transformations. Use the following approach.

- (a) Show that $W_1 = A W_1 A'$ and $W_2 = A W_2 A'$.
- (b) Show that minimizing $\|W\|$ is not invariant.
- (c) Show that minimizing $\|W\|$ is invariant.
- (d) Show that maximizing $\|W^{-1}W\|$ is invariant.

14.6 Verify the statement in Section 14.4.2a that in $\mathbf{g}_{AB}$, $i = 1, 2, \dots, p$; $\Sigma_i i = 1, 2, \dots, p$; and n_{ij} , $i = 1, 2, \dots, p$; $j = 1, 2, \dots, p$; the total number of parameters is given by $\frac{1}{2}p(p+1)(p+2)-1$ as in (14.35).

14.7 Use the runs here data of Table 1.6. Carry out the following cluster methods and compare to the principal-component plot in Figure 1.5.

- (a) Find a two-cluster solution using the single linkage method.
- (b) Find a two-cluster solution using the average linkage method and compare to the result in (a). Which seems better?

- (c) Carry out a cluster analysis using the Ward's, complete linkage, centroid, and median methods.
- (d) Use the flexible beta method with $\beta = -.25$, $\beta = -.5$, and $\beta = -.75$.

14.8 Use the hematology data of Table 4.1.

- (a) Carry out a cluster analysis using the centroid method and find the distance between the centroids of the two-cluster solution.
- (b) Carry out a cluster analysis using the average linkage method. How many clusters are indicated in the dendrogram?
- (c) Using the two-cluster solution from part (b), label observations from one cluster as group 1 and the observations from the other cluster as group 2. Calculate and plot the discriminant function, as in Example 9.2. Do the two clusters overlap?

14.9 Use all the variables of the Swiss data of Table 7.1.

- (a) Find the three-cluster solution using four single linkage, complete linkage, average linkage, centroid, median, and Ward's methods. Which observation appears to be an outlier? Which cluster is the same in all six solutions?
- (b) Using the cluster found in part (a) to be common to all solutions as group 1 and the rest of the observations as group 2, calculate and plot the discriminant function, as in Problem 14.8(c). Do the two clusters overlap?

14.10 Use the first 10 observations of the temperature data of Table 7.2. Standardize the variables (columns) before doing the following:

- (a) Carry out a k -means cluster analysis using as initial seeds the five observations that are mutually farthest apart. Plot the first two discriminant functions using the five clusters as groups.
- (b) Repeat part (a) using the first five observations as initial seeds.
- (c) Repeat part (a) using as initial seeds the centroids of the five-cluster solution found using Ward's method. Plot the dendrogram resulting from Ward's method.
- (d) Repeat part (c) using average linkage instead of Ward's method. Compare the results with those in part (c).
- (e) Plot the first and second principal components and the second and third components. Which cluster solutions found in parts (a)–(d) seem to agree most with the principal-component plots?
- (f) Repeat parts (c) and (d) using three initial seeds instead of five. How do the cluster solutions compare?
- (g) Repeat part (c) using three initial seeds instead of five. How does the cluster solution compare to your answer in part (f)?

14.11 Table 14.12 contains air pollution data from 41 U.S. cities (Roth and Roth 1981, p. 519). The variables are as follows:

Table 14.13. Air Pollution Levels in U.S. Cities

City	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Phoenix	18	78.2	212	282	6.0	7.08	36
Little Rock	13	68.8	91	132	8.3	48.32	100
San Francisco	12	86.2	433	738	8.7	26.86	67
Denver	17	81.8	484	818	9.0	12.98	80
San Diego	26	89.3	412	128	9.0	43.27	127
Washington	28	94.8	80	80	9.0	40.22	114
Washington	29	97.3	434	923	9.3	38.89	111
Jacksonville	14	68.4	136	328	8.8	54.47	190
Miami	18	73.3	387	323	9.0	29.80	128
Atlanta	24	61.3	368	487	9.1	48.34	112
Chicago	108	80.8	2344	2889	10.4	34.44	120
Indianapolis	28	82.3	383	786	9.7	38.78	123
San Antonio	17	89.8	384	204	11.2	30.83	938
Richland	8	58.8	123	272	12.3	30.38	82
Louisville	28	95.8	283	283	8.3	43.31	123
New Orleans	9	68.3	384	383	8.4	35.77	113
Baltimore	47	93.8	623	983	9.8	41.31	111
Detroit	39	88.8	1864	1833	10.3	30.96	120
Minneapolis-St. Paul	29	83.8	888	784	10.8	28.84	127
Kansas City	14	84.8	383	883	10.0	27.00	98
St. Louis	28	93.8	773	622	8.3	33.88	105
Omaha	14	91.3	383	347	10.8	30.38	98
Albuquerque	10	58.8	46	344	8.8	7.77	58
Albany	46	47.8	44	136	8.8	33.86	125
Buffalo	10	47.3	383	463	12.4	35.31	100
Cincinnati	23	94.8	463	463	7.3	38.04	120
Cleveland	68	88.2	1087	783	10.8	34.88	138
Columbus	28	91.3	288	340	8.8	27.84	134
Philadelphia	68	94.8	1882	1880	9.8	38.83	113
Pittsburgh	63	98.4	347	328	8.4	38.22	147
Portland	14	58.8	343	278	10.0	42.25	125
Memphis	18	61.8	277	624	8.2	48.30	105
Nashville	18	95.4	273	448	7.4	48.00	110
Dallas	9	68.2	643	844	10.8	33.84	78
Houston	18	88.8	723	1233	10.8	48.38	108
San Jose City	28	81.8	127	278	8.7	13.77	88
Seattle	33	98.3	88	388	10.0	44.88	110
Richmond	28	97.8	187	288	7.8	42.88	113
San Jose	28	91.3	278	333	8.4	38.76	104
Charlotte	33	93.3	23	73	6.3	40.73	148
Milwaukee	18	49.2	888	777	11.8	28.07	128

- y_1 = SO_2 content of air in micrograms per cubic meter
 y_2 = Average annual temperature in $^{\circ}\text{F}$
 y_3 = Number of manufacturing enterprises employing 25 or more workers
 y_4 = Population size (1970) measured in thousands
 y_5 = Average annual wind speed in miles per hour
 y_6 = Average annual precipitation in inches
 y_7 = Average number of days with precipitation per year

Standardize each variable to mean 0 and standard deviation 1. Carry out a cluster analysis using the density estimation method with k equal to 2, 3, 4, 5 and values of r ranging from .1 to .7 by increments of .1 for each value of k . What is the maximum value of k that produces a two-cluster solution?

14.12 Table 14.14 gives the yields of winter wheat in each of the years 1970–1973 at 12 different sites in England (Hartley et al. 1994, p. 37).

- (a) Carry out a cluster analysis using the density estimation method with $k = 2, 3, 4$ and $r = .2, .4, \dots, .7$.
 (b) Plot the first two discriminant functions from the three-cluster solution obtained with $k = 3$ and $r = 1$.
 (c) Plot the first two principal components and compare with the plot in part (b).
 (d) Repeat part (b) using a two-cluster solution obtained with $k = 3$ and $r = 1$. Which two clusters of the three-cluster solution found in part (b) merged into one cluster?

Table 14.14. Yields of Winter Wheat (kg per unit area)

Site	Year			
	1970	1971	1972	1973
Cambridge	46.81	55.46	55.64	52.01
Chichester	46.49	54.05	43.06	41.03
Happes, Adams	44.03	41.03	40.12	40.23
Hindley Hall	52.24	50.19	47.09	50.56
Morley	50.23	43.06	38.05	45.17
Myerscough	54.88	49.71	40.80	50.08
Rowmaston	56.34	47.67	45.48	50.60
Salis-Horne	43.67	57.56	45.48	50.50
Spauldine	43.67	46.87	58.78	47.40
Stanton Farmington	54.44	49.34	54.48	48.04
Tunstallton	54.93	52.09	50.91	50.13
Wye	48.94	48.63	51.69	50.72

Graphical Procedures

In Sections 15.1, 15.2, and 15.3, we consider three graphical techniques: multidimensional scaling, correspondence analysis, and biplots. These methods are designed to reduce dimensionality and portray relationships among observations or variables.

15.1 MULTIDIMENSIONAL SCALING

15.1.1 Introduction

In the dimension reduction technique called *multidimensional scaling*, we begin with the distances d_{ij} between each pair of items. We wish to represent the n items in a low-dimensional coordinate system, in which the distances d_{ij} between items closely match the original distances d_{ij} , that is,

$$d_{ij} \approx d_{ij} \quad \text{for all } i, j.$$

The final distances d_{ij} are usually Euclidean. The original distances d_{ij} may be actual measured distances between observations $\mathbf{y}_i$ and $\mathbf{y}_j$ in p dimensions, such as

$$d_{ij} = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j)^{1/2}. \quad (15.1)$$

On the other hand, the distances d_{ij} may be only a proximity or similarity based on human judgments—for example, the perceived degree of similarity between all pairs of brands of a certain type of appliance (for a discussion of similarities and dissimilarities, see Section 14.2). The goal of multidimensional scaling is a plot that exhibits information about how the items relate to each other or provides some other meaningful interpretation of the data. For example, the aim may be correlation or ranking. If the points lie close to a curve in two dimensions, then the ordering of points along the curve is used to rank the points.

If the observation vectors $\mathbf{y}_i$, $i = 1, 2, \dots, n$, are available and we calculate distances using (15.1) or a similar measure, or if the original $\mathbf{y}_i$'s are not available, but we have actual distances between items, then the process of reduction to a lower-dimensional geometric representation is called *metric multidimensional scaling*. If

the original distances are only similarities based on judgment, the process is called *nonmetric multidimensional scaling*, and the final spatial representation preserves only the rank order among the similarities. We consider metric scaling in Section 15.1.2 and nonmetric scaling in Section 15.1.3. For useful discussions of various aspects of multidimensional scaling, see Davidson (1983), Gordon (1988, Sections 6.2 and 6.5), Kruskal and Wish (1978), Marascuio, Kani, and Fabbri (1979, Chapter 14), Seber (1984, Section 5.7), Young (1987), Johnson (1992, Section 10.3), Shepard, Romney, and Harlow (1972), and Romney, Shepard, and Harlow (1972).

15.1.2 Metric Multidimensional Scaling

In this section, we consider metric multidimensional scaling, which is also known as the *classical solution* and as *principal coordinate analysis*. We begin with an $n \times n$ distance matrix $\mathbf{D} = (d_{ij})$. Our goal is to find n points in k dimensions such that the interpoint distances d_{ij} in the k dimensions are approximately equal to the values of d_{ij} in $\mathbf{D}$. Typically, we use $k = 2$ for plotting purposes, but $k = 1$ or k may also be useful.

The points are found as follows:

1. Construct the $n \times n$ matrix $\mathbf{A} = (a_{ij}) = +\frac{1}{2} d_{ij}^2$, where a_{ij} is the (i,j) -element of $\mathbf{A}$.
2. Construct the $n \times n$ matrix $\mathbf{B} = (b_{ij})$, with elements $b_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}$, where $\bar{a}_i = \sum_{j=1}^n a_{ij}/n$, $\bar{a}_j = \sum_{i=1}^n a_{ij}/n$, $\bar{a} = \sum_{i,j=1}^n a_{ij}/n^2$. The matrix $\mathbf{B}$ can be written as

$$\mathbf{B} = \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \mathbf{A} \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right), \quad (15.2)$$

It can be shown that there exists a q -dimensional real point set $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ with interpoint distances $a_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ ($\mathbf{x}_i = \mathbf{x}_j$) such that $a_{ij} = d_{ij}$ if and only if $\mathbf{B}$ is positive semidefinite of rank q (Belenberg 1958; Young and Householder 1968; Gower 1966; Seber 1984, p. 236).

3. Since $\mathbf{B}$ is a symmetric matrix, we can use the spectral decomposition in (1.10) to write $\mathbf{B}$ in the form

$$\mathbf{B} = \mathbf{B} \mathbf{\Lambda} \mathbf{B}', \quad (15.3)$$

where $\mathbf{B}$ is the matrix of eigenvectors of $\mathbf{B}$ and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{B}$. If $\mathbf{B}$ is positive semidefinite of rank q , there are q positive eigenvalues, and the remaining $n - q$ eigenvalues are zero. If $\mathbf{\Lambda}_1 = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_q)$ contains the positive eigenvalues and $\mathbf{B}_1 = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_q)$ contains the corresponding eigenvectors, then we can express (15.3)

in the form

$$\begin{aligned}\mathbf{B} &= \mathbf{V}_1 \mathbf{A}_1 \mathbf{V}_1' \\ &= \mathbf{V}_1 \mathbf{A}_1^{1/2} \mathbf{A}_1^{1/2} \mathbf{V}_1' \\ &= \mathbf{Z} \mathbf{Z}',\end{aligned}$$

where

$$\mathbf{Z} = \mathbf{V}_1 \mathbf{A}_1^{1/2} = (\sqrt{\lambda_1} \mathbf{v}_1, \sqrt{\lambda_2} \mathbf{v}_2, \dots, \sqrt{\lambda_p} \mathbf{v}_p) = \begin{pmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{p1} \end{pmatrix}. \quad (15.4)$$

4. The rows $z_{11}, z_{21}, \dots, z_{p1}$ of $\mathbf{Z}$ in (15.4) are the points whose interpoint distances $d_{ij} = (\mathbf{z}_i - \mathbf{z}_j)'(\mathbf{z}_i - \mathbf{z}_j)$ match the λ_{ij} 's in the original distance matrix $\mathbf{B}$, as noted following (15.2).
5. Since p in (15.4) will typically be too large to be of practical interest and we would prefer a smaller dimension k for plotting, we can use the first k eigenvalues and corresponding eigenvectors in (15.4) to obtain k points whose interpoint distances d_{ij} are approximately equal to the corresponding λ_{ij} 's.
6. If $\mathbf{B}$ is not positive semidefinite, but its first k eigenvalues are positive and relatively large, then these eigenvalues and associated eigenvectors may be used in (15.4) to construct points that give reasonably good approximations to the λ_{ij} 's.

Note that the method used to obtain $\mathbf{Z}$ from $\mathbf{B}$ closely resembles principal component analysis. Note also that the solution $\mathbf{Z}$ in (15.4) is not unique, since a shift in origin or a rotation will not change the distances d_{ij} . For example, if $\mathbf{C}$ is a $p \times q$ orthogonal matrix producing a rotation [see (2.10)], then

$$\begin{aligned}C\mathbf{z}_i - C\mathbf{z}_j &= C(\mathbf{z}_i - \mathbf{z}_j) = (\mathbf{z}_i - \mathbf{z}_j)'C' C(\mathbf{z}_i - \mathbf{z}_j) \\ &= (\mathbf{z}_i - \mathbf{z}_j)'(\mathbf{z}_i - \mathbf{z}_j) \quad \text{[see (2.10)]}.\end{aligned}$$

Thus the rotated points $C\mathbf{z}_i$ have the same interpoint distances d_{ij} .

Example 15.1.2(a). To illustrate the first four steps of the above algorithm for matrix multi-dimensional scaling, consider the 5×5 distance matrix

$$\mathbf{B} = (\lambda_{ij}) = \begin{pmatrix} 0 & 2\sqrt{2} & 2\sqrt{2} & 3\sqrt{2} & 2\sqrt{2} \\ 2\sqrt{2} & 8 & 4 & 4\sqrt{2} & 4 \\ 2\sqrt{2} & 4 & 0 & 4 & 4\sqrt{2} \\ 3\sqrt{2} & 4\sqrt{2} & 4 & 0 & 4 \\ 2\sqrt{2} & 4 & 4\sqrt{2} & 4 & 0 \end{pmatrix}.$$

The matrix $A = (1 - \frac{1}{2}J_3^T)$ in step 1 is given by

$$A = - \begin{pmatrix} 8 & 4 & 4 & 4 & 4 \\ 4 & 8 & 8 & 16 & 8 \\ 4 & 8 & 8 & 8 & 16 \\ 4 & 16 & 8 & 0 & 8 \\ 4 & 8 & 16 & 8 & 0 \end{pmatrix}.$$

For the moment, we have $d_1 = d_2 = -16/3$, $d_3 = d_4 = -32/3$, $i = 1, \dots, 3$, $d_5 = -32/3$. With $n = 5$, the matrix B in step 2 is given by

$$B = \left(1 - \frac{1}{5}J\right)A\left(1 - \frac{1}{5}J\right) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & -8 & 0 \\ 0 & 0 & 8 & 0 & -8 \\ 0 & -8 & 0 & 8 & 0 \\ 0 & 0 & -8 & 0 & 8 \end{pmatrix}.$$

The rank of B is clearly 2. For step 3, the (nonsymmetric) eigenvalues and corresponding eigenvectors of B are given by $\lambda_1 = 16$, $\lambda_2 = 16$,

$$v_1 = \begin{pmatrix} 0 \\ \frac{1}{2}\sqrt{2} \\ 0 \\ -\frac{1}{2}\sqrt{2} \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 8 \\ 8 \\ \frac{1}{2}\sqrt{2} \\ 8 \\ -\frac{1}{2}\sqrt{2} \end{pmatrix}.$$

Then for step 4 we have, by (15.4),

$$X = \left(\sqrt{\lambda_1}v_1, \sqrt{\lambda_2}v_2\right) = \begin{pmatrix} 8 & 8 \\ 2\sqrt{2} & 2\sqrt{2} \\ 8 & 8 \\ -2\sqrt{2} & -2\sqrt{2} \\ 8 & -2\sqrt{2} \end{pmatrix}.$$

It can be shown (step 4) that the distance matrix for these five points is B . The five points constitute a square with each side of length 4 and a center point at the origin. The five points (rows of X) are plotted in Figure 15.1. $\square$

Example 15.1.264. For another example of matrix multidimensional scaling, we use all the distances between 18 U.S. cities, as given in Table 15.1 (Grossman and Wolk, 1978, pp. 3–6). The points given by matrix multidimensional scaling are plotted in Figure 15.2. Notice that north and south have been reversed; the eigenvectors v_i in (15.4) are normalized but are subject to multiplication by -1 . $\square$

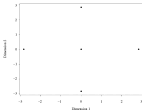


Figure 15.4. Plot of the five points found in Example 15.1.2(a)

15.1.2 Nonmetric Multidimensional Scaling

Suppose the $m \times m = O(n^2)$ dissimilarities d_{ij} cannot be measured as in (15.1) but can be ranked in order

$$d_{i_1 j_1} < d_{i_2 j_2} < \cdots < d_{i_m j_m}. \quad (15.5)$$

Table 15.1. Airline Distances between Ten U.S. Cities

City	1	2	3	4	5	6	7	8	9	10
1	0	567	1212	761	1055	606	748	2105	1881	541
2	567	0	938	540	1145	1188	713	1898	1177	809
3	1212	938	0	479	851	1728	1891	949	1051	1494
4	761	540	479	0	1374	998	1428	1643	1891	1150
5	1055	1145	851	1374	0	2039	2491	347	489	2800
6	606	1188	1728	998	2039	0	1892	2394	2134	923
7	748	713	1891	1428	2491	1892	0	2371	2408	308
8	2105	1898	949	1643	347	2394	2371	0	678	1841
9	1881	1177	1051	1891	489	2134	2408	678	0	2520
10	541	809	1494	1150	2800	923	308	1841	1820	0

Cities: (1) Atlanta, (2) Chicago, (3) Denver, (4) Houston, (5) Los Angeles, (6) Miami, (7) New York, (8) San Francisco, (9) Seattle, (10) Washington, D.C.

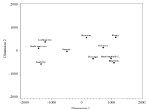


Figure 15.2. Plot of the points found in Example 15.1.2.

where $r_{1,2}$ indicates the pair of items with the smallest dissimilarity and $r_{n,n}$ represents the pair with greatest dissimilarity. In acoustic multidimensional scaling, we seek a low-dimensional representation of the points such that the rankings of the distances

$$d_{1,2} < d_{2,3} < \cdots < d_{n,n} \quad (15.6)$$

match exactly the ordering of dissimilarities in (15.5). Thus, although metric scaling uses the magnitudes of the d_{ij} 's, nonmetric scaling is based only on the rank-order of the d_{ij} 's.

For a given set of points with distances d_{ij} , a plot of d_{ij} versus b_{ij} may not be monotonic; that is, the ordering in (15.6) may not match exactly the ordering in (15.5). A lack of monotonicity of this type is illustrated in Figure 15.3.

In Figure 15.3, the dashed line and open circles show some values of $\hat{d}_{ij}$ that are estimated in such a way that the plot becomes monotonic. Suitable $\hat{d}_{ij}$'s can be estimated by monotonic regression, in which we seek values of $\hat{d}_{ij}$ to minimize the scaled sum of squared differences

$$S^2 = \frac{\sum_{i,j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2} \quad (15.7)$$

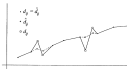


Figure B.3. Plot of distance d versus dissimilarity D illustrating lack of monotonicity. The dashed line represents best fit by monotonic regression.

subjected to the constraint

$$\hat{d}_{r_1r_2} \leq \hat{d}_{r_2r_3} \leq \cdots \leq \hat{d}_{r_{k-1}r_k},$$

where $r_1, r_2, \dots, r_{k-1}, r_k$ are defined as in (23.5) and (23.6) (Kruskal 1964a, 1964b). The minimum value of D^2 for a given dimension, k , is called the STRESS. Note that the $\hat{d}_{ij}$'s are not distances. They are merely numbers used as a reference to measure the monotonicity of the d_{ij} 's. The $\hat{d}_{ij}$'s are sometimes called dissimilarities.

The minimum value of the STRESS over all possible configurations of points can be found using the following algorithm.

1. Rank the $m = n(n-1)/2$ distances or dissimilarities d_{ij} as in (23.5).
2. Choose a value of k and an initial configuration of points in k dimensions. The initial configuration could be a points chosen at random from a uniform or normal distribution, a evenly spaced points in k -dimensional space, or the matrix solution obtained by treating the ordinal measurements as continuous and using the algorithm in Section 15.1.2.
3. For the initial configuration of points, find the interitem distances d_{ij} . Find the corresponding $\hat{d}_{ij}$'s by monotonic regression as defined above using (15.7).
4. Choose a new configuration of points whose distances d_{ij} minimize D^2 in (15.7) with respect to the $\hat{d}_{ij}$'s found in step 3. One approach is to use an iterative gradient technique such as the method of steepest descent or the Newton-Raphson method.
5. Using monotonic regression, find new $\hat{d}_{ij}$'s for the d_{ij} 's found in step 4. This gives a new value of STRESS.
6. Repeat steps 4 and 5 until STRESS converges to a minimum over all possible k -dimensional configurations of points.

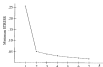


Figure 15.4. Ideal plot of minimum STRESS versus k .

- Using the preceding six steps, calculate the minimum STRESS for values of k starting at $k = 1$ and plot them. As k increases, the curve will decrease, with occasional exceptions due to round off or numerical anomalies in the search procedure for minimum STRESS. We look for a discernible bend in the plot, following which the curve is low and relatively flat. An ideal plot is shown in Figure 15.4. The curve levels off after $k = 2$, which is convenient for plotting the resulting points in 2 dimensions.

There is a possibility that the minimum value of STRESS found by the above seven steps for a given value of k may be a local minimum rather than the global minimum. Such an anomaly may show up in the plot of minimum STRESS versus k . This possibility of a local minimum can be checked by repeating the procedure, starting with a different initial configuration.

As was the case with metric scaling, the final configuration of points from a non-metric scaling is invariant to a rotation of axes.

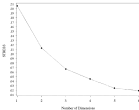
Example 15.13. The voting records for 15 congressmen from New Jersey on 19 environmental bills are given in Table 15.2 in the form of a dissimilarity matrix (Hand et al., 1994, p. 233). The congressmen are identified by party: R_k for Republican k , D_k for Democrat k , etc. Each entry shows how often the indicated congressmen voted differently from each of the other 14.

Using an initial configuration of points from a multivariate normal distribution with mean vector $\mu = 0$ and $\Sigma = \mathbf{I}$, we find six "optimal" configurations of points for each of $k = 1, 2, \dots, 5$. A plot of the STRESS is given in Figure 15.5.

From the plot of STRESS vs. number of dimensions, we see that either two or three dimensions will be sufficient. For plotting purposes, we choose two dimensions, which has a STRESS value of .113. The plot of the first two dimensions is given in Figure 15.6. It is apparent that the plot separates the Republicans from the Democrats except for Republican 6, who voted much the same as the Democrats.

Table 15.1. Descriptive Means for Voting Records of 33 Congressmen

	R_1	R_2	D_1	D_2	A_1	A_2	R_3	R_4	D_3	A_3	A_4	D_4	R_5	R_6	A_5	A_6
R_1	8	8	15	15	10	7	7	15	16	14	15	16	7	11	15	15
R_2	8	6	17	12	13	15	12	16	17	15	16	17	13	12	16	16
D_1	18	21	8	8	16	12	18	8	8	8	8	8	11	18	7	7
D_2	18	21	8	8	14	12	13	18	8	8	8	8	18	18	7	7
R_3	18	14	16	14	6	8	8	12	16	12	12	12	16	11	11	11
R_4	8	14	12	12	8	8	7	12	11	10	9	18	8	8	10	10
R_5	7	12	15	15	8	7	8	17	16	15	14	15	16	11	13	13
D_3	15	16	5	18	14	12	17	8	8	5	5	5	12	7	6	6
D_4	16	21	5	8	14	11	16	4	8	5	5	1	13	7	5	5
A_1	14	15	6	8	11	18	17	5	5	0	1	2	11	4	6	6
A_2	18	16	8	8	12	9	14	8	2	1	0	1	12	8	8	8
R_6	16	21	4	8	12	18	13	3	1	2	1	8	12	8	4	4
R_7	7	14	11	18	10	8	16	12	18	11	12	12	6	9	13	13
R_8	11	12	16	18	11	6	11	7	7	4	5	6	8	8	9	9
D_5	14	16	7	7	11	18	12	8	5	6	5	8	12	9	0	0

Figure 15.1. Plot of STRESS for each value of k for the voting data in Table 15.1.

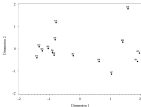


Figure 12.4. Plot of points found using initial points from a multivariate normal distribution.

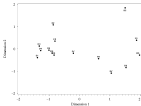


Figure 12.5. Plot of points found using initial points from a uniform distribution.

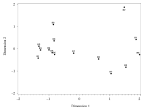


Figure 15.6. Plot of points found using initial points from a metric solution.

We now use a different initial configuration of points drawn from a uniform distribution. The resulting plot is given in Figure 15.7. The results are very similar, with the exception of R_1 and R_9 .

We next use a third initial configuration of points resulting from the metric solution, as described in Section 15.1.3. The resulting plot is given in Figure 15.8. All three plots are very similar, indicating a good fit. $\square$

15.2 CORRESPONDENCE ANALYSIS

15.2.1 Introduction

Correspondence analysis is a graphical technique for representing the information in a two-way contingency table, which contains the counts (frequencies) of items for a cross-classification of two categorical variables. With correspondence analysis, we construct a plot that shows the interaction of the two categorical variables along with the relationship of the rows to each other and of the columns to each other. In Sections 15.2.2–15.2.4, we consider correspondence analysis for ordinary two-way contingency tables. In Section 15.2.5 we consider multiple correspondence analysis for three-way and higher-order contingency tables. Useful treatments of correspondence analysis have been given by Greenacre (1984), Johnson (1992, Section 9.4), Khuriyar and Walk (1999, Chapter 7), Gower and Hand (1996, Chapters 4 and 9), and Benedetti (1991).

To test for significance of association of the two categorical variables in a contingency table, we could use a chi-square test or a log-linear model, both of which represent an asymptotic approach. Since correspondence analysis is associated with the chi-square approach, we will review it in Section 15.2.3. If a contingency table has some cell frequencies that are small or zero, the chi-square approximation is not very satisfactory. In this case, some categories can be combined to increase the cell frequencies. Correspondence analysis may be useful in identifying the categories that are similar, which we may thereby wish to combine.

In correspondence analysis, we plot a point for each row and a point for each column of the contingency table. These points are, in effect, projections of the rows and columns of the contingency table onto a two-dimensional Euclidean space. The goal is to preserve as far as possible the relationship of the rows (or columns) to each other in a two-dimensional space. If two row points are close together, the profiles of the two rows (across the columns) are similar. Likewise, two column points that are close together represent columns with similar profiles across the rows (see Section 15.2.2 for a definition of profiles). If a row point is close to a column point, this combination of categories of the two variables occurs more frequently than would occur by chance if the two variables were independent. Another output of a correspondence analysis is the inertia, or amount of information in each of the two dimensions in the plot (see Section 15.2.4).

15.2.2 Rows and Column Profiles

A contingency table with a rows and b columns is represented in Table 15.3. The entries n_{ij} are the counts or frequencies for every two-way combination of row and column (every cell). The marginal totals are shown using the familiar dot notation: $n_{i\cdot} = \sum_{j=1}^b n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^a n_{ij}$. The overall total frequency is denoted by n instead of $n_{\cdot\cdot}$, for simplicity: $n = \sum_{i,j} n_{ij}$.

The frequency n_{ij} in a contingency table can be converted to relative frequencies p_{ij} by dividing by n : $p_{ij} = n_{ij}/n$. The matrix of relative frequencies is called the *correspondence matrix* and is denoted by $\mathbf{P}$:

Table 15.3. Contingency Table with a Rows and b Columns

		Column				Row Total
		1	2	...	b	
Rows	1	n_{11}	n_{12}	...	n_{1b}	$n_{1\cdot}$
	2	n_{21}	n_{22}	...	n_{2b}	$n_{2\cdot}$
	...	...	...	...	...	...
	...	...	...	...	...	...
	a	n_{a1}	n_{a2}	...	n_{ab}	$n_{a\cdot}$
Column Total		$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot b}$	n

Table 15.4. Correspondence Matrix of Relative Frequencies

		Column				Row Total
		1	2	...	b	
Row	1	p_{11}	p_{12}	...	p_{1b}	$p_{1.}$
	2	p_{21}	p_{22}	...	p_{2b}	$p_{2.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	a	p_{a1}	p_{a2}	...	p_{ab}	$p_{a.}$
Column Total		$p_{.1}$	$p_{.2}$	...	$p_{.b}$	1

$$\mathbf{P} = (p_{ij}) = (n_{ij}/N) \quad (15.8)$$

In Table 15.4 we show the contingency table in Table 15.3 converted to a correspondence matrix.

The last column of Table 15.4 contains the row sums $p_{i.} = \sum_{j=1}^b p_{ij}$. This column vector is denoted by $\mathbf{r}$ and can be obtained as

$$\mathbf{r} = \mathbf{P}\mathbf{j} = (p_{1.}, p_{2.}, \dots, p_{a.})' = (n_{1.}/n, n_{2.}/n, \dots, n_{a.}/n)', \quad (15.9)$$

where $\mathbf{j}$ is an $a \times 1$ vector of 1's. Similarly, the last row of Table 15.4 contains the column sums $p_{.j} = \sum_{i=1}^a p_{ij}$. This row vector is denoted by $\mathbf{c}'$ and can be obtained as

$$\mathbf{c}' = \mathbf{j}'\mathbf{P} = (p_{.1}, p_{.2}, \dots, p_{.b}) = (n_{.1}/n, n_{.2}/n, \dots, n_{.b}/n), \quad (15.10)$$

where $\mathbf{j}'$ is a $1 \times b$ vector of 1's. The elements of the vectors $\mathbf{r}$ and $\mathbf{c}$ are sometimes referred to as row and column masses. The correspondence matrix and marginal totals in Table 15.4 can be expressed as

$$\left(\begin{array}{c} \mathbf{P} \\ \mathbf{c}' \end{array} \right) = \left(\begin{array}{cccc|c} p_{11} & p_{12} & \cdots & p_{1b} & p_{1.} \\ p_{21} & p_{22} & \cdots & p_{2b} & p_{2.} \\ \vdots & \vdots & & \vdots & \vdots \\ p_{a1} & p_{a2} & \cdots & p_{ab} & p_{a.} \\ \hline p_{.1} & p_{.2} & \cdots & p_{.b} & 1 \end{array} \right). \quad (15.11)$$

We now convert each row and column of $\mathbf{P}$ to a profile. The i th row profile $\mathbf{r}'_i$, $i = 1, 2, \dots, a$, is obtained by dividing the i th row of either Table 15.3 or 15.4 by its marginal total:

$$\mathbf{r}'_i = \left(\frac{p_{i1}}{p_{i.}}, \frac{p_{i2}}{p_{i.}}, \dots, \frac{p_{ib}}{p_{i.}} \right) = \left(\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \dots, \frac{n_{ib}}{n_{i.}} \right). \quad (15.12)$$

The elements in each r_i^j are relative frequencies, and therefore they sum to 1:

$$r_i^j = \sum_{j=1}^J \frac{n_{ij}}{n_{i.}} = \frac{n_{i.}}{n_{i.}} = 1. \quad (15.13)$$

By defining

$$R_i = \text{diag}(r_i) = \begin{pmatrix} r_{i1} & 0 & \cdots & 0 \\ 0 & r_{i2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & r_{iJ} \end{pmatrix} \quad (15.14)$$

and using (2.58), the matrix R of row profiles can be expressed as

$$R = R_c^{-1} P = \begin{pmatrix} r_1^j \\ r_2^j \\ \vdots \\ r_i^j \end{pmatrix} = \begin{pmatrix} \frac{p_{11}}{p_{1.}} & \frac{p_{12}}{p_{1.}} & \cdots & \frac{p_{1J}}{p_{1.}} \\ \frac{p_{21}}{p_{2.}} & \frac{p_{22}}{p_{2.}} & \cdots & \frac{p_{2J}}{p_{2.}} \\ \vdots & \vdots & & \vdots \\ \frac{p_{i1}}{p_{i.}} & \frac{p_{i2}}{p_{i.}} & \cdots & \frac{p_{iJ}}{p_{i.}} \end{pmatrix}. \quad (15.15)$$

Similarly, the j th column profile r_j , $j = 1, 2, \dots, J$, is defined by dividing the j th column of either Table 15.3 or Table 15.4 by its marginal total:

$$r_j = \left(\frac{p_{1j}}{p_{.j}}, \frac{p_{2j}}{p_{.j}}, \dots, \frac{p_{ij}}{p_{.j}} \right)' = \left(\frac{n_{1j}}{n_{.j}}, \frac{n_{2j}}{n_{.j}}, \dots, \frac{n_{ij}}{n_{.j}} \right)'. \quad (15.16)$$

The elements in each r_j are relative frequencies, and therefore they sum to 1:

$$r_j = \sum_{i=1}^I \frac{n_{ij}}{n_{.j}} = \frac{n_{.j}}{n_{.j}} = 1. \quad (15.17)$$

By defining

$$R_j = \text{diag}(r_j) = \begin{pmatrix} r_{j1} & 0 & \cdots & 0 \\ 0 & r_{j2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & r_{jI} \end{pmatrix} \quad (15.18)$$

and using (2.58), the matrix C of column profiles can be expressed as

$$\mathbf{C} = \mathbf{P}\mathbf{Q}_c^{-1} = (v_1, v_2, \dots, v_n) = \begin{pmatrix} \frac{P_{11}}{P_{11}} & \frac{P_{12}}{P_{11}} & \dots & \frac{P_{1n}}{P_{11}} \\ \frac{P_{21}}{P_{11}} & \frac{P_{22}}{P_{11}} & \dots & \frac{P_{2n}}{P_{11}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{P_{n1}}{P_{11}} & \frac{P_{n2}}{P_{11}} & \dots & \frac{P_{nn}}{P_{11}} \end{pmatrix}. \quad (15.18)$$

The vector v is defined in (15.18) as the column vector of row sums of $\mathbf{P}$. It can also be expressed as a weighted average of the column profiles:

$$v = \sum_{j=1}^n p_{.j} v_j. \quad (15.20)$$

Similarly, v' in (15.19) is the row vector of column sums of $\mathbf{P}$ and can be expressed as a weighted average of the row profiles:

$$v' = \sum_{i=1}^n p_i v'_i. \quad (15.21)$$

Note that $\sum_{j=1}^n p_{.j} = \sum_{i=1}^n p_i = 1/n$

$$j\mathbf{v} = v\mathbf{j} = \mathbf{1}, \quad (15.22)$$

where the first $\mathbf{j}$ is $n \times 1$ and the second is $1 \times n$. Therefore, the $p_{.j}$'s and p_i 's serve as appropriate weights in the weighted averages (15.20) and (15.21).

Example 15.2.2. In Table 15.3 (Hsieh et al. 1994, p. 12) we have the number of piston ring failures in each of three laps in four compressors found in the same building (all four compressors are oriented in the same direction). We obtain the correspondence matrix in Table 15.4 by dividing each element of Table 15.3 by $n = \sum_{ij} x_{ij} = 180$.

Table 15.3. Piston Ring Failures

Compressor	Lap			Row Total
	North	Center	South	
1	17	17	12	46
2	11	9	13	33
3	11	6	19	36
4	14	7	28	49
Column Total	53	49	72	180

Table 13.6. Correspondence Matrix Obtained from Table 13.5

Comparison	Log			Row Total
	North	Center	South	
1	.402	.402	.081	.273
2	.066	.064	.088	.099
3	.066	.040	.114	.126
4	.065	.041	.109	.129
Column Total	.419	.447	.444	1.000

The vectors of row and column sums (marginal totals) in Table 13.6 are given by (13.8) and (13.9) as

$$\mathbf{r} = \begin{pmatrix} .273 \\ .099 \\ .126 \\ .129 \end{pmatrix},$$

$$\mathbf{c} = (.419, .447, .444).$$

The matrix of row profiles is given by (13.10) as

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P} = \begin{pmatrix} .370 & .370 & .260 \\ .333 & .279 & .388 \\ .290 & .311 & .399 \\ .285 & .345 & .370 \end{pmatrix}.$$

The matrix of column profiles is given by (13.11) as

$$\mathbf{C} = \mathbf{P}\mathbf{D}_c^{-1} = \begin{pmatrix} .311 & .415 & .187 \\ .268 & .229 & .180 \\ .298 & .195 & .284 \\ .264 & .171 & .348 \end{pmatrix}. \quad \square$$

13.1.3 Testing Independence

In Section 13.2.1, we noted that the data in a contingency table can be used to check for associations of two categorical variables. If the two variables are denoted by x and y , then the assumption of independence can be expressed in terms of probability as

$$P(x_i y_j) = P(x_i)P(y_j), \quad i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J, \quad (13.22)$$

where x_i and y_j correspond to the i th row and j th column of the contingency table. Using the notation in Table 13.4, we can estimate (13.22) as

$$\hat{P}_{ij} = \hat{p}_{i.}\hat{p}_{.j}, \quad i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J. \quad (13.24)$$

The usual chi-square statistic for testing independence of x and y (comparing p_{ij} with $p_{i\cdot} p_{\cdot j}$ for all i, j) is given by

$$\chi^2 = n \sum_{i=1}^I \sum_{j=1}^J \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}, \quad (15.25)$$

which is approximately (asymptotically) distributed as a chi-square random variable with $(I-1)(J-1)$ degrees of freedom. The statistic in (15.25) can also be written in terms of the frequencies n_{ij} rather than the relative frequencies p_{ij} :

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}}. \quad (15.26)$$

Two other alternative forms of (15.25) are

$$\chi^2 = \sum_{i=1}^I n p_{i\cdot} \sum_{j=1}^J \left[\left(\frac{p_{ij}}{p_{i\cdot}} - p_{\cdot j} \right)^2 / p_{\cdot j} \right], \quad (15.27)$$

$$\chi^2 = \sum_{j=1}^J n p_{\cdot j} \sum_{i=1}^I \left[\left(\frac{p_{ij}}{p_{\cdot j}} - p_{i\cdot} \right)^2 / p_{i\cdot} \right]. \quad (15.28)$$

In vector and matrix form, (15.27) and (15.28) can be written as

$$\chi^2 = \sum_{i=1}^I n p_{i\cdot} (r_i - r)^T \mathbf{B}_1^{-1} (r_i - r), \quad (15.29)$$

$$\chi^2 = \sum_{j=1}^J n p_{\cdot j} (s_j - s)^T \mathbf{B}_2^{-1} (s_j - s), \quad (15.30)$$

where $r_i, s_j, n p_{i\cdot}, n p_{\cdot j}$, $\mathbf{B}_1$, and $\mathbf{B}_2$ are defined in (15.9), (15.10), (15.12), (15.16), (15.14), and (15.18), respectively. Thus, in (15.29) we compare r_i to r for each i , and in (15.30) we compare s_j to s for each j . Either of these is equivalent to testing independence by comparing p_{ij} to $p_{i\cdot} p_{\cdot j}$ for all i, j , since all the definitions of χ^2 in (15.25)–(15.30) are equal. Thus, the following three statements of independence are equivalent (for simplicity, we express the three statements in terms of sample quantities rather than their theoretical counterparts):

1. $p_{ij} = p_{i\cdot} p_{\cdot j}$ for all i, j (or $\mathbf{P} = \mathbf{P}^I$).
2. All rows r_i^T of $\mathbf{R}$ in (15.15) are equal (and equal to their weighted average, r^T).
3. All columns s_j of $\mathbf{C}$ in (15.16) are equal (and equal to their weighted average, s).

Thus, if the variables x and y were independent, we would expect the rows of the contingency table to have similar profiles, or equivalently, the columns to have similar profiles. We can compare the row profiles to each other by comparing each row profile $\mathbf{r}_i^*$ to the weighted average $\mathbf{r}^*$ of the row profiles defined in (15.23). This comparison is made in the χ^2 statistic (15.29). Similarly, we compare column profiles in (15.30).

The chi-square statistic in (15.29) can be expressed in vector and matrix terms as

$$\chi^2 = n(\mathbf{D}_y^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}')\mathbf{D}_y^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}')) \quad (15.31)$$

$$= n \sum_{i=1}^b \lambda_i^2 \quad \text{[eq. (2.187)].} \quad (15.32)$$

where $\lambda_1^2, \lambda_2^2, \dots, \lambda_b^2$ are the nonzero eigenvalues of $\mathbf{D}_y^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}')\mathbf{D}_y^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}')$ and

$$k = \text{rank}[\mathbf{D}_y^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}')\mathbf{D}_y^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}')] = \text{rank}(\mathbf{P} - \mathbf{r}\mathbf{r}'). \quad (15.33)$$

The rank of $\mathbf{P} - \mathbf{r}\mathbf{r}'$ is obviously $k = \min\{a - 1, b - 1\}$. It is clear that the rank is less than $\min\{a, b\}$ since

$$(\mathbf{P} - \mathbf{r}\mathbf{r}')\mathbf{j} = (\mathbf{P}\mathbf{j} - \mathbf{r}\mathbf{j}) = \mathbf{0} - \mathbf{0} = \mathbf{0} \quad (15.34)$$

[see (15.9) and (15.72)].

Example 15.13. In order to test independence of the rows (compressors) and columns (logs) of Table 15.5 in Example 15.1.1, we perform a chi-square test. Using (15.29) or (15.30), we obtain $\chi^2 = 11.722$, with 4 degrees of freedom, for which the p -value is .0085. There is some evidence of lack of independence between log and compressor. $\square$

15.1.6 Coordinates for Plotting Row and Column Profiles

We now obtain coordinates of the row points and column points for the best two-dimensional representation of the data in a contingency table. As we will see, the matrix for the row points and column points is the same, and the two sets of points can therefore be superimposed on the same plot.

In multidimensional scaling in Section 15.3, we transformed the distance matrix and then factored it by a spectral decomposition to obtain coordinates for plotting. In correspondence analysis, the matrix $\mathbf{P} - \mathbf{r}\mathbf{r}'$ is not symmetric, and we therefore use a singular value decomposition to obtain coordinates.

We first scale $\mathbf{P} - \mathbf{r}\mathbf{r}'$ to obtain

$$\mathbf{Z} = \mathbf{D}_y^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{r}')\mathbf{D}_y^{-1/2}, \quad (15.35)$$

whose elements are

$$z_{ij} = \frac{P_{ij} - P_i P_j}{\sqrt{P_i P_j}}, \quad (15.36)$$

as in (15.35). The $n \times k$ matrix $\mathbf{Z}$ has rank $k = \min(n-1, p-1)$, the rank of $\mathbf{P} - \mathbf{p}\mathbf{p}'$. We factor $\mathbf{Z}$ using the singular value decomposition (2.117):

$$\mathbf{Z} = \mathbf{U}\mathbf{A}\mathbf{V}', \quad (15.37)$$

The columns of the $n \times k$ matrix $\mathbf{U}$ are (normalized) eigenvectors of $\mathbf{Z}\mathbf{Z}'$; the columns of the $k \times k$ matrix $\mathbf{V}$ are (normalized) eigenvectors of $\mathbf{Z}'\mathbf{Z}$; and $\mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$, where $\lambda_1^2, \lambda_2^2, \dots, \lambda_k^2$ are the nonzero eigenvalues of $\mathbf{Z}\mathbf{Z}'$ and of $\mathbf{Z}'\mathbf{Z}$. The eigenvectors in $\mathbf{U}$ and $\mathbf{V}$ correspond to the eigenvalues $\lambda_1^2, \lambda_2^2, \dots, \lambda_k^2$. Since the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$. The values $\lambda_1, \lambda_2, \dots, \lambda_k$ in $\mathbf{A}$ are called the singular values of $\mathbf{Z}$. Note that, by (15.35),

$$\begin{aligned} \mathbf{Z}\mathbf{Z}' &= \mathbf{U}\mathbf{A}^{-1/2}(\mathbf{P} - \mathbf{p}\mathbf{p}')\mathbf{U}_1^{-1/2}\mathbf{U}_2^{-1/2}(\mathbf{P} - \mathbf{p}\mathbf{p}')'\mathbf{U}_2^{-1/2} \\ &= \mathbf{U}\mathbf{A}^{-1/2}(\mathbf{P} - \mathbf{p}\mathbf{p}')\mathbf{U}_1^{-1}(\mathbf{P} - \mathbf{p}\mathbf{p}')'\mathbf{U}_2^{-1/2}. \end{aligned} \quad (15.38)$$

The nonzero eigenvalues of $\mathbf{Z}\mathbf{Z}'$ in (15.38) are the same as those of

$$\mathbf{U}_2^{-1/2}\mathbf{U}_2^{-1/2}(\mathbf{P} - \mathbf{p}\mathbf{p}')\mathbf{U}_1^{-1}(\mathbf{P} - \mathbf{p}\mathbf{p}')' \quad (15.39)$$

(see Section 2.1.3). The matrix expression in (15.39) is the same as that in (23.34). We have therefore denoted the eigenvalues as $\lambda_1^2, \lambda_2^2, \dots, \lambda_k^2$ as in (15.32).

We can obtain a decomposition of $\mathbf{P} - \mathbf{p}\mathbf{p}'$ by equating the right-hand sides of (15.39) and (15.37) and solving for $\mathbf{P} - \mathbf{p}\mathbf{p}'$:

$$\begin{aligned} \mathbf{U}_1^{-1/2}(\mathbf{P} - \mathbf{p}\mathbf{p}')\mathbf{U}_2^{-1/2} &= \mathbf{U}\mathbf{A}\mathbf{V}', \\ \mathbf{P} - \mathbf{p}\mathbf{p}' &= \mathbf{U}_1^{1/2}\mathbf{U}\mathbf{A}\mathbf{V}'\mathbf{U}_2^{1/2} \\ &= \mathbf{A}\mathbf{V}\mathbf{B}' = \sum_{i=1}^k \lambda_i \mathbf{a}_i \mathbf{b}_i', \end{aligned} \quad (15.40)$$

where $\mathbf{A} = \mathbf{U}_1^{1/2}\mathbf{U}$, $\mathbf{B} = \mathbf{U}_2^{1/2}\mathbf{V}$, $\mathbf{a}_i$ and $\mathbf{b}_i$ are columns of $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$.

Since $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$, $\mathbf{A}$ and $\mathbf{B}$ in (15.40) are scaled so that $\mathbf{A}'\mathbf{B}'^{-1}\mathbf{A} = \mathbf{B}\mathbf{U}_2^{-1}\mathbf{B} = \mathbf{I}$. With this scaling, the decomposition in (15.40) is often called the *generalized singular value decomposition*.

In (15.40) the rows of $\mathbf{P} - \mathbf{r}\mathbf{c}'$ are expressed as linear combinations of the rows of $\mathbf{B}'$, which are the columns of $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k)$. The coordinates (coefficients) for the rows of $\mathbf{P} - \mathbf{r}\mathbf{c}'$ are found in the i th row of $\mathbf{A}\mathbf{A}'$. In like manner, the coordinates for the columns of $\mathbf{P} - \mathbf{r}\mathbf{c}'$ are given by the columns of $\mathbf{A}\mathbf{B}'$, since the columns of $\mathbf{A}\mathbf{B}'$ provide coefficients for the columns of $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L)$ in (15.40).

To find coordinates for the row deviations $e'_i = e'_i$ in $\mathbf{B} - \mathbf{j}\mathbf{r}'$ and the column deviations $e_j = e_j$ in $\mathbf{C} - \mathbf{r}\mathbf{j}'$, we express the two matrices as functions of $\mathbf{P} - \mathbf{r}\mathbf{c}'$ (see Problem 15.30):

$$\mathbf{B} - \mathbf{j}\mathbf{r}' = \mathbf{B}_r^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}'), \quad (15.41)$$

$$\mathbf{C} - \mathbf{r}\mathbf{j}' = \mathbf{B}_c^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}'). \quad (15.42)$$

Thus the coordinates for the row deviations in $\mathbf{B} - \mathbf{j}\mathbf{r}'$ with respect to the axes provided by $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k$ are given by the columns of

$$\mathbf{X} = \mathbf{B}_r^{-1}\mathbf{A}\mathbf{A}'. \quad (15.43)$$

Similarly, the coordinates for the column deviations in $\mathbf{C} - \mathbf{r}\mathbf{j}'$ with respect to the axes $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L$ are given by the columns of

$$\mathbf{Y} = \mathbf{B}_c^{-1}\mathbf{B}\mathbf{A}. \quad (15.44)$$

Therefore, to plot the coordinates for the row profile deviations $e'_i = e'_i$, $i = 1, 2, \dots, L$, in two dimensions, we plot the rows of the first two columns of $\mathbf{X}$:

$$\mathbf{X}_1 = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{L1} & x_{L2} \end{pmatrix}.$$

Similarly, to plot the coordinates for the column profile deviations $e_j = e_j$, $j = 1, 2, \dots, k$, in two dimensions, we plot the rows of the first two columns of $\mathbf{Y}$:

$$\mathbf{Y}_1 = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{k1} & y_{k2} \end{pmatrix}.$$

Both plots can be superimposed on the same graph because $\mathbf{A}$ and $\mathbf{B}$ in (15.40) share the same singular values $\lambda_1, \lambda_2, \dots, \lambda_d$ in $\mathbf{A}$. Distances between row points and distances between column points are meaningful. For example, the distance between two row points is related to the chi-square metric implicit in (15.38). The

chi-square distance between two row profiles r_i and r_j is given by

$$d_{ij}^2 = (r_i - r_j)' D_r^{-1} (r_i - r_j).$$

If two row points (or two column points) are close, the two rows (or two columns) could be combined into a single category if necessary to improve the chi-square approximation.

The distance between a row point and a column point is not meaningful, but the proximity of a row point and a column-point has meaning as noted in Section 15.2.1, namely, that these two categories of the two variables occur more frequently than would be expected to happen by chance if the two variables were independent.

The weighted average (weighted by p_i) of the chi-square distances $(r_i - c)' D_r^{-1} (r_i - c)$ between the row profiles r_i and their mean c [see (15.21)] is called the total inertia. By (15.28) this can be expressed as x^2/n :

$$\text{Total inertia} = \frac{x^2}{n} = \sum_{i=1}^p p_i (r_i - c)' D_r^{-1} (r_i - c). \quad (15.45)$$

As noted following (15.21), $\sum_i p_i = 1$, and therefore the p_i 's serve as appropriate weights.

By (15.32), we can write (15.45) as

$$\frac{x^2}{n} = \sum_{i=1}^p \lambda_i^2 z_i^2. \quad (15.46)$$

Therefore, the contribution of each of the first two dimensions (axes) of our plot to the total inertia is $(\lambda_1^2 / \sum_{i=1}^p \lambda_i^2)$ and $(\lambda_2^2 / \sum_{i=1}^p \lambda_i^2)$. The combined contribution of the two dimensions is

$$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{i=1}^p \lambda_i^2}. \quad (15.47)$$

If (15.47) is large, then the points in the plane of the first two dimensions account for nearly all the variation in the data, including the associations. The total inertia in (15.45) and (15.46) can also be described in terms of the columns by using (15.30):

$$\text{Total inertia} = \frac{x^2}{n} = \sum_{j=1}^q p_{.j} (C_j - c)' D_c^{-1} (C_j - c) = \sum_{j=1}^q \lambda_j^2 z_j^2. \quad (15.48)$$

Since the inertia associated with the axes for columns is the same as that for rows, the row and column points can be plotted on the same axes.

Some computer programs use a singular value decomposition of $\mathbf{P}$ rather than of $\mathbf{P} = \mathbf{r}\mathbf{r}'$. The results are the same if the first singular value (which is 1) is discarded along with the first column of $\mathbf{A}$ (which is $\mathbf{r}$) and the first column of $\mathbf{B}$ (which is $\mathbf{c}$).

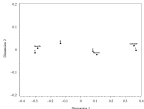


Figure 15.8. Row points $r_1, 2, 3, 4$ and column points c_1, c_2, c_3, c_4 .

Example 15.1.4. We continue the analysis of the piston ring data of Table 15.5. A correspondence analysis is performed and a plot of the row and column points is given in Figure 15.8. Row points do not lie near other row points and column points do not lie near column points. However, component 1 seems to be closely associated with the center leg, component 2 with the north leg, and component 4 with the south leg. These associations illustrate the lack of independence between component and leg position.

Singular values and inertia are given in Table 15.7. Most of the variation is due to the first dimension, and the first two dimensions explain all the variation because $\text{rank}(\mathbf{X}) = \min(4 - 1, 4 - 1) = \min(3 - 1, 3 - 1) = 2$, where $\mathbf{X}$ is defined in (15.55).

□

Table 15.7. Singular Values (λ_i), Inertia (λ_i^2), Chi-Square ($\lambda_i^2 N$), and Percent ($\lambda_i^2 / \sum_{j=1}^4 \lambda_j^2$) for the Data in Table 15.5

Singular Value	Inertia	Chi-Square	Percent
.50826	.00257	11.6819	59.65
.01500	.00024	.3484	.36
Total	.00280	11.7223	100

15.1.5 Multiple Correspondence Analysis

Correspondence analysis of a two-way contingency table can be extended to a three-way or higher-order multi-way table. By the method of multiple correspondence analysis, we obtain a two-dimensional graphical display of the information in the multi-way contingency table. The method involves a correspondence analysis of an indicator matrix $\mathbf{G}$, in which there is a row for each item. Thus the number of rows of $\mathbf{G}$ is the total number of items in the sample. The number of columns of $\mathbf{G}$ is the total number of categories in all variables. The elements of $\mathbf{G}$ are 1's and 0's. In each row, an element is 1 if the item belongs in the corresponding category of the variable; otherwise, the element is 0. Thus the number of 1's in a row of $\mathbf{G}$ is the number of variables, for a four-way contingency table, for example, there would be four 1's in each row of $\mathbf{G}$.

We illustrate a four-way classification with the (continued) data in Table 15.8. There are $n = 12$ items (people) and $p = 4$ categorical variables. The four variables and their categories are listed in Table 15.9. The indicator matrix $\mathbf{G}$ for the data in Table 15.8 is given in Table 15.10.

A correspondence analysis on $\mathbf{G}$ is equivalent to a correspondence analysis on $\mathbf{G}\mathbf{G}'$, which is called the *flattened* matrix. This equivalence can be justified as follows. In the singular value decomposition $\mathbf{G} = \mathbf{U}\mathbf{A}\mathbf{V}'$, the matrix $\mathbf{V}$ contains eigenvectors

Table 15.8. A List of 12 People and Their Categories on Four Variables

Person	Gender	Age	Marital Status	Hair Color
1	Male	Young	Single	Brown
2	Male	Old	Single	Red
3	Female	Middle	Married	Blond
4	Male	Old	Single	Black
5	Female	Middle	Married	Black
6	Female	Middle	Single	Brown
7	Male	Young	Married	Red
8	Male	Old	Married	Blond
9	Male	Middle	Single	Brown
10	Female	Young	Married	Black
11	Female	Old	Single	Brown
12	Male	Young	Married	Blond

Table 15.9. The Categories for the Four Variables in Table 15.8

Variable	Levels
Gender	Male, female
Age	Young, middle-aged, old
Marital status	Single, married
Hair color	Blond, brown, black, red

Table 13.10. Indicator Matrix $\mathbf{Q}$ for the Data in Table 13.8

Person	Gender	Age	Marital Status	Hair Color
1	1 0	1 0 0	1 0	0 1 0 0
2	1 0	0 0 1	1 0	0 0 0 1
3	0 1	0 1 0	0 1	1 0 0 0
4	1 0	0 0 1	1 0	0 0 1 0
5	0 1	0 1 0	0 1	0 0 1 0
6	0 1	0 1 0	1 0	0 1 0 0
7	1 0	1 0 0	0 1	0 0 0 1
8	1 0	0 0 1	0 1	1 0 0 0
9	1 0	0 1 0	1 0	1 0 0 0
10	0 1	1 0 0	0 1	0 0 1 0
11	0 1	0 0 1	1 0	0 1 0 0
12	1 0	1 0 0	0 1	1 0 0 0

of $\mathbf{Q}\mathbf{G}$. The same matrix $\mathbf{V}$ would be used in the spectral decomposition of $\mathbf{Q}\mathbf{G}$. Thus the columns of $\mathbf{V}$ are used in plotting r -coordinates for the columns of $\mathbf{Q}$ or the columns of $\mathbf{Q}\mathbf{G}$. If $\mathbf{Q}$ is $n \times p$ with $p > n$, then $\mathbf{Q}\mathbf{G}$ would be smaller in size than $\mathbf{Q}$.

The best matrix $\mathbf{Q}\mathbf{G}$ has a square block on the diagonal for each variable and a rectangular block off-diagonal for each pair of variables. Each diagonal block is a diagonal matrix showing the frequencies for the categories in the corresponding variable. Each off-diagonal block is a two-way contingency table for the corresponding pair of variables. In Table 13.11, we show the $\mathbf{Q}\mathbf{G}$ matrix for the $\mathbf{Q}$ matrix in Table 13.10.

A correspondence analysis of $\mathbf{Q}\mathbf{G}$ yields only column coordinates. A point is plotted for each column of $\mathbf{Q}$ (or of $\mathbf{Q}\mathbf{G}$). Thus, each point represents a category attribute of one of the variables.

Table 13.11. Best Matrix $\mathbf{Q}\mathbf{G}$ for the Matrix $\mathbf{Q}$ in Table 13.10

Gender		Age			Marital Status		Hair Color			
		3	1	3	4	0	3	0	1	0
0	0	0	0	1	2	0	1	2	2	0
0	1	4	0	0	0	0	1	0	1	1
1	0	0	4	0	2	2	0	0	1	0
0	1	0	0	4	0	1	1	0	1	1
0	2	0	2	0	0	0	1	0	1	1
0	3	0	0	1	0	0	0	0	2	1
0	1	0	2	1	0	0	4	0	0	0
1	0	0	1	1	0	0	0	0	0	0
1	2	0	1	1	0	2	0	0	0	0
0	0	0	0	1	0	1	0	0	0	0

Table 15.12. Singular Values (λ_i), Inertia (λ_i^2), and Chi-Square (λ_i^2) for the Best Matrix G/G in Table 15.11

Singular Value	Inertia	Chi-Square	Percent
.99905	.47338	31.551	27.85
.87431	.44491	30.324	26.88
.81495	.26813	17.652	15.58
.50000	.12500	10.000	8.89
.41941	.11360	11.754	10.53
.10259	.10294	7.981	7.03
.14591	.01980	1.123	1.00
Total	1.75000	116.038	100.00

Distances between points are not as meaningful as in correspondence analysis, but points in the same quadrant or approximate vicinity indicate an association. If two close points represent attributes of the same variable, the two attributes may be combined into a single attribute.

Since the Best matrix G/G has only two-way contingency tables, three-way and higher-order interactions are not represented in the plot. The various two-way tables are analyzed simultaneously, however.

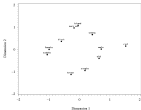


Figure 15.1b. Plot of points representing the 11 columns of Table 15.10 or 15.11.

Example 15.1.5(a). We continue the illustration in this section. A correspondence analysis of the first matrix $\mathbf{W}^{(1)}$ in Table 15.11 yields the singular values, inertia, and chi-squared values in Table 15.12. The first two singular values account for only 53.05% of the total variation. A plot of the first two dimensions for the 11 columns in Table 15.10 or 15.11 is given in Figure 15.13. It appears that married and blended families have a greater association that would be expected by chance alone. Another association is that between female and middle age. $\square$

Example 15.1.5(b). Table 15.13 (Gibson and Kainer 1986) is a two-way contingency table of employed men between the ages of 18 and 45 who were asked whether they themselves carried out repair work on their home, as opposed to hiring a carpenter to do the job. The two categorical variables are as follows:

Work: independent, skilled, unskilled, other.

Tenure: rent, own.

Age: under 30, 30–45, over 45.

Accommodation type: apartment, house.

Response to repair question: yes, no.

A multiple correspondence analysis produced the inertia and singular values in Table 15.14. The plot of the first two dimensions is given in Figure 15.14.

Table 15.13. Do-It-Yourself Data

Work	Tenure	Response	Accommodation Type					
			Apartment			House		
			Age			Age		
			under 30	30–45	over 45	under 30	30–45	over 45
Skilled	Rent	Yes	38	19	8	34	10	3
		No	19	13	8	28	4	8
	Own	Yes	3	3	1	50	30	23
		No	3	1	1	13	23	8
Unskilled	Rent	Yes	17	18	19	29	3	3
		No	34	17	18	44	13	18
	Own	Yes	2	8	3	23	32	69
		No	3	2	8	9	23	21
Other	Rent	Yes	20	23	21	22	13	21
		No	23	19	40	23	10	12
	Own	Yes	8	3	1	34	193	182
		No	4	2	2	19	70	81

Table 15.14. Singular Values (λ_i), Inertia (α_i^2), and Chi-Square (α_i^2/λ_i) for the Do-It-Yourself Data in Table 15.13

Singular Value	Inertia	Chi-Square	Percent
.60007	.36013	3,446.9	26.52
.46477	.21590	2,389.4	17.46
.43864	.19243	1,943.6	14.89
.42704	.18237	1,793.9	13.83
.40540	.16433	1,593.2	11.79
.36040	.13177	1,458.2	11.08
.27771	.07713	728.3	5.56
Total	1.40000	13,882.6	100

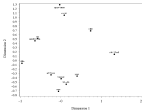


Figure 15.11. Plot of points representing the 12 categories in Table 15.13.

Unskilled employment has a high association with not doing one's own repairs. Doing one's own repairs is associated with owning a house, age between 31 and 45, and doing office work. Living in an apartment is associated with renting. □

15.3 BIPLOTS

15.3.1 Introduction

A *biplot* is a two-dimensional representation of a data matrix $\mathbf{Y}$ [see (3.17)] showing a point for each of the n observation vectors (rows of $\mathbf{Y}$) along with a point for each of the p variables (columns of $\mathbf{Y}$). The prefix *bi* refers to the two kinds of points, not to the dimensionality of the plot. The method presented here could, in fact, be generalized to a three-dimensional (or higher-order) biplot. Biplots were introduced by Gabriel (1971) and have been discussed at length by Gower and Hand (1996); see also Kshirsree and Malik (2000), Jacoby (1998, Chapter 1), and Fisher (1984, pp. 264–272).

If $p = 2$, a simple scatter plot, as in Section 3.3, has both kinds of information, namely, a point for each observation and the line axis representing the variables. We can see at a glance the placement of the points relative to each other and relative to the variables.

When $p > 2$, we can obtain a two-dimensional plot of the observations by plotting the first two principal components of $\mathbf{X}$ as in Section 12.4. We can then add a representation of the p variables to the plot of principal components to obtain a biplot. The principal component approach is discussed in Section 15.3.2. A method based on the singular value decomposition is presented in Section 15.3.3, and other methods are reviewed in Section 15.3.5.

15.3.2 Principal Component Plots

A principal component is given by $\mathbf{z} = \mathbf{a}'\mathbf{y}$, where $\mathbf{a}$ is an eigenvector of $\mathbf{S}$, the sample covariance matrix, and $\mathbf{y}$ is a $p \times 1$ observation vector (see Section 12.2). There are p eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p$, and thus there are p principal components $z_1, z_2, \dots, z_p$ for each observation vector $\mathbf{y}_i, i = 1, 2, \dots, n$. Hence (using the centered form) the observation vectors are transformed to $\mathbf{z}_{ij} = \mathbf{a}_j'(\mathbf{y}_i - \bar{\mathbf{y}}) = (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{a}_j, i = 1, 2, \dots, n, j = 1, 2, \dots, p$. Each $p \times 1$ observation vector $\mathbf{y}_i$ is transformed to a $p \times 1$ vector of principal components,

$$\begin{aligned} \mathbf{z}_i' &= (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{a}_1, (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{a}_2, \dots, (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{a}_p \\ &= (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{A}, \mathbf{a}_1, \dots, \mathbf{a}_p = (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{A}, \quad i = 1, 2, \dots, n, \end{aligned} \quad (15.46)$$

where $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p)$ is the $p \times p$ matrix whose columns are (normalized) eigenvectors of $\mathbf{S}$. (Note that the matrix $\mathbf{A}$ in (15.46) is the transpose of $\mathbf{A}$ in (12.3).) With $\mathbf{Z}$ and $\mathbf{Y}_c$ defined as

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_n' \end{pmatrix}, \quad \mathbf{Y}_c = \begin{pmatrix} \mathbf{y}_1 - \bar{\mathbf{y}} \\ \mathbf{y}_2 - \bar{\mathbf{y}} \\ \vdots \\ \mathbf{y}_n - \bar{\mathbf{y}} \end{pmatrix} \quad (15.48)$$

from (15.13), we can express the principal components in (15.46) as

$$\mathbf{Z} = \mathbf{Y}_1 \mathbf{A}, \quad (15.50)$$

Since the eigenvalues α_j of the symmetric matrix $\mathbf{B}$ are mutually orthogonal (see Section 2.11.4), $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p)$ is an orthogonal matrix, and $\mathbf{A}\mathbf{A}' = \mathbf{I}$. Multiplying (15.51) on the right by $\mathbf{A}'$, we obtain

$$\mathbf{Y}_1 = \mathbf{Z}\mathbf{A}'. \quad (15.52)$$

The best two-dimensional representation of $\mathbf{Y}_1$ is given by taking the first two columns of $\mathbf{Z}$ and the first two columns of $\mathbf{A}$. If the resulting matrices are denoted by $\mathbf{Z}_2$ and $\mathbf{A}_2$, we have

$$\mathbf{Y}_1 \approx \mathbf{Z}_2 \mathbf{A}_2'. \quad (15.53)$$

The fit in (15.53) is best in a least squares sense. If the left side of (15.53) is represented by $\mathbf{Y}_1 = \mathbf{B} = (b_{ij})$ and the right side by $\mathbf{Z}_2 \mathbf{A}_2' = \mathbf{C} = (c_{ij})$, then $\sum_{i=1}^n \sum_{j=1}^2 (b_{ij} - c_{ij})^2$ is minimized (Folter 1964, p. 205).

The coordinates for the n observations are the rows of $\mathbf{Z}_2$, and the coordinates for the p variables are the rows of $\mathbf{A}_2$ (columns of $\mathbf{A}_2'$). The coordinates are discussed further in Section 15.3.4.

The adequacy of the fit in (15.53) can be evaluated by examining the first two eigenvalues λ_1 and λ_2 of $\mathbf{B}$. Thus a large value close to 1 of

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^p \lambda_i}$$

would indicate that $\mathbf{Y}_1$ is represented well visually in the plot.

15.3.3 Singular Value Decomposition Plot

We can also obtain $\mathbf{Y}_1 = \mathbf{Z}\mathbf{A}'$ in (15.52) by means of the singular value decomposition of $\mathbf{Y}_1$. By (2.1.17), we have

$$\mathbf{Y}_1 = \mathbf{U}\mathbf{A}\mathbf{V}', \quad (15.54)$$

where $\mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ is a diagonal matrix containing square roots of the nonzero eigenvalues $\lambda_1^2, \lambda_2^2, \dots, \lambda_p^2$ of $\mathbf{Y}_1'\mathbf{Y}_1$ (and of $\mathbf{Y}_1\mathbf{Y}_1'$), the columns of $\mathbf{U}$ are the corresponding eigenvectors of $\mathbf{Y}_1'\mathbf{Y}_1$, and the columns of $\mathbf{V}$ are the corresponding eigenvectors of $\mathbf{Y}_1\mathbf{Y}_1'$.

The product $\mathbf{U}\mathbf{A}$ in (15.54) is equal to $\mathbf{Z}$, the matrix of principal component scores in (15.51). To see this we multiply (15.54) by $\mathbf{V}$, which is orthogonal because it contains the (normalized) eigenvectors of the symmetric matrix $\mathbf{Y}_1'\mathbf{Y}_1$ (see Section 2.11.4). This gives

$$\mathbf{Y}_1\mathbf{V} = \mathbf{U}\mathbf{A}\mathbf{V}'\mathbf{V} = \mathbf{U}\mathbf{A}. \quad (15.55)$$

By (1.37), $\mathbf{Y}_c \mathbf{Y}_c$ is equal to $-\mathbf{a} - \mathbf{1}\mathbf{b}$. By (2.66), eigenvectors of $-\mathbf{a} - \mathbf{1}\mathbf{b}$ are also eigenvectors of $\mathbf{S}$. Thus $\mathbf{V}$ is the same as $\mathbf{A}$ in (15.51), which contains eigenvectors of $\mathbf{S}$. Hence, $\mathbf{Y}_c \mathbf{V}$ in (15.59) becomes

$$\begin{aligned}\mathbf{Y}_c \mathbf{V} &= \mathbf{Y}_c \mathbf{A} \\ &= \mathbf{Z} \quad \text{by (15.51)} \\ &= \mathbf{U} \mathbf{A} \quad \text{by (15.59)}.\end{aligned}$$

We can therefore write (15.54) as

$$\mathbf{Y}_c = \mathbf{U} \mathbf{A} \mathbf{V}' = \mathbf{Z} \mathbf{V}' = \mathbf{Z} \mathbf{A}', \quad (15.59)$$

Thus the singular value decomposition of $\mathbf{Y}_c$ gives the same factoring as the expression in (15.52) based on principal components.

15.3.4 Coordinates

In this section, we consider the coordinates for the methods of Sections 15.3.2 and 15.3.5. Let us return to (15.53), the two-dimensional representation of $\mathbf{Y}_c$ based on principal components (which is the same representation as that based on the singular value decomposition).

$$\mathbf{Y}_c = \mathbf{Z}_2 \mathbf{A}_2' = \begin{pmatrix} z_{21} & z_{22} \\ z_{31} & z_{32} \\ \vdots & \vdots \\ z_{n1} & z_{n2} \end{pmatrix} \begin{pmatrix} a_{21} & a_{22} & \cdots & a_{2p} \\ a_{31} & a_{32} & \cdots & a_{3p} \end{pmatrix}, \quad (15.57)$$

The elements of (15.57) are of the form

$$y_{ij} = \beta_j \{i\} a_{2j} + \gamma_j \{i\} a_{3j}, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, p.$$

Thus each observation is represented as a linear combination, the coefficients being the elements of the vector $(\gamma_{2j}, \gamma_{3j})$ and the axes being the elements of the vector (a_{2j}, a_{3j}) . We therefore plot the points $(\gamma_{2i}, \gamma_{3i})$, $i = 1, 2, \dots, n$, and the points (a_{2j}, a_{3j}) , $j = 1, 2, \dots, p$. To distinguish them and to show relationship of the points to the axes, the points $(\gamma_{2i}, \gamma_{3i})$ are connected to the origin with a straight line forming an arrow. If necessary, the scale of the points (a_{2j}, a_{3j}) could be adjusted to be compatible with that of the principal components $(\gamma_{2i}, \gamma_{3i})$.

The Euclidean distance between two points $(\gamma_{2i}, \gamma_{3i})$ and $(\gamma_{2j}, \gamma_{3j})$ is approximately equal to the distance between the corresponding points (rows) $\mathbf{y}_i'$ and $\mathbf{y}_j'$ in the data matrix $\mathbf{Y}$. If all of the principal components were used, as in (15.50) and

(15.32), the distance would be the same, but, with only two principal components, the distance is an approximation.

The cosine of the angle between the arrows (lines) drawn to each pair of axis points (x_{1j}, x_{2j}) and (x_{1k}, x_{2k}) shows the correlation between the two corresponding variables [see (3.14) and (3.15)]. Thus, a small angle between two vectors indicates that the two variables are highly correlated, two variables whose vectors form a 90° angle are uncorrelated, and an angle greater than 90° indicates that the variables are negatively correlated.

The values of the p variables in the i th observation vector $\mathbf{y}_i$ (corrected for means) are related to the perpendicular projections of the point (x_{1i}, x_{2i}) on the p vectors from the origin to the points (x_{1j}, x_{2j}) representing variations. The further from the origin a projection falls on an arrow, the larger the value of the observation on that variable. Hence the vectors will be oriented toward the observations that have larger values on the corresponding variables.

Example 18.3.8. Using the city crime data of Table 18.1, we illustrate the principal component approach. The first two eigenvalues of the sample covariance matrix $\mathbf{S}$ are given by

$$\lambda_1 = \begin{pmatrix} .802 & .088 \\ .017 & .014 \\ .192 & .086 \\ .164 & .221 \\ .747 & -.240 \\ .612 & -.189 \\ .155 & .636 \end{pmatrix}.$$

The matrix of the first two principal components is given by

$$\mathbf{X}_2 = \mathbf{Y}_1 \mathbf{A}_2 = \begin{pmatrix} -317.2 & -186.1 \\ -491.8 & 352.4 \\ -650.0 & 127.6 \\ 341.7 & -138.8 \\ 343 & -69.3 \\ 312.1 & 164.1 \\ -318.7 & -160.4 \\ 88.6 & -180.7 \\ -38.9 & 15.8 \\ 75.7 & 40.1 \\ 878.1 & 7.1 \\ -352.4 & 165.4 \\ 342.7 & 164.8 \\ 238.1 & -360.8 \\ -345.8 & -184.3 \\ 150.4 & 200.8 \end{pmatrix}.$$

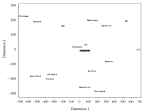


Figure B.11: Principal component plot for city crime data in Table 14.1.

The coordinates for the 19 cities are found in $\mathbf{X}_2$, and the coordinates for the 7 variables are found in $\mathbf{A}_2$. The plot of the city and variable points is given in Figure 15.12. The observation points are spread out, whereas the variable points are clustered tightly around the origin. Suitable scaling of the eigenvectors in $\mathbf{A}_2$ would enable the arrows representing the variables to pass through the points (see Example 15.5.7).

□

B.5.5 Other Methods

The singular value decomposition of $\mathbf{Y}_r$ is given in (15.54) as

$$\mathbf{Y}_r = \mathbf{U}\mathbf{A}\mathbf{V}^T. \quad (15.58)$$

In Section 15.5.3, it was shown that $\mathbf{U}\mathbf{A} = \mathbf{Z}$ and $\mathbf{V} = \mathbf{A}$ (see (15.56)), so that (15.58) can be written as

$$\mathbf{Y}_r = (\mathbf{U}\mathbf{A})\mathbf{V}^T = \mathbf{Z}\mathbf{A}^T,$$

which is equivalent to the principal component solution $\mathbf{Y}_r = \mathbf{Z}\mathbf{A}^T$ in (15.52). Alternative factorings may be of interest. Two that have been considered are

$$\mathbf{U}_1 = \mathbf{U}(\mathbf{A}\mathbf{N}'), \quad (15.59)$$

$$\mathbf{U}_1 = (\mathbf{U}\mathbf{A}^{1/2})_1\mathbf{A}^{1/2}\mathbf{N}', \quad (15.60)$$

If we denote the submatrix consisting of the first two columns of $\mathbf{U}$ and $\mathbf{V}$ as $\mathbf{U}_2$ and $\mathbf{V}_2$, respectively, and define $\mathbf{A}_2 = \text{diag}(\lambda_1, \lambda_2)$, then the two-dimensional representations of (15.59) and (15.60) are

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{U}_2(\mathbf{A}_2\mathbf{N}')_2 \\ &= \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ \vdots & \vdots \\ u_{p1} & u_{p2} \end{pmatrix} \begin{pmatrix} \lambda_1 v_{11} & \lambda_1 v_{12} & \cdots & \lambda_1 v_{1p} \\ \lambda_2 v_{21} & \lambda_2 v_{22} & \cdots & \lambda_2 v_{2p} \end{pmatrix}, \end{aligned} \quad (15.61)$$

$$\begin{aligned} \mathbf{Y}_1 &= (\mathbf{U}_2\mathbf{A}_2^{1/2})_1(\mathbf{A}_2^{1/2}\mathbf{N}')_2 \\ &= \begin{pmatrix} \sqrt{\lambda_1}u_{11} & \sqrt{\lambda_1}u_{12} \\ \sqrt{\lambda_1}u_{21} & \sqrt{\lambda_1}u_{22} \\ \vdots & \vdots \\ \sqrt{\lambda_1}u_{p1} & \sqrt{\lambda_1}u_{p2} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_2}v_{11} & \sqrt{\lambda_2}v_{12} & \cdots & \sqrt{\lambda_2}v_{1p} \\ \sqrt{\lambda_2}v_{21} & \sqrt{\lambda_2}v_{22} & \cdots & \sqrt{\lambda_2}v_{2p} \end{pmatrix}. \end{aligned} \quad (15.62)$$

For the biplot corresponding to (15.61), we plot the set of points (u_{i1}, u_{i2}) , $i = 1, 2, \dots, n$, and the set of points $(\lambda_1 v_{j1}, \lambda_2 v_{j2})$, $j = 1, 2, \dots, p$, with the latter points connected to the origin by an arrow to show the axes. For the biplot arising from (15.62), we plot the set of points $(u_{i1}\sqrt{\lambda_1}, u_{i2}\sqrt{\lambda_1})$, $i = 1, 2, \dots, n$, and the set of points $(\sqrt{\lambda_2}v_{j1}, \sqrt{\lambda_2}v_{j2})$, $j = 1, 2, \dots, p$, with the latter points connected to the origin with an arrow.

The presence of λ_1 and λ_2 in (15.61) and (15.62) provides scaling that is absent in (15.57). For many data sets the scaling in (15.61) will be adequate with no further adjustment.

If we write (15.59) in the form

$$\mathbf{Y}_1 = \mathbf{U}(\mathbf{A}\mathbf{N}') = \mathbf{U}(\mathbf{U}\mathbf{A})' = \mathbf{U}\mathbf{U}'\mathbf{A}, \quad (15.63)$$

then

$$\mathbf{U}\mathbf{U}' = \left(\frac{1}{n-1} \right) \mathbf{U}_2\mathbf{A}_2^{-1}\mathbf{U}_2', \quad (15.64)$$

$$\mathbf{U}\mathbf{U}' = \mathbf{I}_n - \mathbf{1}\mathbf{1}' \quad (15.65)$$

(see Problem 15.10). With suitable scaling of the eigenvalues in $\mathbf{U}$ and $\mathbf{V}$, we could eliminate the coefficients involving $n-1$ from (15.64) and (15.65).

By (15.64) (with scaling to eliminate $n = T$), the (Euclidean) distance $(\mathbf{a}_i - \mathbf{a}_j)'(\mathbf{a}_i - \mathbf{a}_j)$ between two points $\mathbf{a}_i$ and $\mathbf{a}_j$ is equal to the Mahalanobis distance $(\mathbf{z}_i - \mathbf{z}_j)' \mathbf{S}^{-1} (\mathbf{z}_i - \mathbf{z}_j)$ between the corresponding points $\mathbf{z}_i$ and $\mathbf{z}_j$ in the data matrix $\mathbf{Z}$:

$$(\mathbf{a}_i - \mathbf{a}_j)'(\mathbf{a}_i - \mathbf{a}_j) = (\mathbf{z}_i - \mathbf{z}_j)' \mathbf{S}^{-1} (\mathbf{z}_i - \mathbf{z}_j) \quad (15.66)$$

(see Problem 15.11). By (15.65), the covariance r_{ij} between the j th and i th variables (columns of $\mathbf{Z}$) is given by

$$r_{ij} = \mathbf{h}_j' \mathbf{h}_i, \quad (15.67)$$

where $\mathbf{h}_j$ and $\mathbf{h}_i$ are rows of $\mathbf{H}$. By (3.14) and (3.15), this can be converted to the correlation

$$r_{ij} = \cos \theta = \frac{\mathbf{h}_j' \mathbf{h}_i}{\sqrt{(\mathbf{h}_j' \mathbf{h}_j)(\mathbf{h}_i' \mathbf{h}_i)}}, \quad (15.68)$$

so that the angle between the two vectors $\mathbf{h}_j$ and $\mathbf{h}_i$ is related to the correlation between the j th and i th variables.

The two-dimensional representation of $\mathbf{a}_i$ and $\mathbf{a}_j$ in (15.61) has the approximate Mahalanobis distance and correlation properties discussed earlier.

Example 15.3.6. Using the city crime data of Table 14.1, we illustrate the singular value decomposition method with the factorings in (15.61) and (15.62). The matrices $\mathbf{U}_2$, $\mathbf{A}_2$, and $\mathbf{V}_2$ are

$$\mathbf{U}_2 = \begin{pmatrix} -.211 & -.250 \\ -.327 & .288 \\ -.432 & .355 \\ .094 & -.197 \\ .327 & -.302 \\ .208 & .342 \\ -.342 & -.345 \\ .098 & -.388 \\ -.018 & .058 \\ .058 & .059 \\ .451 & .040 \\ -.128 & .341 \\ .381 & .287 \\ .155 & -.369 \\ .028 & -.372 \\ .108 & .258 \end{pmatrix}, \quad \mathbf{V}_2 = \begin{pmatrix} .002 & .008 \\ .043 & .048 \\ .182 & .089 \\ .104 & .121 \\ .707 & -.240 \\ .612 & -.378 \\ .153 & .639 \end{pmatrix},$$

$$\mathbf{A}_2 = \text{diag}\{1503.694, 678.615\}.$$

By (15.61), the two-dimensional representation is given by plotting the rows of V_2 and the rows of V_2A_2 (or the columns of $A_2A_2^T$). For V_2A_2 we have

$$V_2A_2 = \begin{pmatrix} 3.0 & 5.1 \\ 25.4 & 9.5 \\ 273.6 & 407.5 \\ 156.3 & 159.1 \\ 1128.4 & -162.6 \\ 926.0 & -74.1 \\ 229.3 & 452.9 \end{pmatrix}.$$

The plot of the observation points and variable points is given in Figure 15.13.

For (15.62), we obtain

$$V_2A_2^{1/2} = \begin{pmatrix} -6.38 & -5.88 \\ -12.68 & 7.79 \\ -16.76 & 9.74 \\ 3.65 & -5.04 \\ 3.82 & -2.88 \\ 8.05 & 6.36 \\ -13.27 & -6.38 \\ 1.51 & -9.26 \\ -0.93 & .99 \\ 1.05 & 1.54 \\ 17.49 & .27 \\ -4.96 & 6.27 \\ 15.88 & 7.48 \\ 6.81 & -10.24 \\ -9.87 & -7.88 \\ 3.88 & 7.71 \end{pmatrix}, \quad V_2A_2^{1/2} = \begin{pmatrix} .88 & .26 \\ .88 & .37 \\ 7.88 & 17.99 \\ 4.63 & 3.76 \\ 28.87 & -6.24 \\ 15.73 & -1.89 \\ 5.81 & 16.62 \end{pmatrix}.$$

The plot of these coordinates is given in Figure 15.14. For this data set, the factoring given by (15.62) in Figure 15.14 is preferred because it plots both observations and variable points on the same scale. The factorings shown in Figures 15.12 and 15.13 would need an adjustment in scaling. $\square$

PROBLEMS

- 15.1 In step 2 of the algorithm for metric scaling in Section 15.1.2, the matrix $B = (b_{ij})$ is defined in terms of $\mathbf{A} = (a_{ij})$ as $b_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$. Show that b_{ij} in $B = (I - \frac{1}{n}J)A(I - \frac{1}{n}J)$ in (15.2) is equivalent to $b_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$.
- 15.2 Verify the result stated in step 2 of the algorithm in Section 15.1.2, namely, that there exists a q -dimensional configuration $x_1, x_2, \dots, x_n$ such that $a_{ij} = b_{ij}$ if and only if B is positive semidefinite of rank q . Use the following approach.

- (a) Assuming the existence of $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ such that $\mathbf{b}_i^T = \mathbf{a}_i^T - \mathbf{a}_1^T/(\mathbf{a}_1 - \mathbf{a}_1)$, show that $\mathbf{B}$ is positive semidefinite.
- (b) Assuming $\mathbf{B}$ is positive semidefinite, show that there exist $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ such that $\mathbf{b}_i^T = (\mathbf{a}_i - \mathbf{a}_1)^T/(\mathbf{a}_i - \mathbf{a}_1) = \mathbf{a}_i^T$.
- 15.3 (a) Show that $\mathbf{r} = \sum_{j=1}^p \mathbf{r}_j \mathbf{r}_j^T$ in (15.20) is the same as $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_p)^T$ in (15.19).
- (b) Show that $\mathbf{r}^T = \sum_{j=1}^p \mathbf{r}_j^T \mathbf{r}_j$ in (15.21) is equivalent to $\mathbf{r}^T = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_p)$ in (15.18).
- 15.4 Show that $\mathbf{f}^T \mathbf{r} = \mathbf{r}^T \mathbf{f} = 1$ as in (15.22).
- 15.5 Show that the chi-square statistic in (15.26) is equal to that in (15.25).
- 15.6 (a) Show that the chi-square statistic in (15.27) is equal to that in (15.25).
- (b) Show that the chi-square statistic in (15.28) is equal to that in (15.25).
- 15.7 (a) Show the chi-square statistic in (15.29) is equal to that in (15.27).
- (b) Show the chi-square statistic in (15.30) is equal to that in (15.28).
- 15.8 (a) Show that $\mathbf{B} = \mathbf{J}\mathbf{r}^T = \mathbf{B}_1^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}^T)$ as in (15.41).
- (b) Show that $\mathbf{C} = \mathbf{r}\mathbf{f}^T = \mathbf{B}_1^{-1}(\mathbf{P} - \mathbf{r}\mathbf{r}^T)$ as in (15.42).
- 15.9 Show that if all the principal components were used, the distance between $\mathbf{a}_1$ and $\mathbf{a}_2$ would be the same as between $\mathbf{y}_1$ and $\mathbf{y}_2$, as noted in Section 15.3.4.
- 15.10 (a) Show that $\mathbf{B}\mathbf{U}^T = \mathbf{Y}_1\mathbf{B}^{-1}\mathbf{Y}_1^T/(\alpha - 1)$ as in (15.60).
- (b) Show that $\mathbf{B}\mathbf{H} = (\alpha - 1)\mathbf{B}$ as in (15.60).
- 15.11 Show that $(\mathbf{a}_1 - \mathbf{a}_2)^T(\mathbf{a}_1 - \mathbf{a}_2) = (\mathbf{y}_1 - \mathbf{y}_2)^T\mathbf{B}^{-1}(\mathbf{y}_1 - \mathbf{y}_2)$ as in (15.66).
- 15.12 In Table 15.15, we have road distances between major UK towns if hard et al. (1994, p. 346). The towns are as follows:
 $A =$ Aberdeen, $B =$ Birmingham, $C =$ Brighton, $D =$ Bristol, $E =$ Cardiff,
 $F =$ Carlisle, $G =$ Dover, $H =$ Edinburgh, $I =$ Fort William, $J =$ Glasgow,
 $K =$ Holyhead, $L =$ Hull, $M =$ Inverness, $N =$ Leeds, $O =$ Liverpool,
 $P =$ London, $Q =$ Manchester, $R =$ Newcastle, $S =$ Norwich, $T =$ Nottingham,
 $U =$ Penzance, $V =$ Plymouth, $W =$ Sheffield.
- (a) Find the matrix $\mathbf{B}$ as in (15.7).
- (b) Using the spectral decomposition, find the first two columns of the matrix $\mathbf{Z}$ as in (15.4).
- (c) Create a metric multidimensional scaling plot of the first two dimensions. What do you notice about the positions of the cities?
- 15.13 Zhang, Holander, and Drury (1996) analysed a 41 × 41 similarity matrix for 41 descriptors of comfort, such as *sofas*, *legging*, *stuffed*, etc. For the similarity matrix, see the Wiley site (Appendix C).
- (a) Carry out a metric multidimensional scaling analysis and plot the first two dimensions. What pattern is seen in the plot?

1. The first step is to identify the problem or question that needs to be addressed. This involves understanding the context and the specific requirements of the task.

1	天
2	地
3	人
4	物
5	事
6	理
7	法
8	道
9	德
10	義
11	禮
12	智
13	信
14	忠
15	孝
16	悌
17	友
18	愛
19	敬
20	恭
21	肅
22	靜
23	安
24	寧
25	和
26	平
27	泰
28	康
29	寧
30	壽
31	福
32	祿
33	壽
34	考
35	終
36	祭
37	祀
38	宗
39	廟
40	社
41	稷
42	土
43	穀
44	神
45	農
46	工
47	商
48	戰
49	國
50	家
51	身
52	心
53	意
54	志
55	德
56	行
57	言
58	動
59	靜
60	安
61	寧
62	和
63	平
64	泰
65	康
66	寧
67	壽
68	福
69	祿
70	壽
71	考
72	終
73	祭
74	祀
75	宗
76	廟
77	社
78	稷
79	土
80	穀
81	神
82	農
83	工
84	商
85	戰
86	國
87	家
88	身
89	心
90	意
91	志
92	德
93	行
94	言
95	動
96	靜
97	安
98	寧
99	和
100	平

Table 15.16. Dissimilarity Matrix for World War II Estimates

Person	Holst	Morrell	Churchill	Eisenhower
Holst	8	5	11	15
Morrell	5	9	14	16
Churchill	11	14	9	7
Eisenhower	15	16	7	8
Stalin	8	15	11	16
Aches	17	18	11	16
Truman	5	5	12	14
De Gaulle	18	11	8	8
Mac-Ten	16	18	15	17
Trotsky	17	18	9	8
Chamberlain	12	14	10	7
Tito	18	17	8	12
	Stalin	Aches	Truman	De-Gaulle
Holst	8	17	8	16
Morrell	12	18	9	15
Churchill	11	11	12	5
Eisenhower	16	16	16	8
Stalin	8	15	15	11
Aches	17	9	18	12
Truman	13	16	9	9
De Gaulle	18	12	9	8
Mac-Ten	12	18	17	12
Trotsky	14	12	18	8
Chamberlain	16	9	19	11
Tito	12	18	12	5
	Mac-Ten	Truman	Chamberlain	Tito
Holst	16	17	12	16
Morrell	18	18	14	17
Churchill	16	8	10	8
Eisenhower	17	9	7	12
Stalin	12	16	16	12
Aches	18	12	9	13
Truman	17	16	10	12
De Gaulle	13	9	11	5
Mac-Ten	8	12	17	16
Trotsky	12	9	9	11
Chamberlain	17	9	9	18
Tito	18	11	15	8

- (b) For an alternative approach, carry out a cluster analysis of the configuration of points found in part (a), using Ward's method. Create a dendrogram of the cluster solution. How many clusters are indicated?
- 12.18 Use the political data of Table 12.16 (Russett 1967, Table 6.3). Two subjects assessed the degree of dissimilarity between World War II politicians. The data matrix represents the sum of the dissimilarities between the two subjects.
- (a) For $d = 4$, create an initial configuration of points by choosing 12 random observations taken from a multivariate normal distribution with mean vector $\bar{\mathbf{0}}$ and covariance matrix $\mathbf{I}_n$.
- (b) Carry out a metric multidimensional scaling analysis using the seeds found in part (a). Find the value of the STRESS statistic.
- (c) Repeat parts (a) and (b) for $d = 1, \dots, 5$. Plot the STRESS values against the values of d . How many dimensions should be kept? Plot the final configuration of points with two dimensions.
- (d) Repeat parts (a)–(c) using as initial configuration of points from a multivariate normal with different mean vector and covariance matrix from those in part (a). How many dimensions should be kept? Plot the final configuration of points with two dimensions. How does this solution compare to that in part (c)?
- (e) Repeat parts (a)–(c) using as initial configuration of points from a uniform distribution over $(0, 1)$. How many dimensions should be kept? Plot the final configuration of points with two dimensions.
- (f) Repeat part (a) using as initial configuration of points the metric multidimensional scaling solution found by treating the original measurements as continuous. How many dimensions should be kept? Plot the final configuration of points with two dimensions.

Table 12.17. Birth and Death Months of 1281 People

[Table not available in this electronic edition.]

15.15 In Table 15.17 we have the months of birth and death for 1281 people (Anderson and Hirschberg 1985, Table T1.2).

- Find the correspondence matrix P as in (15.8).
- Find the matrices $\mathbf{B}$ and $\mathbf{C}$, as in (15.10) and (15.10).
- Perform a chi-square test for independence between birth and death months.
- Plot the row and column deviations as in Example 15.2.5(a).

15.16 In Table 15.18, we have a cross-classification of crimes in Norway in 1984 categorized by type and site (Clausen 1998, p. 8).

Table 15.18. Crimes by Type and Site

Part of Country	Burglary	Theft	Vandalism	Total
Oslo-area	182	2408	1138	4688
Mid-Norway	147	103	916	1256
North-Norway	694	327	1447	2468
Total	1223	2938	4021	8192

- Find the correspondence matrix P as in (15.8).
- Find the matrices $\mathbf{B}$ and $\mathbf{C}$ as in (15.10) and (15.10).
- Perform a chi-square test for independence between type of crime and site.
- Plot the row and column deviations as in Example 15.2.4.

15.17 In Table 15.19, we have a six-way contingency table (Anderson and Hirschberg 1985, Table 34.1). Carry out a multiple correspondence analysis.

- Set up an indicator matrix $\mathbf{G}$ and find the Bart matrix $\mathbf{G}^*\mathbf{G}$.
- Perform a correspondence analysis on the Bart matrix found in part (a) and plot the coordinates.
- What associations are present?

15.18 Use the protein consumption data of Table 14.7.

- Create a biplot using the principal component approach in (15.23) or (15.27).
- Create a biplot using the singular value-decomposition approach with the factoring as in (15.61).
- Create a biplot using the singular value-decomposition approach with the factoring as in (15.62).
- Which of the three biplots best represents the data?

15.19 Use the perception data of Table 15.1.

[Table not available in this electronic edition.]

[Table not available in this electronic edition.]

- (a) Create a biplet using the principal component approach in (15.53) or (15.57).
- (b) Create a biplet using the singular value decomposition approach with the factoring, as in (15.61).
- (c) Create a biplet using the singular value decomposition approach with the factoring, as in (15.62).
- (d) Which of the three biplets best represents the data?

15.10 Use the word data of Table 6.21.

- (a) Create a biplet using the principal component approach in (15.53) or (15.57).
- (b) Create a biplet using the singular value decomposition approach with the factoring, as in (15.61).
- (c) Create a biplet using the singular value decomposition approach with the factoring, as in (15.62).
- (d) Which of the three biplets best represents the data?

APPENDIX A

Tables

Table A.1. Upper Percentiles for χ^2_k

$$\chi^2_k = \frac{\sqrt{k} \sum_{i=1}^k (Z_i - \bar{Z})^2}{\left(\sum_{i=1}^k (Z_i - \bar{Z})^2 \right)^{1/2}}$$

The sampling distribution of χ^2_k is symmetric about zero, and lower percentage points corresponding to negative deviations are given by the negative of the table values. Reject the hypothesis of normality if χ^2_k is greater than the table value or less than the negative of the table value.

n	Upper Percentiles					
	10	5	2.5	1	.5	.1
4	.821	.949	1.076	1.120	1.157	1.151
5	.821	1.080	1.207	1.257	1.296	1.454
6	.793	1.060	1.239	1.429	1.533	1.671
7	.763	1.043	1.266	1.487	1.599	1.797
8	.733	1.026	1.288	1.533	1.663	1.866
9	.704	.997	1.304	1.576	1.726	1.898
10	.675	.978	1.323	1.617	1.778	1.926
11	.646	.959	1.334	1.657	1.823	1.959
12	.617	.940	1.348	1.695	1.871	1.982
13	.588	.920	1.363	1.732	1.917	1.999
14	.560	.899	1.380	1.768	1.966	1.992
15	.531	.879	1.399	1.803	1.998	1.990
16	.503	.858	1.418	1.837	1.992	1.973
17	.475	.837	.992	1.870	1.983	1.944
18	.447	.816	.976	1.902	1.976	1.914
19	.419	.795	.960	1.935	1.964	1.885
20	.391	.773	.942	1.967	1.956	1.857
21	.363	.752	.925	1.994	1.947	1.828
22	.336	.730	.908	1.994	1.939	1.800
23	.309	.709	.894	1.996	1.931	1.773
24	.282	.687	.880	1.998	1.923	1.746
25	.255	.666	.866	1.999	1.915	1.720

Table A.2. Coefficients for Transforming $\sqrt{E}$ to a standard Normal

n	d	$1/L$	n	d	$1/L$
8	3.863	.3030	62	3.589	1.0400
9	4.260	.2686	64	3.620	1.0449
10	4.734	.2434	66	3.650	1.0494
11	5.247	.2239	68	3.680	1.0540
12	5.790	.2081	70	3.709	1.0581
13	6.351	.1950	72	3.740	1.0621
14	6.939	.1843	74	3.769	1.0659
15	7.554	.1753	76	3.799	1.0695
16	8.196	.1678	78	3.828	1.0730
17	8.865	.1616	80	3.857	1.0763
18	9.561	.1563	82	3.886	1.0794
19	1.028	.1518	84	3.915	1.0823
20	1.104	.1479	86	3.944	1.0854
21	1.184	.1444	88	3.972	1.0882
22	1.264	.1413	90	3.999	1.0909
23	1.344	.1385	92	4.026	1.0934
24	1.424	.1359	94	4.052	1.0959
25	1.504	.1335	96	4.078	1.0983
26	1.584	.1313	98	4.103	1.1006
27	1.664	.1292	100	4.128	1.1028
28	1.744	.1272	102	4.152	1.1050
29	1.824	.1253	110	4.276	1.1128
30	1.904	.1235	115	4.340	1.1171
31	1.984	.1218	120	4.400	1.1211
32	2.064	.1202	125	4.457	1.1250
33	2.144	.1187	130	4.510	1.1287
34	2.224	.1173	135	4.558	1.1323
35	2.304	.1159	140	4.603	1.1358
36	2.384	.1146	145	4.645	1.1391
37	2.464	.1133	150	4.685	1.1423
38	2.544	.1121	155	4.724	1.1454
39	2.624	.1109	160	4.760	1.1483
40	2.704	.1098	165	4.795	1.1511
41	2.784	.1087	170	4.828	1.1538
42	2.864	.1076	175	4.860	1.1564
43	2.944	.1066	180	4.891	1.1589
44	3.024	.1056	185	4.921	1.1613
45	3.104	.1046	190	4.950	1.1636
46	3.184	.1037	195	4.978	1.1658
47	3.264	.1028	200	5.005	1.1679
48	3.344	.1019	205	5.032	1.1699
49	3.424	.1010	210	5.058	1.1718
50	3.504	.1001	215	5.083	1.1736
51	3.584	.0992	220	5.107	1.1754
52	3.664	.0983	225	5.130	1.1771

Table A.1. *Continued*

52	5.237	1.0508	238	5.583	1.1383
54	5.258	1.0514	240	5.581	1.1389
56	5.258	1.0515	248	5.584	1.1394
58	5.529	1.0593	249	5.581	1.1394
60	5.539	1.0548	250	5.588	1.1394

Values of t and $1/t$ are values $t_{\alpha, \nu} / F_{\alpha, \nu} = t_{\alpha, \nu}^{*} / F_{\alpha, \nu}^{*}$ as given in Appendix F(1), (3).

Table A.3. Percentiles for b_1

Upper and lower percentiles for

$$b_1 = \frac{n \sum_{i=1}^n x_i^2 y_i - \bar{y}^2}{\left[\sum_{i=1}^n x_i^2 y_i - \bar{y} \right]^2},$$

for sample coefficient of kurtosis. Report the hypothesis of normality if b_1 is greater than its upper percentile or less than lower percentile.

Sample size	Percentiles											
	1	2	2.5	5	10	20	50	90	95	97.5	98	99
7	1.25	1.30	1.34	1.41	1.50	1.70	2.09	3.28	3.55	3.85	3.93	4.25
8	1.31	1.37	1.40	1.46	1.56	1.75	2.04	3.30	3.70	4.09	4.30	4.55
9	1.35	1.42	1.45	1.51	1.60	1.80	2.08	3.40	3.80	4.28	4.41	4.65
10	1.38	1.45	1.48	1.56	1.66	1.85	2.04	3.50	3.90	4.40	4.53	4.75
12	1.46	1.52	1.56	1.64	1.76	1.93	2.08	3.69	4.08	4.58	4.73	4.90
15	1.55	1.61	1.66	1.73	1.86	2.01	2.13	3.82	4.19	4.69	4.83	5.00
20	1.65	1.71	1.74	1.82	1.95	2.10	2.21	3.98	4.35	4.85	4.97	5.16
25	1.72	1.78	1.82	1.91	2.03	2.28	2.35	3.98	4.35	4.85	4.92	5.10
30	1.78	1.85	1.88	1.96	2.10	2.30	2.35	3.98	4.35	4.75	4.75	5.21
35	1.84	1.91	1.95	2.03	2.14	2.31	2.35	3.98	4.35	4.75	4.68	5.15
40	1.89	1.96	1.98	2.07	2.19	2.34	2.38	3.97	4.35	4.68	4.63	5.04
45	1.93	2.00	2.03	2.11	2.22	2.37	2.38	3.99	4.35	4.59	4.63	4.94
50	1.95	2.03	2.06	2.13	2.25	2.40	2.38	3.92	3.99	4.59	4.49	4.88

Table A.4. Percentiles for Shapiro-Wilk Test for Normality

Upper and lower percentiles for the statistic

$$T = \frac{\sqrt{n(n-1)} \left| \hat{r}_n \right|}{\sqrt{12} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

where

$$\hat{r}_n = \frac{\sum_{i=1}^n \left(i - \frac{1}{2} \right) \left(x_{(i)} - x_{(n-i+1)} \right)}{\sqrt{n^3 \sum_{i=1}^n (x_i - \bar{x})^2}}$$

and the observations $x_1, x_2, \dots, x_n$ are ordered as $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Reject the hypothesis of normality if T is greater than an upper percentile or less than a lower percentile.

Percentiles of T										
n	2	1.8	1.5	1	.5	.05	.01	.005	.001	.0005
10	-4.05	-4.05	-3.25	-2.62	-2.09	1.49	1.55	1.69	1.79	1.85
12	-4.05	-4.05	-3.25	-2.66	-2.14	1.57	1.59	1.73	1.83	1.89
14	-4.07	-4.07	-3.26	-2.71	-2.19	1.65	1.66	1.80	1.90	1.96
16	-4.10	-4.10	-3.31	-2.76	-2.24	1.67	1.69	1.83	1.93	1.99
18	-4.17	-4.17	-3.38	-2.83	-2.31	1.77	1.78	1.91	1.99	2.05
20	-4.41	-4.41	-3.64	-2.94	-2.44	1.82	1.83	1.96	2.04	2.10
22	-4.55	-4.55	-3.78	-3.01	-2.51	1.87	1.88	1.99	2.08	2.14
24	-4.55	-4.55	-3.75	-3.06	-2.56	1.90	1.91	2.02	2.10	2.16
26	-4.57	-4.57	-3.76	-3.07	-2.57	1.91	1.92	2.03	2.11	2.17
28	-4.58	-4.58	-3.77	-3.08	-2.58	1.92	1.93	2.04	2.12	2.18
30	-4.59	-4.59	-3.78	-3.09	-2.59	1.93	1.94	2.05	2.13	2.19
32	-4.59	-4.59	-3.78	-3.10	-2.60	1.94	1.95	2.06	2.14	2.20
34	-4.61	-4.61	-3.79	-3.10	-2.61	1.95	1.96	2.07	2.15	2.21
36	-4.63	-4.63	-3.81	-3.12	-2.63	1.96	1.97	2.08	2.16	2.22
38	-4.65	-4.65	-3.83	-3.13	-2.65	1.97	1.98	2.09	2.17	2.23
40	-4.65	-4.65	-3.83	-3.14	-2.66	1.98	1.99	2.10	2.18	2.24
42	-4.66	-4.66	-3.84	-3.15	-2.67	1.99	2.00	2.11	2.19	2.25
44	-4.68	-4.67	-3.78	-3.14	-2.68	1.99	2.01	2.11	2.19	2.25
46	-4.69	-4.68	-3.77	-3.13	-2.67	1.99	2.01	2.11	2.19	2.25
48	-4.69	-4.68	-3.77	-3.13	-2.67	1.99	2.01	2.11	2.19	2.25
50	-4.69	-4.68	-3.77	-3.13	-2.67	1.99	2.01	2.11	2.19	2.25
55	-4.73	-4.73	-3.81	-3.17	-2.71	2.03	2.04	2.15	2.23	2.29
60	-4.77	-4.77	-3.85	-3.21	-2.75	2.07	2.08	2.19	2.27	2.33
65	-4.81	-4.81	-3.89	-3.25	-2.79	2.11	2.12	2.23	2.31	2.37
70	-4.85	-4.85	-3.93	-3.29	-2.83	2.15	2.16	2.27	2.35	2.41
75	-4.89	-4.89	-3.97	-3.33	-2.87	2.19	2.20	2.31	2.39	2.45
80	-4.93	-4.93	-4.01	-3.37	-2.91	2.23	2.24	2.35	2.43	2.49
85	-4.97	-4.97	-4.05	-3.41	-2.95	2.27	2.28	2.39	2.47	2.53
90	-5.01	-5.01	-4.09	-3.45	-2.99	2.31	2.32	2.43	2.51	2.57
100	-5.07	-5.14	-4.14	-3.47	-3.07	2.39	2.44	2.55	2.63	2.69
150	-5.609	-5.609	-4.612	-3.684	-3.038	2.89	2.93	3.03	3.11	3.16
200	-5.802	-5.822	-4.811	-3.880	-3.241	3.33	3.38	3.48	3.56	3.61
250	-5.977	-5.994	-4.988	-4.026	-3.471	3.69	3.74	3.84	3.92	3.97

Table A.4: Upper Percentiles for $k_{\lambda, \mu}$ and Upper and Lower Percentiles for $k_{\lambda, \mu}^*$

Upper Percentiles for $k_{\lambda, \mu}$										Upper and Lower Percentiles for $k_{\lambda, \mu}^*$									
$\mu = 2$										$\mu = 2$									
Percentiles										Percentiles									
α	99	97.5	95	92.5	90	87.5	85	82.5	80	α	1	1.1	5	10	50	95	97.5	99	α
2	2.07	2.26	2.59	3.04	3.74	4.64	5.89	7.64	10.04	32	4.86	4.73	4.07	3.07	1.86	0.95	0.78	0.70	10.99
12	2.83	3.04	3.50	4.08	4.98	6.34	8.24	10.74	14.04	33	4.73	4.60	3.93	2.93	1.72	0.80	0.63	0.55	10.99
14	2.97	3.19	3.66	4.26	5.18	6.56	8.48	11.00	14.34	34	4.62	4.49	3.81	2.81	1.60	0.76	0.59	0.51	11.00
16	3.12	3.34	3.82	4.44	5.38	6.78	8.72	11.24	14.64	35	4.51	4.38	3.70	2.70	1.48	0.72	0.55	0.47	11.00
18	3.28	3.50	3.99	4.62	5.58	6.99	8.94	11.44	14.94	36	4.40	4.27	3.59	2.59	1.36	0.69	0.52	0.44	11.00
20	3.44	3.66	4.16	4.80	5.78	7.20	9.16	11.64	15.24	37	4.30	4.17	3.48	2.48	1.24	0.66	0.49	0.41	11.00
25	3.84	4.08	4.60	5.24	6.24	7.68	9.64	11.84	15.54	38	4.20	4.07	3.38	2.38	1.12	0.63	0.46	0.38	11.00
30	4.24	4.49	5.02	5.68	6.68	8.14	10.10	12.04	15.84	39	4.10	3.97	3.28	2.28	1.00	0.60	0.43	0.35	11.00
40	4.84	5.10	5.64	6.32	7.34	8.80	10.76	12.74	16.14	40	4.00	3.87	3.18	2.18	0.88	0.57	0.40	0.32	11.00
50	5.44	5.70	6.24	6.94	7.98	9.44	11.40	13.34	16.44	45	3.70	3.57	2.88	1.88	0.76	0.53	0.37	0.29	11.00
60	6.04	6.30	6.84	7.54	8.58	10.04	12.00	14.04	16.74	50	3.40	3.27	2.58	1.58	0.64	0.50	0.34	0.26	11.00
70	6.64	6.90	7.44	8.14	9.18	10.64	12.60	14.64	17.04	60	3.00	2.87	2.18	1.18	0.52	0.48	0.32	0.24	11.00
80	7.24	7.50	8.04	8.74	9.78	11.24	13.20	15.24	17.34	70	2.60	2.47	1.78	0.78	0.40	0.46	0.30	0.22	11.00
90	7.84	8.10	8.64	9.34	10.38	11.84	13.80	15.84	17.64	80	2.20	2.07	1.38	0.38	0.28	0.44	0.28	0.20	11.00
95	8.24	8.50	9.04	9.74	10.78	12.24	14.20	16.24	17.94	90	1.80	1.67	0.98	0.18	0.24	0.40	0.24	0.16	11.00
97.5	8.54	8.80	9.34	10.04	11.08	12.54	14.50	16.54	18.24	95	1.40	1.27	0.58	0.08	0.14	0.36	0.20	0.12	11.00
99	8.84	9.10	9.64	10.34	11.38	12.84	14.80	16.84	18.54	97.5	1.00	0.87	0.18	0.00	0.00	0.32	0.16	0.08	11.00

(continued)

Table A.3. (Continued)

$\hat{p} = 1$										$\hat{p} = 2$									
Upper Boundaries for $\hat{\theta}_{1,2}$										Upper and Lower Boundaries for $\hat{\theta}_{1,2}$									
Percentiles										Percentiles									
n	99	97.5	95	92.5	90	99	1	2.5	5	10	90	95	97.5	99					
500	1.00	1.00	.99	.98	.96	100	2.364	2.573	2.811	3.086	3.401	3.750	4.134	4.542					
1000	1.00	1.00	.99	.98	.95	1500	2.365	2.433	2.504	2.583	2.676	2.785	2.909	3.050					
1500	.99	.99	.98	.98	.94	2000	2.366	2.397	2.433	2.473	2.516	2.563	2.613	2.673					
2000	.99	.98	.97	.97	.93	2500	2.367	2.389	2.426	2.466	2.508	2.553	2.601	2.661					
3000	.98	.97	.95	.95	.91	3000	2.368	2.381	2.418	2.458	2.500	2.543	2.589	2.647					
4000	.97	.96	.94	.94	.90	3500	2.369	2.372	2.409	2.449	2.489	2.532	2.577	2.634					
5000	.96	.95	.93	.93	.89	4000	2.370	2.363	2.399	2.439	2.479	2.521	2.565	2.621					
						5000	2.370	2.348	2.376	2.414	2.455	2.500	2.544	2.599					

$\hat{p} = 3$										$\hat{p} = 3$									
Upper Boundaries for $\hat{\theta}_{1,2}$										Upper and Lower Boundaries for $\hat{\theta}_{1,2}$									
Percentiles										Percentiles									
n	99	97.5	95	92.5	90	99	1	2.5	5	10	90	95	97.5	99					
10	4.8	4.5	4.3	4.1	3.8	10	18.8	18.2	18.4	18.7	14.0	14.6	14.8	15.6					
12	5.2	4.8	4.4	4.1	3.8	12	18.2	18.4	18.7	11.0	16.7	15.2	15.9	16.4					
14	5.6	5.1	4.7	4.4	4.1	14	18.4	18.6	18.8	11.3	16.1	15.8	16.5	17.1					
16	6.0	5.5	5.1	4.8	4.5	16	18.5	18.8	11.1	11.5	15.4	16.1	16.8	17.5					
18	6.3	5.8	5.4	5.1	4.8	18	18.7	19.0	11.3	11.6	15.5	16.4	17.1	17.8					
20	6.6	6.1	5.7	5.4	5.1	20	18.8	19.1	11.4	11.8	15.7	16.5	17.2	18.0					
25	7.0	6.4	6.0	5.6	5.3	25	19.1	19.4	11.6	12.1	15.9	16.7	17.6	18.3					
30	7.4	6.8	6.3	5.9	5.5	30	19.5	19.8	11.8	12.3	16.0	16.7	17.8	18.5					
40	7.9	7.2	6.6	6.2	5.8	40	19.7	20.0	11.9	12.4	16.1	16.7	17.9	18.6					
50	8.1	7.4	6.8	6.4	6.0	50	19.8	20.2	11.9	12.5	16.1	16.7	17.9	18.6					

(Continued)

Table A.1. (Continued)

$p = 3$									
Upper Percentiles for $h_{3,2}$					Upper and Lower Percentiles for $h_{3,2}$				
Percentiles					Percentiles				
α	99	97.5	95	92.5	90	88	85	82.5	80
50	1.6	1.6	1.6	1.6	1.6	15.1	13.1	10.1	7.9
75	1.4	1.4	1.4	1.4	1.4	13.6	11.6	8.6	6.7
100	1.3	1.3	1.3	1.3	1.3	12.6	10.6	7.6	5.7
150	1.08	1.08	1.08	1.08	1.08	10.2	8.2	5.2	3.3
200	.92	.92	.92	.92	.92	8.8	6.8	3.8	1.9
250	.85	.85	.85	.85	.85	8.0	6.0	3.0	1.4
300	.80	.80	.80	.80	.80	7.4	5.4	2.4	1.0
350	.75	.75	.75	.75	.75	6.9	4.9	1.9	.75
400	.71	.71	.71	.71	.71	6.5	4.5	1.5	.68
450	.67	.67	.67	.67	.67	6.1	4.1	1.1	.63
500	.64	.64	.64	.64	.64	5.8	3.8	1.0	.60
550	.61	.61	.61	.61	.61	5.5	3.5	.96	.57
600	.59	.59	.59	.59	.59	5.2	3.2	.92	.55
650	.57	.57	.57	.57	.57	5.0	3.0	.89	.53
700	.55	.55	.55	.55	.55	4.8	2.8	.86	.51
750	.54	.54	.54	.54	.54	4.6	2.6	.84	.49
800	.53	.53	.53	.53	.53	4.5	2.5	.83	.48
850	.52	.52	.52	.52	.52	4.4	2.4	.82	.47
900	.51	.51	.51	.51	.51	4.3	2.3	.81	.46
950	.50	.50	.50	.50	.50	4.2	2.2	.80	.45
1000	.49	.49	.49	.49	.49	4.1	2.1	.79	.44
1050	.48	.48	.48	.48	.48	4.0	2.0	.78	.43
1100	.47	.47	.47	.47	.47	3.9	1.9	.77	.42
1150	.46	.46	.46	.46	.46	3.8	1.8	.76	.41
1200	.45	.45	.45	.45	.45	3.7	1.7	.75	.40
1250	.44	.44	.44	.44	.44	3.6	1.6	.74	.39
1300	.43	.43	.43	.43	.43	3.5	1.5	.73	.38
1350	.42	.42	.42	.42	.42	3.4	1.4	.72	.37
1400	.41	.41	.41	.41	.41	3.3	1.3	.71	.36
1450	.40	.40	.40	.40	.40	3.2	1.2	.70	.35
1500	.39	.39	.39	.39	.39	3.1	1.1	.69	.34
1550	.38	.38	.38	.38	.38	3.0	1.0	.68	.33
1600	.37	.37	.37	.37	.37	2.9	.99	.67	.32
1650	.36	.36	.36	.36	.36	2.8	.98	.66	.31
1700	.35	.35	.35	.35	.35	2.7	.97	.65	.30
1750	.34	.34	.34	.34	.34	2.6	.96	.64	.29
1800	.33	.33	.33	.33	.33	2.5	.95	.63	.28
1850	.32	.32	.32	.32	.32	2.4	.94	.62	.27
1900	.31	.31	.31	.31	.31	2.3	.93	.61	.26
1950	.30	.30	.30	.30	.30	2.2	.92	.60	.25
2000	.29	.29	.29	.29	.29	2.1	.91	.59	.24
2050	.28	.28	.28	.28	.28	2.0	.90	.58	.23
2100	.27	.27	.27	.27	.27	1.9	.89	.57	.22
2150	.26	.26	.26	.26	.26	1.8	.88	.56	.21
2200	.25	.25	.25	.25	.25	1.7	.87	.55	.20
2250	.24	.24	.24	.24	.24	1.6	.86	.54	.19
2300	.23	.23	.23	.23	.23	1.5	.85	.53	.18
2350	.22	.22	.22	.22	.22	1.4	.84	.52	.17
2400	.21	.21	.21	.21	.21	1.3	.83	.51	.16
2450	.20	.20	.20	.20	.20	1.2	.82	.50	.15
2500	.19	.19	.19	.19	.19	1.1	.81	.49	.14
2550	.18	.18	.18	.18	.18	1.0	.80	.48	.13
2600	.17	.17	.17	.17	.17	.99	.79	.47	.12
2650	.16	.16	.16	.16	.16	.98	.78	.46	.11
2700	.15	.15	.15	.15	.15	.97	.77	.45	.10
2750	.14	.14	.14	.14	.14	.96	.76	.44	.09
2800	.13	.13	.13	.13	.13	.95	.75	.43	.08
2850	.12	.12	.12	.12	.12	.94	.74	.42	.07
2900	.11	.11	.11	.11	.11	.93	.73	.41	.06
2950	.10	.10	.10	.10	.10	.92	.72	.40	.05
3000	.09	.09	.09	.09	.09	.91	.71	.39	.04
3050	.08	.08	.08	.08	.08	.90	.70	.38	.03
3100	.07	.07	.07	.07	.07	.89	.69	.37	.02
3150	.06	.06	.06	.06	.06	.88	.68	.36	.01
3200	.05	.05	.05	.05	.05	.87	.67	.35	.00
3250	.04	.04	.04	.04	.04	.86	.66	.34	.00
3300	.03	.03	.03	.03	.03	.85	.65	.33	.00
3350	.02	.02	.02	.02	.02	.84	.64	.32	.00
3400	.01	.01	.01	.01	.01	.83	.63	.31	.00
3450	.00	.00	.00	.00	.00	.82	.62	.30	.00
3500	.00	.00	.00	.00	.00	.81	.61	.29	.00
3550	.00	.00	.00	.00	.00	.80	.60	.28	.00
3600	.00	.00	.00	.00	.00	.79	.59	.27	.00
3650	.00	.00	.00	.00	.00	.78	.58	.26	.00
3700	.00	.00	.00	.00	.00	.77	.57	.25	.00
3750	.00	.00	.00	.00	.00	.76	.56	.24	.00
3800	.00	.00	.00	.00	.00	.75	.55	.23	.00
3850	.00	.00	.00	.00	.00	.74	.54	.22	.00
3900	.00	.00	.00	.00	.00	.73	.53	.21	.00
3950	.00	.00	.00	.00	.00	.72	.52	.20	.00
4000	.00	.00	.00	.00	.00	.71	.51	.19	.00
4050	.00	.00	.00	.00	.00	.70	.50	.18	.00
4100	.00	.00	.00	.00	.00	.69	.49	.17	.00
4150	.00	.00	.00	.00	.00	.68	.48	.16	.00
4200	.00	.00	.00	.00	.00	.67	.47	.15	.00
4250	.00	.00	.00	.00	.00	.66	.46	.14	.00
4300	.00	.00	.00	.00	.00	.65	.45	.13	.00
4350	.00	.00	.00	.00	.00	.64	.44	.12	.00
4400	.00	.00	.00	.00	.00	.63	.43	.11	.00
4450	.00	.00	.00	.00	.00	.62	.42	.10	.00
4500	.00	.00	.00	.00	.00	.61	.41	.09	.00
4550	.00	.00	.00	.00	.00	.60	.40	.08	.00
4600	.00	.00	.00	.00	.00	.59	.39	.07	.00
4650	.00	.00	.00	.00	.00	.58	.38	.06	.00
4700	.00	.00	.00	.00	.00	.57	.37	.05	.00
4750	.00	.00	.00	.00	.00	.56	.36	.04	.00
4800	.00	.00	.00	.00	.00	.55	.35	.03	.00
4850	.00	.00	.00	.00	.00	.54	.34	.02	.00
4900	.00	.00	.00	.00	.00	.53	.33	.01	.00
4950	.00	.00	.00	.00	.00	.52	.32	.00	.00
5000	.00	.00	.00	.00	.00	.51	.31	.00	.00

(continued)

Table A.4. (Continued)

$\beta = 4$										$\beta = 4$									
Upper bound for $\delta_{1,2}$										Upper and lower Perceptron for $\delta_{1,2}$									
Perceptron										Perceptron									
α	0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	α	1	2.0	3	4	5	6	7	8	9
14	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	14	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	15	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	16	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	17	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	18	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	19	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	20	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
21	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	21	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	22	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	23	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	24	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	25	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	26	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
27	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	27	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	28	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	29	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	30	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	31	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	32	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
33	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	33	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	34	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	35	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	36	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
37	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	37	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
38	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	38	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
39	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	39	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
40	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	40	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
41	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	41	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
42	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	42	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
43	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	43	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
44	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	44	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
45	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	45	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
46	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	46	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
47	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	47	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
48	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	48	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	49	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	50	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table A.6. Upper Percentiles for Test of Single Multivariate Normal Outlier

Upper percentage points for the test statistic

$$J_{n,p}^2 = \max_{1 \leq i \leq n} (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{W}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}})$$

This test for a single outlier is a sample of size n from a multivariate normal distribution. Reject and conclude that the outlier is significant if $J_{n,p}^2$ exceeds the table value.

n	$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
5	6.17	8.19						
6	6.80	8.11	4.34	4.36				
7	6.71	4.95	3.04	3.30	3.32	3.14		
8	5.30	5.30	5.77	5.97	6.01	6.09	6.11	6.12
9	5.85	6.37	6.43	6.76	6.88	6.97	7.38	7.68
10	6.50	6.97	7.01	7.47	7.56	7.79	7.82	7.96
12	7.19	8.00	7.96	8.76	8.67	8.98	9.19	9.37
14	7.76	8.84	8.78	9.71	9.61	10.07	10.29	10.60
16	8.27	9.54	9.44	10.56	10.39	11.06	11.26	12.00
18	8.79	10.15	10.00	11.28	11.06	12.28	11.86	12.68
20	9.12	10.67	10.49	11.91	11.64	12.89	12.62	13.81
25	9.96	11.75	11.48	13.38	12.78	14.48	13.86	15.47
30	10.58	12.54	12.24	14.34	13.67	15.51	14.95	16.75
35	11.19	13.20	12.85	14.93	14.37	16.48	15.75	17.75
40	11.59	13.74	13.36	15.36	14.96	17.13	16.41	18.55
45	11.88	14.20	13.80	15.76	15.46	17.74	16.97	19.24
50	12.20	14.60	14.18	16.16	15.89	18.27	17.46	19.83
100	14.22	16.95	16.43	19.26	18.43	21.08	20.26	23.17
200	15.89	18.94	18.42	21.47	20.39	23.72	22.59	25.82
500	18.12	21.22	20.75	23.93	23.06	26.57	25.21	28.62

Table A.2. Upper Percentile Points of Marshall's β^2 Distribution

Degrees of Freedom, ν	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
1	36.411									
2	30.135	25.000								
3	27.505	24.472	14.868							
4	4.628	17.541	16.562	12.148						
5	3.987	13.607	20.491	73.475	26.646					
6	3.381	12.053	22.179	44.718	126.117	405.679				
7	3.138	10.009	19.529	33.138	62.503	149.059	349.612			
8	3.117	10.005	18.768	31.362	45.416	81.592	186.612	467.156		
10	4.593	9.449	15.348	21.545	36.565	59.465	116.146	193.913	467.517	1089.774
11	4.644	9.095	14.615	21.085	33.568	47.175	79.888	133.945	204.005	
12	4.747	8.660	13.958	19.178	31.046	39.768	69.697	111.012	161.947	399.426
13	4.897	8.148	13.274	18.088	28.148	36.941	64.573	111.698	161.695	399.426
14	4.995	8.077	12.714	17.089	25.293	34.488	62.881	79.432	86.979	152.962
15	4.943	8.001	11.888	16.248	23.045	29.855	58.475	31.972	70.979	146.488
16	4.944	7.955	11.495	15.451	20.505	25.638	53.117	45.612	65.546	80.178
17	4.451	7.752	11.175	15.115	19.762	25.465	51.588	41.173	54.841	75.155
18	4.434	7.606	10.651	14.465	19.071	24.795	49.395	38.395	49.050	69.268
19	4.381	7.504	10.273	14.383	18.171	23.146	38.575	36.083	45.055	58.509
20	4.331	7.444	10.003	13.612	17.626	22.623	37.623	34.658	41.958	56.463
24	4.123	7.166	10.079	13.465	17.146	21.868	34.458	33.468	38.448	54.143
25	4.305	7.268	10.125	13.465	20.605	21.578	33.578	30.693	37.617	52.762
26	4.401	7.300	10.065	13.464	20.605	20.605	32.578	29.598	36.826	52.076
28	4.346	7.146	9.957	13.465	20.605	19.958	31.578	28.571	35.971	50.976
30	4.352	7.060	9.854	13.465	20.605	19.465	31.578	28.571	35.971	50.976

Table A.2. (Continued)

Degree of Freedom ν	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
$\alpha = .05$										
1	98.501					200.000	279.150	321.805	352.321	374.512
2	94.176	37.455				440.571	581.176	675.816	745.518	799.145
3	91.196	81.171	344.457			555.235	725.935	837.885	921.835	975.512
4	88.258	43.008	147.563	493.454		625.932	825.932	957.882	1071.832	1135.512
5	85.258	33.857	78.123	128.674	448.469					
6	82.258	25.491	36.552	119.416	129.416	200.000				
7	81.259	21.821	26.178	71.909	111.219	440.571	579.176			
8	80.261	18.468	17.358	34.952	58.952	555.235	725.935	837.885		
9	79.264	15.625	13.465	44.016	44.016	625.932	825.932	957.882	1071.832	
10	78.266	13.631	11.671	36.111	36.111	675.816	875.816	997.816	1111.816	1185.512
11	77.268	12.411	10.671	30.111	30.111	725.932	925.932	1047.832	1161.832	1235.512
12	76.269	11.712	10.008	24.211	24.211	775.816	975.816	1097.816	1211.816	1285.512
13	75.274	11.008	9.361	21.171	21.171	825.932	1025.932	1147.832	1261.832	1335.512
14	74.282	10.411	8.824	18.071	18.071	875.816	1075.816	1197.816	1311.816	1385.512
15	73.293	9.898	8.367	15.967	15.967	925.932	1125.932	1247.832	1361.832	1435.512
16	72.311	9.458	7.956	14.052	14.052	975.816	1175.816	1297.816	1411.816	1485.512
17	71.336	9.111	7.478	12.416	12.416	1025.932	1225.932	1347.832	1461.832	1535.512
18	70.365	8.743	7.053	11.071	11.071	1075.816	1275.816	1397.832	1511.832	1585.512
19	69.395	8.454	6.708	10.052	10.052	1125.932	1325.932	1447.832	1561.832	1635.512
20	68.428	8.149	6.411	9.171	9.171	1175.816	1375.816	1497.832	1611.832	1685.512
21	67.463	7.828	6.111	8.416	8.416	1225.932	1425.932	1547.832	1661.832	1735.512
22	66.501	7.491	5.828	7.767	7.767	1275.816	1475.816	1597.832	1711.832	1785.512
23	65.541	7.148	5.558	7.211	7.211	1325.932	1525.932	1647.832	1761.832	1835.512
24	64.583	6.798	5.298	6.741	6.741	1375.816	1575.816	1697.832	1811.832	1885.512
25	63.628	6.441	5.048	6.361	6.361	1425.932	1625.932	1747.832	1861.832	1935.512
26	62.676	6.078	4.808	6.061	6.061	1475.816	1675.816	1797.832	1911.832	1985.512
27	61.726	5.711	4.578	5.771	5.771	1525.932	1725.932	1847.832	1961.832	2035.512
28	60.778	5.341	4.358	5.471	5.471	1575.816	1775.816	1897.832	2011.832	2085.512
29	59.833	4.968	4.148	5.171	5.171	1625.932	1825.932	1947.832	2061.832	2135.512
30	58.891	4.591	3.948	4.871	4.871	1675.816	1875.816	1997.832	2111.832	2185.512
31	57.951	4.211	3.758	4.571	4.571	1725.932	1925.932	2047.832	2161.832	2235.512
32	57.013	3.828	3.578	4.271	4.271	1775.816	1975.816	2097.832	2211.832	2285.512
33	56.078	3.441	3.408	3.971	3.971	1825.932	2025.932	2147.832	2261.832	2335.512
34	55.145	3.051	3.248	3.671	3.671	1875.816	2075.816	2197.832	2311.832	2385.512
35	54.214	2.658	3.098	3.371	3.371	1925.932	2125.932	2247.832	2361.832	2435.512
36	53.285	2.261	2.958	3.071	3.071	1975.816	2175.816	2297.832	2411.832	2485.512
37	52.358	1.861	2.828	2.771	2.771	2025.932	2225.932	2347.832	2461.832	2535.512
38	51.433	1.458	2.708	2.471	2.471	2075.816	2275.816	2397.832	2511.832	2585.512
39	50.511	1.051	2.598	2.171	2.171	2125.932	2325.932	2447.832	2561.832	2635.512
40	49.591	0.641	2.498	1.871	1.871	2175.816	2375.816	2497.832	2611.832	2685.512

[illegible]

Table A.8. Bonferroni t -Values, $t_{\alpha/2, n-1, \alpha}$, $n = 10$

α	t									
	1	2	3	4	5	6	7	8	9	10
	10000%									
1	3.0000	2.5000	1.6667	1.2500	1.0000	.8555	.7643	.6940	.6396	.5900
2	4.3027	4.2050	3.4488	2.8600	2.4248	2.0859	1.8190	1.5997	1.3894	1.2090
3	5.3824	5.1750	4.0367	3.3447	2.8489	2.4243	2.0707	1.7832	1.5149	1.2811
4	6.1764	5.8754	4.6088	4.0447	3.4840	3.0540	2.6679	2.2611	1.9066	1.5970
5	6.7760	6.4504	5.1344	4.5900	4.0320	3.5716	3.1618	2.7271	2.3050	1.9311
6	7.2449	6.9007	5.5879	5.0512	4.4874	3.9950	3.5571	3.1152	2.7089	2.3168
7	7.6146	7.2442	6.0376	5.5555	4.9903	4.4558	3.9727	3.5352	3.0867	2.6901
8	7.9169	7.5119	6.3198	5.8380	5.2634	4.7180	4.1944	3.6946	3.2166	2.8324
9	8.1622	7.7350	6.5333	6.1139	5.5398	4.9842	4.4619	3.9619	3.4719	3.0897
10	8.3681	7.9198	6.7381	6.3082	5.7343	5.1788	4.6482	4.1477	3.6582	3.2814
11	8.5410	8.0694	6.9289	6.4809	5.9070	5.3515	4.8209	4.3202	3.8306	3.4000
12	8.6888	8.1900	7.0793	6.6345	6.0545	5.4997	4.9597	4.4589	3.9694	3.4984
13	8.8164	8.3026	7.2459	6.8091	6.2313	5.6700	5.1301	4.6293	4.1397	3.6115
14	8.9288	8.3950	7.3776	6.9540	6.3768	5.8068	5.2664	4.7655	4.2757	3.7257
15	9.0274	8.4809	7.4937	7.0866	6.5067	5.9368	5.3918	4.8907	4.3946	3.8360
16	9.1149	8.5579	7.6038	7.2111	6.6308	6.0603	5.5153	5.0144	4.5135	3.9430
17	9.1938	8.6281	7.6999	7.3298	6.7482	6.1820	5.6369	5.1359	4.6350	4.0484
18	9.2659	8.6920	7.7891	7.4435	6.8704	6.3027	5.7576	5.2566	4.7556	4.1530
19	9.3318	8.7504	7.8721	7.5530	6.9889	6.4189	5.8736	5.3728	4.8726	4.2571
20	9.3916	8.8034	7.9496	7.6584	7.1043	6.5307	5.9858	5.4850	4.9797	4.3594
21	9.4464	8.8516	8.0221	7.7595	7.2164	6.6371	6.0929	5.5921	5.0879	4.4611
22	9.4973	8.8955	8.0903	7.8564	7.3258	6.7405	6.1959	5.6954	5.1917	4.5618
23	9.5444	8.9355	8.1546	7.9500	7.4327	6.8405	6.2953	5.7954	5.2927	4.6610
24	9.5884	8.9719	8.2156	8.0403	7.5375	6.9373	6.3911	5.8921	5.3909	4.7584
25	9.6294	9.0056	8.2736	8.1274	7.6404	7.0319	6.4838	5.9866	5.4869	4.8544
26	9.6674	9.0366	8.3286	8.2116	7.7416	7.1245	6.5736	6.0791	5.5809	4.9491
27	9.7024	9.0649	8.3806	8.2930	7.8404	7.2153	6.6616	6.1696	5.6729	5.0424
28	9.7344	9.0906	8.4296	8.3716	7.9374	7.3045	6.7476	6.2581	5.7639	5.1344
29	9.7634	9.1139	8.4756	8.4474	8.0324	7.3921	6.8321	6.3446	5.8539	5.2254
30	9.7894	9.1349	8.5186	8.5204	8.1254	7.4781	6.9151	6.4291	5.9429	5.3154

Table A.8. (Continued)

ν	ξ									
	1	2	3	4	5	6	7	8	9	10
	1000/2									
ν	5.0000	5.5000	6.0000	6.5000	7.0000	7.500	8.000	8.500	9.000	9.500
27	2.853.8	2.8734	2.8939	2.9143	2.9347	2.9551	2.9755	2.9959	3.0163	3.0366
28	2.9484	2.9685	2.9885	2.9985	3.0185	3.0385	3.0585	3.0785	3.0985	3.1185
29	3.0432	3.0632	3.0832	3.0932	3.1132	3.1332	3.1532	3.1732	3.1932	3.2132
30	3.1380	3.1580	3.1780	3.1880	3.2080	3.2280	3.2480	3.2680	3.2880	3.3080
35	3.3428	3.3628	3.3828	3.3928	3.4128	3.4328	3.4528	3.4728	3.4928	3.5128
40	3.5476	3.5676	3.5876	3.5976	3.6176	3.6376	3.6576	3.6776	3.6976	3.7176
45	3.7524	3.7724	3.7924	3.8024	3.8224	3.8424	3.8624	3.8824	3.9024	3.9224
50	3.8972	3.9172	3.9372	3.9472	3.9672	3.9872	4.0072	4.0272	4.0472	4.0672
55	4.0420	4.0620	4.0820	4.0920	4.1120	4.1320	4.1520	4.1720	4.1920	4.2120
60	4.1868	4.2068	4.2268	4.2368	4.2568	4.2768	4.2968	4.3168	4.3368	4.3568
65	4.2816	4.3016	4.3216	4.3316	4.3516	4.3716	4.3916	4.4116	4.4316	4.4516
70	4.3764	4.3964	4.4164	4.4264	4.4464	4.4664	4.4864	4.5064	4.5264	4.5464
75	4.4712	4.4912	4.5112	4.5212	4.5412	4.5612	4.5812	4.6012	4.6212	4.6412
80	4.5660	4.5860	4.6060	4.6160	4.6360	4.6560	4.6760	4.6960	4.7160	4.7360
85	4.6608	4.6808	4.7008	4.7108	4.7308	4.7508	4.7708	4.7908	4.8108	4.8308
90	4.7556	4.7756	4.7956	4.8056	4.8256	4.8456	4.8656	4.8856	4.9056	4.9256
95	4.8504	4.8704	4.8904	4.9004	4.9204	4.9404	4.9604	4.9804	5.0004	5.0204
100	4.9452	4.9652	4.9852	4.9952	5.0152	5.0352	5.0552	5.0752	5.0952	5.1152
110	5.0400	5.0600	5.0800	5.0900	5.1100	5.1300	5.1500	5.1700	5.1900	5.2100
120	5.1348	5.1548	5.1748	5.1848	5.2048	5.2248	5.2448	5.2648	5.2848	5.3048
130	5.2296	5.2496	5.2696	5.2796	5.2996	5.3196	5.3396	5.3596	5.3796	5.3996
140	5.3244	5.3444	5.3644	5.3744	5.3944	5.4144	5.4344	5.4544	5.4744	5.4944
150	5.4192	5.4392	5.4592	5.4692	5.4892	5.5092	5.5292	5.5492	5.5692	5.5892
160	5.5140	5.5340	5.5540	5.5640	5.5840	5.6040	5.6240	5.6440	5.6640	5.6840
170	5.6088	5.6288	5.6488	5.6588	5.6788	5.6988	5.7188	5.7388	5.7588	5.7788
180	5.7036	5.7236	5.7436	5.7536	5.7736	5.7936	5.8136	5.8336	5.8536	5.8736
190	5.7984	5.8184	5.8384	5.8484	5.8684	5.8884	5.9084	5.9284	5.9484	5.9684
200	5.8932	5.9132	5.9332	5.9432	5.9632	5.9832	6.0032	6.0232	6.0432	6.0632
210	5.9880	6.0080	6.0280	6.0380	6.0580	6.0780	6.0980	6.1180	6.1380	6.1580
220	6.0828	6.1028	6.1228	6.1328	6.1528	6.1728	6.1928	6.2128	6.2328	6.2528
230	6.1776	6.1976	6.2176	6.2276	6.2476	6.2676	6.2876	6.3076	6.3276	6.3476
240	6.2724	6.2924	6.3124	6.3224	6.3424	6.3624	6.3824	6.4024	6.4224	6.4424
250	6.3672	6.3872	6.4072	6.4172	6.4372	6.4572	6.4772	6.4972	6.5172	6.5372
260	6.4620	6.4820	6.5020	6.5120	6.5320	6.5520	6.5720	6.5920	6.6120	6.6320
270	6.5568	6.5768	6.5968	6.6068	6.6268	6.6468	6.6668	6.6868	6.7068	6.7268
280	6.6516	6.6716	6.6916	6.7016	6.7216	6.7416	6.7616	6.7816	6.8016	6.8216
290	6.7464	6.7664	6.7864	6.7964	6.8164	6.8364	6.8564	6.8764	6.8964	6.9164
300	6.8412	6.8612	6.8812	6.8912	6.9112	6.9312	6.9512	6.9712	6.9912	7.0112
310	6.9360	6.9560	6.9760	6.9860	7.0060	7.0260	7.0460	7.0660	7.0860	7.1060
320	7.0308	7.0508	7.0708	7.0808	7.1008	7.1208	7.1408	7.1608	7.1808	7.2008
330	7.1256	7.1456	7.1656	7.1756	7.1956	7.2156	7.2356	7.2556	7.2756	7.2956
340	7.2204	7.2404	7.2604	7.2704	7.2904	7.3104	7.3304	7.3504	7.3704	7.3904
350	7.3152	7.3352	7.3552	7.3652	7.3852	7.4052	7.4252	7.4452	7.4652	7.4852
360	7.4100	7.4300	7.4500	7.4600	7.4800	7.5000	7.5200	7.5400	7.5600	7.5800
370	7.5048	7.5248	7.5448	7.5548	7.5748	7.5948	7.6148	7.6348	7.6548	7.6748
380	7.5996	7.6196	7.6396	7.6496	7.6696	7.6896	7.7096	7.7296	7.7496	7.7696
390	7.6944	7.7144	7.7344	7.7444	7.7644	7.7844	7.8044	7.8244	7.8444	7.8644
400	7.7892	7.8092	7.8292	7.8392	7.8592	7.8792	7.8992	7.9192	7.9392	7.9592
410	7.8840	7.9040	7.9240	7.9340	7.9540	7.9740	7.9940	8.0140	8.0340	8.0540
420	7.9788	7.9988	8.0188	8.0288	8.0488	8.0688	8.0888	8.1088	8.1288	8.1488
430	8.0736	8.0936	8.1136	8.1236	8.1436	8.1636	8.1836	8.2036	8.2236	8.2436
440	8.1684	8.1884	8.2084	8.2184	8.2384	8.2584	8.2784	8.2984	8.3184	8.3384
450	8.2632	8.2832	8.3032	8.3132	8.3332	8.3532	8.3732	8.3932	8.4132	8.4332
460	8.3580	8.3780	8.3980	8.4080	8.4280	8.4480	8.4680	8.4880	8.5080	8.5280
470	8.4528	8.4728	8.4928	8.5028	8.5228	8.5428	8.5628	8.5828	8.6028	8.6228
480	8.5476	8.5676	8.5876	8.5976	8.6176	8.6376	8.6576	8.6776	8.6976	8.7176
490	8.6424	8.6624	8.6824	8.6924	8.7124	8.7324	8.7524	8.7724	8.7924	8.8124
500	8.7372	8.7572	8.7772	8.7872	8.8072	8.8272	8.8472	8.8672	8.8872	8.9072
510	8.8320	8.8520	8.8720	8.8820	8.9020	8.9220	8.9420	8.9620	8.9820	9.0020
520	8.9268	8.9468	8.9668	8.9768	8.9968	9.0168	9.0368	9.0568	9.0768	9.0968
530	9.0216	9.0416	9.0616	9.0716	9.0916	9.1116	9.1316	9.1516	9.1716	9.1916
540	9.1164	9.1364	9.1564	9.1664	9.1864	9.2064	9.2264	9.2464	9.2664	9.2864
550	9.2112	9.2312	9.2512	9.2612	9.2812	9.3012	9.3212	9.3412	9.3612	9.3812
560	9.3060	9.3260	9.3460	9.3560	9.3760	9.3960	9.4160	9.4360	9.4560	9.4760
570	9.4008	9.4208	9.4408	9.4508	9.4708	9.4908	9.5108	9.5308	9.5508	9.5708
580	9.4956	9.5156	9.5356	9.5456	9.5656	9.5856	9.6056	9.6256	9.6456	9.6656
590	9.5904	9.6104	9.6304	9.6404	9.6604	9.6804	9.7004	9.7204	9.7404	9.7604
600	9.6852	9.7052	9.7252	9.7352	9.7552	9.7752	9.7952	9.8152	9.8352	9.8552
610	9.7800	9.8000	9.8200	9.8300	9.8500	9.8700	9.8900	9.9100	9.9300	9.9500
620	9.8748	9.8948	9.9148	9.9248	9.9448	9.9648	9.9848	10.0048	10.0248	10.0448
630	9.9696	9.9896	10.0096	10.0196	10.0396	10.0596	10.0796	10.0996	10.1196	10.1396
640	10.0644	10.0844	10.1044	10.1144	10.1344	10.1544	10.1744	10.1944	10.2144	10.2344
650	10.1592	10.1792	10.1992	10.2092	10.2292	10.2492	10.2692	10.2892	10.3092	10.3292
660	10.2540	10.2740	10.2940	10.3040	10.3240	10.3440	10.3640	10.3840	10.4040	10.4240
670	10.3488	10.3688	10.3888	10.3988	10.4188	10.4388	10.4588	10.4788	10.4988	10.5188
680	10.4436	10.4636	10.4836	10.4936	10.5136	10.5336	10.5536	10.5736	10.5936	10.6136
690	10.5384	10.5584	10.5784	10.5884	10.6084	10.6284	10.6484	10.6684	10.6884	10.7084
700	10.6332	10.6532	10.6732	10.6832	10.7032	10.7232	10.7432	10.7632	10.7832	10.8032
710	10.7280	10.7480	10.7680	10.7780	10.7980	10.8180	10.8380	10.8580	10.8780	10.8980
720	10.8228	10.8428	10.8628	10.8728	10.8928	10.9128	10.9328	10.9528	10.9728	10.9928
730	10.9176	10.9376	10.9576	10.9676	10.9876	11.0076	11.0276	11.0476	11.0676	11.0876
740	11.0124	11.0324	11.0524	11.0624	11.0824	11.1024	11.1224	11.1424	11.1624	11.1824
750	11.1072	11.1272	11.1472	11.1572	11.1772	11.1972	11.2172	11.2372	11.2572	11.2772
760	11.2020	11.2220	11.2420	11.2520	11.2720	11.2920	11.3120	11.3320	11.3520	11.3720
770	11.2968	11.3168	11.3368	11.3468	11.3668	11.3868	11.4068	11.4268	11.4468	11.4668
780	11.3916	11.4116	11.4316	11.4416	11.4616	11.4816	11.5016	11.5216	11.5416	11.5616
790	11.4864	11.5064	11.5264	11.5364	11.5564	11.5764	11.5964	11.6164	11.6364	11.6564
800	11.5812	11.6012	11.6212	11.6312	11.6512	11.6712	11.6912	11.7112	11.7312	11.7512
810	11.6760	11.6960	11.7160	11.7260	11.7460	11.7660	11.7860	11.8060	11.8260	11.8460
820	11.7708	11.7908	11.8108	11.8208	11.8408	11.8608	11.8808	11.9008	11.9208	11.9408
830	11.8656	11.8856	11.9056	11.9156	11.9356	11.9556	11.9756	11.9956	12.0156	12.0356
840	11.9604	11.9804	12.0004	12.0104	12.0304	12.0504	12.0704	12.0904	12.1104	12.1304
850	12.0552	12.0752	12.0952	12.1052	12.1252	12.1452	12.1652	12.1852	12.2052	12.2252
860	12.1500	12.1700	12.1900	12.2000	12.2200	12.2400	12.2600	12.2800	12.3000	12.32

Table A.8. *Continued*

v	λ								
	11	12	13	14	15	16	17	18	19
	$\lambda_{\text{obs}}/\text{\AA}$								
2	4595	4647	4648	4771	4773	4775	4843	4778	4851
3	4679.8	4680.5	46808	46889.3	47277.2	473466	483984	484341	494331
4	47568	47698	481623	483738	483732	487676	488211	488264	490004
4	57465	58053	63054	63388	63541	63643	64093	63697	65059
5	4888.9	48925	53764	53844	53474	53239	54003	54713	53363
6	44867	44878	45612	46317	46576	47604	48396	48756	49265
7	41368	41743	42388	42989	43773	44684	44586	45052	45514
8	38669	38618	40161	40734	41334	41693	42137	42886	43643
9	3774.3	38079	38682	38688	38942	38669	40371	40732	41114
10	36388	36913	37681	37832	38273	38669	39261	39364	39728
11	35508	36089	36462	36887	37283	37634	38004	38333	38648
12	34838	35274	35769	36112	36489	36842	37173	37487	37783
13	34023	34674	35091	35476	35838	36176	36493	36793	37086
14	33576	34179	34676	34949	35296	35621	35926	36214	36487
15	33029	33549	34139	34681	34937	35193	35447	35713	35986
16	32873	33388	33769	34138	34443	34749	35066	35366	35662
17	32667	33178	33548	33783	34082	34408	34686	34944	35193
18	32369	32796	33156	33462	33804	34093	34369	34626	34870
19	32163	32558	32886	33235	33546	33826	34094	34347	34585
20	31952	32335	32683	33006	33366	33687	33956	34266	34511
21	31764	32139	32483	32802	33107	33373	33621	33876	34106
22	31595	31964	32364	32668	32969	33261	33536	33806	34066
23	31461	31827	32142	32451	32739	33007	33264	33509	33739
24	31333	31693	31994	32308	32604	32849	33087	33311	33531
25	31179	31532	31839	32162	32443	32709	32966	33181	33400
26	31058	31402	31706	32035	32313	32572	32819	33044	33260

Table A.8. (Continued)

ν	ξ								
	11	12	13	14	15	16	17	18	19
	1000 μ								
ν	.444	.467	.488	.511	.533	.545	.561	.578	.595
27	5.0911	5.1383	5.1822	5.2234	5.2634	5.2851	5.2991	5.2918	5.2732
28	5.0832	5.1304	5.1747	5.2161	5.2564	5.2784	5.2777	5.2688	5.2493
29	5.0760	5.1233	5.1679	5.2095	5.2498	5.2720	5.2671	5.2564	5.2364
30	5.0695	5.1167	5.1616	5.2034	5.2438	5.2664	5.2573	5.2454	5.2250
35	5.0535	5.1008	5.1460	5.1881	5.2287	5.2514	5.2371	5.2235	5.2030
40	5.0379	5.0853	5.1308	5.1734	5.2144	5.2374	5.2195	5.1984	5.1761
45	5.0215	5.0691	5.1150	5.1581	5.1994	5.2220	5.2005	5.1764	5.1536
50	5.0054	5.0531	5.1004	5.1441	5.1858	5.2080	5.1831	5.1565	5.1321
55	4.9894	5.0373	5.0850	5.1293	5.1714	5.1934	5.1651	5.1361	5.1091
60	4.9736	5.0217	5.0698	5.1145	5.1570	5.1788	5.1471	5.1157	5.0861
65	4.9580	5.0063	5.0548	5.1000	5.1430	5.1648	5.1301	5.0951	5.0631
70	4.9425	5.0009	5.0498	5.0955	5.1380	5.1598	5.1221	5.0841	5.0501
75	4.9271	4.9857	5.0350	5.0813	5.1243	5.1460	5.1051	5.0641	5.0281
80	4.9118	4.9706	5.0203	5.0673	5.1110	5.1328	5.0881	5.0441	5.0051
85	4.8966	4.9556	5.0058	5.0535	5.0978	5.1198	5.0721	5.0251	4.9831
90	4.8815	4.9407	4.9914	5.0400	5.0850	5.1072	5.0561	5.0061	4.9611
100	4.8665	4.9259	4.9770	5.0263	5.0720	5.0945	5.0391	4.9861	4.9381
110	4.8516	4.9113	4.9629	5.0129	5.0593	5.0822	5.0231	4.9671	4.9161
120	4.8368	4.8968	4.9489	5.0000	5.0470	5.0704	5.0081	4.9501	4.8961
130	4.8221	4.8824	4.9350	4.9867	5.0343	5.0582	4.9921	4.9321	4.8761
140	4.8075	4.8681	4.9213	4.9737	5.0220	5.0465	4.9771	4.9151	4.8571
150	4.7930	4.8539	4.9076	4.9607	5.0098	5.0348	4.9621	4.9001	4.8411
160	4.7786	4.8398	4.8940	4.9470	4.9969	5.0225	4.9461	4.8821	4.8221
170	4.7643	4.8258	4.8805	4.9333	4.9841	5.0104	4.9311	4.8651	4.8041
180	4.7501	4.8119	4.8671	4.9200	4.9716	5.0000	4.9181	4.8501	4.7881
190	4.7360	4.7981	4.8538	4.9065	4.9580	4.9880	4.9031	4.8341	4.7711
200	4.7220	4.7844	4.8406	4.8933	4.9456	4.9765	4.8891	4.8191	4.7551
250	4.6735	4.7363	4.7931	4.8463	4.8980	4.9300	4.8391	4.7671	4.7021
300	4.6250	4.6882	4.7456	4.8000	4.8513	4.8840	4.7891	4.7151	4.6491
350	4.5765	4.6402	4.6982	4.7533	4.8053	4.8390	4.7411	4.6651	4.5981
400	4.5280	4.5921	4.6507	4.7065	4.7593	4.7940	4.6931	4.6151	4.5471
450	4.4795	4.5441	4.6034	4.6600	4.7135	4.7490	4.6451	4.5651	4.4961
500	4.4310	4.4961	4.5560	4.6133	4.6675	4.7040	4.5981	4.5161	4.4461
550	4.3825	4.4481	4.5087	4.5665	4.6215	4.6590	4.5511	4.4671	4.3961
600	4.3340	4.4001	4.4614	4.5200	4.5758	4.6140	4.5041	4.4191	4.3471
650	4.2855	4.3521	4.4141	4.4733	4.5300	4.5690	4.4571	4.3711	4.3001
700	4.2370	4.3041	4.3668	4.4265	4.4840	4.5240	4.4111	4.3241	4.2531
750	4.1885	4.2561	4.3194	4.3800	4.4380	4.4790	4.3641	4.2761	4.2051
800	4.1400	4.2081	4.2720	4.3333	4.3920	4.4340	4.3171	4.2281	4.1571
850	4.0915	4.1601	4.2246	4.2865	4.3460	4.3890	4.2701	4.1801	4.1091
900	4.0430	4.1121	4.1772	4.2400	4.2990	4.3430	4.2221	4.1311	4.0601
950	3.9945	4.0641	4.1300	4.1933	4.2530	4.2980	4.1751	4.0831	3.9921
1000	3.9460	4.0161	4.0826	4.1465	4.2070	4.2530	4.1291	4.0361	3.9451
1100	3.8975	3.9681	4.0354	4.1000	4.1610	4.2080	4.0821	3.9881	3.8971
1200	3.8490	3.9201	3.9880	4.0533	4.1150	4.1630	4.0351	3.9401	3.8491
1300	3.8005	3.8721	3.9406	4.0065	4.0690	4.1180	4.0001	3.9031	3.8121
1400	3.7520	3.8241	3.8932	3.9600	4.0230	4.0730	3.9531	3.8551	3.7641
1500	3.7035	3.7761	3.8460	3.9133	3.9770	4.0280	3.9061	3.8071	3.7161
1600	3.6550	3.7281	3.7986	3.8665	3.9310	3.9830	3.8601	3.7601	3.6691
1700	3.6065	3.6801	3.7514	3.8200	3.8850	3.9380	3.8141	3.7131	3.6221
1800	3.5580	3.6321	3.7040	3.7733	3.8390	3.8930	3.7681	3.6661	3.5751
1900	3.5095	3.5841	3.6566	3.7265	3.7930	3.8480	3.7221	3.6191	3.5281
2000	3.4610	3.5361	3.6092	3.6800	3.7470	3.8030	3.6761	3.5721	3.4811
2500	3.3630	3.4381	3.5120	3.5833	3.6510	3.7080	3.5801	3.4751	3.3841
3000	3.2650	3.3401	3.4146	3.4865	3.5550	3.6130	3.4841	3.3781	3.2871
3500	3.1670	3.2421	3.3172	3.3890	3.4580	3.5170	3.3871	3.2801	3.1891
4000	3.0690	3.1441	3.2198	3.2925	3.3630	3.4230	3.2921	3.1841	3.0931
4500	2.9710	3.0461	3.1224	3.1960	3.2670	3.3280	3.1961	3.0871	2.9961
5000	2.8730	2.9481	3.0250	3.0990	3.1710	3.2330	3.1001	2.9901	2.9001
5500	2.7750	2.8501	2.9276	3.0025	3.0750	3.1380	3.0041	2.8931	2.8031
6000	2.6770	2.7521	2.8302	2.9050	2.9780	3.0420	2.9071	2.7951	2.7051
6500	2.5790	2.6541	2.7328	2.8080	2.8820	2.9470	2.8111	2.6981	2.6081
7000	2.4810	2.5561	2.6354	2.7110	2.7860	2.8520	2.7151	2.6021	2.5121
7500	2.3830	2.4581	2.5380	2.6140	2.6890	2.7560	2.6181	2.5041	2.4141
8000	2.2850	2.3601	2.4406	2.5170	2.5930	2.6610	2.5221	2.4071	2.3171
8500	2.1870	2.2621	2.3432	2.4200	2.4970	2.5660	2.4261	2.3111	2.2211
9000	2.0890	2.1641	2.2458	2.3230	2.4000	2.4700	2.3291	2.2141	2.1241
9500	1.9910	2.0661	2.1484	2.2260	2.3030	2.3750	2.2331	2.1181	2.0281
10000	1.8930	1.9681	2.0510	2.1290	2.2070	2.2800	2.1371	2.0221	1.9321

Table A.3. Lower Critical Values of F-Dist λ_0 , $\alpha = .05$

$\lambda_0 = \frac{B_0}{B_0 + B_1} = \prod_{i=1}^p \frac{1}{1 + \lambda_i}$													
where $\lambda_1, \lambda_2, \dots, \lambda_p$ are eigenvalues of $K^{-1}H$. Reject H_0 if $\lambda_0 \leq$ table value. ¹ Multiply entry by 10^{-3} .													
v_2	v_1												
	1	2	3	4	5	6	7	8	9	10	11	12	13
	$p = 1$												
1	6.59 ²	1.89 ²	1.49 ²	1.19 ²	.999 ²	1.09 ²	.999 ²	.999 ²	.999 ²	.999 ²	.999 ²	.999 ²	.999 ²
2	.999	.899	.804	.725	.650	.577	.505	.433	.361	.289	.217 ²	.145 ²	.073 ²
3	.729	.636	.547	.476	.403	.330	.258	.185	.112	.039	.006	.000	.000
4	.540	.449	.363	.295	.223	.150	.078	.005	.000	.000	.000	.000	.000
5	.413	.322	.238	.170	.100	.028	.000	.000	.000	.000	.000	.000	.000
6	.336	.246	.163	.095	.024	.000	.000	.000	.000	.000	.000	.000	.000
7	.286	.200	.118	.050	.011	.000	.000	.000	.000	.000	.000	.000	.000
8	.250	.167	.085	.030	.000	.000	.000	.000	.000	.000	.000	.000	.000
9	.223	.143	.061	.020	.000	.000	.000	.000	.000	.000	.000	.000	.000
10	.200	.122	.049	.010	.000	.000	.000	.000	.000	.000	.000	.000	.000
11	.180	.105	.040	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
12	.163	.090	.034	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
13	.148	.078	.029	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
14	.135	.065	.020	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
15	.123	.057	.018	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
16	.113	.050	.016	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
17	.105	.045	.015	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
18	.097	.041	.014	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
19	.091	.037	.013	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
20	.085	.034	.012	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
21	.080	.031	.011	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
22	.076	.028	.010	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
23	.072	.026	.009	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
24	.068	.024	.008	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
25	.065	.022	.007	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
26	.062	.020	.006	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
27	.060	.018	.005	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
28	.057	.017	.004	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
29	.055	.016	.004	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
30	.053	.015	.003	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
31	.051	.014	.003	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
32	.049	.013	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
33	.047	.012	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
34	.046	.011	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
35	.044	.010	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
36	.043	.009	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
37	.041	.008	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
38	.040	.007	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
39	.039	.006	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
40	.038	.005	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
41	.037	.004	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
42	.036	.003	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
43	.035	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
44	.034	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
45	.033	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
46	.032	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
47	.031	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
48	.030	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
49	.029	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
50	.028	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
51	.027	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
52	.026	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
53	.025	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
54	.024	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
55	.023	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
56	.022	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
57	.021	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
58	.020	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
59	.019	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
60	.018	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

(continued)

1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26

[illegible]

100

Figure 1

n	m											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00 ¹	0.00 ²	0.00 ³	0.00 ⁴	0.00 ⁵	0.00 ⁶	0.00 ⁷
3	1.79 ⁸	1.68 ⁹	1.79 ¹⁰	1.17 ¹¹	1.16 ¹²	0.00 ¹³	0.00 ¹⁴	0.00 ¹⁵	0.00 ¹⁶	0.00 ¹⁷	0.00 ¹⁸	0.00 ¹⁹
4	0.04	0.00	0.00	0.02	0.01	0.00	0.00 ²⁰	0.00 ²¹	0.00 ²²	0.00 ²³	0.00 ²⁴	0.00 ²⁵
5	0.00	0.00	0.00	0.00	0.00	0.00 ²⁶	0.00 ²⁷	0.00 ²⁸	0.00 ²⁹	0.00 ³⁰	0.00 ³¹	0.00 ³²
6	0.00	0.04	0.00	0.04	0.04	0.04	0.00	0.04	0.04	0.00 ³³	0.00 ³⁴	0.00 ³⁵
7	0.00	0.00	0.00	0.04	0.00	0.02	0.00	0.02	0.00 ³⁶	0.00 ³⁷	0.00 ³⁸	0.00 ³⁹
8	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.04	0.04	0.04	0.00 ⁴⁰
9	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.04	0.00	0.04	0.04
10	0.00	0.04	0.04	0.00	0.04	0.00	0.00	0.00	0.04	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.04	0.00	0.04	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
46	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

1000

100

1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26

η_2	η_3											
	i	j	k	l	m	n	o	p	q	r	s	t
$p = 1$												
1 000	000	000	000	000	000	000	000	000	000	000	000	000
2 000	000	000	000	000	000	000	000	000	000	000	000	000
3 000	000	000	000	000	000	000 ¹	000 ¹	000 ¹	000 ¹	000 ¹	000 ¹	000 ¹
4 1 00 ¹	200 ¹	120 ¹	080 ¹	050 ¹	030 ¹	020 ¹	010 ¹	000 ¹	000 ¹	000 ¹	000 ¹	000 ¹
5 000	4 00 ²	2 10 ²	1 10 ²	0 60 ²	0 30 ²	0 20 ²	0 10 ²	0 00 ²	0 00 ²	0 00 ²	0 00 ²	0 00 ²
6 000	000	000	000	0 00 ³	0 00 ³	0 00 ³	0 00 ³	0 00 ³	0 00 ³	0 00 ³	0 00 ³	0 00 ³
7 1 00	000	004	014	034	0 70 ⁴	0 40 ⁴	0 20 ⁴	0 10 ⁴	0 00 ⁴	0 00 ⁴	0 00 ⁴	0 00 ⁴
8 000	000	004	004	004	0 00 ⁵	0 00 ⁵	0 00 ⁵	0 00 ⁵	0 00 ⁵	0 00 ⁵	0 00 ⁵	0 00 ⁵
9 0 00	1 00	000	000	000	0 00 ⁶	0 00 ⁶	0 00 ⁶	0 00 ⁶	0 00 ⁶	0 00 ⁶	0 00 ⁶	0 00 ⁶
10 000	000	004	004	004	0 00 ⁷	0 00 ⁷	0 00 ⁷	0 00 ⁷	0 00 ⁷	0 00 ⁷	0 00 ⁷	0 00 ⁷
11 0 00	000	000	1 00	000	0 00 ⁸	0 00 ⁸	0 00 ⁸	0 00 ⁸	0 00 ⁸	0 00 ⁸	0 00 ⁸	0 00 ⁸
12 0 00	000	000	000	1 00	0 00 ⁹	0 00 ⁹	0 00 ⁹	0 00 ⁹	0 00 ⁹	0 00 ⁹	0 00 ⁹	0 00 ⁹
13 0 00	000	000	000	000	0 00 ¹⁰	0 00 ¹⁰	0 00 ¹⁰	0 00 ¹⁰	0 00 ¹⁰	0 00 ¹⁰	0 00 ¹⁰	0 00 ¹⁰
14 0 00	000	000	000	000	0 00 ¹¹	0 00 ¹¹	0 00 ¹¹	0 00 ¹¹	0 00 ¹¹	0 00 ¹¹	0 00 ¹¹	0 00 ¹¹
15 0 00	000	000	000	000	0 00 ¹²	0 00 ¹²	0 00 ¹²	0 00 ¹²	0 00 ¹²	0 00 ¹²	0 00 ¹²	0 00 ¹²
16 0 00	000	000	000	000	0 00 ¹³	0 00 ¹³	0 00 ¹³	0 00 ¹³	0 00 ¹³	0 00 ¹³	0 00 ¹³	0 00 ¹³
17 0 00	000	000	000	000	0 00 ¹⁴	0 00 ¹⁴	0 00 ¹⁴	0 00 ¹⁴	0 00 ¹⁴	0 00 ¹⁴	0 00 ¹⁴	0 00 ¹⁴
18 0 00	000	000	000	000	0 00 ¹⁵	0 00 ¹⁵	0 00 ¹⁵	0 00 ¹⁵	0 00 ¹⁵	0 00 ¹⁵	0 00 ¹⁵	0 00 ¹⁵
19 0 00	000	000	000	000	0 00 ¹⁶	0 00 ¹⁶	0 00 ¹⁶	0 00 ¹⁶	0 00 ¹⁶	0 00 ¹⁶	0 00 ¹⁶	0 00 ¹⁶
20 0 00	000	000	000	000	0 00 ¹⁷	0 00 ¹⁷	0 00 ¹⁷	0 00 ¹⁷	0 00 ¹⁷	0 00 ¹⁷	0 00 ¹⁷	0 00 ¹⁷
21 0 00	000	000	000	000	0 00 ¹⁸	0 00 ¹⁸	0 00 ¹⁸	0 00 ¹⁸	0 00 ¹⁸	0 00 ¹⁸	0 00 ¹⁸	0 00 ¹⁸
22 0 00	000	000	000	000	0 00 ¹⁹	0 00 ¹⁹	0 00 ¹⁹	0 00 ¹⁹	0 00 ¹⁹	0 00 ¹⁹	0 00 ¹⁹	0 00 ¹⁹
23 0 00	000	000	000	000	0 00 ²⁰	0						

1000

Figure 1

[illegible]

1000

Figure 1. The effect of the number of trials on the number of correct responses.

[illegible]

100

Table A.18. Upper Critical Values for Roy's Test, $\alpha = .05$

Roy's test statistic is given by

$$F = \frac{\lambda_1}{1 + \lambda_1},$$

where λ_1 is the largest eigenvalue of $\mathbf{B}^{-1}\mathbf{B}$. The parameters are

$$J = \min(p, p'), \quad m = \frac{(p_1 - p) - 1}{2}, \quad N = \frac{(p_2 - p) - 1}{2}.$$

Reject H_0 if $F >$ table value.

m										
p	0	1	2	3	4	5	7	10	15	
J = 2										
5	.500	.601	.700	.740	.770	.790	.808	.820	.834	
10	.394	.455	.514	.561	.598	.629	.659	.682	.700	
15	.299	.349	.402	.440	.483	.515	.550	.577	.596	
20	.221	.281	.324	.360	.404	.436	.468	.500	.520	
25	.164	.206	.238	.264	.288	.309	.328	.344	.358	
30	.117	.150	.171	.184	.200	.210	.220	.230	.237	
40	.070	.090	.105	.118	.130	.140	.148	.156	.162	
50	.049	.060	.071	.080	.088	.095	.101	.106	.110	
60	.034	.040	.046	.050	.054	.057	.060	.063	.065	
80	.024	.028	.032	.035	.037	.039	.041	.043	.044	
100	.016	.019	.021	.023	.024	.025	.026	.027	.028	
200	.002	.003	.003	.004	.004	.004	.005	.005	.005	
J = 3										
5	.609	.720	.770	.800	.822	.840	.857	.869	.880	
10	.412	.507	.560	.605	.636	.663	.685	.700	.714	
15	.302	.402	.450	.500	.540	.569	.595	.615	.630	
20	.250	.340	.380	.427	.460	.488	.510	.530	.545	
25	.200	.284	.324	.367	.400	.428	.450	.470	.484	
30	.160	.238	.274	.312	.340	.365	.385	.405	.419	
40	.100	.160	.201	.234	.260	.280	.298	.314	.328	
50	.070	.110	.142	.168	.189	.207	.222	.236	.249	
60	.050	.080	.105	.127	.144	.158	.170	.181	.191	
100	.020	.035	.048	.060	.070	.078	.085	.090	.094	
200	.004	.006	.007	.008	.009	.009	.010	.010	.011	

Tabelle A.18. *Kontingenztabelle*

<i>i</i>	<i>ω_j</i>									
	0	1	2	3	4	5	7	8	9	
<i>J</i> = 4										
5	359	392	413	436	454	468	489	511	533	
10	547	601	641	674	708	725	759	796	840	
15	654	742	813	876	947	982	1054	1101	1156	
20	764	865	941	1014	1093	1138	1221	1289	1364	
25	876	994	1080	1152	1239	1294	1387	1469	1554	
30	991	1116	1214	1286	1373	1438	1541	1623	1711	
40	1207	1340	1449	1521	1608	1673	1786	1869	1961	
50	1371	1519	1634	1707	1794	1859	2001	2093	2191	
60	1485	1640	1761	1834	1921	1986	2149	2251	2351	
80	1612	1775	1901	1974	2061	2126	2309	2421	2521	
120	1877	2049	2181	2254	2341	2406	2609	2731	2831	
200	2469	2649	2781	2854	2941	3006	3229	3361	3461	
<i>J</i> = 5										
5	369	401	421	444	462	476	497	519	541	
10	557	611	651	684	718	735	769	806	849	
15	664	752	823	886	957	992	1064	1111	1166	
20	774	875	951	1024	1103	1148	1231	1299	1374	
25	886	994	1080	1152	1239	1294	1387	1469	1554	
30	991	1116	1214	1286	1373	1438	1541	1623	1711	
40	1207	1340	1449	1521	1608	1673	1786	1869	1961	
50	1371	1519	1634	1707	1794	1859	2001	2093	2191	
60	1485	1640	1761	1834	1921	1986	2149	2251	2351	
80	1612	1775	1901	1974	2061	2126	2309	2421	2521	
120	1877	2049	2181	2254	2341	2406	2609	2731	2831	
200	2469	2649	2781	2854	2941	3006	3229	3361	3461	

(continued)

Abstract

Table 1									
J = 0									
R	0	1	2	3	4	5	6	7	8
5	825	858	890	921	955	984	1018	1048	1075
10	885	892	921	944	964	981	998	1018	1031
15	937	956	969	983	998	1009	1016	1020	1025
20	984	991	998	1001	1005	1006	1007	1008	1009
25	1021	1028	1030	1031	1032	1033	1034	1035	1036
30	1051	1056	1058	1059	1060	1061	1062	1063	1064
40	1088	1092	1094	1095	1096	1097	1098	1099	1100
50	1121	1126	1128	1129	1130	1131	1132	1133	1134
60	1150	1155	1157	1158	1159	1160	1161	1162	1163
80	1185	1190	1192	1193	1194	1195	1196	1197	1198
100	1208	1213	1215	1216	1217	1218	1219	1220	1221
200	1285	1290	1292	1293	1294	1295	1296	1297	1298

Table 2									
J = 1									
R	0	1	2	3	4	5	6	7	8
5	831	872	907	944	980	1017	1054	1091	1128
10	881	926	970	1011	1050	1087	1124	1161	1198
15	939	983	1024	1063	1100	1137	1174	1211	1248
20	994	1038	1077	1114	1151	1187	1224	1261	1298
25	1041	1083	1121	1158	1195	1231	1268	1305	1342
30	1081	1121	1158	1195	1231	1268	1305	1342	1379
40	1124	1163	1200	1237	1274	1311	1348	1385	1422
50	1165	1203	1240	1277	1314	1351	1388	1425	1462
60	1201	1238	1275	1312	1349	1386	1423	1460	1497
80	1245	1282	1319	1356	1393	1430	1467	1504	1541
100	1281	1318	1355	1392	1429	1466	1503	1540	1577
200	1371	1408	1445	1482	1519	1556	1593	1630	1667
500	1581	1618	1655	1692	1729	1766	1803	1840	1877
800	1681	1718	1755	1792	1829	1866	1903	1940	1977
1000	1746	1783	1820	1857	1894	1931	1968	2005	2042

Tabelle A.18. Koeffizienten

A	m								
	0	1	2	3	4	5	6	7	8

$j = 0$									
5	874	888	902	915	928	942	955	968	980
10	736	754	772	789	806	823	840	857	873
15	615	643	670	697	724	752	779	806	833
20	511	553	595	636	678	719	760	801	842
25	426	483	540	597	654	711	768	825	882
30	354	425	496	567	638	709	779	850	920
40	289	375	461	547	633	719	805	891	976
50	240	339	438	537	636	735	834	933	1031
60	200	309	418	527	636	745	854	963	1071
70	167	279	390	501	612	723	834	945	1055
80	140	254	366	477	588	699	810	921	1031
100	104	219	332	443	554	665	776	887	997
200	60	164	278	391	504	617	730	843	956
300	40	130	245	360	474	588	702	816	930
400	28	98	214	330	445	560	675	790	905
500	19	79	196	312	428	544	660	776	892
600	13	64	182	298	414	530	646	762	878
700	9	54	172	288	404	520	636	752	868
800	6	46	164	280	396	512	628	744	860
1000	4	39	158	276	392	508	624	740	856

$j = 1$									
5	891	904	917	930	943	956	969	982	995
10	756	778	799	820	841	862	883	904	925
15	647	674	699	724	750	775	801	826	851
20	560	593	624	655	686	717	748	779	810
25	497	535	570	605	640	675	710	745	780
30	445	487	528	569	610	651	692	733	774
40	380	427	473	519	565	611	657	703	749
50	320	371	419	467	514	561	608	655	702
60	274	328	378	428	478	528	578	628	678
70	237	294	346	397	448	499	550	601	652
80	207	267	321	374	427	480	533	586	639
100	177	240	305	368	431	494	557	620	683
200	109	184	261	338	415	492	569	646	723
300	80	158	237	316	395	474	553	632	711
400	60	139	220	300	380	460	540	620	700
500	48	129	212	292	372	452	532	612	692
600	39	121	205	285	365	445	525	605	685
700	32	114	199	279	359	439	519	599	679
800	26	108	194	274	354	434	514	594	674
1000	21	103	190	270	350	430	510	590	670

$j = 10$									
5	908	918	928	938	948	958	968	978	987
10	780	799	817	836	854	873	891	910	928
15	685	709	732	755	778	801	824	847	870
20	612	643	673	703	733	763	793	823	853
25	550	586	621	656	691	726	761	796	831
30	497	538	578	618	658	698	738	778	818
40	420	466	511	556	601	646	691	736	781
50	350	400	449	498	547	596	645	694	743
60	290	343	395	447	499	551	603	655	707
70	240	296	350	404	458	512	566	620	674
80	200	259	315	370	425	480	535	590	645
100	160	221	278	334	390	446	502	558	614
200	104	167	226	285	344	403	462	521	580
300	81	146	207	268	329	390	451	512	573
400	64	131	194	257	320	383	446	509	572
500	53	124	189	252	315	378	441	504	567
600	45	118	185	248	311	374	437	500	563
700	39	113	181	244	307	370	433	496	559
800	34	108	177	240	303	366	429	492	555
1000	29	103	173	236	299	362	425	488	551

Table A.16. Upper Critical Values of Fisher's Statistic $F_{\alpha; \nu_1, \nu_2}$ in (A.16)

		F_{α}															
		$\frac{F_{\alpha}}{1 + F_{\alpha}}$															
		$\frac{F_{\alpha}}{1 + F_{\alpha}} \times 100$															
ν_2	ν_1	0	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30
		0	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30
0	1	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99	15.99
1	1	161.45	159.92	158.58	157.32	156.14	155.02	153.95	152.93	151.95	151.01	150.10	149.22	148.37	147.54	146.73	145.94
2	1	19.00	18.51	18.04	17.59	17.15	16.72	16.30	15.89	15.48	15.08	14.69	14.30	13.92	13.54	13.17	12.81
3	1	16.69	16.19	15.72	15.27	14.83	14.40	13.98	13.56	13.15	12.75	12.35	11.96	11.58	11.20	10.83	10.47
4	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
5	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
6	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
7	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
8	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
9	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
10	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
12	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
15	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
20	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
25	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77
30	1	15.99	15.49	15.02	14.57	14.13	13.70	13.28	12.86	12.45	12.05	11.65	11.26	10.88	10.50	10.13	9.77

周安周惠周壽周壽周壽

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

周安周周周周周周周周

100

[illegible]

Table A.11. *Continued*

n	k										N
	0	1	2	3	4	5	6	7	8	9	
10	3.50	3.53	3.56	3.58	3.60	3.62	3.64	3.66	3.68	3.70	10
11	3.52	3.55	3.58	3.60	3.62	3.64	3.66	3.68	3.70	3.72	11
12	3.54	3.57	3.60	3.62	3.64	3.66	3.68	3.70	3.72	3.74	12
13	3.56	3.59	3.62	3.64	3.66	3.68	3.70	3.72	3.74	3.76	13
14	3.58	3.61	3.64	3.66	3.68	3.70	3.72	3.74	3.76	3.78	14
15	3.60	3.63	3.66	3.68	3.70	3.72	3.74	3.76	3.78	3.80	15
16	3.62	3.65	3.68	3.70	3.72	3.74	3.76	3.78	3.80	3.82	16
17	3.64	3.67	3.70	3.72	3.74	3.76	3.78	3.80	3.82	3.84	17
18	3.66	3.69	3.72	3.74	3.76	3.78	3.80	3.82	3.84	3.86	18
19	3.68	3.71	3.74	3.76	3.78	3.80	3.82	3.84	3.86	3.88	19
20	3.70	3.73	3.76	3.78	3.80	3.82	3.84	3.86	3.88	3.90	20
21	3.72	3.75	3.78	3.80	3.82	3.84	3.86	3.88	3.90	3.92	21
22	3.74	3.77	3.80	3.82	3.84	3.86	3.88	3.90	3.92	3.94	22
23	3.76	3.79	3.82	3.84	3.86	3.88	3.90	3.92	3.94	3.96	23
24	3.78	3.81	3.84	3.86	3.88	3.90	3.92	3.94	3.96	3.98	24
25	3.80	3.83	3.86	3.88	3.90	3.92	3.94	3.96	3.98	4.00	25
26	3.82	3.85	3.88	3.90	3.92	3.94	3.96	3.98	4.00	4.02	26
27	3.84	3.87	3.90	3.92	3.94	3.96	3.98	4.00	4.02	4.04	27
28	3.86	3.89	3.92	3.94	3.96	3.98	4.00	4.02	4.04	4.06	28
29	3.88	3.91	3.94	3.96	3.98	4.00	4.02	4.04	4.06	4.08	29
30	3.90	3.93	3.96	3.98	4.00	4.02	4.04	4.06	4.08	4.10	30
31	3.92	3.95	3.98	4.00	4.02	4.04	4.06	4.08	4.10	4.12	31
32	3.94	3.97	4.00	4.02	4.04	4.06	4.08	4.10	4.12	4.14	32
33	3.96	3.99	4.02	4.04	4.06	4.08	4.10	4.12	4.14	4.16	33
34	3.98	4.01	4.04	4.06	4.08	4.10	4.12	4.14	4.16	4.18	34
35	4.00	4.03	4.06	4.08	4.10	4.12	4.14	4.16	4.18	4.20	35
36	4.02	4.05	4.08	4.10	4.12	4.14	4.16	4.18	4.20	4.22	36
37	4.04	4.07	4.10	4.12	4.14	4.16	4.18	4.20	4.22	4.24	37
38	4.06	4.09	4.12	4.14	4.16	4.18	4.20	4.22	4.24	4.26	38
39	4.08	4.11	4.14	4.16	4.18	4.20	4.22	4.24	4.26	4.28	39
40	4.10	4.13	4.16	4.18	4.20	4.22	4.24	4.26	4.28	4.30	40
41	4.12	4.15	4.18	4.20	4.22	4.24	4.26	4.28	4.30	4.32	41
42	4.14	4.17	4.20	4.22	4.24	4.26	4.28	4.30	4.32	4.34	42
43	4.16	4.19	4.22	4.24	4.26	4.28	4.30	4.32	4.34	4.36	43
44	4.18	4.21	4.24	4.26	4.28	4.30	4.32	4.34	4.36	4.38	44
45	4.20	4.23	4.26	4.28	4.30	4.32	4.34	4.36	4.38	4.40	45
46	4.22	4.25	4.28	4.30	4.32	4.34	4.36	4.38	4.40	4.42	46
47	4.24	4.27	4.30	4.32	4.34	4.36	4.38	4.40	4.42	4.44	47
48	4.26	4.29	4.32	4.34	4.36	4.38	4.40	4.42	4.44	4.46	48
49	4.28	4.31	4.34	4.36	4.38	4.40	4.42	4.44	4.46	4.48	49
50	4.30	4.33	4.36	4.38	4.40	4.42	4.44	4.46	4.48	4.50	50

Table A.13. Upper Critical Values for the Ljung-Box-Boxing Test Statistic, $\alpha = .05$

		m															
		2	3	4	5	6	8	10	12	14	16	18	20	25	30	40	50
		$p = 1$															
2	0.0001	10.000	10.000	11.500	11.500	11.500	11.600	11.600	11.600	11.600	11.600	11.600	11.600	11.600	11.600	11.600	11.600
3	0.0001	9.400	9.400	10.900	10.900	10.900	11.000	11.000	11.000	11.000	11.000	11.000	11.000	11.000	11.000	11.000	11.000
4	0.0001	8.800	8.800	10.300	10.300	10.300	10.400	10.400	10.400	10.400	10.400	10.400	10.400	10.400	10.400	10.400	10.400
5	0.0001	8.200	8.200	9.700	9.700	9.700	9.800	9.800	9.800	9.800	9.800	9.800	9.800	9.800	9.800	9.800	9.800
6	0.0001	7.600	7.600	9.100	9.100	9.100	9.200	9.200	9.200	9.200	9.200	9.200	9.200	9.200	9.200	9.200	9.200
7	0.0001	7.000	7.000	8.500	8.500	8.500	8.600	8.600	8.600	8.600	8.600	8.600	8.600	8.600	8.600	8.600	8.600
8	0.0001	6.400	6.400	7.900	7.900	7.900	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000
10	0.0001	5.800	5.800	7.300	7.300	7.300	7.400	7.400	7.400	7.400	7.400	7.400	7.400	7.400	7.400	7.400	7.400
12	0.0001	5.200	5.200	6.700	6.700	6.700	6.800	6.800	6.800	6.800	6.800	6.800	6.800	6.800	6.800	6.800	6.800
14	0.0001	4.600	4.600	6.100	6.100	6.100	6.200	6.200	6.200	6.200	6.200	6.200	6.200	6.200	6.200	6.200	6.200
16	0.0001	4.000	4.000	5.500	5.500	5.500	5.600	5.600	5.600	5.600	5.600	5.600	5.600	5.600	5.600	5.600	5.600
18	0.0001	3.400	3.400	4.900	4.900	4.900	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000
20	0.0001	2.800	2.800	4.300	4.300	4.300	4.400	4.400	4.400	4.400	4.400	4.400	4.400	4.400	4.400	4.400	4.400
25	0.0001	2.200	2.200	3.700	3.700	3.700	3.800	3.800	3.800	3.800	3.800	3.800	3.800	3.800	3.800	3.800	3.800
30	0.0001	1.600	1.600	3.100	3.100	3.100	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200
35	0.0001	1.000	1.000	2.500	2.500	2.500	2.600	2.600	2.600	2.600	2.600	2.600	2.600	2.600	2.600	2.600	2.600
40	0.0001	0.400	0.400	1.900	1.900	1.900	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
50	0.0001	0.200	0.200	1.300	1.300	1.300	1.400	1.400	1.400	1.400	1.400	1.400	1.400	1.400	1.400	1.400	1.400
60	0.0001	0.100	0.100	0.700	0.700	0.700	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800
70	0.0001	0.050	0.050	0.300	0.300	0.300	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400
80	0.0001	0.020	0.020	0.150	0.150	0.150	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200
100	0.0001	0.010	0.010	0.070	0.070	0.070	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
200	0.0001	0.002	0.002	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
∞	0.0001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The test statistic is $m \log L^{(m)} / m$, where $L^{(m)}$ is the Ljung-Boxing statistic. Report $L^{(m)} / m$ in table entries.

λ	λ_{rest}	λ_{obs}	z	λ_{rest}	λ_{obs}	z	λ_{rest}	λ_{obs}	z	λ_{rest}	λ_{obs}	z
3	31.08	30.96	0.0035	36.12	36.27	-0.0042	41.16	41.32	-0.0039	46.24	46.41	-0.0037
4	31.09	30.95	0.0036	36.13	36.28	-0.0041	41.17	41.33	-0.0038	46.25	46.42	-0.0036
5	31.10	30.94	0.0040	36.14	36.30	-0.0040	41.18	41.35	-0.0037	46.26	46.44	-0.0036
6	31.11	30.93	0.0044	36.15	36.32	-0.0039	41.19	41.37	-0.0036	46.27	46.46	-0.0035
7	31.12	30.92	0.0048	36.16	36.34	-0.0038	41.20	41.39	-0.0035	46.28	46.48	-0.0034
8	31.13	30.91	0.0052	36.17	36.36	-0.0037	41.21	41.41	-0.0034	46.29	46.50	-0.0033
9	31.14	30.90	0.0056	36.18	36.38	-0.0036	41.22	41.43	-0.0033	46.30	46.52	-0.0032
10	31.15	30.89	0.0060	36.19	36.40	-0.0035	41.23	41.45	-0.0032	46.31	46.54	-0.0031
11	31.16	30.88	0.0064	36.20	36.42	-0.0034	41.24	41.47	-0.0031	46.32	46.56	-0.0030
12	31.17	30.87	0.0068	36.21	36.44	-0.0033	41.25	41.49	-0.0030	46.33	46.58	-0.0029
13	31.18	30.86	0.0072	36.22	36.46	-0.0032	41.26	41.51	-0.0029	46.34	46.60	-0.0028
14	31.19	30.85	0.0076	36.23	36.48	-0.0031	41.27	41.53	-0.0028	46.35	46.62	-0.0027
15	31.20	30.84	0.0080	36.24	36.50	-0.0030	41.28	41.55	-0.0027	46.36	46.64	-0.0026
16	31.21	30.83	0.0084	36.25	36.52	-0.0029	41.29	41.57	-0.0026	46.37	46.66	-0.0025
17	31.22	30.82	0.0088	36.26	36.54	-0.0028	41.30	41.59	-0.0025	46.38	46.68	-0.0024
18	31.23	30.81	0.0092	36.27	36.56	-0.0027	41.31	41.61	-0.0024	46.39	46.70	-0.0023
19	31.24	30.80	0.0096	36.28	36.58	-0.0026	41.32	41.63	-0.0023	46.40	46.72	-0.0022
20	31.25	30.79	0.0100	36.29	36.60	-0.0025	41.33	41.65	-0.0022	46.41	46.74	-0.0021
21	31.26	30.78	0.0104	36.30	36.62	-0.0024	41.34	41.67	-0.0021	46.42	46.76	-0.0020
22	31.27	30.77	0.0108	36.31	36.64	-0.0023	41.35	41.69	-0.0020	46.43	46.78	-0.0019
23	31.28	30.76	0.0112	36.32	36.66	-0.0022	41.36	41.71	-0.0019	46.44	46.80	-0.0018
24	31.29	30.75	0.0116	36.33	36.68	-0.0021	41.37	41.73	-0.0018	46.45	46.82	-0.0017
25	31.30	30.74	0.0120	36.34	36.70	-0.0020	41.38	41.75	-0.0017	46.46	46.84	-0.0016
26	31.31	30.73	0.0124	36.35	36.72	-0.0019	41.39	41.77	-0.0016	46.47	46.86	-0.0015
27	31.32	30.72	0.0128	36.36	36.74	-0.0018	41.40	41.79	-0.0015	46.48	46.88	-0.0014
28	31.33	30.71	0.0132	36.37	36.76	-0.0017	41.41	41.81	-0.0014	46.49	46.90	-0.0013
29	31.34	30.70	0.0136	36.38	36.78	-0.0016	41.42	41.83	-0.0013	46.50	46.92	-0.0012
30	31.35	30.69	0.0140	36.39	36.80	-0.0015	41.43	41.85	-0.0012	46.51	46.94	-0.0011
31	31.36	30.68	0.0144	36.40	36.82	-0.0014	41.44	41.87	-0.0011	46.52	46.96	-0.0010
32	31.37	30.67	0.0148	36.41	36.84	-0.0013	41.45	41.89	-0.0010	46.53	46.98	-0.0009
33	31.38	30.66	0.0152	36.42	36.86	-0.0012	41.46	41.91	-0.0009	46.54	47.00	-0.0008
34	31.39	30.65	0.0156	36.43	36.88	-0.0011	41.47	41.93	-0.0008	46.55	47.02	-0.0007
35	31.40	30.64	0.0160	36.44	36.90	-0.0010	41.48	41.95	-0.0007	46.56	47.04	-0.0006
36	31.41	30.63	0.0164	36.45	36.92	-0.0009	41.49	41.97	-0.0006	46.57	47.06	-0.0005
37	31.42	30.62	0.0168	36.46	36.94	-0.0008	41.50	41.99	-0.0005	46.58	47.08	-0.0004
38	31.43	30.61	0.0172	36.47	36.96	-0.0007	41.51	42.01	-0.0004	46.59	47.10	-0.0003
39	31.44	30.60	0.0176	36.48	36.98	-0.0006	41.52	42.03	-0.0003	46.60	47.12	-0.0002
40	31.45	30.59	0.0180	36.49	37.00	-0.0005	41.53	42.05	-0.0002	46.61	47.14	-0.0001
41	31.46	30.58	0.0184	36.50	37.02	-0.0004	41.54	42.07	-0.0001	46.62	47.16	0.0000
42	31.47	30.57	0.0188	36.51	37.04	-0.0003	41.55	42.09	0.0000	46.63	47.18	0.0001
43	31.48	30.56	0.0192	36.52	37.06	-0.0002	41.56	42.11	0.0001	46.64	47.20	0.0002
44	31.49	30.55	0.0196	36.53	37.08	-0.0001	41.57	42.13	0.0002	46.65	47.22	0.0003
45	31.50	30.54	0.0200	36.54	37.10	0.0000	41.58	42.15	0.0003	46.66	47.24	0.0004
46	31.51	30.53	0.0204	36.55	37.12	0.0001	41.59	42.17	0.0004	46.67	47.26	0.0005
47	31.52	30.52	0.0208	36.56	37.14	0.0002	41.60	42.19	0.0005	46.68	47.28	0.0006
48	31.53	30.51	0.0212	36.57	37.16	0.0003	41.61	42.21	0.0006	46.69	47.30	0.0007
49	31.54	30.50	0.0216	36.58	37.18	0.0004	41.62	42.23	0.0007	46.70	47.32	0.0008
50	31.55	30.49	0.0220	36.59	37.20	0.0005	41.63	42.25	0.0008	46.71	47.34	0.0009
51	31.56	30.48	0.0224	36.60	37.22	0.0006	41.64	42.27	0.0009	46.72	47.36	0.0010
52	31.57	30.47	0.0228	36.61	37.24	0.0007	41.65	42.29	0.0010	46.73	47.38	0.0011
53	31.58	30.46	0.0232	36.62	37.26	0.0008	41.66	42.31	0.0011	46.74	47.40	0.0012
54	31.59	30.45	0.0236	36.63	37.28	0.0009	41.67	42.33	0.0012	46.75	47.42	0.0013
55	31.60	30.44	0.0240	36.64	37.30	0.0010	41.68	42.35	0.0013	46.76	47.44	0.0014
56	31.61	30.43	0.0244	36.65	37.32	0.0011	41.69	42.37	0.0014	46.77	47.46	0.0015
57	31.62	30.42	0.0248	36.66	37.34	0.0012	41.70	42.39	0.0015	46.78	47.48	0.0016
58	31.63	30.41	0.0252	36.67	37.36	0.0013	41.71	42.41	0.0016	46.79	47.50	0.0017
59	31.64	30.40	0.0256	36.68	37.38	0.0014	41.72	42.43	0.0017	46.80	47.52	0.0018
60	31.65	30.39	0.0260	36.69	37.40	0.0015	41.73	42.45	0.0018	46.81	47.54	0.0019
61	31.66	30.38	0.0264	36.70	37.42	0.0016	41.74	42.47	0.0019	46.82	47.56	0.0020
62	31.67	30.37	0.0268	36.71	37.44	0.0017	41.75	42.49	0.0020	46.83	47.58	0.0021
63	31.68	30.36	0.0272	36.72	37.46	0.0018	41.76	42.51	0.0021	46.84	47.60	0.0022
64	31.69	30.35	0.0276	36.73	37.48	0.0019	41.77	42.53	0.0022	46.85	47.62	0.0023
65	31.70	30.34	0.0280	36.74	37.50	0.0020	41.78	42.55	0.0023	46.86	47.64	0.0024
66	31.71	30.33	0.0284	36.75	37.52	0.0021	41.79	42.57	0.0024	46.87	47.66	0.0025
67	31.72	30.32	0.0288	36.76	37.54	0.0022	41.80	42.59	0.0025	46.88	47.68	0.0026
68	31.73	30.31	0.0292	36.77	37.56	0.0023	41.81	42.61	0.0026	46.89	47.70	0.0027
69	31.74	30.30	0.0296	36.78	37.58	0.0024	41.82	42.63	0.0027	46.90	47.72	0.0028
70	31.75	30.29	0.0300	36.79	37.60	0.0025	41.83	42.65	0.0028	46.91	47.74	0.0029
71	31.76	30.28	0.0304	36.80	37.62	0.0026	41.84	42.67	0.0029	46.92	47.76	0.0030
72	31.77	30.27	0.0308	36.81	37.64	0.0027	41.85	42.69	0.0030	46.93	47.78	0.0031
73	31.78	30.26	0.0312	36.82	37.66	0.0028	41.86	42.71	0.0031	46.94	47.80	0.0032
74	31.79	30.25	0.0316	36.83	37.68	0.0029	41.87	42.73	0.0032	46.95	47.82	0.0033
75	31.80	30.24	0.0320	36.84	37.70	0.0030	41.88	42.75	0.0033	46.96	47.84	0.0034
76	31.81	30.23	0.0324	36.85	37.72	0.0031	41.89	42.77	0.0034	46.97	47.86	0.0035
77	31.82	30.22	0.0328	36.86	37.74	0.0032	41.90	42.79	0.0035	46.98	47.88	0.0036
78	31.83	30.21	0.0332	36.87	37.76	0.0033	41.91	42.81	0.0036	46.99	47.90	0.0037
79	31.84	30.20	0.0336	36.88	37.78	0.0034	41.92	42.83	0.0037	47.00	47.92	0.0038
80	31.85	30.19	0.0340	36.89	37.80	0.0035	41.93	42.85	0.0038	47.01	47.94	0.0039
81	31.86	30.18	0.0344	36.90	37.82	0.0036	41.94	42.87	0.0039	47.02	47.96	0.0040
82	31.87	30.17	0.0348	36.91	37.84	0.0037	41.95	42.89	0.0040	47.03	47.98	0.0041
83	31.88	30.16	0.0352	36.92	37.86	0.0038	41.96	42.91	0.0041	47.04	48.00	0.0042
84	31.89	30.15	0.0356	36.93	37.88	0.0039	41.97	42.93	0.0042	47.05	48.02	0.0043
85	31.90	30.14	0.0360	36.94	37.90	0.0040	41.98	42.95	0.0043	47.06	48.04	0.0044
86	31.91	30.13	0.0364	36.95	37.92	0.0041	41.99	42.97	0.0044	47.07	48.06	0.0045
87	31.92	30.12	0.0368	36.96	37.94	0.0042	42.00	42.99	0.0045	47.08	48.08	0.0046
88	31.93	30.11	0.0372	36.97	37.96	0.0043	42.01	43.01	0.0046	47.09	48.10	0.0047
89	31.94	30.10	0.0376	36.98	37.98	0.0044	42.02	43.03	0.0047	47.10	48.12	0.0048
90	31.95	30.09	0.0380	36.99	38.00	0.0045	42.03	43.05	0.0048	47.11	48.14	0.0049
91	31.96	30.08	0.0384	37.00	38.02	0.0046	42.04	43.07	0.0049	47.12	48.16	0.0050
92	31.97	30.07	0.0388	37.01	38.04	0.0047	42.05					

Table A.13. (Continued)

ν	4	5	6	8	10	12	15	20	25	30	40	50
$\beta = 1$												
4	49.994	61.364	81.034	101.112	109.096	104.398	104.70	111.7	104.6	—	—	—
6	1.89944	2.6012	2.0046	2.0007	2.0112	2.0128	2.0109	2.0111	2.0111	2.009	—	—
8	32.715	60.999	64.491	60.901	60.902	61.18	62.889	62.973	62.966	62.11	—	—
10	37.740	66.679	66.125	66.414	66.964	67.962	64.405	64.326	64.076	60.75	—	—
12	39.516	71.868	71.411	70.614	70.405	70.178	71.672	71.608	71.411	67.54	—	—
15	47.875	77.79	76.628	75.434	75.096	75.075	75.949	75.995	75.750	71.03	67.62	67.62
16	44.978	74.648	73.606	72.67	72.714	72.975	73.731	73.695	73.461	68.747	64.861	64.861
18	43.465	72.602	71.680	71.246	71.675	71.778	71.573	71.666	71.534	68.026	64.105	64.105
20	41.208	69.868	69.048	69.114	69.608	69.683	69.378	69.481	69.362	64.908	60.633	60.633
25	40.008	67.004	66.313	66.670	67.167	67.165	66.778	66.899	66.785	61.775	57.129	57.129
30	38.967	64.664	64.113	64.606	65.106	65.098	64.621	64.758	64.645	59.006	54.488	54.488
35	38.075	62.675	62.263	62.861	63.366	63.357	62.887	63.031	62.918	56.845	52.435	52.435
40	37.398	60.986	60.706	61.311	61.824	61.816	61.349	61.493	61.380	54.981	50.595	50.595
50	36.664	57.125	57.011	57.628	58.146	58.138	57.673	57.817	57.704	51.724	47.338	47.338
60	36.047	54.062	54.072	54.698	55.218	55.210	54.754	54.898	54.785	49.008	44.622	44.622
80	35.698	49.896	49.894	50.526	51.046	51.038	50.582	50.726	50.613	46.295	41.906	41.906
100	35.544	46.724	46.726	47.358	47.878	47.870	47.414	47.558	47.445	43.581	39.192	39.192
150	35.471	43.549	43.552	44.181	44.699	44.691	44.235	44.379	44.266	40.967	36.577	36.577
200	35.460	41.375	41.378	41.999	42.516	42.508	42.052	42.196	42.083	38.753	34.363	34.363
300	35.458	40.001	40.002	40.619	41.135	41.127	40.670	40.814	40.701	37.537	33.147	33.147

Table A.12. *(Continued)*

ν	8	9	10	11	12	13	14	15	16	17
$\beta = 2$										
10	45 772	44 987	44 095	43 107	42 123	41 143	40 166	39 193	38 223	37 255
12	26 566	26 111	25 562	24 920	24 287	23 663	23 048	22 442	21 845	21 257
14	22 101	21 600	21 111	20 633	20 167	19 713	19 270	18 838	18 417	18 007
16	18 628	18 265	17 918	17 587	17 262	16 943	16 630	16 323	16 022	15 727
18	15 755	15 475	15 211	14 961	14 725	14 493	14 265	14 041	13 821	13 605
20	13 535	13 299	13 076	12 865	12 665	12 476	12 289	12 104	11 921	11 740
25	15 295	15 085	14 886	14 697	14 518	14 339	14 161	13 984	13 808	13 633
30	12 550	12 359	12 178	11 997	11 816	11 635	11 454	11 273	11 093	10 913
35	11 623	11 461	11 300	11 139	10 978	10 817	10 656	10 495	10 334	10 173
40	11 009	10 871	10 732	10 593	10 454	10 315	10 176	10 037	9 898	9 759
50	10 402	10 282	10 162	10 042	9 922	9 802	9 682	9 562	9 442	9 322
60	10 304	10 200	10 096	9 992	9 888	9 784	9 680	9 576	9 472	9 368
70	9 9915	9 8965	9 8015	9 7065	9 6115	9 5165	9 4215	9 3265	9 2315	9 1365
80	9 9914	9 9005	9 8096	9 7187	9 6278	9 5369	9 4460	9 3551	9 2642	9 1733
100	9 9998	9 9990	9 9982	9 9974	9 9966	9 9958	9 9950	9 9942	9 9934	9 9926
200	8 0000	8 0000	8 0000	8 0000	8 0000	8 0000	8 0000	8 0000	8 0000	8 0000
∞	8 0000	8 0000	7 9999	7 9999	7 9999	7 9999	7 9999	7 9999	7 9999	7 9999

Each entry in this table is the sum of two.

Table A.13. Orthogonal Polynomial Coefficients

p	Polynomial	Variable										$\phi(p)$
		1	2	3	4	5	6	7	8	9	10	
2	Linear	-1	0	1								2
	Quadratic	1	-2	1								6
4	Linear	-3	-1	1	3							20
	Quadratic	1	-1	-1	1							4
	Cubic	-1	3	-3	1							20
6	Linear	-5	-1	0	1	3						40
	Quadratic	3	-1	-3	-1	3						14
	Cubic	-1	2	0	-2	1						10
	Quartic	1	-4	6	-4	1						70
8	Linear	-7	-3	-1	1	3	5					70
	Quadratic	5	-1	-4	-4	-1	5					64
	Cubic	-5	7	4	-4	-7	5					180
	Quartic	1	-5	2	2	-5	1					28
	Quintic	-1	6	-10	10	-6	1					280
7	Linear	-9	-2	-1	0	1	2	3				28
	Quadratic	3	0	-6	-4	-3	0	3				88
	Cubic	-1	1	1	0	-1	-1	1				6
	Quartic	3	-7	1	0	1	-7	3				174
	Quintic	-1	4	-5	0	5	-4	1				84
	Sextic	1	-6	15	-20	15	-6	1				924
8	Linear	-7	-5	-3	-1	1	3	5	7			198
	Quadratic	7	1	-6	-8	-3	-3	1	7			198
	Cubic	-7	8	7	3	-3	-7	-8	7			264
	Quartic	7	-12	-9	9	9	-12	-9	7			940
	Quintic	-7	22	-17	-15	15	17	-22	7			2,184
	Sextic	1	-5	9	-5	-5	9	-5	1			264
	Septic	-1	7	-21	35	-35	21	-7	1			34,320
9	Linear	-4	-3	-2	-1	0	1	2	3	4		60
	Quadratic	28	7	-6	-17	-20	-17	-6	7	28		2,772
	Cubic	-14	7	10	9	0	-9	-10	-7	14		490
	Quartic	14	-21	-11	9	18	9	-11	-21	14		2,002
	Quintic	-4	11	-4	-9	0	9	4	-11	4		498
	Sextic	4	-17	22	1	-20	1	22	-17	4		1,240
	Septic	-1	6	-14	14	0	-14	14	-6	1		858
	Octic	1	-6	26	-50	70	-50	26	-6	1		12,870
10	Linear	-3	-7	-5	-5	-1	1	3	5	7	9	500
	Quadratic	9	3	-1	-5	-4	-4	-3	-5	3	9	110
	Cubic	-42	18	36	18	12	-12	-31	-39	-34	42	8,580
	Quartic	18	-22	-17	3	18	18	3	-22	-22	18	2,280
	Quintic	-6	14	-1	-11	-6	6	11	1	-14	6	780
	Sextic	3	-11	19	0	-9	-9	9	19	0	3	660
	Septic	-3	47	-60	90	-50	-50	-42	60	-47	3	20,172
	Octic	1	-7	29	-28	14	14	-28	29	-7	1	2,880
	Nonic	-1	9	-36	64	-126	126	-84	36	-9	1	48,520

Note: Under columns $\phi(p)$, where $(p-1) \times p$ results if illustrated in (A.13) in Section A.10.1.

Table A.14. Test for Equal Covariance Matrices, $\alpha = .05$

ν	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
$p = 2$									
5	12.68	16.70	14.91	16.09	15.45	48.68	45.61	50.67	45.66
6	10.70	16.68	11.65	15.97	16.67	16.76	48.48	46.37	45.64
8	9.97	11.68	16.71	11.97	16.23	14.76	38.26	41.67	48.05
9	9.56	11.62	16.97	14.66	16.18	13.61	37.95	40.22	46.13
7	9.28	14.62	16.46	14.93	16.46	12.83	37.68	41.26	45.40
8	9.04	14.14	16.30	11.62	17.06	12.26	36.44	40.57	44.64
9	8.86	14.11	16.81	11.86	17.62	11.64	35.36	40.05	44.06
10	8.76	11.64	16.61	11.63	17.33	11.58	35.61	39.65	43.64
11	8.67	11.61	16.44	11.63	17.06	11.25	35.52	39.55	43.50
12	8.56	11.70	16.30	11.66	16.66	11.30	35.68	39.37	43.00
13	8.52	11.60	16.19	11.54	16.79	10.65	36.67	39.68	42.56
14	8.47	11.51	16.10	11.62	16.60	10.76	36.71	39.66	42.56
15	8.41	11.46	16.01	11.59	16.56	10.57	36.57	39.50	42.48
16	8.36	11.40	15.94	11.54	16.46	10.45	36.40	39.35	42.13
17	8.35	11.35	15.87	11.57	16.36	10.35	36.52	39.34	42.10
18	8.32	11.30	15.81	11.66	16.23	10.27	36.25	39.15	41.99
19	8.26	11.26	15.71	11.64	16.06	10.18	36.14	39.04	41.88
20	8.26	11.21	15.51	11.66	16.06	10.12	36.07	37.96	41.79
25	8.17	11.10	15.31	11.59	15.67	10.66	35.76	37.66	41.44
30	8.11	11.01	15.44	11.63	15.72	10.66	35.59	37.62	41.21
$p = 3$									
4	10.41	15.05	46.36	35.66	66.56	79.11	69.65	56.54	110.11
5	10.19	16.52	46.63	36.63	66.66	79.26	79.66	66.56	96.17
6	17.57	16.54	36.66	47.46	56.67	65.66	76.56	65.56	95.06
7	16.50	16.64	36.54	41.57	54.56	62.66	71.46	76.60	88.50
8	16.06	16.60	31.50	41.63	62.64	66.66	69.65	77.67	83.71
9	15.66	15.22	34.54	41.66	50.53	56.62	67.76	75.66	81.67
10	15.11	16.71	33.59	41.53	56.42	56.57	66.62	76.56	82.66
11	14.65	16.31	33.66	41.56	46.56	57.56	65.71	75.57	81.56
12	14.64	15.66	32.67	41.66	46.53	57.11	64.67	72.55	80.45
13	14.45	15.73	32.53	46.66	46.65	56.56	64.56	72.56	79.75
14	14.26	15.50	32.65	46.56	46.56	56.11	63.66	71.55	79.11
15	14.15	15.51	31.61	46.47	45.62	56.79	63.66	71.66	78.60
16	14.04	15.56	31.66	44.52	45.63	55.66	63.66	70.64	78.14
17	13.94	15.62	31.43	44.56	45.56	55.11	62.79	70.57	77.56
18	13.85	15.66	31.56	44.51	45.56	54.66	62.65	69.67	77.41
19	13.79	15.76	31.51	44.53	46.66	54.64	62.51	69.56	77.11
20	13.72	15.66	31.61	46.66	46.76	54.44	61.56	69.45	76.64
25	13.66	15.11	36.55	46.44	46.65	53.76	61.55	68.54	75.64
30	13.55	15.10	36.55	46.66	45.73	53.22	60.62	67.64	75.16

Table A.14. H -contours

r	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
$p = 4$									
5	55.50	56.50	75.00	93.97	115.17	138.11	167.41	195.09	220.80
6	55.26	48.62	66.60	82.60	98.09	113.83	130.64	148.69	167.51
7	57.11	44.69	60.89	76.36	91.90	108.68	125.90	143.71	161.39
8	57.61	42.34	57.77	72.77	87.80	103.94	120.23	136.67	153.70
9	58.45	40.71	55.62	70.07	84.42	99.46	115.32	130.88	146.74
10	59.50	39.34	54.06	68.26	82.19	95.99	109.46	123.19	136.59
11	60.76	38.41	52.86	66.81	80.69	93.95	105.27	118.46	131.57
12	62.46	37.67	51.99	65.66	79.14	92.40	103.54	115.95	128.45
13	64.66	37.09	51.15	64.73	78.04	91.15	101.32	113.69	125.74
14	67.33	36.69	50.50	63.99	77.13	89.12	99.47	111.69	123.35
15	70.47	36.47	49.97	63.30	76.37	87.26	97.49	110.09	121.34
16	74.08	35.82	49.54	62.76	75.73	85.56	95.14	110.87	120.10
17	78.09	35.32	49.12	62.28	75.10	83.87	92.92	112.87	118.22
18	82.60	35.20	48.76	61.86	74.58	82.30	90.80	112.17	116.40
19	87.70	35.02	48.47	61.56	74.15	80.82	88.25	111.56	115.79
20	93.50	34.82	48.25	61.37	73.87	79.38	86.73	111.32	115.19
25	109.06	34.66	47.79	60.98	73.47	76.78	84.04	109.61	113.99
30	130.71	33.89	46.61	59.21	71.80	73.74	81.79	107.71	111.87
$p = 3$									
6	51.11	61.99	110.92	158.08	199.54	233.71	259.66	287.37	316.90
7	49.40	71.00	97.89	122.22	149.99	176.34	195.49	215.47	236.30
8	50.20	65.15	89.45	113.83	136.89	159.84	181.08	200.14	219.49
9	50.71	61.39	86.62	107.17	130.50	150.17	171.60	190.27	207.504
10	51.04	58.76	81.25	103.66	124.46	143.64	160.56	187.57	204.00
11	51.62	56.83	78.79	100.62	120.00	140.54	155.96	180.26	200.37
12	52.62	55.77	76.60	97.68	116.11	136.98	150.38	173.23	196.08
13	54.03	54.79	75.03	95.82	113.96	133.89	145.34	170.19	191.71
14	55.79	53.23	73.85	94.29	111.50	130.89	141.21	172.49	189.68
15	57.90	52.44	73.08	93.82	110.00	128.87	139.29	170.56	189.66
16	60.22	51.70	72.14	93.04	111.41	126.60	140.00	168.50	187.30
17	62.85	51.19	71.50	91.83	110.54	125.38	140.25	168.59	185.01
18	65.74	50.69	70.79	90.23	109.50	124.29	141.01	167.65	184.10
19	68.90	50.36	70.17	89.34	108.87	123.59	141.97	166.69	182.81
20	72.37	49.88	69.67	88.43	107.93	122.92	141.02	165.98	181.66
25	86.08	48.48	67.69	86.70	105.21	123.50	141.62	164.60	177.49
30	107.48	47.61	66.71	85.29	103.50	121.68	139.47	159.22	174.87

Note: Table contains upper percentage points for

$$-2 \ln M = n \left(4 \ln(R) - \sum_{i=1}^4 \ln(R_i) \right)$$

for k samples, each with n degrees of freedom. Subject $R_1 : R_2 : R_3 : \dots : R_4$ if $-2 \ln M >$ table value.

Table A.15. Test for Independence of p Variables

Upper percentage points for

$$\chi^2 = - \left(p - \frac{2p+3}{n} \right) \ln \left(\frac{96}{v_{1,2} \cdots v_{p,p}} \right) = - \left(p - \frac{2p+3}{n} \right) \ln (R),$$

where n is the degrees of freedom of $\mathbf{Y}$ or $\mathbf{R}$. Reject hypothesis of independence if χ^2 is greater than table value. The $\chi^2_{0.05}$ values are shown for comparison, since χ^2 is approximately χ^2 distributed with $f = \lfloor p(p-1)/2 \rfloor$ degrees of freedom.

	n	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$	$p=10$
		$\alpha = .05$							
	4	9.000							
	5	7.834	15.01						
	6	7.834	15.47	24.86					
	7	7.811	15.95	26.44	34.80				
	8	7.811	12.85	19.45	26.71	46.05			
	9	7.811	12.76	19.62	27.71	50.41	66.26		
	10	7.812	12.71	19.66	26.77	56.06	66.62	73.74	
	11	7.812	12.66	19.67	27.66	54.91	66.22	67.76	80.02
	12	7.813	12.66	19.58	27.71	54.26	66.67	67.68	75.43
	13	7.813	12.65	19.52	27.53	56.89	63.76	67.63	70.43
	14	7.813	12.64	19.46	27.44	56.05	63.23	64.46	67.67
	15	7.813	12.65	19.45	27.36	55.44	62.62	63.69	65.34
	16	7.814	12.62	19.40	27.30	55.51	62.55	63.53	65.13
	17	7.814	12.62	19.40	27.28	55.59	62.64	63.77	64.68
	18	7.814	12.62	19.40	27.21	55.12	62.19	62.68	64.12
	19	7.814	12.61	19.38	27.19	55.06	62.66	62.56	63.73
	20	7.814	12.61	19.37	27.16	55.01	62.37	62.68	63.43
$\chi^2_{0.05}$		7.815	12.59	19.35	27.05	52.07	61.24	61.65	61.66
		$\alpha = .01$							
	4	11.79							
	5	11.41	25.18						
	6	11.36	18.27	32.18					
	7	11.34	17.54	26.86	44.63				
	8	11.34	17.26	24.92	36.64	56.61			
	9	11.34	17.10	24.29	33.63	47.05	74.16		
	10	11.34	17.01	23.95	32.34	45.59	59.16	66.67	
	11	11.34	16.96	23.75	31.93	45.00	54.80	73.63	100.13
	12	11.34	16.95	23.62	31.60	44.15	52.79	67.37	66.06
	13	11.34	16.95	23.61	31.50	43.89	51.49	64.64	61.50
	14	11.34	16.89	23.47	31.30	43.26	50.75	63.66	77.63
	15	11.34	16.87	23.42	31.08	40.97	50.22	62.63	73.64
	16	11.34	16.86	23.39	31.00	40.79	49.60	61.56	74.56
	17	11.34	16.85	23.36	30.94	40.65	49.58	60.86	73.00
	18	11.34	16.85	23.34	30.88	40.54	49.46	60.49	71.61
	19	11.34	16.85	23.32	30.84	40.46	49.22	60.21	72.12
	20	11.34	16.84	23.31	30.81	40.40	49.09	59.94	72.13
$\chi^2_{0.01}$		11.34	16.81	23.28	30.76	39.95	48.26	58.77	66.95

Answers and Hints to Problems

CHAPTER 1

$$2.1 \text{ (a) } A + B = \begin{pmatrix} 7 & 8 & 7 \\ 13 & 14 & 3 \end{pmatrix}, \quad A - B = \begin{pmatrix} 1 & 4 & -1 \\ 1 & -4 & 13 \end{pmatrix}$$

$$\text{(b) } A^t A = \begin{pmatrix} 65 & 45 & 68 \\ 45 & 29 & 45 \\ 68 & 45 & 75 \end{pmatrix}, \quad A A^t = \begin{pmatrix} 29 & 62 \\ 62 & 134 \end{pmatrix}$$

$$2.2 \text{ (a) } (A + B)^t = \begin{pmatrix} 3 & 15 \\ 8 & 14 \\ 7 & 3 \end{pmatrix}, \quad A^t + B^t = \begin{pmatrix} 3 & 15 \\ 8 & 14 \\ 7 & 3 \end{pmatrix}$$

$$\text{(b) } A^t = \begin{pmatrix} 4 & 7 \\ 2 & 3 \\ 3 & 8 \end{pmatrix}, \quad (A^t)^t = \begin{pmatrix} 4 & 2 & 3 \\ 7 & 3 & 8 \end{pmatrix} = A$$

$$2.3 \text{ (a) } AB = \begin{pmatrix} 5 & 15 \\ 3 & -5 \end{pmatrix}, \quad BA = \begin{pmatrix} 2 & 8 \\ 11 & -2 \end{pmatrix}$$

$$\text{(b) } (AB)^t = -B^t A^t, \quad (A)^t = -B, \quad (B)^t = A$$

$$2.4 \text{ (a) } A + B = \begin{pmatrix} 3 & 5 \\ 2 & 4 \end{pmatrix}, \quad \operatorname{tr}(A + B) = 7$$

$$\text{(b) } \operatorname{tr}(A) = 5, \quad \operatorname{tr}(B) = 7$$

$$2.5 \text{ (a) } AB = \begin{pmatrix} 4 & 1 \\ 3 & -3 \end{pmatrix}, \quad BA = \begin{pmatrix} -1 & 8 & 7 \\ 2 & 4 & 6 \\ 1 & -3 & -2 \end{pmatrix}$$

$$\text{(b) } \operatorname{tr}(AB) = 1, \quad \operatorname{tr}(BA) = 1$$

$$2.6 \text{ (b) } x = (-1 \ 1 \ 1 \ -1)^t$$

$$2.7 \text{ (a) } Bx = (13, 6, 39)^t \quad \text{(b) } y^t B = (23, -1, 17) \quad \text{(c) } x^t A x = 18$$

$$\text{(d) } x^t B y = 42 \quad \text{(e) } x^t x = 6 \quad \text{(f) } x^t y = 3$$

$$\text{(g) } Bx^t = \begin{pmatrix} 8 & -3 & 2 \\ -3 & 1 & -2 \\ 2 & -2 & 4 \end{pmatrix}$$

$$(b) \mathbf{xy}' = \begin{pmatrix} 3 & 2 & 1 \\ -3 & -2 & -1 \\ 6 & 4 & 2 \end{pmatrix}$$

$$(c) \mathbf{BFB} = \begin{pmatrix} 61 & 7 & 22 \\ 7 & 14 & 7 \\ 21 & 3 & 40 \end{pmatrix}$$

$$2.8 \text{ (a) } \mathbf{x} + \mathbf{y} = (3, 3, 3)', \mathbf{x} - \mathbf{y} = (-2, -3, 1)'$$

$$(b) (\mathbf{x} - \mathbf{y})' \mathbf{A}(\mathbf{x} - \mathbf{y}) = -31$$

$$2.9 \quad \mathbf{Bx} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + b_3\mathbf{x}_3$$

$$= (1) \begin{pmatrix} 3 \\ 3 \\ 2 \end{pmatrix} + (-1) \begin{pmatrix} -2 \\ 1 \\ 3 \end{pmatrix} + (2) \begin{pmatrix} 4 \\ 0 \\ 5 \end{pmatrix} = \begin{pmatrix} 13 \\ 6 \\ 9 \end{pmatrix}$$

$$2.10 \text{ (a) } (\mathbf{AB})' = \begin{pmatrix} 7 & 16 \\ 8 & 4 \\ 7 & 11 \end{pmatrix}, \quad \mathbf{B}'\mathbf{A}' = \begin{pmatrix} 7 & 16 \\ 8 & 4 \\ 7 & 11 \end{pmatrix} \quad (b) |\mathbf{A}| = 5$$

$$2.11 \text{ (a) } \mathbf{a}'\mathbf{b} = 5, \quad \mathbf{a}'\mathbf{b}\mathbf{b}' = 25$$

$$(b) \mathbf{bab}' = \begin{pmatrix} 4 & 2 & 6 \\ 2 & 1 & 3 \\ 6 & 3 & 9 \end{pmatrix}, \quad \mathbf{a}'(\mathbf{bab}')\mathbf{a} = 25$$

$$2.12 \quad \mathbf{Bx} = \begin{pmatrix} ax & 2bx & 3cx \\ 4b & 5b & 6b \\ 7c & 8c & 9c \end{pmatrix}, \quad \mathbf{A}\mathbf{B} = \begin{pmatrix} a & 2b & 3c \\ 4c & 5b & 6c \\ 7a & 8b & 9c \end{pmatrix},$$

$$\mathbf{B}\mathbf{A}\mathbf{B} = \begin{pmatrix} a^2 & 2ab & 3ac \\ 4ab & 5b^2 & 6bc \\ 7ac & 8bc & 9c^2 \end{pmatrix}$$

$$2.13 \quad \mathbf{A}\mathbf{B} = \left(\begin{array}{ccc|c} 8 & 9 & 5 & 6 \\ 7 & 3 & 5 & 4 \\ 3 & 4 & 2 & 2 \end{array} \right)$$

$$2.14 \quad \mathbf{A}\mathbf{B} = \begin{pmatrix} 3 & 3 \\ 1 & 4 \end{pmatrix}, \quad \mathbf{C}\mathbf{B} = \begin{pmatrix} 3 & 3 \\ 1 & 4 \end{pmatrix}$$

$$2.15 \text{ (a) } \mathbf{v}'\mathbf{A} = 5, \quad \mathbf{v}'\mathbf{B} = 2$$

$$(b) \mathbf{A} + \mathbf{B} = \begin{pmatrix} 6 & 4 & 3 \\ 2 & -2 & 1 \\ 4 & 0 & 6 \end{pmatrix}, \quad \mathbf{v}'(\mathbf{A} + \mathbf{B}) = 20$$

$$(c) |\mathbf{A}| = 8, \quad |\mathbf{B}| = 2$$

$$(d) \mathbf{A}\mathbf{B} = \begin{pmatrix} 9 & 12 & 17 \\ 3 & -1 & 3 \\ 6 & 12 & 12 \end{pmatrix}, \quad \det(\mathbf{A}\mathbf{B}) = 0$$

$$2.16 \text{ (a) } \det(A) = 70 \quad \text{(b) } T = \begin{pmatrix} 1.7521 & 2.3008 & 1.7821 \\ 8.0 & 1.8350 & 1.2347 \\ 8.0 & 8.8 & 2.1213 \end{pmatrix}$$

$$2.17 \text{ (a) } \det(A) = 1 \quad \text{(b) } T = \begin{pmatrix} 1.7521 & -2.8868 & -.5774 \\ 8.0 & 2.1682 & -.7765 \\ 8.0 & 8.0 & .2829 \end{pmatrix}$$

$$2.18 \text{ (a) } C = \begin{pmatrix} .4082 & -.5774 & .7071 \\ .5193 & .5774 & .0000 \\ .4082 & -.5774 & -.7071 \end{pmatrix}$$

$$2.19 \text{ (a) Eigenvalues: } 2, 1, -1$$

$$\text{Eigenvectors } \begin{pmatrix} .3662 \\ .9328 \\ .5009 \end{pmatrix}, \begin{pmatrix} .7008 \\ .3588 \\ .2884 \end{pmatrix}, \begin{pmatrix} .7071 \\ 0 \\ .7071 \end{pmatrix}$$

$$\text{(b) } \det(A) = 2, \quad \det(A) = -2$$

$$2.20 \text{ (a) } C = \begin{pmatrix} .8930 & .5714 & -.8085 \\ -.7671 & -.5714 & -.4882 \\ .7671 & -.5714 & -.4882 \end{pmatrix}$$

$$\text{(b) } C^TAC = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{pmatrix} \quad \text{or } C^TAC = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix} = A$$

$$2.21 \text{ Eigenvalues } 1, 3, \quad C = \begin{pmatrix} -.7071 & -.7071 \\ -.7071 & .7071 \end{pmatrix}$$

$$A^{1/2} = C D^{1/2} C^T = \begin{pmatrix} 1.3660 & -.3660 \\ -.3660 & 1.3660 \end{pmatrix}$$

$$2.22 \text{ (a) The spectral decomposition of } A \text{ is given by } A = C D C^T, \text{ where}$$

$$C = \begin{pmatrix} .455 & -.580 & .675 \\ .586 & .645 & -.531 \\ .378 & .815 & .511 \end{pmatrix} \text{ and } D = \text{diag}(23.542, 3.835, -1.477).$$

(b) The spectral decomposition of A^2 is given by $A^2 = C D^2 C^T$, where C is the same as in part (a) and $D = \text{diag}(233.338, 15.486, 6.135)$. Note that the diagonal elements of D are the squares of the diagonal elements of D in part (a).

(c) The spectral decomposition of A^{-1} is given by $A^{-1} = C D C^T$, where

$$C = \begin{pmatrix} -.580 & .455 & .675 \\ .645 & .586 & -.531 \\ .815 & .378 & .511 \end{pmatrix} \text{ and } D = \text{diag}(.254, .024, -.404).$$

The diagonal elements of $\mathbf{B}$ are the reciprocals of those of $\mathbf{D}$ in part (a). The first two columns of $\mathbf{C}$ have been interchanged to match the interchange of the corresponding elements of $\mathbf{D}$; that is, $\mathbf{D} = (1/3, 2, 1/3, 1/3, 1/3)$.

$$2.23 \quad \mathbf{A} = \mathbf{UDV}^T, \text{ where } \mathbf{D} = \text{diag}((1/5), 7, 80, 1/45),$$

$$\mathbf{U} = \begin{pmatrix} .382 & -.730 & .424 \\ .394 & -.346 & .184 \\ -.728 & .468 & .886 \\ .721 & .831 & -.640 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} .836 & -.825 & .343 \\ -.195 & .840 & .284 \\ .494 & .324 & -.807 \end{pmatrix}$$

$$2.24 \text{ (a) } f(\mathbf{x}) = (1)a_1 + (1)a_2 + \cdots + (1)a_n = \sum_1^n a_i = n$$

$$\begin{aligned} \text{(b) } f(\mathbf{A}) &= [(1)a_{11} + (1)a_{22} + \cdots + (1)a_{nn}, \dots, (1)a_{1p} \\ &\quad + (1)a_{2p} + \cdots + (1)a_{np}] \\ &= (\sum_1^n a_{1i}, \sum_1^n a_{2i}, \dots, \sum_1^n a_{pi}) \end{aligned}$$

$$\text{(c) } A_j = \begin{pmatrix} (1)a_{11} + (1)a_{22} + \cdots + (1)a_{np} \\ (1)a_{12} + (1)a_{22} + \cdots + (1)a_{np} \\ \vdots \\ (1)a_{1n} + (1)a_{2n} + \cdots + (1)a_{np} \end{pmatrix} = \begin{pmatrix} \sum_1^n a_{1i} \\ \sum_1^n a_{2i} \\ \vdots \\ \sum_1^n a_{pi} \end{pmatrix}$$

$$\begin{aligned} 2.25 \quad (x-y)(x-y) &= (x-y)(x-y) = x(x-y) - y(x-y) \\ &= x^2x - xyx + y^2x \end{aligned}$$

$$2.26 \text{ By (2.23), } (A'A)' = A'(A')'. \text{ By (2.8), } (A')' = A. \text{ Thus, } (A'A)' = A'A.$$

$$\begin{aligned} 2.27 \text{ (a) } \sum_1^n a^i a_i &= a^1 a_1 + a^2 a_2 + \cdots + a^n a_n \\ &= a^i (a_1 + a_2 + \cdots + a_n) \quad [\text{by (2.21)}] \\ &= a^i \sum_1^n a_i \end{aligned}$$

$$\begin{aligned} \text{(b) } \sum_1^n Aa_i &= Aa_1 + Aa_2 + \cdots + Aa_n \\ &= A(a_1 + a_2 + \cdots + a_n) \quad [\text{by (2.20)}] \\ &= A \sum_1^n a_i \end{aligned}$$

$$\begin{aligned} \text{(c) } \sum_1^n (a^i a_i)^2 &= \sum_1^n a^i (a_i a_i^i) \quad [\text{by (2.40)}] \\ &= a^i (\sum_1^n a_i a_i^i) \quad [\text{by (2.20)}] \end{aligned}$$

$$\text{(d) } \sum_1^n Aa_i (Aa_i)' = \sum_1^n Aa_i a_i^i A' = A (\sum_1^n a_i a_i^i) A' \quad [\text{by (2.20)}]$$

$$2.28 \text{ (a) } Ax = \begin{pmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \end{pmatrix} x = \begin{pmatrix} a_1^1 x \\ a_2^1 x \\ a_3^1 x \end{pmatrix}$$

$$\begin{aligned} \text{by } ABA' &= \begin{pmatrix} a_1' \\ a_2' \end{pmatrix} S(a_1, a_2) = \begin{pmatrix} a_1' \\ a_2' \end{pmatrix} (Sa_1, Sa_2) \quad [\text{by (2.48)}] \\ &= \begin{pmatrix} a_1' Sa_1 & a_2' Sa_1 \\ a_1' Sa_2 & a_2' Sa_2 \end{pmatrix} \end{aligned}$$

$$\text{2.29 (a) If } A = \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_n' \end{pmatrix}, \text{ then by (2.48), } A' = (a_1, a_2, \dots, a_n) \text{ and}$$

$$A'A = (a_1, a_2, \dots, a_n) \begin{pmatrix} a_1' \\ a_2' \\ \vdots \\ a_n' \end{pmatrix} = a_1 a_1' + a_2 a_2' + \cdots + a_n a_n' \quad [\text{by (2.46)}].$$

$$\text{2.30 } A^{-1}A = I$$

$$(A^{-1}A)' = I' = I$$

$$A'(A^{-1})' = I$$

$$(A')^{-1}A'(A^{-1})' = (A')^{-1}I = (A')^{-1}$$

$$(A^{-1})' = (A')^{-1}.$$

$$\begin{aligned} \text{2.31 } & \frac{1}{2} \begin{pmatrix} bA_{11}^{-1} + A_{11}^{-1}a_2a_2'a_1^{-1} & -A_{11}^{-1}a_2 \\ -a_2'a_1^{-1} & 1 \end{pmatrix} \begin{pmatrix} A_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} A + A_{11}^{-1}a_2a_2' - A_{11}^{-1}a_2a_2' & bA_{11}^{-1}a_{12} + A_{11}^{-1}a_2a_2'a_1^{-1}a_{12} - A_{11}^{-1}a_2a_{12} \\ -a_2' + a_2' & -a_2'A_{11}^{-1}a_{12} + a_{12} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} H & b \\ a' & a \end{pmatrix}, \text{ where } b = a_{22} - a_2'a_1^{-1}a_1 \\ &= \begin{pmatrix} 1 & 0 \\ 0' & 1 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \text{2.32 } & (B + \alpha I) \left(B^{-1} - \frac{B^{-1}\alpha I B^{-1}}{1 + \alpha B^{-1}\alpha} \right) \\ &= I - \frac{\alpha I B^{-1}}{1 + \alpha B^{-1}\alpha} + \alpha I B^{-1} - \frac{\alpha I B^{-1}\alpha I B^{-1}}{1 + \alpha B^{-1}\alpha} \quad [\text{by (2.25)}] \\ &= I - \alpha I B^{-1} \left(\frac{1 + I B^{-1}\alpha}{1 + \alpha B^{-1}\alpha} \right) + \alpha I B^{-1} = I \end{aligned}$$

$$\begin{aligned}
 2.33 \quad |cA| &= |cIA| \\
 &= |cI| |A| \quad [\text{by (2.89)}] \\
 &= c^n |A| \quad [\text{by (2.84)}]
 \end{aligned}$$

$$\begin{aligned}
 2.34 \quad AA^{-1} &= I \\
 |AA^{-1}| &= |I| \\
 |A| |A^{-1}| &= 1 \quad [\text{by (2.87)}] \\
 |A^{-1}| &= \frac{1}{|A|}
 \end{aligned}$$

2.35 In (2.90) and (2.94), let $A_{11} = B$, $A_{22} = a$, $A_{21} = -a'$, and $A_{12} = 1$. Then equate the right-hand sides of (2.93) and (2.94) to obtain (2.95).

$$\begin{aligned}
 2.36 \quad \text{By (2.82), } |c(AA^T)| &= \sum_{i=1}^n a_i^2 a_i = \sum_{i=1}^n a_i^2 (a_1^2 + a_2^2 + \cdots + a_n^2) \\
 &= \sum_{i=1}^n \sum_{j=1}^n a_i^2 a_j^2.
 \end{aligned}$$

2.37 Show that $|C| \neq 0$ by taking the determinant of both sides of $CC^T = I$. Thus C is nonsingular, and C^{-1} exists. Multiply $CC^T = I$ on the right by C^{-1} and on the left by C .

2.38 Multiply $AAx = \lambda x$ on the left by B . Then b is an eigenvalue of BA , and Bx is an eigenvector.

$$\begin{aligned}
 2.39 \quad (a) \quad (A^{1/2})^T J^2 &= (C(C^T)^{-1}C^T)^T J^2 = (C^T)^{1/2} C^T (C^T)^{1/2} C^T \\
 &= C^T C^T \quad [\text{by (2.101)}] \\
 &= A \quad [\text{by (2.100)}]
 \end{aligned}$$

$$(b) \quad \text{By (2.116), } A^{1/2} A^{1/2} = A. \quad \text{By (2.89),}$$

$$|A^{1/2} A^{1/2}| = |A|$$

$$|A^{1/2}| |A^{1/2}| = |A|$$

$$|A^{1/2}|^2 = |A|$$

$$(c) \quad \text{Since } A \text{ is positive definite, we have, from part (b), } |A^{1/2}| = |A|^{1/2}.$$

CHAPTER 3

3.1 $\bar{x} = \sum_{i=1}^n w_i/x = \sum_{i=1}^n w_i/x = w_1/x + \cdots + w_n/x$. Now factor a out of the sum.

3.2 The numerator of s_y^2 is $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - x_i)^2 = \sum_{i=1}^n (a + b(x_i - \bar{x}))^2$.

$$3.3 \quad \bar{x} = 4, \quad \bar{y} = 4;$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
2	3	-2	-2	4
3	4	-1	-1	1
3	6	-1	1	-1
4	2	0	-2	0
4	4	0	-1	0
4	6	0	1	0
5	3	1	-2	-2
5	4	1	-1	-1
5	6	1	1	1
				Sum = 0

$$2.4 \quad \mathbf{x} - \bar{x} \mathbf{1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \bar{x} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \end{pmatrix} = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$$

$$2.6 \quad \mathbf{y}_1 - \bar{y} \mathbf{1} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \end{pmatrix} - \begin{pmatrix} \bar{y}_1 \\ \bar{y}_1 \\ \bar{y}_1 \end{pmatrix} = \begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ y_{13} - \bar{y}_1 \end{pmatrix}$$

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n \begin{pmatrix} x_i - \bar{x}_1 \\ x_i - \bar{x}_2 \\ x_i - \bar{x}_3 \end{pmatrix} (y_i - \bar{y}_1, y_i - \bar{y}_2, y_i - \bar{y}_3) \\ &= \sum_{i=1}^n \begin{pmatrix} (x_i - \bar{x}_1)^2 & (x_i - \bar{x}_1)(x_i - \bar{x}_2) & (x_i - \bar{x}_1)(x_i - \bar{x}_3) \\ (x_i - \bar{x}_2)(x_i - \bar{x}_1) & (x_i - \bar{x}_2)^2 & (x_i - \bar{x}_2)(x_i - \bar{x}_3) \\ (x_i - \bar{x}_3)(x_i - \bar{x}_1) & (x_i - \bar{x}_3)(x_i - \bar{x}_2) & (x_i - \bar{x}_3)^2 \end{pmatrix} \end{aligned}$$

2.6 $\bar{z} = \sum_{i=1}^n z_i/n = \sum_{i=1}^n (w'y_i)/n = w'y_1 + \cdots + w'y_n/n$. Now factor out w' on the left. See also (2.42).

2.7 The numerator of s_y^2 is $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (w'y_i - w'\bar{y})^2 = \sum_{i=1}^n (w'y_i - w'\bar{y})(w'y_i - w'\bar{y})$. The scalar $w'y_i$ is equal to its transpose, as in (2.36). Thus $w'y_i = (w'y_i)' = y_i'w$, and $\sum_{i=1}^n (w'y_i - w'\bar{y})(w'y_i - w'\bar{y}) = \sum_{i=1}^n (y_i'w - w'\bar{y})(y_i'w - w'\bar{y})$. By (2.33) and (2.34), this becomes $\sum_{i=1}^n w'y_i = \bar{y}'y_1 = \bar{y}'(n)$. Now factor out w' on the left and n on the right. See also (2.44).

2.8 By (3.43) and (3.44),

$$A\mathbf{S}\mathbf{S}' = \begin{pmatrix} a_1' s_{11} & a_1' s_{12} & \cdots & a_1' s_{1n} \\ a_2' s_{21} & a_2' s_{22} & \cdots & a_2' s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_p' s_{p1} & a_p' s_{p2} & \cdots & a_p' s_{pn} \end{pmatrix},$$

from which the result follows immediately.

$$\begin{aligned}
 3.9 \quad \cos(\mu) &= \cos[(\Sigma^{1/2})^{-1}\tilde{\mu}] = (\Sigma^{1/2})^{-1}\mu \\
 &= (\Sigma^{1/2})^{-1} \cos(\tilde{\mu})(\Sigma^{1/2})^{-1} \quad \text{by (3.35)} \\
 &= (\Sigma^{1/2})^{-1} \left(\frac{\Sigma}{n} \right) (\Sigma^{1/2})^{-1} \\
 &= \frac{1}{n} (\Sigma^{1/2})^{-1} \Sigma (\Sigma^{1/2})^{-1} \quad \text{by (2.14)} \\
 &= \frac{1}{n} \mathbf{I}
 \end{aligned}$$

3.10 Answers are given in Examples 3.4 and 3.1.

$$3.11 \quad \text{(a) } \bar{y} = 439.856 \quad \text{(b) } s(\bar{y}) = 213.843$$

$$3.12 \quad \text{(a) } \bar{y} = 27,236,568 \quad \text{(b) } s(\bar{y}) = 292,898$$

$$3.13 \quad \mathbf{R} = \begin{pmatrix} 1.880 & .644 & .797 & .376 & .413 \\ .814 & 1.000 & .547 & .750 & .548 \\ .797 & .547 & 1.880 & .505 & .592 \\ .376 & .750 & .505 & 1.000 & .524 \\ .413 & .548 & .592 & .524 & 1.880 \end{pmatrix}$$

$$3.14 \quad \bar{t} = 65.258, \quad s_{\bar{t}}^2 = 1048.629$$

$$3.15 \quad r_{yx} = -.5108$$

$$3.16 \quad y_1 = (1, 3, 6)y = \mathbf{A}y, \quad \frac{1}{2}(y_2 + y_3) = (3, \frac{1}{2}, \frac{1}{2})y = \mathbf{B}y. \text{ Use (3.27) to obtain } r_{yx} = .4873.$$

$$3.17 \quad \text{(a) } \bar{t} = \begin{pmatrix} 38.369 \\ 40.825 \\ -24.727 \end{pmatrix}, \quad \mathbf{R}_t = \begin{pmatrix} 313.64 & 19.29 & -460.88 \\ 19.29 & 588.67 & 164.67 \\ -460.88 & 164.67 & 686.27 \end{pmatrix}$$

$$\text{(b) } \mathbf{R}_c = \begin{pmatrix} 1.0000 & .6441 & -.7781 \\ .6441 & 1.0000 & .3657 \\ -.7781 & .3657 & 1.0000 \end{pmatrix}$$

$$3.18 \quad \text{(a) } \bar{y} = \begin{pmatrix} 48.833 \\ 49.825 \\ 30.570 \\ 31.445 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} 6.3368 & 6.3894 & 3.7770 & 3.3548 \\ 6.3894 & 6.4405 & 6.1934 & 5.9097 \\ 3.7770 & 6.1934 & 6.6480 & 6.9287 \\ 3.3548 & 5.9097 & 6.9287 & 5.4333 \end{pmatrix},$$

$$\mathbf{R} = \begin{pmatrix} 1.0000 & .6487 & .3750 & .3669 \\ .6487 & 1.0000 & .9112 & .8730 \\ .3750 & .9112 & 1.0000 & .9689 \\ .3669 & .8730 & .9689 & 1.0000 \end{pmatrix}$$

$$\text{(b) } \bar{y} = 1.0863, \quad s(\bar{y}) = 27.1308$$

$$3.19 \quad \text{(a) } \bar{t} = 44.1400, \quad s_{\bar{t}}^2 = 21.2308, \quad \bar{w} = 103.8826, \quad s_{\bar{w}}^2 = 30.8164$$

$$\text{(b) } r_{yx} = 0.5338, \quad r_{yz} = .2553$$

$$3.20 \quad \mathbf{r} = \begin{pmatrix} -481.48 \\ -47.59 \\ 120.48 \end{pmatrix}, \quad \mathbf{R}_1 = \begin{pmatrix} 308.33 & -44.33 & 148.33 \\ -44.33 & 12.36 & -18.80 \\ 148.33 & -18.80 & 59.46 \end{pmatrix},$$

$$\mathbf{R}_0 = \begin{pmatrix} 1.00 & -.63 & .86 \\ -.63 & 1.00 & -.42 \\ .86 & -.42 & 1.00 \end{pmatrix}$$

$$3.21 \quad \mathbf{w} \left(\begin{pmatrix} x \\ y \\ z \end{pmatrix} \right) = \begin{pmatrix} 183.72 \\ 121.12 \\ 183.88 \\ 149.24 \end{pmatrix}$$

$$3.22 \quad \mathbf{S} = \begin{pmatrix} 98.29 & 81.87 & 89.66 & 46.11 \\ 82.87 & 94.56 & 91.31 & 35.03 \\ 89.66 & 91.31 & 186.82 & 36.24 \\ 46.11 & 35.03 & 36.24 & 45.82 \end{pmatrix}$$

$$3.23 \quad \begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} 70.68 \\ 75.84 \\ 75.10 \\ \frac{308.68}{308.24} \\ 308.68 \end{pmatrix}.$$

$$\mathbf{S} = \begin{pmatrix} 95.54 & 17.68 & 12.38 & 60.92 & 23.08 & 82.84 \\ 17.68 & 73.19 & 14.25 & 5.73 & 61.28 & -1.66 \\ 12.38 & 14.25 & 96.77 & 46.73 & 81.93 & 68.86 \\ 60.92 & 5.73 & 46.73 & 808.63 & 330.58 & 227.36 \\ 23.08 & 61.28 & 81.93 & 330.58 & 585.86 & 167.55 \\ 82.84 & -1.66 & 68.84 & 227.36 & 167.55 & 588.71 \end{pmatrix}$$

CHAPTER 4

- 4.1 $|\Sigma_1| = 1$, $\text{tr}(\Sigma_1) = 2$, $|\Sigma_2| = 4$, $\text{tr}(\Sigma_2) = 15$. Thus $\text{tr}(\Sigma_1) > \text{tr}(\Sigma_2)$, but $|\Sigma_1| < |\Sigma_2|$. When converted to correlations, we have

$$\mathbf{P}_{\Sigma_1} = \begin{pmatrix} 1.00 & .98 & .80 \\ .98 & 1.00 & .88 \\ .80 & .88 & 1.00 \end{pmatrix}, \quad \mathbf{P}_{\Sigma_2} = \begin{pmatrix} 1.00 & .87 & .41 \\ .87 & 1.00 & .71 \\ .41 & .71 & 1.00 \end{pmatrix}.$$

As noted at the end of Section 4.1.3, a decrease in intercorrelations or an increase in the variances will lead to a larger $|\Sigma|$. In this case, the decrease in correlations from Σ_1 to Σ_2 outweighed the increase in the variances (the increase in trace).

$$\begin{aligned}
4.2 \quad E(\mathbf{y}) &= \langle \mathbf{T} \rangle^{-1} E(\mathbf{y}) = \boldsymbol{\mu} \quad \text{[by (3.75)]} \\
&= \langle \mathbf{T} \rangle^{-1} \boldsymbol{\mu} = \boldsymbol{\mu}, \\
\text{var}(\mathbf{y}) &= \langle \mathbf{T} \rangle^{-1} \boldsymbol{\Sigma} \langle \mathbf{T} \rangle^{-1} \quad \text{[by (3.76)]} \\
&= \langle \mathbf{T} \rangle^{-1} \mathbf{T} \mathbf{T}^{-1} \quad \text{[by (2.75) and (2.76)]} \\
&= \mathbf{I}
\end{aligned}$$

4.3 By the last expression in Section 2.3.1,

$$\prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\mathbf{\Sigma}|^{1/2}} = \frac{1}{(2\pi)^{np/2} |\mathbf{\Sigma}|^{n/2}}.$$

The same in the exponent of (4.11) follows from the basic algebra of exponents.

4.4 Since $(\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ is a scalar, we have $E[(\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})] = E\{E[(\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) | \mathbf{y} - \boldsymbol{\mu}]\} = E\{E[\mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})'] | \mathbf{y} - \boldsymbol{\mu}]\} = E[\mathbf{\Sigma}^{-1} \mathbf{\Sigma}] = E(\mathbf{I}_p) = \mathbf{I}_p$.

4.5 The other two terms are of the form $\{ \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \}$, which is equal to $n[(\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})]$. This vanishes because $\sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu}) = \mathbf{a}\mathbf{y} - \mathbf{a}\boldsymbol{\mu} = \mathbf{0}$.

4.6 We replace y_1 in $\sqrt{b_1}$ by $z_1 = ay_1 + b$. By our extension of (3.34), $E = aE + b$. Then (4.18) becomes

$$\begin{aligned}
\frac{\sqrt{b_1} \sum_{i=1}^n (z_i - E)^2}{(\sum_{i=1}^n (z_i - E)^2)^{p/2}} &= \frac{\sqrt{b_1} \sum_{i=1}^n (ay_i + b - aE - b)^2}{(\sum_{i=1}^n (ay_i + b - aE - b)^2)^{p/2}} \\
&= \frac{\sqrt{b_1} \sum_{i=1}^n (ay_i - E)^2}{b^{p/2} \sum_{i=1}^n (y_i - E)^2)^{p/2}} = \frac{\sqrt{b_1} \sum_{i=1}^n (ay_i - E)^2}{(\sum_{i=1}^n (y_i - E)^2)^{p/2}} = \sqrt{b_1}.
\end{aligned}$$

Similarly, if (4.19) is expressed in terms of $z_1 = ay_1 + b$, it reduces to b_1 in terms of y_1 .

4.7 $b_{1,p} = E[(\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})^2]$ by (4.33). But when $\mathbf{y}$ is $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, $r = (\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ is distributed as $\chi^2(p)$ by property 3 in Section 4.2. Then $E(r^2) = \text{var}(r) + [E(r)]^2$.

4.8 To show that $b_{1,p}$ and $b_{2,p}$ are invariant under the transformation $\mathbf{z} = A\mathbf{y} + \mathbf{h}$, where A is nonsingular, it is sufficient to show that $\mu_{1,p}(\mathbf{z}) = (\mathbf{z}_1 - \bar{\mathbf{z}})' \mathbf{\Sigma}^{-1} (\mathbf{z}_1 - \bar{\mathbf{z}})$. By (3.67) and (3.68), $\mathbf{z} = A\mathbf{y} + \mathbf{h}$ and $\mathbf{\Sigma}_z = A\mathbf{\Sigma}_y A'$. Then $\mu_{1,p}$ for $\mathbf{z}$ becomes

$$\begin{aligned}
\mu_{1,p}(\mathbf{z}) &= (\mathbf{z}_1 - \bar{\mathbf{z}})' \mathbf{\Sigma}_z^{-1} (\mathbf{z}_1 - \bar{\mathbf{z}}) \\
&= (A\mathbf{y}_1 + \mathbf{h} - A\bar{\mathbf{y}} - \mathbf{h})' (A\mathbf{\Sigma}_y A')^{-1} (A\mathbf{y}_1 + \mathbf{h} - A\bar{\mathbf{y}} - \mathbf{h}) \\
&= (\mathbf{y}_1 - \bar{\mathbf{y}})' A' (A')^{-1} \mathbf{\Sigma}_y^{-1} A^{-1} A (\mathbf{y}_1 - \bar{\mathbf{y}}) \\
&= (\mathbf{y}_1 - \bar{\mathbf{y}})' \mathbf{\Sigma}_y^{-1} (\mathbf{y}_1 - \bar{\mathbf{y}}) = \mu_{1,p}(\mathbf{y}).
\end{aligned}$$

4.9 Let $l = (x)$ in (4.44); then solve for $D_{xx}^2 l$ in (4.43) and substitute into (4.44) to obtain $\tilde{P}_{xx}$ in terms of x , as in (4.45).

4.10 (a) $\mathbf{a}' = (5, -1, 3)$, $\mathbf{z} = \mathbf{a}'\mathbf{y}$ in $N(11, 20)$

$$(b) \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix}, \quad \mathbf{x} = \mathbf{A}\mathbf{y}$$
 in $N_2\left[\begin{pmatrix} 8 \\ 18 \end{pmatrix}, \begin{pmatrix} 28 & -2 \\ -1 & 8 \end{pmatrix}\right]$

(c) By property 4b in Section 4.2, y_2 is $N(1, 12)$.

$$(d) \text{ By property 4a in Section 4.2, } \begin{pmatrix} 8 \\ y_2 \end{pmatrix} \text{ is } N_2\left[\begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 6 & -1 \\ -1 & 4 \end{pmatrix}\right]$$

$$(e) \mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \quad \mathbf{A}\mathbf{y}$$
 in $N_3\left[\begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 6 & -2 & 3.5 \\ -2 & 4 & 1 \\ 3.5 & 1 & 3.25 \end{pmatrix}\right]$

$$4.11 (a) \mathbf{x} = \begin{pmatrix} .468 & 0 & 0 \\ -.047 & .279 & 0 \\ .285 & -.167 & .731 \end{pmatrix} \begin{pmatrix} y-3 \\ y-1 \\ y-4 \end{pmatrix}$$

$$(b) \mathbf{x} = \begin{pmatrix} .468 & -.079 & .170 \\ -.079 & .326 & -.366 \\ .170 & -.168 & .692 \end{pmatrix} \begin{pmatrix} y-3 \\ y-1 \\ y-4 \end{pmatrix}$$

(c) By (4.6), $(\mathbf{y} - \boldsymbol{\mu})'\mathbf{K}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is distributed as χ_3^2 .

4.12 (a) $\mathbf{a}' = (4, -2, 1, -3)$, $\mathbf{z} = \mathbf{a}'\mathbf{y}$ in $N(-35, 1234)$

$$(b) \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -2 & 3 & 1 & -2 \end{pmatrix}, \quad \mathbf{x} = \mathbf{A}\mathbf{y}$$
 in $N_2\left[\begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 17 & -79 \\ -79 & 364 \end{pmatrix}\right]$

$$(c) \mathbf{A} = \begin{pmatrix} 3 & 1 & -4 & -1 \\ -1 & -3 & 1 & -2 \\ 2 & 2 & 4 & -3 \end{pmatrix},$$

$$\mathbf{x} = \mathbf{A}\mathbf{y}$$
 in $N_3\left[\begin{pmatrix} -4 \\ -18 \\ -27 \end{pmatrix}, \begin{pmatrix} 23 & -18 & -6 \\ -18 & 46 & 14 \\ -6 & 14 & 55 \end{pmatrix}\right]$

(d) By property 4b in Section 4.2, y_2 is $N(-1, 2)$.

$$(e) \text{ By property 4a in Section 4.2, } \begin{pmatrix} 30 \\ y_2 \end{pmatrix} \text{ is } N_2\left[\begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 9 & -6 \\ -6 & 6 \end{pmatrix}\right]$$

$$(f) \mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{A}\mathbf{y}$$
 in $N_4\left[\begin{pmatrix} -2 \\ .5 \\ 0 \\ 1.25 \end{pmatrix}, \begin{pmatrix} 11 & 1.5 & 2 & 3.75 \\ 1.5 & 1 & .67 & .875 \\ 2 & .67 & .67 & 1 \\ 3.75 & .875 & 1 & 1.688 \end{pmatrix}\right]$

$$6.12 \text{ (a) } \mathbf{z} = \begin{pmatrix} .302 & 0 & 0 & 0 \\ .408 & .501 & 0 & 0 \\ -.087 & .351 & 1.045 & 0 \\ -.858 & -.349 & -.685 & .972 \end{pmatrix} \begin{pmatrix} y+2 \\ y-3 \\ y+1 \\ y-5 \end{pmatrix}$$

$$\text{(b) } \mathbf{z} = \begin{pmatrix} .318 & .305 & .345 & -.478 \\ .308 & .382 & .349 & -.683 \\ .143 & .348 & 1.033 & -.298 \\ -.448 & -.665 & -.288 & .787 \end{pmatrix} \begin{pmatrix} y+2 \\ y-3 \\ y+1 \\ y-5 \end{pmatrix}$$

(c) $(\mathbf{y} - \boldsymbol{\mu})'\mathbf{Z}'\mathbf{Z}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})'\mathbf{Z}^{-1/2}\mathbf{Z}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{z}'\mathbf{z}$, which is $\chi^2_3(\mathbf{y}) = \chi^2_3(\mathbf{z})$.

6.14 The variables in (b), (c), and (d) are independent.

6.15 The variables in (a), (c), (d), (f), (i), (j), and (h) are independent.

6.16 (a) $E(\mathbf{y}|\mathbf{x}) = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)$

$$= \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -3 & 2 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - 3 \\ x_2 - 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -.3 & .25 \\ .5 & 1.25 \end{pmatrix} \begin{pmatrix} x_1 - 3 \\ x_2 - 1 \end{pmatrix}$$

$$= \begin{pmatrix} 3.25 \\ -3.75 \end{pmatrix} + \begin{pmatrix} -.3 & .25 \\ .5 & 1.25 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

(b) $\text{cov}(\mathbf{y}|\mathbf{x}) = \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{10}$

$$= \begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix} - \begin{pmatrix} -3 & 2 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}^{-1} \begin{pmatrix} -3 & 8 \\ 2 & 4 \end{pmatrix}$$

$$= \begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix} - \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 2 & 4 \end{pmatrix}$$

6.17 (a) $E(\mathbf{y}|\mathbf{x}) = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)$

$$= \begin{pmatrix} 3 \\ -2 \end{pmatrix} + \begin{pmatrix} 15 & 0 & 3 \\ 0 & 6 & -2 \end{pmatrix} \begin{pmatrix} 20 & 8 & 3 \\ 0 & 4 & 0 \\ 3 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - 4 \\ x_2 + 3 \\ x_3 - 5 \end{pmatrix}$$

$$= \begin{pmatrix} 3 \\ -2 \end{pmatrix} + \begin{pmatrix} 15 \\ -24.8 \end{pmatrix} + \begin{pmatrix} 8 & 0 & 3 \\ .87 & .187 & -.373 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$= \begin{pmatrix} -12 \\ 22.8 \end{pmatrix} + \begin{pmatrix} 0 & 8 & 3 \\ .87 & .187 & -.373 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

(b) $\text{cov}(\mathbf{y}|\mathbf{x}) = \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{10}$

$$= \begin{pmatrix} 18 & -8 \\ -8 & 18 \end{pmatrix} - \begin{pmatrix} 15 & 0 & 3 \\ 0 & 6 & -2 \end{pmatrix} \begin{pmatrix} 20 & 8 & 3 \\ 0 & 4 & 0 \\ 3 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 12 & 8 \\ 0 & 8 \\ 3 & -2 \end{pmatrix}$$

$$= \begin{pmatrix} 18 & -8 \\ -8 & 18 \end{pmatrix} - \begin{pmatrix} 9 & -6 \\ -6 & 17 \end{pmatrix} = \begin{pmatrix} 9 & -2 \\ -2 & 1 \end{pmatrix}$$

4.18 (a) By the central limit theorem in Section 4.3.1, $\sqrt{n}(\bar{Y} - \mu)$ is approximately $N(\mu, \Sigma)$.

(b) $\bar{Y}$ is approximately $N(\mu, \Sigma/n)$.

4.19 (a) The plots show almost no deviation from normality.

Variable	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
$\sqrt{n}\bar{Y}_1$	.000	.1111	.0443	.007
$\bar{y}_1$	1.632	1.671	1.792	1.879

The values of $\sqrt{n}\bar{Y}_1$ show a small amount of positive skewness, but none exceeds the upper 1.5% critical value for $\sqrt{n}\bar{Y}_1$ given in Table A.1 as .842. The values of $\bar{y}_1$ show negative kurtosis. For $\bar{y}_4$, the kurtosis is significant, since $\bar{y}_4 < 1.34$, the lower 2.5 percentile in Table A.1.

Variable	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
$\bar{y}$	.2648	.2641	.2690	.2679
F	.4020	.2534	.0750	.4491

From Table A.4, the lower 2.5 percentile for F is -3.64 and the upper 97.5 percentile is $.828$. We reject the hypothesis of normality only for $\bar{y}_4$.

(b) z defined in (4.34) is approximately $N(0, 1/n)$. To obtain a $N(0, 1)$ statistic, we calculate $z^* = z\sqrt{n}/\sqrt{2}$.

Variable	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
z^*	-.1866	-.0000	-.0717	-.0006

4.20 (a)	i	1	2	3	4	5	6	7	8	9	10
	$\bar{y}_i^2$	1.00	1.00	1.34	1.34	0.61	.61	.81	2.67	.85	1.78

(b) The .05 critical value from Table A.8 is 7.01. $\bar{y}_{\text{avg}}^2 = 7.94 > 7.01$.

(c)	i	1	2	3	4	5	6	7	8	9	10
	$\alpha_{i,2}$	.08	.10	.12	.13	.20	.50	.44	.47	.57	.95
	α_i	.07	.11	.18	.23	.28	.34	.40	.45	.55	.68

The plot of $(\alpha_i, \alpha_{i,2})$ shows some evidence of nonlinearity and an outlier.

(d) $\bar{y}_{1,2} = 7.255$, $\bar{y}_{2,2} = 14.405$. Both (barely) exceed upper .05 critical values in Table A.8.

4.21 (a)	Variable	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$
	$\sqrt{n}\bar{Y}_1$	.2076	.0037	.1864	-.0027	-.1772
	$\bar{y}_1$	2.879	1.681	2.580	1.794	2.450

None of the values of $\sqrt{n}\bar{Y}_1$ exceeds 1.154 (from Table A.1) or is less than -1.154 . None of the values of $\bar{y}_1$ is less than 1.53 (from Table A.3). Thus there is no significant departure from normality.

(d) Variable	y_1	y_2	y_3	y_4	y_5
$\bar{y}$	.276	.260	.279	.263	.276
$\bar{r}$	-.893	-1.099	-.893	-.114	-.699

(e) $\bar{r}^2 = .2149/375$, where $\bar{r}$ is defined in (4.24).

Variable	y_1	y_2	y_3	y_4	y_5
$\bar{r}^2$	-.4946	-1.7165	-1.3627	.8061	.5686

4.22 (a)	i	1	2	3	4	5	6	7	8	9	10	11
	$\bar{y}_i^2$	5.10	1.15	1.58	5.34	3.34	1.75	5.21	5.90	3.72	6.02	2.56
(a)	i	1	2	3	4	5	6	7	8	9	10	11
	$\bar{u}_{1i}$	.38	.28	.28	.30	.37	.37	.39	.44	.43	.46	.38
	$\bar{u}_i$	.38	.27	.34	.39	.43	.36	.33	.44	.46	.73	.42

The plot shows a sharp break from the fourth to the fifth points.

(b) $\bar{r}_{1,p} = 12.985$, $\bar{r}_{1,p} = 29.872$

4.23 (a) The $\bar{Q}$ - $\bar{Q}$ plots for y_1 and y_2 show little departure from normality. The $\bar{Q}$ - $\bar{Q}$ plots for y_3 and y_4 show some evidence of heavier tails than the normal. The $\bar{Q}$ - $\bar{Q}$ plots for y_5 and y_6 show some evidence of positive skewness.

(d) Variable	y_1	y_2	y_3	y_4	y_5	y_6
$\sqrt{V_1}$	.8928	.8862	.9427	1.8627	.2208	.4974
$\bar{u}_1$	31.68	3.253	2.712	6.875	2.176	4.328

(d) Variable	y_1	y_2	y_3	y_4	y_5	y_6
$\bar{y}$	.276	.261	.279	.260	.266	.271
$\bar{r}$	-0.469	-1.045	-1.873	-5.289	.889	-2.741

(d) Variable	y_1	y_2	y_3	y_4	y_5	y_6
$\bar{r}^2$	-0.448	-.692	-2.883	-10.661	-.878	-2.466

4.24 (a) $\bar{y}_i^2 = 7.816, 1.640, 1.736, \dots, 6.403$

(b) $\bar{y}_{1,p}^2 = 25.028$. By interpolation in Table A.5, the $\bar{y}$ -critical value for $p = 5$ is approximately 19. Thus we reject the hypothesis of multivariate normality.

(c) $\bar{u}_{1,1}, \bar{u}_{1,2} = (.621, .654, 1.029, .028), \dots, (.506, .521)$. The plot shows nonlinearity for the last 4 points.

(d) $\bar{r}_{1,p} = 16.287$, $\bar{r}_{2,p} = 28.371$. By interpolation to $p = 5$ in Table A.5, both appear to exceed their critical values.

CHAPTER 5

5.1 By (3.8), we have

$$\begin{aligned} \bar{y} - \mu_0 / \left(\frac{2}{n} \right)^{-1} (\bar{y} - \mu_0) &= (\bar{y} - \mu_0) / \left(\frac{1}{n} \right)^{-1} S^{-1} (\bar{y} - \mu_0) \\ &= n (\bar{y} - \mu_0) / S^{-1} (\bar{y} - \mu_0). \end{aligned}$$

5.2 From (3.9), we have

$$\begin{aligned} \frac{n\mu_1}{n_1 + n_2} (\bar{y}_1 - \bar{y}_2) / S_{\mu_1}^{-1} (\bar{y}_1 - \bar{y}_2) &= (\bar{y}_1 - \bar{y}_2) / \left(\frac{n_1 + n_2}{n_1 n_2} \right)^{-1} S_{\mu_1}^{-1} (\bar{y}_1 - \bar{y}_2) \\ &= (\bar{y}_1 - \bar{y}_2) / \left(\frac{n_1 + n_2}{n_1 n_2} S_{\mu_1} \right)^{-1} (\bar{y}_1 - \bar{y}_2) \\ &= (\bar{y}_1 - \bar{y}_2) / \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{\mu_1} \right]^{-1} (\bar{y}_1 - \bar{y}_2). \end{aligned}$$

5.3 By (3.12) and (3.14),

$$J^2(\mu) = \frac{[\mu'(\bar{y}_1 - \bar{y}_2)]^2}{(n_1 + n_2)(n_1 n_2) S_{\mu_1} \mu} = \frac{n_1 n_2}{n_1 + n_2} \frac{(\bar{y}_1 - \bar{y}_2)' S_{\mu_1}^{-1} (\bar{y}_1 - \bar{y}_2)}{S_{\mu_1} S_{\mu_1}^{-1} (\bar{y}_1 - \bar{y}_2)}.$$

5.4 It is assumed that y and x have a bivariate normal distribution. Let $y_0 = \begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix}$. Then y_0 can be expressed as $y_0 = y_1 + y_2 = \mu' y_0$, where $\mu' = (1, -1)$. By property 1a in Section 4.2, y_0 is $N(\mu' \mu, \mu' \Sigma \mu)$. Show that $\mu' \bar{y} = \bar{y} - \bar{x}$, $\mu' \Sigma \mu = s_y^2 - 2s_{xy} + s_x^2 = s_D^2$, and that $J^2 = n(\mu' \bar{y})' (\mu' \Sigma \mu)^{-1} (\mu' \bar{y})$ is the square of $J = \bar{D}'(y_0) / \sqrt{s_D}$.

$$\begin{aligned} \bar{D} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n (y_i - x_i) = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n x_i = \bar{y} - \bar{x}, \\ s_D^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{D})^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - x_i - \bar{y} + \bar{x})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{y}) - (x_i - \bar{x})]^2. \end{aligned}$$

When this is expanded, we obtain $s_D^2 = s_y^2 + s_x^2 - 2s_{xy}$.

5.6 The solution is similar to that for Problem 5.1.

5.7 By (3.7), $(y - \mu + (y - \mu) D_{\mu}^T D_{\mu}^{-1}) = P_{D, y - \mu + 1}$. By (3.29), $(y - \mu)' (P_{D, y - \mu}^T y - P_{D, y - \mu}^T y) = P_{D, y - \mu}^T y$. Replacing y by y and x by $x - \mu$ in (3.7), we see that $\frac{(x - \mu)' (x - \mu)}{D_{\mu}^T D_{\mu}} (x - \mu) = \frac{P_{D, x - \mu}^T (x - \mu)}{D_{\mu}^T D_{\mu}}$ is $P_{D, x - \mu + 1}$.

5.8 Under H_{02} , we have $C\mu_1 = 0$ and $C\mu_2 = 0$. Then

$$E(C\bar{y}) = CE(\bar{y}) = CE\left(\frac{n\bar{y}_1 + n\bar{y}_2}{n_1 + n_2}\right) = \frac{n_1 C\mu_1 + n_2 C\mu_2}{n_1 + n_2} = 0.$$

Since $\tilde{y}_1$ and $\tilde{y}_2$ are independent,

$$\begin{aligned}\text{cov}(\tilde{\mathbf{y}}) &= \text{cov}\left(\frac{y_1\tilde{y}_1 + y_2\tilde{y}_2}{y_1 + y_2}\right) = \frac{\pi^2(\tilde{y}_1^2/y_1 + \pi^2\tilde{y}_2^2/y_2)}{(y_1 + y_2)^2} \\ &= \frac{(y_1 + y_2)\tilde{\Sigma}}{(y_1 + y_2)^2}.\end{aligned}$$

5.10 $(\Sigma_y\pi^2)/(y_1 + y_2)$ is the sample covariance matrix of $\tilde{\mathbf{y}}$. Hence the equation immediately above (5.8) exhibits the characteristic form of the F^2 -distribution.

5.11 $F^2 = .081$

5.12 (a) $F^2 = 45.3317$

(b) $t_1 = 1.5836, t_2 = .3665, t_3 = -2.5157, t_4 = .6618, t_5 = .3681$

5.13 $F^2 = 30.3868$

5.14 (a) $F^2 = 1.4193$

(b) $t_1 = 1.1643, t_2 = 1.1605, t_3 = .9852, t_4 = .7295$. None of these is significant. In fact, ordinarily they would not have been examined because the F^2 -test in part (a) did not reject H_0 .

5.15 $F^2 = 79.8513$

5.16 (a) $F^2 = 135.4875$

(b) $t_1 = 3.4876, t_2 = -3.4652, t_3 = -3.4611, t_4 = -3.4435$

(c) $\mathbf{a}' = (-345, -130, -106, -143)$

(d) $F^2 = 135.4875$

(e) $R^2 = .782913, F^2 = 135.4875$

(f) By (5.12), $r^2(y_1/y_2, y_3, y_4) = .553158, r^2(y_1/y_2, y_3, y_4) = .57984, r^2(y_1/y_2, y_3, y_4) = .57949, r^2(y_4/y_2, y_3, y_4) = .82595$

(g) $R^2(3.7\%, F^2(y_1, y_4/y_2, y_3) = 12.5286, F(y_1, y_4/y_2, y_3) = 6.8814$

5.17 By (5.34), the test for parallelism gives $F^2 = 152.6859$. The discriminant function coefficient vector is given by (5.38) as $\mathbf{a}' = (-.362, -.225, -.137)$.

5.18 (a) $F^2 = 65.6604$

(b) $t_1 = -.6556, t_2 = 1.6439, t_3 = -3.3884, t_4 = -4.6315, t_5 = 1.6874, t_6 = -5.1209$

(c) By (5.12),

$$r^2(y_1/y_2, y_3, y_4, y_5, y_6) = .8738, \quad r^2(y_1/y_2, y_3, y_4, y_5, y_6) = .84513,$$

$$r^2(y_1/y_2, y_3, y_4, y_5, y_6) = .85514, \quad r^2(y_4/y_2, y_3, y_5, y_6, y_1) = .82089,$$

$$r^2(y_1/y_2, y_3, y_4, y_5, y_6) = .85582, \quad r^2(y_4/y_2, y_3, y_5, y_6, y_1) = .82628.$$

(d) By (5.12), $F^2(y_1, y_3, y_4/y_2, y_5, y_6) = 27.547$.

$$5.19 \text{ (a) } T^2 = 70.5679 \quad \text{(b) } T^2(y_1, y_2 | x_1, x_2) = 15.1317$$

$$\text{(c) } T^2(y_1, y_2 | y_3, y_4, y_5, y_6) = 9.5362$$

$$5.20 \text{ (a) } T^2 = 18.4625 \quad \text{(b) } \mathbf{a}' = (-.053, -.303, -.362, -.371)$$

$$\text{(c) } \operatorname{Re}(5.32),$$

$$t^2(y_1 | y_2, y_3, y_4) = 3.3343, \quad t^2(y_2 | y_1, y_3, y_4) = .0702,$$

$$t^2(y_3 | y_1, y_2, y_4) = 1.4923, \quad t^2(y_4 | y_1, y_2, y_3) = .0813.$$

$$5.21 \text{ (a) } T^2 = 15.1812 \quad \text{(b) } \mathbf{a}' = (-.056, .048)$$

$$\text{(c) } r_1 = -3.6576, r_2 = -2.4562$$

$$5.22 \quad T^2 = 22.3288$$

$$5.23 \text{ (a) } T^2 = 208.1188$$

$$\text{(b) } t^2(a_1 | a_2, a_3) = 79.0370, \quad t^2(a_2 | a_1, a_3) = 35.4507, \quad t^2(a_3 | a_1, a_2) = 80.3349$$

CHAPTER 6

6.1 (a) Using $\bar{y}_i = y_i/n$, we have

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^k (y_{ij} - \bar{y}_i)^2 &= \sum_{i=1}^k (x_i^2 - 2x_i\bar{y}_i + \bar{y}_i^2) \\ &= \sum_{i=1}^k x_i^2 - \sum_{i=1}^k \bar{y}_i \sum_{j=1}^k y_{ij} + n \sum_{i=1}^k \bar{y}_i^2 \\ &= \sum_{i=1}^k x_i^2 - 2 \sum_{i=1}^k \frac{\bar{y}_i}{n} n_i + n \sum_{i=1}^k \left(\frac{\bar{y}_i}{n}\right)^2 \\ &= \sum_{i=1}^k x_i^2 - 2 \sum_{i=1}^k \frac{\bar{y}_i^2}{n} + \sum_{i=1}^k \frac{\bar{y}_i^2}{n}. \end{aligned}$$

$$6.2 \quad \frac{\mathbf{E}^{-1}(\mathbf{E})}{(\mathbf{E}^{-1}(\mathbf{E} + \mathbf{I}))} = \frac{\mathbf{E}^{-1}\mathbf{E}}{\mathbf{E}^{-1}(\mathbf{E} + \mathbf{I})} = \frac{\mathbf{I}}{\mathbf{I} + \mathbf{E}^{-1}\mathbf{E}} = \frac{1}{\prod_{i=1}^n (1 + \lambda_i)^2},$$

see Section 2.11.2.

$$\begin{aligned} 6.3 \quad & (\mathbf{E}^{-1}\mathbf{I} - \mathbf{I})\mathbf{a} = \mathbf{0} \\ & (\mathbf{E}^{-1/2}\mathbf{E}^{1/2}\mathbf{E}^{-1}\mathbf{I} - \mathbf{I})\mathbf{a} = \mathbf{0} \\ & (\mathbf{E}^{-1/2}\mathbf{E}_1^{-1}\mathbf{E}^{1/2}\mathbf{E}_1^{-1}\mathbf{I} - \mathbf{I})\mathbf{a} = \mathbf{0} \\ & (\mathbf{E}^{-1/2}\mathbf{E}_1^{-1}\mathbf{I} - \mathbf{I})\mathbf{E}^{1/2}\mathbf{a} = \mathbf{0} \\ & (\mathbf{E}^{-1/2}\mathbf{E}_1^{-1}\mathbf{I} - \mathbf{I})\mathbf{E}^{1/2}\mathbf{E}_1^{-1}\mathbf{E}^{1/2}\mathbf{a} = \mathbf{0} \\ & (\mathbf{E}^{-1/2}\mathbf{E}_1^{-1}\mathbf{I})\mathbf{E}^{1/2}\mathbf{E}_1^{-1}\mathbf{a} = \mathbf{0} \end{aligned}$$

6.4 We need to show that $(2s + i + 1)(2m + i + 1) = (r_2 - p + i)/s'$. Using the definitions $K = \frac{1}{2}(r_2 - p - 1)$, $m = \frac{1}{2}(r_2 - p) - 1$, $s' = \max(p, m)$,

and $x = \min\{p, v_B\}$, we have $2\theta + x + 1 = 2\lfloor \frac{1}{2}x(v_B - p - 1) + x + 1 \rfloor = v_B - p - 1 + x + 1 = v_B - p + x$. For the denominator, we have $2\theta + x + 1 = 2\lfloor \frac{1}{2}x(v_B - p) - 1 \rfloor + x + 1 = (v_B - p) + x$. Suppose $v_B = p$. Then $(v_B - p) + x = v_B - p + p = v_B = d$. On the other hand, if $v_B < p$, then $(v_B - p) + x = p - v_B + v_B = p = d$.

6.6 If $p \leq v_B$, we have $x = p$ and $(v_B - p) = v_B = p$. Then (6.30) becomes

$$\begin{aligned} \frac{2\alpha(N+1)4^{\theta\theta}}{\theta^2(2\theta+x+1)} &= \frac{2\left[p\binom{1}{1}\right](v_B-p-1)+1}{\theta^2\left[2\binom{1}{1}\right](v_B-p-1)+p+1)} 2^{\theta\theta} \\ &= \frac{1p(v_B-p-1)+2\theta^{\theta\theta}}{\theta^2(v_B-p-1+p+1)} \\ &= \frac{1p(v_B-p-1)+2\theta^{\theta\theta}}{\theta^2v_B}, \end{aligned}$$

which is the same as (6.31) because $p = v$.

6.6 When $x = 1$, we have $\mathbf{Y}^{(2)} = \lambda_1/(1 + \lambda_1)$, $\mathbf{Z}^{(2)} = \lambda_1$, $\mathbf{E} = \mathbf{1}/(1 + \lambda_1)$, and $\mathbf{F} = \lambda_1/(1 + \lambda_1)$. Solving the last of them for λ_1 gives $\lambda_1 = \mathbf{F}/(\mathbf{1} - \mathbf{F})$, and the results in (6.34), (6.35), and (6.36) follow immediately.

6.7 With $\mathbf{F}' = (\alpha_1 + \alpha_2 - 2\mathbf{F})^{(2)}$ and $\mathbf{F}^{(2)} = \mathbf{F}/(1 - \mathbf{F})$, we obtain (6.19). We obtain (6.18) from (6.19) by $\mathbf{F}^{(2)} = \mathbf{F}$. A similar argument leads to (6.16).

6.8 (a) With $\mathbf{F}_1 = \mathbf{F}_1/\alpha_1$ and $\mathbf{F}_2 = \mathbf{F}_2/\alpha_2$, we obtain

$$\begin{aligned} \mathbf{H} &= \sum_{i=1}^n \alpha_i (\mathbf{F}_i - \mathbf{F}_i)(\mathbf{F}_i - \mathbf{F}_i) \\ &= \sum_i \alpha_i (\mathbf{F}_i \mathbf{F}_i - \mathbf{F}_i \mathbf{F}_2 - \mathbf{F}_2 \mathbf{F}_i + \mathbf{F}_i \mathbf{F}_i) \\ &= \sum_i \alpha_i \mathbf{F}_i - \left(\sum_i \alpha_i \mathbf{F}_i \right) \mathbf{F}_2 - \mathbf{F}_2 \sum_i \alpha_i \mathbf{F}_i + \mathbf{F}_2 \mathbf{F}_2 \sum_i \alpha_i \\ &= \sum_i \alpha_i \frac{\mathbf{F}_i \mathbf{F}_i}{\alpha_i^2} - \frac{(\sum_i \alpha_i) \mathbf{F}_2 \mathbf{F}_2}{N} - \frac{\mathbf{F}_2}{N} \sum_i \alpha_i \mathbf{F}_i + \frac{\mathbf{H}_2 \mathbf{F}_2}{N^2} \\ &= \sum_i \alpha_i \frac{\mathbf{F}_i \mathbf{F}_i}{\alpha_i} - \frac{\mathbf{F}_2 \mathbf{F}_2}{N} - \frac{\mathbf{F}_2 \mathbf{F}_2}{N} + \frac{\mathbf{F}_2 \mathbf{F}_2}{N}. \end{aligned}$$

6.9 $\mathbf{F}_1 - \mathbf{F}_2$ becomes

$$\mathbf{F}_1 - \mathbf{F}_2 = \frac{\alpha_1 \mathbf{F}_1 + \alpha_2 \mathbf{F}_2}{\alpha_1 + \alpha_2} = \frac{\alpha_1 \mathbf{F}_1 + \alpha_2 \mathbf{F}_1 - \alpha_1 \mathbf{F}_1 - \alpha_2 \mathbf{F}_2}{\alpha_1 + \alpha_2} = \frac{\alpha_1 (\mathbf{F}_1 - \mathbf{F}_2)}{\alpha_1 + \alpha_2}.$$

The first term in the sum is

$$\frac{\sigma_1 \sigma_2^2}{(\sigma_1 + \sigma_2)^3} (\mathbf{Q}_1 - \mathbf{P}_2)(\mathbf{Q}_1 - \mathbf{P}_2)^T.$$

The second term in the sum is

$$\frac{\sigma_2^3 \sigma_3}{(\sigma_1 + \sigma_2)^3} (\mathbf{Q}_1 - \mathbf{P}_2)(\mathbf{Q}_1 - \mathbf{P}_2)^T.$$

$$6.10 \quad \beta = \frac{\lambda_1}{1 + \lambda_1} = \frac{5511(2)(5511(2))}{1 + 5511(2)(5511(2))} = \frac{5511(2)}{5511(2) + 5511(2)}$$

6.11 From $r_1^2 = \lambda_{\alpha^2}(I) + \lambda_{\beta^2}$, obtain $\lambda_{\beta^2} = r_1^2(I) - r_1^2$. Substitute this into $I(I) + \lambda_{\beta^2}$ to obtain the result.

6.12 Substitute $A_F = I^{(D)}/\alpha$ into (5.35) to obtain (5.36).

6.13 When $\alpha = 1$, (5.34) becomes

$$A_{\text{LR}} = \frac{I^{(D)}}{1 + I^{(D)}}$$

By (5.34), $I^{(D)} = \lambda_D$.

6.14 Substitute $A_{\text{LR}} = I^{(D)}/(\alpha + I^{(D)})$ from (5.34) into (5.32) to obtain P_2 in (5.34).

6.15 To show $\text{cov}(y, \hat{y}_k) = c_k^T \Sigma / \alpha$, use (5.34), $\text{cov}(A_F) = A \Sigma A^T$, with $A = c_k \mathbf{1}$.

6.16 By (5.34),

$$\begin{aligned} \mathbf{H}_1 &= \alpha \sum_{i=1}^K (\mathbf{Q}_i - \mathbf{P}_i)(\mathbf{Q}_i - \mathbf{P}_i)^T \\ &= \alpha \sum_{i=1}^K (\mathbf{Q}_i \mathbf{Q}_i^T - \mathbf{Q}_i \mathbf{P}_i^T - \mathbf{P}_i \mathbf{Q}_i^T + \mathbf{P}_i \mathbf{P}_i^T) \\ &= \alpha \sum_{i=1}^K (\mathbf{Q}_i \mathbf{Q}_i^T - \mathbf{P}_i \mathbf{P}_i^T) (\mathbf{Q}_i \mathbf{Q}_i^T - \mathbf{P}_i \mathbf{P}_i^T)^T \\ &= \alpha \mathbf{C} \left[\sum_{i=1}^K (\mathbf{Q}_i - \mathbf{P}_i)(\mathbf{Q}_i - \mathbf{P}_i)^T \right] \mathbf{C}^T \quad \text{by (2.40)} \end{aligned}$$

6.17 $\mathbf{C}$ is not square.

$$\begin{aligned} 6.18 \quad E(\mathbf{Q}_i \mathbf{Q}_i^T) &= E(\mathbf{P}_i \mathbf{P}_i^T) = E\mathbf{A} \left(\sum_{j=1}^K \mathbf{Q}_j / K \right) \\ &= E \left(\sum_{j=1}^K \mathbf{A}(\mathbf{Q}_j) / K \right) = E \sum_{j=1}^K \mu_j / K \\ &= \Phi \quad \text{[by } \mathbf{H}_0 \text{ in (5.34)]} \end{aligned}$$

$\text{cov}(\mathbf{Q}_i) = E \Sigma \mathbf{C}^T$ (as if there are no differences in the group means, $\mathbf{C} \mu_1, \mathbf{C} \mu_2, \dots, \mathbf{C} \mu_K$). This condition is ensured by $\mathbf{H}_0$ in (5.34).

6.19 For our purposes, it will suffice to show that T^2 has the characteristic form of the F^2 -distribution in (3.6).

6.20 If $\Sigma = \sigma^2 \mathbf{I}$, (3.19) becomes

$$r = \frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{I} - \mathbf{J})\sigma^2(\mathbf{y} - \mathbf{X}\beta)}{(\mathbf{y} - \mathbf{1})(\mathbf{I} - \mathbf{J})\sigma^2(\mathbf{y} - \mathbf{1})} = \frac{(\mathbf{y} - \mathbf{1})'(\mathbf{I} - \mathbf{J})\sigma^2(\mathbf{y} - \mathbf{1})}{\sigma^2(\mathbf{y} - \mathbf{1})'(\mathbf{I} - \mathbf{J})\sigma^2(\mathbf{y} - \mathbf{1})}.$$

Show that $(\mathbf{I} - \mathbf{J})\mathbf{y} = \mathbf{y} - \mathbf{1}$. Then

$$r = \frac{\sigma^2(\mathbf{y} - \mathbf{1})'(\mathbf{y} - \mathbf{1})}{\sigma^2(\mathbf{y} - \mathbf{1})(\mathbf{y} - \mathbf{1})} = \frac{(\mathbf{y} - \mathbf{1})^2}{(\mathbf{y} - \mathbf{1})^2} = 1.$$

6.21 The (univariate) expected mean square corresponding to μ , in a one-way ANOVA, is $\sigma^2 + B\mu^2$. Thus the mean square for β , is tested with MSE. The corresponding multivariate test therefore uses $\mathbf{W}^2$ and $\mathbf{E}$.

6.22 From (5.105) we have

$$A = \frac{(\mathbf{A}\mathbf{E}\mathbf{A})}{(\mathbf{A}\mathbf{E}\mathbf{A} + \mathbf{W}^2\mathbf{A})} = \frac{(\mathbf{A}\mathbf{E}\mathbf{A})}{(\mathbf{A}\mathbf{E}\mathbf{A}) + \mathbf{A}\mathbf{W}^2\mathbf{A}}.$$

Substitute $\mathbf{W}^2 = \lambda_1\mathbf{y}_1\mathbf{y}_1' + \lambda_2\mathbf{y}_2\mathbf{y}_2'$, to obtain

$$A = \frac{(\mathbf{A}\mathbf{E}\mathbf{A})}{(\mathbf{A}\mathbf{E}\mathbf{A}) + \lambda_1^2(\mathbf{y}_1\mathbf{A}\mathbf{E}\mathbf{A}) + \lambda_2^2(\mathbf{y}_2\mathbf{A}\mathbf{E}\mathbf{A})}.$$

Now use (2.93) with $\mathbf{B} = \mathbf{A}\mathbf{E}\mathbf{A}$ and $\mathbf{r} = \sqrt{\lambda_1}\mathbf{y}_1$, to obtain

$$A = \frac{1}{1 + \lambda_1(\mathbf{y}_1'\mathbf{y}_1)(\mathbf{A}\mathbf{E}\mathbf{A})^{-1}(\mathbf{A}\mathbf{y}_1)}.$$

Multiply and divide by v_1 and use (5.103) to obtain (6.106).

6.23 Same as F^2 in (6.106).

6.24 In $C_1(\mathbf{E})$ the rows of C_1 are multiplied by the rows of $\mathbf{E}$. Show that $C_1\mathbf{B}' = \mathbf{0}$.

6.25 As noted, the function $(\mathbf{y} - \mathbf{A}_0\mathbf{\beta})'(\mathbf{S}^{-1}(\mathbf{y} - \mathbf{A}_0\mathbf{\beta}))$ is similar to $\text{RSS} = (\mathbf{y} - \mathbf{X}_0\mathbf{\beta})'(\mathbf{y} - \mathbf{X}_0\mathbf{\beta})$ in (10.4) and (10.5). By an argument similar to that used in Section 10.2.2 to obtain $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, it follows that $\hat{\boldsymbol{\beta}} = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}_0)^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{y}$. An alternative approach (for those familiar with differentiation with respect to a vector) is to expand $(\mathbf{y} - \mathbf{A}_0\mathbf{\beta})'(\mathbf{S}^{-1}(\mathbf{y} - \mathbf{A}_0\mathbf{\beta}))$ to four terms, differentiate with respect to $\hat{\boldsymbol{\beta}}$, and set the result equal to $\mathbf{0}$.

6.26 Expand $\mathbf{r}(\mathbf{y} - \mathbf{A}_0\mathbf{\beta})'(\mathbf{S}^{-1}(\mathbf{y} - \mathbf{A}_0\mathbf{\beta}))$ to four terms and substitute

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}_0)^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{y}$$

into the last one.

$$6.27 \text{ (a) } \mathbf{B} = \begin{pmatrix} 13.42 & 7.32 & 8.68 & 5.86 \\ 7.72 & 8.48 & 7.85 & 6.21 \\ 8.68 & 7.23 & 11.64 & 7.84 \\ 5.86 & 6.21 & 7.04 & 10.27 \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 1.05 & 2.37 & -1.28 & -.78 \\ 2.17 & 4.89 & -2.33 & -1.26 \\ -1.32 & -2.37 & 2.38 & 1.38 \\ -.78 & -1.26 & 1.28 & .81 \end{pmatrix}.$$

$\lambda = .224$, $F^{(3)} = .866$, $D^{(3)} = 1.08$, and $\theta = .247$. All four are significant.

(b) $\eta_{11}^2 = 1 - \lambda = .775$, $\eta_{12}^2 = \theta = .247$, $\lambda_1 = 1 - \lambda^{1/3} = .326$, $\lambda_{12} = .896$, $\lambda_F = F^{(3)}/\eta = .430$.

(c) The eigenvalues of $\mathbf{E}^{-1}\mathbf{B}$ are .20515 and .1273. The essential dimensionality of the space of the mean vectors is 1.

(d) For 1, 2 vs. 3 we have $\lambda = .210$, $F^{(2)} = .380$, $D^{(2)} = 2.502$, and $\theta = .738$. All four are significant. For 1 vs. 2 we obtain $\lambda = .728$, $F^{(2)} = .224$, $D^{(2)} = .377$, and $\theta = .176$. All four are significant.

(e) Variable	β_1	β_2	β_3	β_4
F	1.29	6.88	3.19	1.27

The F 's for β_2 and β_3 are significant. For the discriminant function $z = \mathbf{a}'\mathbf{y}$, where $\mathbf{a}$ is the first eigenvector of $\mathbf{E}^{-1}\mathbf{B}$, we have $\mathbf{a}' = (-.652, -.810, .515, .206)$. Again β_2 and β_3 contribute most to separation of groups.

(f) By (6.177), $\delta(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{y}_3, \mathbf{y}_4) = \delta(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4) / \delta(\mathbf{y}_1, \mathbf{y}_2) = .724 / .908 = .798 < \lambda_{12} = .728$.

(g) By (6.128),

$$\begin{aligned} \delta(\mathbf{y}_1 | \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4) &= \delta(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4) / \delta(\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4) \\ &= .724 / .948 = .764 > \lambda_{12} = .728, \end{aligned}$$

$$\delta(\mathbf{y}_2 | \mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_4) = .724 / .938 = .772 < .819,$$

$$\delta(\mathbf{y}_3 | \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_4) = .724 / .908 = .800 < .819,$$

$$\delta(\mathbf{y}_4 | \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3) = .724 / .943 = .768 > .819.$$

6.28 (a) λ effect: $\lambda = .8063$, $F^{(3)} = 2.237$, $D^{(3)} = 142.808$, $\theta = .896$. All are significant.

F effect: $\lambda = .883$, $F^{(3)} = 1.103$, $D^{(3)} = 11.679$, $\theta = .926$. All are significant.

XY interaction: $\lambda = .138$, $F^{(3)} = 1.321$, $D^{(3)} = 3.456$, $\theta = .726$. All are significant.

(d) Contrast on V comparing 1 vs. 1, 3: $A = .0804$, $F^{(2)} = .520$, $U^{(2)} = 11.443$, $P = .510$. All are significant.

(e) Linear contrast for X : $A = .0875$, $F^{(2)} = .495$, $U^{(2)} = 133.273$, $P = .595$. All are significant.

Quadratic contrast for X : $A = .368$, $F^{(2)} = .831$, $U^{(2)} = 4.056$, $P = .831$. All are significant.

Cubic contrast for X : $A = .315$, $F^{(2)} = .873$, $U^{(2)} = 2.050$, $P = .873$. All are significant.

(f) The ANCOVA F 's for each variable are as follows:

Source	X	Y	Z_1	Z_2
S	990.24	234.26	876.33	71.41
T	261.32	6.47	14.72	21.11
TF	20.37	2.86	3.44	1.08

All F 's are significant except the last one, 1.08.

(g) Test of significance of Z_1 and Z_2 adjusted for Z_1 and Z_2 :

	F	Y	TF
$Ac(Z_1, Z_2, T, S)$	.1228	.1336	.0402

(h) Test of significance of each variable adjusted for the other three:

	F	Y	TF
$Ac(Z_1, Z_2, T, S)$	.1158	.1055	.0365
$Ac(Z_1, Z_2, S, T)$	.0986	.1134	.0403
$Ac(Z_1, Z_2, S, T)$	.1275	.0637	.0624
$Ac(Z_1, Z_2, S, T)$	.0692	.0763	.0693

6.28 V = velocity (Block, L = laboratory (random)).

V effect (using $W_{P,L}$ for error matrix): $A = .0491$, $F^{(2)} = .950$, $U^{(2)} = 19.315$, $P = .351$. With $p = 2$, $v_B = 1$, and $v_T = 3$, $A_{22} = .058$, $F_{22}^{(2)} = .990$, $U_{22}^{(2)} = T_{22}^{(2)}/v_B = 25.08$, $P_{22} = .990$. Thus all four test statistics are significant.

L effect (using B for error matrix): $A = .012$, $F^{(2)} = .364$, $U^{(2)} = .408$, $P = .595$. None is significant.

$V \times L$ interaction (using B for error matrix): $A = .002$, $F^{(2)} = .060$, $U^{(2)} = .073$, $P = .934$. None is significant.

6.29 Source	A	$F^{(2)}$	$U^{(2)}$	P	Significant?
(a) Blocks	.0003	1.126	6.393	.008	Yes
(b) Common 1 vs. 3, 3, 4	.146	.854	3.675	.034	Yes
Subjects	.00000001	2.647	109.137	.001	Yes

6.31 P = properties of filter, T = surface treatment, F = filter.

Source	Δ	$F^{(1)}$	$L^{(2)}$	s	Significant?
P	108	.471	5444	648	No
T	888	.455	11,828	828	No
PT	712	.283	386	274	No
F	318	.885	24,182	388	No
PF	178	.558	1855	288	No
TF	222	.645	1315	345	No
PTF	252	.564	555	172	No

6.32 Δ = period. P , T , and F are defined in Problem 6.31.

Source	Δ	$F^{(1)}$	$L^{(2)}$	s	Significant?
Δ	828	.478	41,568	578	No
ΔP	478	.343	1,058	585	No
ΔT	162	.838	5545	838	No
ΔPT	222	.338	282	288	No
ΔF	888	.654	5458	888	No
ΔTF	822	.381	388	388	No
ΔPF	382	.613	1,588	613	No
ΔPTF	288	.228	252	188	No

For the between-subject factors and interactions, we have

Source	df	F	p -Value
P	2	11.78	< .0001
T	1	78.34	< .0001
PT	2	1.28	.3143
F	1	343.64	< .0001
PF	2	13.78	.0004
TF	1	5.85	.0192
PTF	2	.48	.6264
Error	52		

6.33 For parallelism, we use (6.75) to obtain $\Delta = .2881$. For trends, we use (6.81) and (6.82) to obtain $\Delta = .3651$ and $F = .857$. For factors we use (6.84) to obtain $F^2 = 110.521$.

6.34 (a) B_1 (6.93), $F^2 = 10.7428$, B_2 (6.105) or (6.106), $\Delta = .3655$.

(b) For each row c_i' of C , we use $T_i^2 = n(n)'n(B_1)^{-1}n'$, as in Example 6.4.2: $T_1^2 = 17.0485$, $T_2^2 = .5218$, $T_3^2 = .2714$. This can also be done by Miller's Δ using $\Delta_1 = c_1'(B_1/c_1'(B_1 + B^2)c_1)$, $\Delta_2 = .8137$, $\Delta_3 = .9882$, $\Delta_4 = .9800$.

6.35 The six variables represent two within-subjects factors: y_1 in A_1B_1 , y_2 in A_1B_2 , y_3 in A_2B_1 , y_4 in A_2B_2 , y_5 in A_3B_1 , and y_6 in A_3B_2 . Using linear and quadratic effects (other orthogonal contrasts could be used), the matrices A , B , and G in

(5.181), (5.182), and (5.183) become

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix}, \\ \mathbf{B} &= \begin{pmatrix} 1 & 0 & -1 & 1 & 0 & -1 \\ 1 & -2 & 1 & 1 & -2 & 1 \end{pmatrix}, \\ \mathbf{C} &= \begin{pmatrix} 1 & 0 & -1 & -1 & 0 & 1 \\ 1 & -2 & 1 & -1 & 1 & -1 \end{pmatrix}. \end{aligned}$$

Using these in F^2 as given by (5.181), (5.182), and (5.183), we obtain $F_1^2 = 191.0091$, $F_2^2 = 1.9098$, and $F_{1,2}^2 = 5.8676$. Using MANOVA's tests for the same within-subjects factors, we obtain:

Source	Λ	$F^{(1)}$	$F^{(2)}$	F	Significant?
A	.202	.768	1.641	.768	No
B	.946	.054	.057	.056	No
A:B	.827	.113	.140	.120	No

6.26 MANOVA tests for the within-subjects fixed T (time), and interactions of time with the between-subjects effects C (concent) and G (gender):

Source	Λ	$F^{(1)}$	$F^{(2)}$	F
T	.298	.742	1.878	.792
T:C	.362	.889	1.209	.889
T:G	.929	.071	.077	.071
T:CG	.889	.261	.225	.238

ANOVA F -tests for between-subjects factors and interactions:

Source	df	F	p -value
C	3	4.16	.008
G	1	2.89	.097
C:G	3	.37	.809

6.27 (a) $T^2 = 79.331$

(b) Using $\mathbf{c}_i = \mathbf{c}_i' \mathbf{T} / \sqrt{\mathbf{c}_i' \mathbf{S} \mathbf{c}_i} / s$, where $\mathbf{c}_i'$ is the i th row of $\mathbf{C}$, we obtain $\mathbf{c}_1 = 7.1193$, $\mathbf{c}_2 = -4.45$, $\mathbf{c}_3 = -1.05$.

6.28 (a) $T^2 = 1712.1291$

(b) Using $\mathbf{c}_i = \mathbf{c}_i' \mathbf{T} / \sqrt{\mathbf{c}_i' \mathbf{S} \mathbf{c}_i} / s$, we obtain $\mathbf{c}_1 = 501.358$, $\mathbf{c}_2 = 34.596$, $\mathbf{c}_3 = 2.056$, $\mathbf{c}_4 = 7.637$, $\mathbf{c}_5 = 0.344$, $\mathbf{c}_6 = 0.968$.

6.29 (a) Using $T^2 = N(\mathbf{S} \mathbf{T} \mathbf{T}' \mathbf{T} (\mathbf{S} \mathbf{T} \mathbf{S} \mathbf{T}')^{-1} (\mathbf{T} \mathbf{T}'))$ in (5.112), we obtain $T^2 = 17.582 = F_{3,2,2,2}^2 = 21.332$.

(b) $\mathbf{c}_1 = .891$, $\mathbf{c}_2 = 1.836$, $\mathbf{c}_3 = .127$ [Since the T^2 -test in part (a) did not reject H_0 , these would ordinarily not be calculated.]

- (a) Using $A = (CXC^T)/(C\mathbf{1}\mathbf{1}^T + H\mathbf{1}\mathbf{1}^T)$ in (5.134), we obtain $A = .3487 > \alpha_{.05, 3, 22, 9} = .263$.
- (b) To compute groups using each row of C , we use $\hat{\beta}_i = c_i^T B_0 / c_i^T (\mathbf{1} + H\mathbf{1})$ to obtain $\hat{\beta}_1 = .833$, $\hat{\beta}_2 = .598$, $\hat{\beta}_3 = .850$. [Since the A -test in part (a) did not reject H_0 , we would ordinarily not have calculated these.]
- 6.40 (a) Using $T^2 = N(Cy_1 / (C\hat{\Sigma}_0 C)^{-1} (C\mathbf{1}))$ in (5.112), we obtain $T^2 = 35.802 > F_{.05, 4, 20}^2 = 12.985$.
- (b) Using $\hat{\sigma}_i^2 = N(c_i^T y_1^2) / c_i^T \hat{\Sigma}_0 c_i$, we obtain $\hat{\sigma}_1^2 = .815$, $\hat{\sigma}_2^2 = .593$, $\hat{\sigma}_3^2 = .7236$. Only the cubic effect is significant.
- (c) For an overall test comparing groups, we use (5.134),

$$A = (CXC^T)/(C\mathbf{1}\mathbf{1}^T + H\mathbf{1}\mathbf{1}^T) = .6861.$$

- (d) To compute groups using each row of C , we use

$$\hat{\beta}_i = c_i^T B_0 / c_i^T (\mathbf{1} + H\mathbf{1}) : \hat{\beta}_1 = .594, \hat{\beta}_2 = .794, \hat{\beta}_3 = .590.$$

- 6.41 (a) Using $T^2 = N(Cy_1 / (C\hat{\Sigma}_0 C)^{-1} (C\mathbf{1}))$ in (5.112), we obtain $T^2 = 45.800$.
- (b) Using $\hat{\sigma}_i^2 = N(c_i^T y_1^2) / c_i^T \hat{\Sigma}_0 c_i$, we obtain $\hat{\sigma}_1^2 = 18.413$, $\hat{\sigma}_2^2 = 9.583$, $\hat{\sigma}_3^2 = 3.446$, $\hat{\sigma}_4^2 = .811$, $\hat{\sigma}_5^2 = .098$, $\hat{\sigma}_6^2 = 2.976$.
- (c) For an overall test comparing groups, we use (5.134),

$$A = (CXC^T)/(C\mathbf{1}\mathbf{1}^T + H\mathbf{1}\mathbf{1}^T) = .364.$$

- (d) To compute groups using each row of C , we use

$$\begin{aligned} \hat{\beta}_i &= c_i^T B_0 / c_i^T (\mathbf{1} + H\mathbf{1}) : \hat{\beta}_1 = .895, \hat{\beta}_2 = .929, \hat{\beta}_3 = .711, \\ \hat{\beta}_4 &= .814, \hat{\beta}_5 = .890, \hat{\beta}_6 = .896. \end{aligned}$$

- 6.42 (a) Combined groups (pooled covariance matrix). Using $r = \text{number of units} = 30$, we obtain, by (5.113),

$$\hat{\boldsymbol{\mu}} = (.98, 1, .991, .9418, -.99104, -.99098).$$

By (5.115), we obtain $T^2 = .218$. By (5.116), we have

$$\hat{\boldsymbol{\mu}}' = (.95, 5, .96, 9, .95, 9, .95, 9, .95, 9).$$

- (b) Group 1: $\hat{\boldsymbol{\mu}}_1' = (.06, 7, .898, .940, -.0089, -.00899)$, $T^2 = .0111$, $\hat{\boldsymbol{\mu}}_1 = (.005, 2, .904, .100, .99, 6, .90, 7)$.
- (c) Groups 2-6: $\hat{\boldsymbol{\mu}}_2' = (.97, 6, 1.010, .0089, -.00403, -.00009)$, $T^2 = .2994$, $\hat{\boldsymbol{\mu}}_2 = (.97, 6, .96, 4, .95, 9, .92, 4, .97, 4, .98, 6)$.

6.43 (a) For the control group, the overall test is

$$F^2 = s_2(\mathbf{C}\mathbf{Y}_1)'(\mathbf{C}\mathbf{S}_0\mathbf{C}')^{-1}(\mathbf{C}\mathbf{Y}_1) = 394.349.$$

For each row of $\mathbf{C}$ (linear, quadratic, etc.), we have

$$t_1^2 = s_2(\mathbf{C}\mathbf{Y}_1)'(\mathbf{C}\mathbf{S}_0\mathbf{C}')^{-1}(\mathbf{C}\mathbf{Y}_1) : t_1^2 = 5.714,$$

$$t_2^2 = 39.111, t_3^2 = 39.792, t_4^2 = 3.811, t_5^2 = .588.$$

(b) For the obese group, we obtain $F^2 = s_2(\mathbf{C}\mathbf{Y}_2)'(\mathbf{C}\mathbf{S}_0\mathbf{C}')^{-1}(\mathbf{C}\mathbf{Y}_2) = 128.932$. For the five rows of $\mathbf{C}$, we obtain $t_1^2 = 4.978, t_2^2 = 107.129, t_3^2 = 5.223, t_4^2 = 10.780, t_5^2 = 3.871$.

(c) For the combined group ($\mathbf{S}_{\text{pooled}}$ = pooled covariance matrix), we use $F^2 = N_2(\mathbf{C}\mathbf{Y})'(\mathbf{C}\mathbf{S}_{\text{pooled}})^{-1}(\mathbf{C}\mathbf{Y})$ in (5.112) to obtain $F^2 = 241.0038$. We test for linear, quadratic, etc., trends using the rows of $\mathbf{C}$ in $t_1^2 = N_2(\mathbf{C}\mathbf{Y})'(\mathbf{C}\mathbf{S}_{\text{pooled}})^{-1}(\mathbf{C}\mathbf{Y})$: $t_1^2 = 1.162, t_2^2 = 125.037, t_3^2 = 30.240, t_4^2 = 1.338, t_5^2 = .806$. To compare groups, we use $A = (\mathbf{C}\mathbf{B}\mathbf{B}')/(\mathbf{C}\mathbf{B}\mathbf{B} + \mathbf{H}\mathbf{H}')$ in (5.126) and $A_1 = c_1(\mathbf{K}_1)c_2(\mathbf{K} + \mathbf{H}\mathbf{H})$: $A = .4902, A_1 = .7947, A_2 = .8843, A_3 = .7987, A_4 = .8278, A_5 = .8172$.

6.44 Control group: By (5.115),

$$\hat{\beta}_1' = (3.128, .656, -.283, -.364, .192, .603, -.829).$$

By (5.116), $F^2 = .7033$. By (5.118),

$$\hat{\beta}_1' = (\hat{\beta}_{11}, \hat{\beta}_{12}, \dots, \hat{\beta}_{16}) = (14.13, 3.29, 2.71, 2.71, 3.84, 3.29, 3.24, 3.95).$$

Obese group: $\hat{\beta}_1' = (3.207, -.187, .463, .895, -.192, -.310, .310)$, $F^2 = .3845$, $\hat{\mu}_1' = (4.514, 4.11, 3.81, 3.48, 3.34, 3.37, 3.70, 4.02)$.

Combined group (pooled covariance matrix): $\hat{\beta}_1' = (3.15, .162, .483, -.115, .811, .370, -.032)$, $F^2 = .8158$, $\hat{\mu}' = (4.35, 3.80, 3.36, 3.15, 3.15, 3.32, 3.63, 3.88)$.

6.45 A = activator, T = time, C = group. In (5.101), (5.102), and (5.103), we use

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 2 & -1 & -1 & -1 & -1 & -1 & -1 \\ 8 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix},$$

$$\mathbf{T} = \begin{pmatrix} -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ 1 & -2 & 1 & 1 & -2 & 1 & 1 & -2 & 1 \end{pmatrix},$$

$$\mathbf{C}_0 = \begin{pmatrix} -2 & 8 & 2 & 1 & 8 & -2 & 1 & 0 & -1 \\ 2 & -4 & 2 & -1 & 2 & -2 & -1 & 2 & -1 \\ 8 & 8 & 0 & 1 & 8 & -2 & -1 & 0 & 1 \\ 8 & 8 & 0 & 1 & -2 & 1 & -1 & 2 & -1 \end{pmatrix}.$$

$F_A^2 = 5071.579$, $F_B^2 = 358.085$, $F_{AB}^2 = 345.491$. The same within-sample factors and interaction can be tested with Wilks' Λ using (6.105) and the other three MANOVA tests:

Source	Λ	$F^{(1)}$	$U^{(1)}$	t	Significant?
A	.003	.001	107.08	.007	Yes
B	.026	.044	10.35	.000	Yes
AB	.100	.005	0.27	.600	Yes

The interactions of the within factors with the between factor C are tested with Wilks' Λ (Section 6.4.5) and with the other three MANOVA tests:

Source	Λ	$F^{(1)}$	$U^{(1)}$	t	Significant?
AC	.004	.100	.10	.110	No
BC	.000	.111	.125	.111	No
ABC	.005	.205	.250	.205	No

The between-subjects factor C is tested with an ANOVA F -test: $F = .47$, p -value = .504.

CHAPTER 7

7.1 If Σ_0 is substituted for S in (7.1), we have

$$z = n[\ln(\Sigma_0) - \ln(\Sigma_0) + \ln \mathbf{I}] - p = n(0 + p - p) = 0.$$

$$\begin{aligned} 7.2 \quad -\ln(\hat{\Sigma}_0) - \ln(S) &= -\ln(\hat{\Sigma}_0^{-1}) - \ln(S) \\ &= -\ln(\hat{\Sigma}_0^{-1}) - \ln(S) \quad [\text{by (2.91)}] \\ &= -\ln(S) + \ln(\hat{\Sigma}_0^{-1}) \\ &= -\ln(S\hat{\Sigma}_0^{-1}) \quad [\text{by (2.89)}] \end{aligned}$$

$$7.3 \quad -\ln\left[\prod_{i=1}^p \lambda_i\right] + \sum_{i=1}^p \ln \lambda_i = -\sum_{i=1}^p \ln \lambda_i + \sum_{i=1}^p \ln \lambda_i = \sum_{i=1}^p (\lambda_i - \ln \lambda_i).$$

7.4 As noted in Section 7.3, the likelihood ratio in this case involves the ratio of the determinants of the sample covariance matrices under H_0 and H_1 . Under H_1 , which is essentially unrestricted, the maximum likelihood estimate of Σ (normalized for bias) is given by (3.12) as S . Under H_0 , it is assumed that each of the p y_i 's in $\mathbf{y}$ has variance σ^2 and that all y_i 's are independent. Thus we estimate σ^2 (unbiasedly) in each of the p columns of the Y matrix [see (3.1.7) and (3.23)] and pool the p estimates to obtain

$$s^2 = \frac{n}{p} \sum_{j=1}^p \frac{(Y_{0j} - \bar{Y}_j)^2}{(n-1)p},$$

Show that by (3.22) and (3.23) this is equal to

$$d^2 = \sum_{j=1}^p \frac{d_j}{p} = \frac{\mathbf{1}'\mathbf{D}\mathbf{1}}{p}.$$

Thus the likelihood ratio is

$$\text{LR} = \left(\frac{|\mathbf{D}|}{(p^2/\mathbf{1}'\mathbf{1})} \right)^{n/2} = \left(\frac{|\mathbf{D}|}{(1 + n(p-1)/p)} \right)^{n/2}.$$

Show that by (2.85) this becomes

$$\text{LR} = \left(\frac{|\mathbf{D}|}{(p\mathbf{D}'p/p^2)} \right)^{n/2}.$$

7.6 If $k_1 = k_2 = \dots = k_p = k$, say, then by (7.5),

$$a = \frac{p^2 \prod_{i=1}^p k_i}{(\sum_{i=1}^p k_i)^p} = \frac{p^2 p^p}{(p^2 p^p)} = 1.$$

$$\begin{aligned} 7.6 \quad (1-p)\mathbf{I} + p\mathbf{J} &= \begin{pmatrix} 1-p & 0 & \dots & 0 \\ 0 & 1-p & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1-p \end{pmatrix} + \begin{pmatrix} p & p & \dots & p \\ p & p & \dots & p \\ \vdots & \vdots & \ddots & \vdots \\ p & p & \dots & p \end{pmatrix} \\ &= \begin{pmatrix} 1 & p & \dots & p \\ p & 1 & \dots & p \\ \vdots & \vdots & \ddots & \vdots \\ p & p & \dots & 1 \end{pmatrix} \end{aligned}$$

7.7 (a) Substitute $\mathbf{J} = \mathbf{J}'$ and $\mathbf{u} = \mathbf{j}$ into $\mathbf{B}\mathbf{u} = \lambda\mathbf{u}$ to obtain $\mathbf{B}\mathbf{j} = \lambda\mathbf{j}$, which gives $\lambda\mathbf{j} = \lambda\mathbf{j}$.

$$(b) \quad B_0 = p^2(1-r)\mathbf{I} + r\mathbf{J} = p^2(1-r)\left(\mathbf{I} + \frac{r}{1-r}\mathbf{J}\right)$$

(c) By (2.85) and (2.106), we have

$$\begin{aligned} |B_0| &= \left| p^2(1-r)\left(\mathbf{I} + \frac{r}{1-r}\mathbf{J}\right) \right| = p^{2p}p^p(1-r)^p \left| \mathbf{I} + \frac{r}{1-r}\mathbf{J} \right| \\ &= (p^2)^p(1-r)^p \prod_{i=1}^p (1+k_i) = (p^2)^p(1-r)^p \left(1 + \frac{2R}{1-r}\right) \\ &= (p^2)^p(1-r)^{p-1}(1-r+rp) = (p^2)^p(1-r)^{p-1}[1+(p-1)r]. \end{aligned}$$

$$7.8 \quad \bar{M} = \frac{(\mathbf{R}_1^T \mathbf{e}) (\mathbf{R}_2^T \mathbf{e})^T \dots (\mathbf{R}_p^T \mathbf{e})^T}{\|\mathbf{R}_1\|_2 \dots \|\mathbf{R}_p\|_2} = \frac{(\mathbf{R}_1^T \mathbf{e}) (\mathbf{R}_2^T \mathbf{e})^T \dots (\mathbf{R}_p^T \mathbf{e})^T}{\|\mathbf{R}_1\|_2 \|\mathbf{R}_2\|_2 \dots \|\mathbf{R}_p\|_2}.$$

$$7.9 \quad \text{(a)} \quad \bar{M} = \mathbf{J} \mathbf{R} \mathbf{J}^T \quad \quad \quad \text{(b)} \quad \bar{M} = \mathbf{J} \mathbf{J}^T \mathbf{R}.$$

$$\begin{aligned} 7.10 \quad A &= \frac{|\mathbf{R}|}{\|\mathbf{R}_{22}\| \|\mathbf{R}_{33}\|} = \frac{(\mathbf{R}_{11}) (\mathbf{R}_{22} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21})}{\|\mathbf{R}_{22}\| \|\mathbf{R}_{33}\|} \\ &= (\mathbf{R}_{11}^{-1}) (\mathbf{R}_{22} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}) && \text{by (2.10)} \\ &= (\mathbf{R}_{11}^{-1}) (\mathbf{R}_{22} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}) && \text{by (2.10')} \\ &= (\mathbf{I} - \mathbf{R}_{11}^{-1} \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}) \\ &= \left[\prod_{i=1}^p (1 - r_i^2) \right] && \text{by (2.10B)}, \end{aligned}$$

where the r_i^2 's are the nonzero eigenvalues of $\mathbf{R}_{11}^{-1} \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}$. It was shown in Section 2.11.2 that $1 - \lambda_i$ is an eigenvalue of $\mathbf{I} - \mathbf{A}$, where λ_i is an eigenvalue of $\mathbf{A}$.

7.11 When all $p_i = 1$, we have $k = p$, and the substitutions in the denominators of (7.10) and (7.14) reduce to $\mathbf{R}_{jj} = r_{jj}$, $j = 1, 2, \dots, p$, and $\mathbf{R}_{jj} = 1$, $j = 1, 2, \dots, p$.

7.12 When all $p_i = 1$, we have $k = p$ and

$$\begin{aligned} a_0 &= p^2 - \sum_{i=1}^p p_i^2 = p^2 - p, & a_1 &= p^2 - p, \\ c &= 1 - \frac{1}{12(p)} (2m + 3n) \\ &= 1 - \frac{1}{6(p^2 - p)} [3(p^2 - p) + 3(p^2 - p)] \\ &= 1 - \frac{1}{6(p - 1)^2} [2(p^2 - 1) + 3(p - 1)] \\ &= 1 - \frac{1}{6(p - 1)^2} [2(p - 1)(p + 1) + 3(p - 1)] \\ &= 1 - \frac{1}{6p} (2p + 5). \end{aligned}$$

7.13 A_k noted below (7.13), the degree of freedom for the p^k -approximation is the total number of parameters minus the number estimated under $\mathbf{R}_0$. The number of distinct parameters in $\mathbf{I}$ is $p + \binom{p}{2} = \frac{1}{2}p(p + 1)$. The number of parameters estimated under $\mathbf{R}_0$ is p . The difference is $\frac{1}{2}p(p + 1) - p = \frac{1}{2}p(p - 1)$.

7.14 By (7.1) and (7.2), $n = 11.094$ and $n' = 10.658$.

7.15 By (7.3), $n = .0002594$. By (7.5), $n' = 11.518$. For $\mathbf{R}_0: \mathbf{C}(\mathbf{C}) = \mathbf{e}^T \mathbf{I}$, $n = .411$ and $n' = 2.650$.

7.16 For $\mathbf{R}_0: \mathbf{I} = \mathbf{e}^T \mathbf{I}$, $n = .00319$ and $n' = 10.972$. For $\mathbf{R}_0: \mathbf{C}(\mathbf{C}) = \mathbf{e}^T \mathbf{I}$, $n = .129$ and $n' = 36.276$.

7.17 For A_0 : $\Sigma = \sigma^2 \mathbf{I}$, $\alpha = .00471$ and $\sigma' = 135.790$. For A_0 : $\text{Cov}^* = \sigma^2 \mathbf{I}$, $\alpha = .747$ and $\sigma' = 3.465$.

7.18 By (7.18), $\sigma' = 8.5523$ with 13 degrees of freedom. The F -approximation is $F = .4852$ with 13 and 1147 degrees of freedom.

7.19 $\sigma' = 21.488$, $F = 2.511$ with 8 and 217 degrees of freedom.

7.20 $\sigma' = 35.793$, $F = 4.455$ with 8 and 495 degrees of freedom.

7.21 $\alpha = 8.7657$, $F = .8750$ with 10 and 650 degrees of freedom.

7.22 $(\mathbf{B}_1) = 2.620 \times 10^{12}$, $(\mathbf{B}_2) = 2.419 \times 10^{12}$, $(\mathbf{B}_3) = 4.768 \times 10^{12}$, $\lambda = 17.583$, $F = .829$.

7.23 $\ln \mathbf{W} = -85.865$, $\alpha = 156.454$, $a_1 = 21$, $a_2 = 13.793$, $F = 3.4096$.

7.24 $\ln \mathbf{W} = -7.892$, $\alpha = 10.565$, $a_1 = 18$, $a_2 = 1640$, $F = 1.046$.

7.25 $\ln \mathbf{W} = -8.8985$, $\alpha = 14.232$, $a_1 = 26$, $a_2 = 3889$, $F = .707$.

7.26 $\ln \mathbf{W} = -28.917$, $\alpha = 44.835$, $a_1 = 36$, $a_2 = 5238$, $F = .8628$.

7.27 $\ln \mathbf{W} = -142.438$, $\alpha = 174.289$, $a_1 = 119$, $a_2 = 3084$, $F = 1.448$.

7.28 $(\mathbf{B}) = 1.207, 909.5$, $(\mathbf{B}_{11}) = 2385.3$, $(\mathbf{B}_{22}) = 1541.9$, $\lambda = .3771$.

7.29 $(\mathbf{B}) = 4.237 \times 10^{12}$, $(\mathbf{B}_{11}) = 494, 926.6$, $(\mathbf{B}_{22}) = 131, 466, 908$, $\lambda = .6658$.

7.30 $(\mathbf{B}) = 9.876 \times 10^{-6}$, $(\mathbf{B}_{11}) = .02097$, $(\mathbf{B}_{22}) = 9.34 \times 10^{-6}$, $\lambda = .4642$.

7.31 $(\mathbf{B}) = 1.7148 \times 10^{10}$, $(\mathbf{B}_{11}) = 11, 284,907$, $(\mathbf{B}_{22}) = 15, 891.15$, $(\mathbf{B}_{33}) = 28,650,608$, $\lambda_{11} = 52,127.158$, $\lambda_{22} = 114.86$, $\alpha = .80493$, $\sigma' = 274.787$, $r = .46$.

7.32 $(\mathbf{B}) = 459.96$, $a_{11} = 186.56$, $a_{22} = 72.25$, $a_{33} = .280$, $\alpha = .1811$, $\sigma' = 12.149$, $f = 3$.

7.33 $\alpha = .0003376$, $\sigma' = 16.287$.

7.34 $\alpha = .0005735$, $\sigma' = 127.367$.

7.35 $\alpha = .005971$, $\sigma' = 191.236$.

CHAPTER 8

8.1 Using $\mathbf{u} = \mathbf{S}_p^{-1}(\mathbf{Y}_1 - \mathbf{Y}_2)$, we obtain

$$\begin{aligned} \frac{(\mathbf{u}'\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{d}^2}{\mathbf{u}'\mathbf{S}_p\mathbf{u}} &= \frac{(\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{S}_p^{-1}(\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{d}^2}{(\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{S}_p^{-1} \mathbf{S}_p \mathbf{S}_p^{-1}(\mathbf{Y}_1 - \mathbf{Y}_2)} \\ &= \frac{(\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{S}_p^{-1}(\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{d}^2}{(\mathbf{Y}_1 - \mathbf{Y}_2)' \mathbf{S}_p^{-1}(\mathbf{Y}_1 - \mathbf{Y}_2)}. \end{aligned}$$

8.2 You may wish to use the following steps:

- In Section 3.6.2 the grouping variable w is defined as $\alpha_2/(\alpha_1 + \alpha_2)$ for each observation in group 1 and $-\alpha_1/(\alpha_1 + \alpha_2)$ for group 2. Show that with this formulation, $\mathbf{W} = \mathbf{0}$.

(iii) Because $\bar{y} = 0$, there is no intercept, and the fitted model becomes

$$\begin{aligned} \hat{w}_i &= \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_p(x_{ip} - \bar{x}_p) \\ i &= 1, 2, \dots, n_1 + n_2. \end{aligned}$$

Denote the resulting matrix of p values corrected for their means as $\mathbf{Y}_c$ and the vector of w 's as $\mathbf{w}$. Then the least squares estimate $\mathbf{b} = (\beta_1, \beta_2, \dots, \beta_p)'$ is obtained as

$$\mathbf{b} = (\mathbf{Y}_c' \mathbf{Y}_c)^{-1} \mathbf{Y}_c' \mathbf{w}.$$

Using (2.51), show that

$$\begin{aligned} \mathbf{Y}_c' \mathbf{Y}_c &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\bar{y}_{ij} - \bar{y})(\bar{x}_{ij} - \bar{x})' \\ &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\bar{y}_{ij} - \bar{x}_1)(\bar{x}_{ij} - \bar{x}_2)' + \frac{n_1 n_2}{n_1 + n_2} (\bar{y}_1 - \bar{x}_1)(\bar{y}_1 - \bar{x}_2)', \end{aligned}$$

where $\bar{y} = (n_1 \bar{y}_1 + n_2 \bar{y}_2)/(n_1 + n_2)$. It will be helpful to write the first sum above as

$$\sum_{j=1}^{n_2} (\bar{y}_{1j} - \bar{y})(\bar{x}_{1j} - \bar{x})' + \sum_{j=1}^{n_2} (\bar{y}_{2j} - \bar{y})(\bar{x}_{2j} - \bar{x})'$$

and add and subtract $\bar{y}_1$ in the first term and $\bar{y}_2$ in the second.

(iii) Show that

$$\mathbf{Y}_c' \mathbf{w} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\bar{y}_{ij} - \bar{y}) w_{ij} = \frac{n_1 n_2}{n_1 + n_2} (\bar{y}_1 - \bar{y}_2).$$

Again it will be helpful to sum separately over the two groups.

(iv) From steps (ii) and (iii) we have

$$\mathbf{b} = (\mathbf{B} + \mathbf{1} \mathbf{1} \mathbf{1}')^{-1} \mathbf{1} \mathbf{d},$$

where $\mathbf{B} = \sum_{i,j} (\bar{y}_{ij} - \bar{y}_1)(\bar{x}_{ij} - \bar{x}_2)/(n_1 + n_2 - 2)$, $\mathbf{1} = n_1 + n_2 - 2$, $\mathbf{d} = n_1 n_2 / (n_1 + n_2)$, and $\mathbf{1} = \bar{y}_1 - \bar{y}_2$. Use (2.57) for the inverse of a partitioned matrix of the type $\mathbf{B} + \mathbf{1} \mathbf{1} \mathbf{1}'$ to obtain (8.4).

8.3 You may want to use the following steps:

(a) R^2 is defined as (see (9.30))

$$R^2 = \frac{\mathbf{b}'\mathbf{Y}_1\mathbf{a} - n\bar{y}^2}{\mathbf{a}'\mathbf{a} - n\bar{y}^2}.$$

In this case the expression simplifies because $\bar{y} = 0$. Using $\mathbf{Y}_1\mathbf{a}$ in Problem 8.3(iii), show that $R^2 = \mathbf{b}'(\mathbf{Y}_1 - \mathbf{Y}_2)$.

(b) Show that

$$\mathbf{b}'(\mathbf{Y}_1 - \mathbf{Y}_2) = \frac{r^2}{n_1 + n_2 - 2 + 2^2}.$$

8.4 $(\mathbf{x}'(\mathbf{Y}_1 - \mathbf{Y}_2))^2 = \mathbf{x}'(\mathbf{Y} - \mathbf{Y}_1)(\mathbf{x}'(\mathbf{Y}_1 - \mathbf{Y}_2)) = \mathbf{x}'(\mathbf{Y}_1 - \mathbf{Y}_2)(\mathbf{Y}_1 - \mathbf{Y}_2)\mathbf{x}$

8.5 $\mathbf{H}\mathbf{a} - \lambda\mathbf{K}\mathbf{a} = \mathbf{0}$

$$\mathbf{E}^{-1}(\mathbf{H}\mathbf{a} - \lambda\mathbf{K}\mathbf{a}) = \mathbf{E}^{-1}\mathbf{0}$$

$$\mathbf{E}^{-1}\mathbf{H}\mathbf{a} - \lambda\mathbf{E}^{-1}\mathbf{K}\mathbf{a} = \mathbf{0}$$

$$(\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}$$

8.6 Substituting $\mathbf{a}_i' = a_i\mathbf{a}_i$, $i = 1, 2, \dots, p$, into (8.17), we obtain

$$\begin{aligned} v_{11} &= a_1a_1 \frac{r_{11} - \bar{y}_{11}}{v_1} + a_2a_1 \frac{r_{21} - \bar{y}_{11}}{v_2} + \cdots + a_pa_1 \frac{r_{p1} - \bar{y}_{11}}{v_p} \\ &= a_1r_{11} + a_2r_{21} + \cdots + a_p r_{p1} - a_1\bar{y}_{11} - a_2\bar{y}_{11} - \cdots - a_p\bar{y}_{11} \\ &= a_1r_{11} + a_2r_{21} + \cdots + a_p r_{p1} - \mathbf{a}'\bar{\mathbf{y}}_1 \end{aligned}$$

8.7 (a) $\mathbf{a}' = (0.368, -0.930, 2.819, -1.403)$

(b) $v_1 = 5.461, v_2 = 2.607, v_3 = 3.775, v_4 = .688$

(c) The standardized coefficients rank the variables in the order y_3, y_4, y_1, y_2 . The t -tests rank them in the order y_3, y_2, y_1, y_4 .

(d) The partial R^2 's calculated by (8.20) are $F(y_3|y_2, y_1, y_4) = 7.844$, $F(y_2|y_1, y_3, y_4) = 2.612$, $F(y_1|y_1, y_3, y_4) = 40.503$, and $F(y_4|y_1, y_2, y_3) = 0.698$.

8.8 (a) $\mathbf{a}' = (3.845, -1.301, -.598, -1.431)$

(b) $\mathbf{a}' = (4.157, -2.804, -1.188, -2.088)$

(c) $v_1 = 3.888, v_2 = -3.888, v_3 = -5.604, v_4 = -5.643$

(d) $F(y_1|y_2, y_3, y_4) = 23.994$, $F(y_2|y_1, y_3, y_4) = 5.798$, $F(y_3|y_1, y_2, y_4) = 1.775$, $F(y_4|y_1, y_2, y_3) = 0.258$

8.9 (a) $\mathbf{a}' = (-1.431, .051, -.005, -.089, -.007, -.012)$

(b) $\mathbf{a}' = (-1.056, 1.471, -.540, -1.035, -.107, -1.200)$

- (ii) $r_1 = -4.655$, $r_2 = .352$, $r_3 = -4.394$, $r_4 = -5.257$, $r_5 = -4.832$, $r_6 = -6.489$
 (iii) $F(\eta_1|y_1, y_2, y_3, y_4, y_5, y_6) = 8.081$, $F(\eta_2|y_1, y_2, y_3, y_4, y_5, y_6) = 4.286$,
 $F(\eta_3|y_1, y_2, y_3, y_4, y_5, y_6) = 8.135$, $F(\eta_4|y_1, y_2, y_3, y_4, y_5, y_6) = 8.593$,
 $F(\eta_5|y_1, y_2, y_3, y_4, y_5, y_6) = 8.126$, $F(\eta_6|y_1, y_2, y_3, y_4, y_5, y_6) = 9.192$

8.10 (a) $\mathbf{a}' = (.053, .048, .362, .033)$

(b) $\mathbf{a}' = (6.398, .083, 1.628, .031)$

(c) $r_1 = -3.713$, $r_2 = .348$, $r_3 = -3.552$, $r_4 = -.724$

(d) $F(\eta_1|y_1, y_2, y_3) = 3.532$, $F(\eta_2|y_1, y_2, y_3) = .048$,
 $F(\eta_3|y_1, y_2, y_3) = 1.482$, $F(\eta_4|y_1, y_2, y_3) = .031$

8.11 (a) $\mathbf{a}_1' = (.051, .833, -.347, -.138)$, $\mathbf{a}_2' = (-.317, .298, .343, -.026)$

(b) $\lambda_1/(\lambda_1 + \lambda_2) = .958$, $\lambda_3/(\lambda_1 + \lambda_2) = .042$. Using the methods of Section 8.5.2, we have two tests, the first for significance of λ_1 and λ_2 and the second for significance of λ_3 .

Test	Δ	F	p -Value for F
1	.2243	8.1204	<.0001
2	.0971	1.2177	.2869

(c) $\mathbf{a}_1' = (.049, 1.553, -1.182, -.099)$, $\mathbf{a}_2' = (-1.161, .869, .828, -.045)$

(d) $F(\eta_1|y_1, y_2, y_3) = 3.087$, $F(\eta_2|y_1, y_2, y_3) = 30.873$,
 $F(\eta_3|y_1, y_2, y_3) = 9.638$, $F(\eta_4|y_1, y_2, y_3) = 1.228$

(e) In the plot, the first discriminant function separates groups 1 and 2 from group 3, but the second is ineffective in separating group 1 from group 2.

8.12 (a) λ_1 $\lambda_1/(\sum_{i=1}^4 \lambda_i)$

	λ_1	$\lambda_1/(\sum_{i=1}^4 \lambda_i)$	Eigenvalue
1	1.9777	.6428	$\mathbf{a}_1' = (.449, -.263, .653, -.054)$
2	.7619	.2507	$\mathbf{a}_2' = (.176, .188, -.038, .1736)$
3	.1246	.0394	$\mathbf{a}_3' = (-.055, .236, .476, -.885)$
4	.0360	.0089	$\mathbf{a}_4' = (-.364, .475, .316, -.475)$

(b) Test of significance of each eigenvalue and those that follow it:

Test	Δ	Approximate F	p -Value for F
1	.1348	4.087	<.0001
2	.4429	3.199	.0006
3	.7619	1.080	.1481
4	.8747	.545	.5819

(c) $\mathbf{a}_1' = (.266, -.363, 1.333, -.060)$, $\mathbf{a}_2' = (.160, .654, -.1291, 2.333)$,
 $\mathbf{a}_3' = (-.087, .899, .873, -1.117)$, $\mathbf{a}_4' = (-2.044, 1.694, .843, -.628)$

(d) $F(\eta_1|y_1, y_2, y_3) = .399$, $F(\eta_2|y_1, y_2, y_3) = 3.959$,
 $F(\eta_3|y_1, y_2, y_3) = 6.083$, $F(\eta_4|y_1, y_2, y_3) = 4.659$

- (ii) In the plot, the first discriminant function separates groups 1, 4, and 6 from groups 2, 3, and 5. The second function achieves some separation of group 6 from groups 1 and 4 and some separation of group 5 from groups 1 and 3.

8.13 Three variables entered the model in the stepwise selection. The summary table is as follows:

Step	Variable Entered	Overall F	p -Value	Partial F	Partial R^2	p -Value
1	x_1	.4080	<.0001	.4080	.11196	<.0001
2	x_2	.2658	<.0001	.1499	.1448	.0006
3	x_3	.1589	<.0001	.1022	.1284	.0008

8.14 Summary table:

Step	Variable Entered	Overall F	p -Value	Partial F	Partial R^2	p -Value
1	x_1	.8592	<.0001	.8592	.21451	<.0001
2	x_2	.9180	<.0001	.1498	.0184	.0477
3	x_3	.1096	<.0001	.1011	.0140	.0148
4	x_4	.8080	<.0001	.1943	.1798	.0044
5	x_5	.1619	<.0001	.1077	.1070	.0347

In this case, the fifth variable to enter, x_5 , would not ordinarily be included in the subset. The p -value of .0347 is large in this setting, where several tests are run at each step and the variable with smallest p -value is selected.

8.15 Summary table:

Step	Variable Entered	Overall F	p -Value	Partial F	Partial R^2	p -Value
1	x_1	.0007	.9996	.0007	.0000	.9996
2	x_2	.2688	<.0001	.1456	.22473	<.0001

CHAPTER 9

9.1 $\bar{Y}_1 - \bar{Y}_2 = \mu\bar{Y}_1 - \mu\bar{Y}_2 = \mu(\bar{Y}_1 - \bar{Y}_2) = (\bar{Y}_1 - \bar{Y}_2)/N_p^{-1}(\bar{Y}_1 - \bar{Y}_2)$

9.2 $\frac{1}{2}(\bar{Y}_1 + \bar{Y}_2) = \frac{1}{2}(\mu\bar{Y}_1 + \mu\bar{Y}_2) = \frac{1}{2}\mu(\bar{Y}_1 + \bar{Y}_2) = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)/N_p^{-1}(\bar{Y}_1 + \bar{Y}_2)$

9.3 Write (9.2) as follows

$$\frac{f(\bar{y}/G_1)}{f(\bar{y}/G_2)} = \frac{f_2}{f_1}$$

and substitute $f(\bar{y}/G_i) = N_p(\bar{\mu}_i, \Sigma)$ from (4.2) to obtain

$$\frac{f(\bar{y}/G_1)}{f(\bar{y}/G_2)} = e^{-(\bar{y}-\bar{\mu}_1)'(\Sigma^{-1})(\bar{y}-\bar{\mu}_1) - (\bar{y}-\bar{\mu}_2)'(\Sigma^{-1})(\bar{y}-\bar{\mu}_2)/2} \frac{f_2}{f_1}$$

Substitute estimates for μ_1 , μ_2 , and Σ , and take the logarithm of both sides to obtain (9.8). Note that if $x = 0$, then $\ln x = -\infty$.

9.4 Maximizing $p_1 f(y|G_1)$ is equivalent to maximizing $\ln\{p_1 f(y|G_1)\}$. Use $f(y|G_1) = N(y|\mu_1, \Sigma)$ from (4.1) and take the logarithm to obtain

$$\ln\{p_1 f(y|G_1)\} = \ln p_1 - \frac{1}{2}y' \ln(2\pi) - \frac{1}{2}(\mathbf{y} - \mu_1)' \Sigma^{-1}(\mathbf{y} - \mu_1).$$

Expand the last term, delete terms common to all groups (terms that do not involve $\mathbf{y}$), and substitute estimates of μ_1 and Σ to obtain (9.11).

9.6 Use $f(y|G_1) = N(y|\mu_1, \Sigma_1)$ in $\ln\{p_1 f(y|G_1)\}$, delete $-(y)'(2)\ln(2\pi)$, and substitute $\hat{\mu}_1$ and $\hat{\Sigma}_1$ for μ_1 and Σ_1 .

9.6 (a) $\hat{\mu}' = (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}_2^{-1} = (1.65, -1.30, -.108, -.141)$.

$$\frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2)' = -15.8054$$

(b)	Actual Group	Number of Observations	Predicted Group	
			1	2
	1	18	18	0
	2	28	1	18

$$\text{Error rate} = \frac{1}{46} = .0215$$

(c) Using the k nearest neighbor method with $k = 3$, we obtain the same classification table as in part (b). With $k = 4$, two observations are misclassified, and the error rate becomes $2/46 = .0435$.

9.7 (a) $\hat{\mu}' = (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}_2^{-1} = (-1.45, .811, -.005, -.896, -.003, -.811)$.

$$\frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2)' = -21.645$$

(b) Linear Classification				
	Actual Group	Number of Observations	Predicted Group	
			1	2
	1	28	27	1
	2	34	8	26

$$\text{Error rate} = (2 + 26)/72 = .375$$

(c) μ_1 and μ_2 Proportional to Sample Means

	Actual Group	Number of Observations	Predicted Group	
			1	2
	1	28	27	1
	2	34	8	26

$$\text{Error rate} = (2 + 26)/72 = .375$$

$$8.8 \text{ (a) } \hat{\alpha}' = (\hat{\beta}_0 - \hat{\beta}_2/\hat{\Sigma}_{xx}^{-1}) = (-.057, -.040, -.242, -.051),$$

$$\frac{1}{2}(C_1 + C_2) = -7.9688$$

(b) Linear Classification

Actual Group	Number of Observations	Predicted Group	
		1	2
1	9	8	1
2	10	1	9

$$\text{Error rate} = \frac{1}{19} = .053$$

(c) Exhaust Method

Actual Group	Number of Observations	Predicted Group	
		1	2
1	9	8	1
2	10	2	8

$$\text{Error rate} = (2 + 1)/19 = .105$$

(d) Kernel Density Estimate with $h = 2$

Actual Group	Number of Observations	Predicted Group	
		1	2
1	9	9	0
2	10	1	9

$$\text{Error rate} = \frac{1}{19} = .053$$

8.9 (a)

Actual Group	Number of Observations	Predicted Group	
		1	2
1	20	18	2
2	20	2	18

$$\text{Error rate} = (2 + 2)/40 = .10$$

(b) Four variables were selected by the stepwise discriminant analysis: y_2 , y_3 , y_4 , and y_5 (see Problem 8.10). With these four variables we obtain the classification table in part (c).

(ii)

Actual Group	Number of Observations	Predicted Group	
		1	2
1	28	28	0
2	28	2	18

Error rate = $(2 + 2)/46 = .108$. The four variables classified the sample as well as did all six variables in part (a).

8.10 (a) By (5.1.10), $\delta_1(\mathbf{x}) = \frac{1}{2}(\mathbf{S}_p^{-1})^T \mathbf{x} - \frac{1}{2}(\mathbf{S}_p^{-1})^T \mathbf{b}_1 = \mathbf{c}_1^T \mathbf{x} + c_{10}$. The vectors $(\frac{1}{2}\mathbf{S}_p^{-1})^T$, $i = 1, 2, 3$, are

Group 1	Group 2	Group 3
-72.79	-65.36	-68.77
.80	3.33	.68
18.18	36.70	2.74
-1.83	-.34	-6.84
18.82	11.88	13.69

(b)

Linear Classification

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	12	9	3	0
2	12	3	7	2
3	12	0	1	11

Error rate = $(3 + 3 + 2 + 1)/36 = .250$

(c)

Quadratic Classification

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	12	10	2	0
2	12	2	8	2
3	12	0	1	11

Error rate = $(2 + 2 + 2 + 1)/36 = .194$

(d)

Linear Classification-Robust Method

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	12	7	5	0
2	12	4	5	3
3	12	0	1	11

Error rate = $(5 + 4 + 3 + 1)/36 = .361$

(ii) **5 Nearest Neighbors with $k = 5$**

Actual Group	Number of Observations	Predicted Group		
		1	2	3
1	11	9	2	0
2	11	2	7	2
3	12	0	1	11

$$\text{Error rate} = (2 + 2 + 2 + 1)/44 = .306$$

9.81 (a) By B.110, $\hat{L}_n(\mathbf{y}) = \frac{1}{n} \mathbf{y}^T \mathbf{K}_n^{-1} \mathbf{y} - \frac{1}{2n} \mathbf{y}^T \mathbf{K}_n^{-1} \mathbf{p}_1 = \hat{\alpha}_n^T \mathbf{y} + c_n$. The vectors $\{\hat{\alpha}_n^T\}$, $i = 1, 2, \dots, 15$, are

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
-.008.8	-.001.2	-.028.3	-.004.8	-.007.9	-.013.8
.006.8	.007.0	.028.0	.007.0	.008.8	.011.3
-.004.4	-.004.8	-.008.2	-.004.4	-.005.8	-.008.1
.009.6	.008.2	.024.0	.007.7	.008.2	.009.0
-.008.2	-.007.8	-.006.4	-.005.4	-.007.8	-.007.3

(b) **Linear Classification**

Actual Group	Number of Observations	Predicted Group					
		1	2	3	4	5	6
1	8	3	0	0	1	0	2
2	8	0	3	2	1	2	0
3	8	0	0	0	1	1	6
4	8	3	0	1	4	0	0
5	8	0	3	1	0	3	1
6	8	2	0	0	0	2	4

$$\text{Correct classification rate} = (3 + 3 + 6 + 4 + 3 + 4)/48 = .521$$

$$\text{Error rate} = 1 - .521 = .479$$

(c) **Quadratic Classification**

Actual Group	Number of Observations	Predicted Group					
		1	2	3	4	5	6
1	8	8	0	0	0	0	0
2	8	0	7	0	1	0	0
3	8	1	0	0	0	1	6
4	8	0	0	1	7	0	0
5	8	0	3	0	0	5	0
6	8	2	0	0	0	1	5

$$\text{Correct classification rate} = (8 + 7 + 6 + 7 + 4 + 5)/48 = .771$$

$$\text{Error rate} = 1 - .771 = .229$$

(8b) k -Nearest Neighbor with $k = 3$

Actual Group	Number of Observations	Predicted Group						True
		1	2	3	4	5	6	
1	8	8	0	0	2	0	0	1
2	8	0	4	0	0	1	0	2
3	8	1	0	0	0	1	0	0
4	8	0	0	0	2	0	0	2
5	8	0	1	0	0	0	1	0
6	8	1	0	0	0	0	5	1

Correct classification rate = $(8 + 0 + 0 + 2 + 0 + 2)/40 = .700$ Error rate = $1 - .700 = .300$ (8c) k -Nearest Neighbor with $k = 1$
(For this data set, larger values of k do much worse.)

Actual Group	Number of Observations	Predicted Group						
		1	2	3	4	5	6	
1	8	8	0	0	0	0	0	8
2	8	0	0	0	0	0	0	0
3	8	1	0	0	0	0	1	0
4	8	1	0	0	7	0	0	0
5	8	0	0	0	0	7	1	1
6	8	1	0	0	0	0	0	0

Correct classification rate = $(8 + 0 + 0 + 1 + 7 + 0)/40 = .800$ Error rate = $1 - .800 = .200$

CHAPTER 10

$$10.1 \quad y - X\hat{\beta} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} \hat{\beta} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} x_1' \hat{\beta} \\ x_2' \hat{\beta} \\ \vdots \\ x_n' \hat{\beta} \end{pmatrix} = \begin{pmatrix} y_1 - x_1' \hat{\beta} \\ y_2 - x_2' \hat{\beta} \\ \vdots \\ y_n - x_n' \hat{\beta} \end{pmatrix}$$

$$\text{By (2.33), } \sum_{i=1}^n (y_i - x_i' \hat{\beta})^2 = (y - X\hat{\beta})'(y - X\hat{\beta}).$$

$$\begin{aligned} 10.2 \quad \sum_{i=1}^n (y_i - \mu)^2 &= \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \mu)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \bar{y})(\bar{y} - \mu) + \sum_{i=1}^n (\bar{y} - \mu)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + (\bar{y} - \mu) \sum_{i=1}^n (y_i - \bar{y}) + n(\bar{y} - \mu)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \quad \left[\text{since } \sum_{i=1}^n (y_i - \bar{y}) = 0 \right] \end{aligned}$$

$$18.3 \quad \sum_{i=1}^n (x_i \bar{x} - E_i \bar{x}) = \bar{x} \sum_{i=1}^n (x_i - E_i) = \bar{x} \sum_{i=1}^n (x_i - \bar{x}) = \bar{x} (n\bar{x} - n\bar{x}) = \bar{x}(n\bar{x} - n\bar{x})$$

$$\begin{aligned} 18.4 \quad E(\bar{Y}_1 - E(Y_1))^2 &= E(\bar{Y}_1 - E(\bar{Y}_1) + E(\bar{Y}_1) - E(Y_1))^2 \\ &= E(\bar{Y}_1 - E(\bar{Y}_1))^2 + 2E(\bar{Y}_1 - E(\bar{Y}_1))(E(\bar{Y}_1) - E(Y_1)) \\ &\quad + E(E(\bar{Y}_1) - E(Y_1))^2 \end{aligned}$$

The second term on the right vanishes because $(E(\bar{Y}_1) - E(Y_1))$ is a constant and $E(\bar{Y}_1 - E(\bar{Y}_1)) = E(\bar{Y}_1) - E(Y_1) = 0$. For the third term, we have $E(E(\bar{Y}_1) - E(Y_1))^2 = (E(\bar{Y}_1) - E(Y_1))^2$, because $(E(\bar{Y}_1) - E(Y_1))^2$ is constant.

18.5 First show that $\text{cov}(\hat{\beta}_p) = \sigma^2(X_p'X_p)^{-1}$. This can be done by noting that $\hat{\beta}_p = (X_p'X_p)^{-1}X_p'y_p = Ay_p$. Then, by (3.74), $\text{cov}(Ay) = A \text{cov}(y)A' = A\sigma^2\mathbb{I}_nA' = \sigma^2AA'$. By substituting $A = (X_p'X_p)^{-1}X_p'$, this becomes $\text{cov}(\hat{\beta}_p) = \sigma^2(X_p'X_p)^{-1}$. Then, by (3.76), $\text{var}(x_{pi}'\hat{\beta}_p) = x_{pi}'\text{cov}(\hat{\beta}_p)x_{pi}$ and the remaining steps follow as indicated.

18.6 By (3.36), $r_p^2 = \text{SSR}_p/(n - p)$. Then by (3.44),

$$\begin{aligned} c_p &= p + (n - p)\frac{r_p^2 - r_0^2}{r_0^2} = p + (n - p)\left(\frac{r_p^2}{r_0^2} - 1\right) \\ &= p + (n - p)\frac{r_p^2}{r_0^2} - (n - p) = (n - p)\frac{\text{SSR}_p/r_p^2}{n - p} - n + 1p \\ &= \frac{\text{SSR}_p}{r_0^2} - n + 1p. \end{aligned}$$

18.7 $(Y - X\hat{\beta})(Y - X\hat{\beta})' = Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y - \hat{\beta}'X'X\hat{\beta}$. Transpose $\hat{\beta} = (X'X)^{-1}X'Y$ from (3.46) and substitute into $\hat{\beta}'X'X\hat{\beta}$.

$$\begin{aligned} 18.8 \quad E(\bar{Y}_1 - E(Y_1)(\bar{Y}_1 - E(Y_1))' &= E(\bar{Y}_1 - E(Y_1) + E(Y_1) - E(Y_1)(\bar{Y}_1 - E(Y_1) \\ &\quad + E(Y_1) - E(Y_1))' \\ &= E(\bar{Y}_1 - E(Y_1)(\bar{Y}_1 - E(Y_1))' \\ &\quad + E(\bar{Y}_1 - E(Y_1))(E(Y_1) - E(Y_1))' \\ &\quad + E(E(Y_1) - E(Y_1)(\bar{Y}_1 - E(Y_1))' \\ &\quad + E(E(Y_1) - E(Y_1))(E(Y_1) - E(Y_1))' \end{aligned}$$

The second and third terms are equal to 0 because $(E(Y_1) - E(Y_1))$ is a constant vector and $E(\bar{Y}_1 - E(Y_1)) = E(\bar{Y}_1) - E(Y_1) = 0$. The fourth term is a constant matrix and the first E can be deleted.

18.9 As in Problem 18.5, we have $\text{cov}(\hat{\beta}_{p(j)}) = \sigma_{jj}^2(X_p'X_p)^{-1}$, where $\sigma_{jj}^2 = \text{var}(y_j)$ is the j th diagonal element of $\Sigma = \text{cov}(y)$. Similarly, $\text{cov}(\hat{\beta}_{p(j)}\hat{\beta}_{p(k)}) = \sigma_{jk}^2(X_p'X_p)^{-1}$, where $\sigma_{jk}^2 = \text{cov}(y_j, y_k)$ is the (j,k) th element of Σ .

The notation $\text{cov}(\hat{\beta}_{p,q_1}, \hat{\beta}_{p,q_2})$ indicates a matrix, containing the covariance of each element of $\hat{\beta}_{p,q_1}$ and each element of $\hat{\beta}_{p,q_2}$. Now for the covariance matrix, $\text{cov}(\hat{\beta}) = \text{cov}(C_{p,q_1}^T \hat{\beta}_{p,q_1}, \dots, C_{p,q_m}^T \hat{\beta}_{p,q_m})$, we need the variance of each of the m random variables and the covariance of each pair. By Problem 18.5 and (3.73), $\text{var}(C_{p,q_1}^T \hat{\beta}_{p,q_1}) = C_{p,q_1}^T \text{cov}(\hat{\beta}_{p,q_1}) C_{p,q_1} = n \alpha C_{p,q_1}^T (X_p^T X_p)^{-1} \alpha_{p,q_1}$. Similarly, $\text{cov}(C_{p,q_1}^T \hat{\beta}_{p,q_1}, C_{p,q_2}^T \hat{\beta}_{p,q_2}) = n \alpha C_{p,q_1}^T (X_p^T X_p)^{-1} \alpha_{p,q_2}$. The other variances and covariances can be obtained in an analogous manner.

18.10 By (18.77), $S_p = E_p/(n-p)$. Thus by (18.55),

$$\begin{aligned} C_p &= pI + (n-p)E_p^{-1}(E_p - E_1) \\ &= pI + (n-p)E_p^{-1} \frac{E_p}{n-p} - (n-p)I \\ &= E_p^{-1}E_p + Cp - nI. \end{aligned}$$

18.11 $E_p^{-1}E_p = (E_p^{-1}E_p) > 0$, because both E_p^{-1} and E_p are positive definite.

18.12 By (18.54), $C_p = E_p^{-1}E_p + Cp - nI$. Using $S_p = E_p/(n-p)$, we obtain

$$\begin{aligned} \left(\frac{E_p}{n-p} \right)^{-1} E_p &= C_p - Cp - nI, \\ (n-p)E_p^{-1}E_p &= C_p + (n-p)I. \end{aligned}$$

18.13 If C_p is replaced by pI in (18.85), we obtain

$$E_p^{-1}E_p = \frac{C_p + (n-p)I}{n-p} = \frac{pI + nI - pI}{n-p} = \frac{(n-p)I}{n-p}.$$

18.14 (a) $\hat{B} = \begin{pmatrix} .0264 & .01243 \\ .0088 & .008 \\ -.0005 & -.013 \\ .0002 & -.004 \end{pmatrix}$

(b) $\lambda_1 = .314$, $\lambda_2^{(0)} = .285$, $\lambda_3^{(0)} = .105$, $\theta = .264$

(c) $\lambda_1 = .1594$, $\lambda_2 = .0565$. The essential rank of B_0 is 2, and the power ranking is $\theta > \theta^{(0)} > \lambda_1 > \lambda_2^{(0)}$.

(d) The Wilks' Λ test of x_2 adjusted for x_1 and x_3 , for example, is given by (18.62) as

$$\Lambda(x_2|x_1, x_3) = \frac{\Lambda(x_1, x_2, x_3)}{\Lambda(x_1, x_3)},$$

which is distributed as $N_{p-1, p-1}$ and has an exact F -transformation. The tests for x_1 and x_3 are similar. For the three tests we obtain the following:

	A	F	p -Value
$x_1(x_1-x_2)$	.431	1.199	.28
$x_2(x_1-x_2)$	.887	2.606	.088
$x_3(x_1-x_2)$	.762	6.417	.008

10.65 (a) $\hat{\beta} = \begin{pmatrix} 34.282 & 38.802 \\ .394 & .345 \\ .528 & .471 \end{pmatrix}$

(b) $A = .371$, $S^{(0)} = .625$, $S^{(10)} = 1.647$, $\rho = .612$

(c) $\lambda_1 = 1.644$, $\lambda_2 = .0029$. The essential rank of $\hat{B}_1$ is 1, and the power ranking is $S > 0^{(10)} > A > S^{(0)}$.

(d)

	A	F	p -Value
x_1x_2	.868	1.027	.367
x_2x_3	.875	1.988	.141

10.66 (a) $\hat{\beta} = \begin{pmatrix} 54.878 & 65.879 & 58.106 \\ .054 & -.048 & .105 \\ -.024 & .305 & .811 \\ .107 & -.096 & .025 \end{pmatrix}$

(b) $A = .665$, $S^{(0)} = .385$, $S^{(10)} = .458$, $\rho = .340$

(c) $\lambda_1 = .3158$, $\lambda_2 = 1.385$, $\lambda_3 = .6045$. The essential rank of $\hat{B}_1$ is 2, and the power ranking is $S^{(10)} > A > S^{(0)} > S$.

(d)

	A	F	p -Value
$x_1(x_1-x_2)$	.542	.863	.407
$x_2(x_1-x_2)$	.847	2.683	.085
$x_3(x_1-x_2)$	.828	3.020	.085

(e)

	A	F	p -Value
$x_1(x_1-x_2)$	.898	1.804	.182
$x_2(x_1-x_2)$	.833	2.982	.088
$x_3(x_1-x_2)$	.872	2.139	.156

10.67 (a) $\hat{\beta} = \begin{pmatrix} -4.140 & 4.958 \\ 1.103 & -.958 \\ .091 & -.002 \\ 1.171 & 1.779 \\ .113 & .048 \\ .667 & -.058 \\ .287 & .489 \\ -.283 & -.009 \\ -.004 & -.004 \end{pmatrix}$

Test of overall regression of (y_1, y_2) on $(x_1, x_2, \dots, x_9)$: $A = .4662$ (with $p = 2$, exact $F = 1.395$, p -value = .301). Tests on subsets (the F 's are exact because $p = 2$).

	A	F	p -Value
(b) $x_1, x_2, x_3, x_4, \dots, x_8$	.886	.828	.827
(c) $x_2, x_3, x_4, x_5, x_6, x_7, x_8$	.874	1.037	.318
(d) $x_1, x_2, x_3(x_4, x_5, \dots, x_8)$	.869	2.178	.099

- 18.88 (a) The overall test of (y_1, y_2) on $(x_1, x_2, \dots, x_8)$ gives $A = .4042$, with (exact) $F = 1.188$ (p -value = .332). Even though this test result is not significant, we give the results of a backward elimination for illustrative purposes:

Partial A-Test on Each x_i Using (18.72)								
Step	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	.729	.949	.817	.819	.821	.949	.934	.947
2	.768		.858	.831	.829	.938	.927	.928
3	.787		.897	.798	.797		.949	.938
4	.875		.892	.821	.794			.923
5	.880		.854	.823	.817			
6	.798			.803	.806			
7	.855			.930				
8	.891							

At each step, the variable deleted was not significant. In fact, the variable remaining at the last step, x_1 , is not a significant predictor of y_1 and y_2 .

- (b) There were no significant x 's, but to illustrate, we will use the three x 's at step-5 and test for each y :

	A	F	p -Value
y_1/x_1	.788	1.548	.829
y_2/x_1	.888	1.984	.142

18.89 (a)
$$B = \begin{pmatrix} 83.708 & 46.793 & 187.823 \\ .019 & -.095 & 1.810 \\ .159 & .085 & -4.855 \\ .264 & .307 & 1.866 \end{pmatrix}$$

$A = .927$, $F^{(2)} = .885$, $D^{(2)} = 4.709$, $\hat{\rho} = .823$

(b)
$$B = \begin{pmatrix} 99.817 & -29.128 & 121.899 \\ -.008 & -.024 & -.007 \\ .047 & 1.152 & 5.775 \\ -.049 & -.442 & -1.798 \\ -.022 & -.020 & -.488 \\ -.159 & 2.128 & 4.387 \\ .084 & -.087 & -.478 \end{pmatrix}$$

$A = .110$, $F^{(2)} = 1.350$, $D^{(2)} = 4.343$, $\hat{\rho} = .769$

$$(c) \quad \hat{\mathbf{B}} = \begin{pmatrix} 730.286 & 123.403 \\ -1.623 & .099 \\ 24.649 & .094 \\ -8.622 & -.304 \\ -8.224 & .082 \\ 23.826 & -.110 \\ 2.862 & .627 \\ -36.186 & -.207 \\ -.208 & .014 \\ -1.160 & -.306 \end{pmatrix}$$

$$\hat{\lambda} = .302, \quad \hat{\Gamma}^{(0)} = 1.236, \quad \hat{\Gamma}^{(1)} = 9.478, \quad \hat{\sigma} = .827$$

19.28 Using a backward elimination based on (11.72), we obtain the following partial R -values:

Step	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	.493	.462	.446	.438	.418	.379	.361	.368	.397
2	.494	.462	.446	.436	.408	.371	.360		.406
3		.460	.400	.434	.412	.375	.361		.406
4		.448	.388	.433	.404	.367			.364
5		.451	.402		.440	.365			.364
6			.400		.380	.346			.345

At step-6, we stop and retain all four x 's because each R has a p -value less than .05.

CHAPTER 11

- 11.1** By (3.36), $\mathbf{S}_{YY} = \mathbf{D}_Y \mathbf{R}_{YY} \mathbf{D}_Y$ and $\mathbf{S}_{XY} = \mathbf{D}_Y \mathbf{R}_{XY} \mathbf{D}_X$, where $\mathbf{D}_Y$ and $\mathbf{D}_X$ are defined below (11.14). Similarly, $\mathbf{S}_{YX} = \mathbf{D}_Y \mathbf{R}_{YX} \mathbf{D}_X$ and $\mathbf{S}_{XX} = \mathbf{D}_X \mathbf{R}_{XX} \mathbf{D}_X$. Substitute these into (11.7), replace $\mathbf{I}$ by $\mathbf{D}_Y^{-1} \mathbf{D}_Y$, and factor out $\mathbf{D}_Y$ on the right.
- 11.2** Multiply (11.7) by $\mathbf{S}_{XX}^{-1} \mathbf{S}_{XY}$ on the left to obtain $(\mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} - r^2 \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY}) \mathbf{b} = \mathbf{0}$. Factor out $\mathbf{S}_{XX}^{-1} \mathbf{S}_{XY}$ on the right to write this in the form $(\mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} - r^2 \mathbf{I}) \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} \mathbf{b} = \mathbf{0}$. Upon comparing this to (11.8), we see that $\mathbf{b} = \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} \mathbf{a}$.
- 11.3** When $p = 1$, r is also 1, and there is only one canonical correlation, which is equal to R^2 from multiple regression (see comments between (11.28) and (11.30)). Thus

$$R = \frac{1 - r_1^2}{1 - r_1^2} = \frac{1 - R_1^2}{1 - R_1^2}.$$

$$\begin{aligned} \text{11.4 } F &= \frac{(1-A)(n-q-1)}{ab} = \frac{(1-\beta)(1-R_1^2)(n-R_1^2)(n-q-1)}{(1-R_1^2)(1-R_1^2)(n)} \\ &= \frac{(1-R_1^2)(1-R_1^2)(n-q-1)}{(1-R_1^2)(n)} \\ &= \frac{(R_1^2-R_1^2)(n-q-1)}{(1-R_1^2)(n)} \end{aligned}$$

11.5 By (11.39),

$$r_1^2 = \frac{2c_1}{1+\lambda_1}, \quad r_1^2 + r_2^2 \lambda_1 = \lambda_1, \quad \lambda_1(1-r_1^2) = r_1^2.$$

11.6 Substitute $\mathbf{K} = (n-1)(\mathbf{S}_{yy} - \mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy})$ and $\mathbf{H} = (n-1)\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}$ from (11.44) and (11.45) into (11.41):

$$\mathbf{H}\mathbf{a} = \lambda\mathbf{H}\mathbf{a},$$

$$(n-1)\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a} = (n-1)\mathbf{a}(\mathbf{S}_{yy} - \mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy})\mathbf{a},$$

$$\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a} = \lambda(\mathbf{S}_{yy} - \mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy})\mathbf{a}.$$

11.7 By (11.42), $\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a} = r^2\mathbf{S}_{yy}\mathbf{a}$. Subtracting $r^2\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a}$ from both sides gives

$$\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a} - r^2\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a} = r^2\mathbf{S}_{yy}\mathbf{a} - r^2\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a},$$

$$(1-r^2)\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\mathbf{a} = r^2(\mathbf{S}_{yy} - \mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy})\mathbf{a}.$$

11.8 (a) $r_1 = .5142$, $r_2 = .1255$

(b)	e_1	e_2	d_1	d_2	
y_1	1.626	-.088	u_1	.486	.828
y_2	-.189	1.029	u_2	-.754	-.419
			u_3	1.081	-.401

(c) $\hat{\beta}$	$\hat{\beta}$	Approximate F	p -Value
1	.7260	2.095	.161
2	.5945	.508	.719

11.9 (a) $r_1 = .7665$, $r_2 = .4667$

(b)	e_1	e_2	d_1	d_2	
y_1	.3111	-1.5954	u_1	.5844	-1.7685
y_2	.3213	1.5384	u_2	.5383	1.7186

(c) $\hat{\beta}$	$\hat{\beta}$	Approximate F	p -Value
1	.3772	8.9872	.0004
2	.4971	.8637	.4611

11.80 (a) $r_1 = .4800, r_2 = .3488, r_3 = .0885$

(b)	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$
x_1	.433	.081	.886	.481	-.262	1.034
x_2	-.824	.816	1.47	-.378	1.028	-.838
x_3	.843	.400	-.699	.865	.215	-.828

(c)	t	S	Approximate F	p -Value
1	.658	2.175	.024	
2	.435	1.952	.394	
3	.466	.171	.881	

11.81 (a) $r_1 = .8151, r_2 = .4135$

(b)	$\hat{\beta}_1$	$\hat{\beta}_2$
x_1	1.038	-.007
x_2	-.498	1.008

	$\hat{\alpha}_1$	$\hat{\alpha}_2$
x_1	1.001	-.754
x_2	.384	-.268
x_3	.842	1.407
x_4	.844	.641
x_5	1.648	-.154
x_6	.219	1.246
x_7	-.883	-.328
x_8	-.843	-.388

(c)	t	S	Approximate F	p -Value
1	.4642	1.1892	.3321	
2	.3183	.8718	.4766	

11.82 (a) $\log(1.186)$

$$h(x_1, x_2, x_3, \dots, x_6) = \frac{\prod_{i=1}^6 (1 - r_i^2)}{\prod_{i=1}^6 (1 - c_i^2)},$$

where r_1^2 and r_2^2 are the squared canonical correlations from the full model, and c_1^2 and c_2^2 are the squared canonical correlations from the reduced model.

$$h(x_1, x_2, x_3, x_4, \dots, x_6) = \frac{(1 - .5208^2 x) - .4947^2)}{(1 - .3681^2 x) - .1761^2)} = \frac{.4643}{.8128} = .5701$$

$$\begin{aligned} \text{(b)} \quad h(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) &= \frac{(1 - .5208^2 x) - .4947^2)}{(1 - .3681^2 x) - .1761^2)} \\ &= \frac{.4643}{.8681} = .5337 \end{aligned}$$

$$\begin{aligned}
 (d) \quad \Delta(r_1, r_2, r_3; a_0, a_1, \dots, a_6) &= \frac{(1 - .8336^2)(2 - .4941^2)}{(1 - .4931^2)(2 - 2.053^2)} \\
 &= \frac{.6643}{.7865} = .8450
 \end{aligned}$$

11.43 (a) $r_1 = .8779$, $r_2 = .5612$, $r_3 = .1893$,

k	β_k	Approximate F	p -Value
1	.0020	17.6776	<.0001
2	.0054	4.8066	.0020
3	.0025	1.1898	.2835

(b) $r_1 = .8779$, $r_2 = .8795$, $r_3 = .5485$,

k	β_k	Approximate F	p -Value
1	.1097	8.818	<.0001
2	.4751	3.627	.061
3	.2050	1.258	.289

(c) $r_1 = .8055$, $r_2 = .8365$,

k	β_k	Approximate F	p -Value
1	.1022	8.2570	<.0001
2	.4941	3.1126	.0769

(d) $r_1 = .8055$, $r_2 = .7567$, $r_3 = .3097$, $r_4 = .3033$, $r_5 = .3094$,

k	β_k	Approximate F	p -Value
1	.0554	4.892	<.0001
2	.5057	2.688	.0007
3	.7547	.829	.6110
4	.8806	.761	.6080
5	.9497	1.24	.2840

CHAPTER 12

12.1 From $\lambda = a^2 \partial a / \partial a$ in (12.7), we obtain $\partial a / \partial a = a \partial a$, which can be factored as $a \cdot a'(\partial a - \lambda a) = 0$. Since $a = 0$ is not a solution to $\lambda = a^2 \partial a / \partial a$, we have $\partial a - \lambda a = 0$.

12.2 $(\mathbf{K} - \lambda \mathbf{I}) = 0$, $\begin{vmatrix} 1-k & r \\ r & 1-k \end{vmatrix} = (1-k)^2 - r^2 = 0$
 $(1-k+r)(1-k-r) = 0$, $k = 1 \pm r$
 With $\lambda_1 = 1 + r$ in $(\mathbf{K} - \lambda_1 \mathbf{I})\mathbf{u}_1 = 0$, we obtain

$$\begin{pmatrix} -r & r \\ r & -r \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which gives $u_{11} = u_{22}$ for any r . Normalizing to $u_1^2 + u_2^2 = 1$, yields $u_{11} = 1/\sqrt{2}$.

- 12.3 (a) By (4.14) and a comment in Section 7.1, the likelihood ratio is given by $L(\mathbf{x}) = (2\pi)^k/(\hat{\Sigma}_0)^{k/2}$, where $\hat{\Sigma}_0$ is the estimate of Σ under H_0 . By (2.108) and (7.6), the test statistic is

$$\begin{aligned} -2 \ln LR &= -2 \ln \left(\frac{(2\pi)^k}{(\hat{\Sigma}_0)^k} \right)^{1/2} = -2 \left(\frac{n}{2} \right) \ln \frac{\prod_{i=1}^{p-k} \lambda_i \prod_{i=p-k+1}^p \lambda_i}{\prod_{i=1}^p \lambda_i \prod_{i=p-k+1}^p \lambda_i} \\ &= -n \ln \frac{\prod_{i=p-k+1}^p \lambda_i}{\lambda^k} = n \left(k \ln \lambda - \sum_{i=p-k+1}^p \ln \lambda_i \right). \end{aligned}$$

In (12.15), the coefficient n is modified to give an improved chi-square approximation.

- 12.4 $\Sigma \mathbf{B}$ is diagonal, then $\lambda_i = n_{ii}$, as in (12.17). Thus

$$\begin{aligned} \hat{\Sigma} \mathbf{B}_1 &= \hat{\Sigma}_0 \mathbf{B}_1 = \hat{\Sigma}_0 \mathbf{B}_0 \\ \begin{pmatrix} r_{11} & 0 & \cdots & 0 \\ 0 & \lambda_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{pp} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} &= \begin{pmatrix} r_{11}a_1 \\ r_{22}a_2 \\ \vdots \\ \lambda_{pp}a_p \end{pmatrix} = \begin{pmatrix} \lambda_{11}a_1 \\ \lambda_{22}a_2 \\ \vdots \\ \lambda_{pp}a_p \end{pmatrix}. \end{aligned}$$

From the first element, we obtain $r_{11}a_1 = \lambda_{11}a_1$ or $\lambda_{11} = a_1^2 M_{11} = 0$. Since $a_1 = a_2 = \cdots = 0$ (unless $i = 1$), we must have $\lambda_{11} = 0$. Thus, $\mathbf{a}_1 = (0, \dots, 0, a_k, 0, \dots, 0)^T$, and normalizing $\mathbf{a}_1$ leads to $\mathbf{a}_1 = \mathbf{e}_k$.

- 12.5 By (10.34) and (1.23),

$$\begin{aligned} R_{1 \times (p-1) \times \dots \times p}^2 &= \frac{\mathbf{e}_{11}^T \mathbf{B}_{11}^{-1} \mathbf{a}_{12}}{\lambda_{11}^2} \\ &= (a_{121} \ a_{122} \ \cdots \ a_{12p}) \begin{pmatrix} a_{11}^2 & 0 & \cdots & 0 \\ 0 & \lambda_{22}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{pp}^2 \end{pmatrix}^{-1} \begin{pmatrix} a_{p12} \\ a_{p22} \\ \vdots \\ a_{pp2} \end{pmatrix} / \lambda_{11}^2. \end{aligned}$$

Show that this is equal to

$$R_{1 \times (p-1) \times \dots \times p}^2 = \sum_{j=2}^p \frac{a_{12j}^2}{\lambda_{11}^2 \lambda_{jj}^2} = \sum_{j=2}^p r_{12j}^2.$$

- 12.6 The variances of y_1 , y_2 , x_1 , x_2 , and x_3 on the diagonal of $\mathbf{S}$ are .045, .705, 11.665, 2381.3, and 2185.8. The eigenvalues of $\mathbf{S}$ and $\mathbf{R}$ are as follows:

S			R		
λ_i	$\lambda_i / \sum_{j=1}^5 \lambda_j$	Cumulative	λ_i	$\lambda_i / \sum_{j=1}^5 \lambda_j$	Cumulative
5499.18	.898807	.8988	1.72	.04	.04
1268.47	.202828	.9009	1.23	.03	.08
894.27	.147186	.9250	.86	.02	.10
69.34	.012176	.9369	.78	.02	.14
.01	.000002	1.0000	.06	.00	1.00

Two principal components of **S** account for 90% of the variance, but it requires three principal components of **R** to reach 78%. For most purposes we would use two components of **R**, although with these we could account for 93% of the variance. However, we show all five eigenvectors below because of the interesting pattern they exhibit. The first principal component is largely a weighted average of the last two variables, x_2 and x_3 , which have the largest variances. The second and third components represent contrasts in the last three variables and could be described as “shape” components. The fourth and fifth components are associated uniquely with x_1 and x_4 , respectively. These components are “variable specific,” as described in the discussion of method 1 in Section 12.6. As expected, the principal components of **R** show an entirely different pattern. All five variables contribute to the first three components of **R**, whereas in **S**, x_1 and x_4 have small variances and contribute almost nothing to the first three components. The eigenvectors of **S** and **R** are as follows:

	S					R				
	u_1	u_2	u_3	u_4	u_5	u_1	u_2	u_3	u_4	u_5
x_1	.0004	-.0008	.0009	.0028	.0009	.42	.53	-.42	-.48	.48
x_2	-.0008	.0005	.0086	.9944	-.0029	.07	.09	.36	.76	-.19
x_3	.0047	.0082	.7333	-.0808	-.0008	.58	.50	.78	-.44	-.26
x_4	.7438	.0279	-.0143	.0166	.0009	.58	-.43	.33	.09	.08
x_5	.0011	-.0007	.4083	.0042	-.0003	.00	-.18	-.40	.18	-.00

$$\begin{aligned}
 \mathbf{S} &= \begin{pmatrix} 65.1 & 31.6 & 47.6 & 36.8 & 25.4 \\ 31.6 & 46.1 & 28.9 & 65.3 & 28.6 \\ 47.6 & 28.9 & 60.7 & 37.4 & 41.1 \\ 36.8 & 65.3 & 37.4 & 62.8 & 21.7 \\ 25.4 & 28.6 & 41.1 & 21.7 & 58.2 \end{pmatrix} \\
 \mathbf{R} &= \begin{pmatrix} 1.00 & .64 & .76 & .58 & .41 \\ .64 & 1.00 & .59 & .79 & .69 \\ .76 & .59 & 1.00 & .61 & .69 \\ .58 & .79 & .61 & 1.00 & .50 \\ .41 & .69 & .69 & .50 & 1.00 \end{pmatrix}
 \end{aligned}
 \tag{12.7}$$

The eigenvalues of **S** and **R** are as follows:

S			R		
λ_i	$\lambda_i / \sum_{i=1}^3 \lambda_i$	Cumulative	λ_i	$\lambda_i / \sum_{i=1}^3 \lambda_i$	Cumulative
2004	.660	.660	1.42	.663	.663
16.1	.120	.880	.81	.324	.886
14.1	.116	.996	.37	.144	.930
15.0	.09	.995	.27	.094	.957
7.4	.025	1.000	.13	.023	1.000

The first three eigenvectors of **S** and **R** are as follows:

S			R		
$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
.47	-.36	-.42	.44	-.30	-.68
.36	-.11	.43	.45	-.43	.33
.49	.30	-.46	.47	.37	-.38
.47	-.32	.62	.48	-.39	.33
.41	.80	-.39	.41	.70	.41

The variances in **S** are nearly identical, and the covariances are likewise similar in magnitude. Consequently, the percent of variance explained by the eigenvalues of **S** and **R** are indistinguishable. The interpretation of the second principal component from **S** is slightly different from that of the second one from **R**, but otherwise there is little to choose between them.

- 12.8** The variances on the diagonal of **S** are 95.3, 73.2, 76.2, 808.6, 908.3, and 508.7. The eigenvalues of **S** and **R** are as follows:

S			R		
λ_i	$\lambda_i / \sum_{i=1}^6 \lambda_i$	Cumulative	λ_i	$\lambda_i / \sum_{i=1}^6 \lambda_i$	Cumulative
1132.0	.537	.537	2.17	.363	.363
564.1	.261	.798	1.08	.180	.543
540.8	.250	.898	.98	.165	.708
91.8	.047	.965	.87	.144	.808
60.8	.028	.978	.55	.089	.862
44.6	.021	1.000	.35	.058	1.000

We could keep either two or three components from **S**. The first three components of **S** account for a larger percent of variance than do the first three from **R**. The first three eigenvectors of **S** and **R** are as follows:

S			R		
a_1	a_2	a_3	a_1	a_2	a_3
.060	.000	-.008	.136	.176	.497
.066	-.048	.308	.186	.843	-.089
.076	.133	-.011	.176	.848	.186
.158	-.446	-.468	.473	-.328	-.108
.499	-.081	.844	.486	.876	-.167
.412	.616	-.147	.471	-.106	.176

As expected, the first three principal components from S are heavily influenced by the last three variables because of their relatively large variances.

- 12.9 The variances on the diagonal of S are .05, .54, 1.038, .0624, .75.3, .56.4, 1.8.1. With the large variance of y_3 , we would expect the first principal component from S to account for most of the variance, and y_3 would essentially constitute that single component. This is indeed the pattern that emerges in the eigenvalues and eigenvectors of S. The principal components from R, on the other hand, are not dominated by y_3 . The eigenvalues of S and R are as follows:

S		R		Cumulative
λ_i	$\lambda_i / \sum_{j=1}^6 \lambda_j$	λ_i	$\lambda_i / \sum_{j=1}^6 \lambda_j$	
2.006760	.444184	2.43	.406	.406
.00	.000000	1.48	.246	.652
.18	.000008	1.83	.311	.963
.7	.000003	.91	.150	.993
.3	.000001	.30	.050	.999
.9	.000000	.00	.004	1.000

Most of the correlations in R are small (only three exceed .34), and its first three principal components account for only 72% of the variance. The first three eigenvalues of S and R are as follows:

S			R		
a_1	a_2	a_3	a_1	a_2	a_3
.00066	.004	-.0136	.414	-.161	-.130
.00081	.017	.0767	.446	-.108	.087
.00098	-.001	-.0002	.168	.187	-.001
.00020	.008	.0174	.414	.167	.146
.00022	-.116	.0195	.168	.423	-.106
.00020	.021	.0002	.013	.009	.100

12.89 Covariance matrix for males:

$$\Sigma_M = \begin{pmatrix} 5.39 & 4.39 & 6.32 & 5.29 \\ 4.39 & 13.18 & 6.35 & 6.27 \\ 6.32 & 6.35 & 28.67 & 14.47 \\ 5.29 & 6.27 & 14.47 & 16.65 \end{pmatrix}$$

Covariance matrix for females:

$$\Sigma_F = \begin{pmatrix} 9.14 & 7.55 & 4.86 & 4.15 \\ 7.55 & 18.60 & 30.22 & 9.43 \\ 4.86 & 18.22 & 30.84 & 13.49 \\ 4.15 & 9.43 & 13.49 & 28.60 \end{pmatrix}$$

The eigenvalues are as follows:

Males			Females		
λ_i	$10^2/\sum_{i=1}^4 \lambda_i$	Cumulative	λ_i	$10^2/\sum_{i=1}^4 \lambda_i$	Cumulative
45.56	.084	.086	40.96	.591	.591
11.14	.773	.658	16.48	.215	.766
6.47	.155	.813	15.56	.158	.944
2.82	.646	1.000	4.62	.685	1.000

The first two eigenvectors are as follows:

Males		Females	
a_1	a_2	a_1	a_2
.24	.25	.22	.27
.34	.65	.39	.62
.36	-.28	.46	.27
.32	.69	.38	-.72

The variances in Σ_M have a slightly wider range (5.18–28.67) than those in Σ_F (7.14–30.84), and this is reflected in the eigenvalues. The first two components account for 85% of the variance from Σ_M , whereas the first two account for 79% from Σ_F .

12.91 Covariance matrix for species 1:

$$\Sigma_1 = \begin{pmatrix} 187.8 & 176.9 & 48.4 & 113.6 \\ 176.9 & 343.4 & 56.9 & 118.8 \\ 48.4 & 56.9 & 86.4 & 16.2 \\ 113.6 & 118.8 & 16.2 & 259.8 \end{pmatrix}$$

Covariance matrix for species 2:

$$\mathbf{B}_2 = \begin{pmatrix} 354.8 & 128.1 & 37.8 & 32.6 \\ 128.1 & 145.0 & 145.4 & 94.4 \\ 37.8 & 145.4 & 167.5 & 66.5 \\ 32.6 & 94.4 & 66.5 & 173.9 \end{pmatrix}$$

The eigenvalues are as follows:

Species 1			Species 2		
λ_i	$\lambda_i / \sum_{j=1}^4 \lambda_j$	Cumulative	λ_i	$\lambda_i / \sum_{j=1}^4 \lambda_j$	Cumulative
504.9	.668	.668	555.7	.664	.664
145.0	.191	.859	145.4	.174	.838
65.5	.086	.945	93.9	.112	.950
45.7	.057	1.000	41.7	.050	1.000

The first two eigenvectors are as follows:

Species 1		Species 2	
$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_1$	$\mathbf{a}_2$
.50	.04	.38	-.39
.72	-.48	.81	-.36
.17	-.33	.43	.14
.48	.83	.39	.84

The variances in $\mathbf{B}_1$ have a wider range than those in $\mathbf{B}_2$, and the first two components of $\mathbf{B}_1$ account for a higher percent of variance.

12.33 The variances on the diagonal of $\mathbf{B}$ in each case are:

- (a) Pooled: 536.0, 39.9, 116.0, 896.4, 268.1, 982.0,
 (b) Unpooled: 518.2, 68.0, 145.2, 1066.4, 264.6, 1069.2.

The eigenvalues are as follows:

Pooled			Unpooled		
λ_i	$\lambda_i / \sum_{j=1}^6 \lambda_j$	Cumulative	λ_i	$\lambda_i / \sum_{j=1}^6 \lambda_j$	Cumulative
1056.6	.368	.368	1122.0	.368	.368
436.3	.148	.516	436.4	.139	.507
396.4	.137	.653	401.4	.127	.634
284.2	.095	.748	264.1	.079	.713
158.1	.053	.801	126.9	.037	.750
41.4	.014	1.000	80.4	.025	1.000

The first three eigenvectors are as follows:

Pooled			Unpooled		
$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
.443	-.190	.894	.212	.380	.888
.043	-.088	.080	-.070	.084	.070
-.039	.081	.143	.080	-.080	.083
.430	.003	-.053	.276	-.036	.083
-.039	-.004	-.004	-.036	.030	.013
.714	-.407	-.471	.380	.080	-.039

- (ii) The pattern in eigenvectors as well as eigenvalues is similar for the pooled and unpooled cases. The first three principal components account for 87.2% of the variance in the unpooled case compared to 84.9% for the pooled case.

12.53 The variances on the diagonal of $\mathbf{S}$ in each case are:

- (a) Pooled: 49.1, 8.1, 12040.8, 136.2, 219.8, 5983.9,
 (b) Unpooled: 81.2, 8.0, 15168.8, 186.8, 250.4, 6660.7.

The eigenvectors are as follows:

Pooled			Unpooled		
λ_i	$\lambda_i / \sum_{j=1}^6 \lambda_j$	Cumulative	λ_i	$\lambda_i / \sum_{j=1}^6 \lambda_j$	Cumulative
12,040.8	.8209	.8209	17,067.8	.8408	.8408
2,405.9	.1582	.9791	2,958.8	.1404	.9804
131.1	.0088	.9879	168.0	.0080	.9907
71.2	.0058	.9937	71.1	.0038	.9974
42.2	.0027	.9964	44.7	.0022	.9996
7.4	.0005	1.0000	7.3	.0004	1.0000

The eigenvectors are as follows:

Pooled			Unpooled	
$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_1$	$\mathbf{a}_2$
-.004	-.000	.013	.017	
-.005	.004	-.006	.004	
.000	-.113	.001	-.001	
-.003	.013	.008	.009	
.130	.040	.130	.020	
.020	.070	.050	.032	

- 12.54 The variances on the diagonal of $\mathbf{S}$ are all less than 1, except $v_{11}^2 = 5.01$ and $v_{22}^2 = 1341.38$. We therefore expect the last variable, x_3 , to dominate the principal components of $\mathbf{X}$. This is the case for $\mathbf{X}$ but not for $\mathbf{R}$. The eigenvalues of $\mathbf{S}$ and $\mathbf{R}$ are as follows:

S		R		Cumulative
λ_i	$\lambda_i / \sum_{j=1}^n \lambda_j$	λ_i	$\lambda_i / \sum_{j=1}^n \lambda_j$	
1541.35	.000275	5.714	.377	.377
4385	.000126	2.868	.186	.563
.48	.000086	1.432	.113	.677
.37	.000076	1.277	.128	.805
.19	.000066	.561	.034	.839
.08	.000046	.475	.047	.886
.02	.000044	.251	.025	.911
.02	.000041	.116	.012	.923
.01	.000038	.104	.010	.934
.02	.000038	.064	.006	.940

The eigenvalues of S and R are as follows:

S		R			
λ_1	λ_2	λ_1	λ_2	λ_3	λ_4
.0008	-.008	.12	.19	.69	.13
.0007	-.004	.06	.32	.34	.26
.0029	-.007	.46	-.06	.07	-.08
.0004	.004	.26	.17	-.18	.49
.0009	-.000	.32	.14	-.04	-.01
-.0159	.040	-.06	-.42	.07	.55
-.0008	-.000	-.11	.43	-.01	-.14
-.0002	-.156	-.15	.34	-.14	-.13
.0008	-.011	.08	.36	-.08	.08
.0006	.014	.30	.11	-.12	-.09

12.15 The variances in the diagonal of S are .53.7, 16.8, 402.7, 25.3, 71.4, 436.3, 1.5, 186.2, 883.6, 11573.5, 714.1.

The eigenvalues of S and R are as follows:

S			R		
λ_i	$\lambda_i / \sum_{j=1}^n \lambda_j$	Cumulative	λ_i	$\lambda_i / \sum_{j=1}^n \lambda_j$	Cumulative
22,861.5	.9478	.9478	6.828	.5470	.5470
150.7	.0034	.9512	1.019	.1087	.6557
556.0	.0146	.9658	1.038	.1075	.7632
664	.0036	.9694	.768	.0890	.8522
26.5	.0010	.9704	.281	.0204	.8726
17.1	.0007	.9711	.284	.0208	.8934
12.7	.0005	.9716	.122	.0110	.9044
.28	.0001	.9717	.118	.0108	.9152
1.9	.0000	.9718	.088	.0074	.9227
.9	.00004	.9718	.642	.0084	.9311
.7	.00003	1.0000	.821	.0010	1.0000

The eigenvectors of $\mathbf{S}$ and $\mathbf{R}$ are as follows:

	$\mathbf{S}$		$\mathbf{R}$			
	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$
$\mathbf{p}_1$	-.0081	.1111	.1804	-.0087	.0000	-.2807
$\mathbf{p}_2$	.0000	.0000	.1542	.1020	.1071	-.2508
$\mathbf{p}_3$	-.0141	.4397	.1613	.0500	.1105	-.1415
$\mathbf{p}_4$	-.0013	.1070	.1620	.0474	.1554	-.1030
$\mathbf{p}_5$	.0101	.0198	.1513	.1505	.0154	-.0710
$\mathbf{p}_6$	.0107	.4295	.1621	.1501	.1198	.1580
$\mathbf{p}_7$	-.0012	-.0072	-.0004	.0213	.7960	.0114
$\mathbf{p}_8$	.0179	-.1044	-.1301	.1023	.0020	-.1500
$\mathbf{p}_9$	.0420	-.0037	-.1111	.1003	.2130	-.2270
$\mathbf{p}_{10}$	.0002	.0146	-.0043	.4085	-.4000	.5000
$\mathbf{p}_{11}$	.0034	.1111	.1177	-.1155	-.1050	.6550

For most purposes, one or two principal components would suffice for $\mathbf{S}$, with 91% or 98% of the variance explained. For $\mathbf{R}$, on the other hand, three components are required to explain 84% of the variance, and seven components are necessary to reach 98%. The reduction to one or two components for $\mathbf{S}$ is due in part to the relatively large variances of γ_1 , γ_2 , γ_3 , and γ_{10} . In the eigenvectors of $\mathbf{S}$, we see that these four variables figure prominently in the first two principal components.

CHAPTER 13

$$\begin{aligned}
 13.1 \quad \text{var}(y_0) &= \text{var}(y_0 - \mu_0) = \text{var}(\lambda_{00}f_0 + \lambda_{01}f_1 + \cdots + \lambda_{0n}f_n + \epsilon_0) \\
 &= \sum_{j=0}^n \lambda_{0j}^2 \text{var}(f_j) + \text{var}(\epsilon_0) + \sum_{j \neq k} \lambda_{0j}\lambda_{0k} \text{cov}(f_j, f_k) \\
 &\quad + \sum_{j=0}^n \lambda_{0j} \text{cov}(f_j, \epsilon_0) \\
 &= \sum_{j=0}^n \lambda_{0j}^2 + \phi_0.
 \end{aligned}$$

The last equality follows by the assumptions $\text{var}(f_j) = 1$, $\text{var}(\epsilon_0) = \phi_0$, $\text{cov}(f_j, f_k) = 0$, and $\text{cov}(f_j, \epsilon_0) = 0$.

$$\begin{aligned}
 13.2 \quad \text{cov}(\mathbf{Y}, \mathbf{E}) &= \text{cov}(\mathbf{A}\mathbf{F} + \mathbf{u}, \mathbf{E}) && \text{[Eq (13.7)]} \\
 &= \text{cov}(\mathbf{A}\mathbf{F}, \mathbf{E}) && \text{[Eq (13.10)]} \\
 &= \mathbf{E}[\mathbf{A}\mathbf{F} - \mathbf{E}(\mathbf{A}\mathbf{F})][\mathbf{E} - \mathbf{E}(\mathbf{E})]' && \text{[by analogy to (3.31)]} \\
 &= \mathbf{E}[\mathbf{A}\mathbf{F} - \mathbf{A}\mathbf{E}(\mathbf{F})][\mathbf{E} - \mathbf{E}(\mathbf{E})]' \\
 &= \mathbf{A}\mathbf{E}[\mathbf{F} - \mathbf{E}(\mathbf{F})][\mathbf{E} - \mathbf{E}(\mathbf{E})]' \\
 &= \mathbf{A} \text{cov}(\mathbf{F}) = \mathbf{A} && \text{[Eq (13.7)]}
 \end{aligned}$$

$$(13.1) \quad E(F) = E(TF) = TE(F) = T\Phi = \Phi,$$

$$\operatorname{cor}(F) = \operatorname{cor}(TF) = T\operatorname{cor}(F) = T\Gamma = \Gamma$$

(13.4) Let $\mathbf{B} = \mathbf{B} - (\hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Phi})$. Then by (2.30), $\operatorname{tr}(\mathbf{B}\mathbf{C}) = \sum_{i=1}^p e_i^2$ by (13.26), $\hat{\Phi} = \operatorname{diag}(\mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}')$, and $\mathbf{B}$ has zeros on the diagonal. This gives the inequality

$$\sum_{i=1}^p e_i^2 \leq \text{sum of squared elements of } \mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}'.$$

By (2.30),

$$\text{Sum of squared elements of } \mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}' = \operatorname{tr}(\mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}')(\mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}').$$

Since $\mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}'$ is symmetric, we have by (13.28), (13.23), and (13.24),

$$\begin{aligned} \mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}' &= \mathbf{C}\mathbf{B}\mathbf{C}' - \mathbf{C}_1\mathbf{B}_1^{1/2}\mathbf{B}_1^{1/2}\mathbf{C}_1' \\ &= \mathbf{C}\mathbf{B}\mathbf{C}' - \mathbf{C}_1\mathbf{B}_1\mathbf{C}_1', \end{aligned}$$

where $\mathbf{C} = (c_0, c_1, \dots, c_p)$ contains normalized eigenvectors of $\mathbf{B}$, $\mathbf{B} = \operatorname{diag}(b_1, b_2, \dots, b_p)$ contains eigenvalues of $\mathbf{B}$, $\mathbf{C}_1 = (c_1, c_2, \dots, c_{p-1})$, and $\mathbf{B}_1 = \operatorname{diag}(b_1, b_2, \dots, b_{p-1})$.

Using the partitioned forms $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2)$ and $\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{pmatrix}$, show that $\mathbf{C}_1'\mathbf{C}_1 = \mathbf{I}_p$, $\mathbf{C}_1'\mathbf{C}_2 = \mathbf{0}$, $\mathbf{C}'\mathbf{C}_1 = \begin{pmatrix} \mathbf{I}_p \\ \mathbf{0} \end{pmatrix}$, $\mathbf{B}\begin{pmatrix} \mathbf{I}_p \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{pmatrix}$, $\mathbf{C}\begin{pmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{pmatrix} = \mathbf{C}_1\mathbf{B}_1$, and $\mathbf{C}\mathbf{B}\mathbf{C}'\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1' = \mathbf{C}_1\mathbf{B}_1^2\mathbf{C}_1'$. Show similarly that $\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1'\mathbf{C}\mathbf{B}\mathbf{C}' = \mathbf{C}_1\mathbf{B}_1^2\mathbf{C}_1'$ and $\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1'\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1' = \mathbf{C}_1\mathbf{B}_1^3\mathbf{C}_1'$. Now by (2.37) $\operatorname{tr}(\mathbf{C}\mathbf{B}\mathbf{C}') = \operatorname{tr}(\mathbf{C}'\mathbf{C}\mathbf{B}) = \operatorname{tr}(\mathbf{B}^2) = \sum_{i=1}^p b_i^2$. Similarly, $\operatorname{tr}(\mathbf{C}_1\mathbf{B}_1^2\mathbf{C}_1') = \sum_{i=1}^{p-1} b_i^2$. Then

$$\begin{aligned} \operatorname{tr}(\mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}')(\mathbf{B} - \hat{\mathbf{A}}\hat{\mathbf{A}}') &= \operatorname{tr}(\mathbf{C}\mathbf{B}\mathbf{C}' - \mathbf{C}_1\mathbf{B}_1\mathbf{C}_1')(\mathbf{C}\mathbf{B}\mathbf{C}' - \mathbf{C}_1\mathbf{B}_1\mathbf{C}_1') \\ &= \operatorname{tr}(\mathbf{C}\mathbf{B}\mathbf{C}'\mathbf{C}\mathbf{B}\mathbf{C}' - \mathbf{C}\mathbf{B}\mathbf{C}'\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1' - \mathbf{C}_1\mathbf{B}_1\mathbf{C}_1'\mathbf{C}\mathbf{B}\mathbf{C}' \\ &\quad + \mathbf{C}_1\mathbf{B}_1\mathbf{C}_1'\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1') \\ &= \sum_{i=1}^p b_i^2 - \sum_{i=1}^{p-1} b_i^2 - \sum_{i=1}^{p-1} b_i^2 + \sum_{i=1}^{p-1} b_i^2 \\ &= \sum_{i=p}^p b_i^2. \end{aligned}$$

$$13.6 \quad \sum_{i=1}^p \sum_{j=1}^m \lambda_{ij}^2 = \sum_{i=1}^p \left[\sum_{j=1}^m \lambda_{ij}^2 \right] = \sum_{i=1}^p \hat{\lambda}_i^2 \quad \text{By (13.28)}$$

By interchanging the order of summation, we have

$$\sum_{i=1}^p \sum_{j=1}^m \lambda_{ij}^2 = \sum_{j=1}^m \sum_{i=1}^p \lambda_{ij}^2 = \sum_{j=1}^m \hat{\lambda}_j^2 \quad \text{By (13.28)}.$$

13.6 We use the covariance matrix to avoid working with standardized variables. The eigenvalues of $\mathbf{B}$ are 38.76, 8.78, .66, .50, and 0. The eigenvector corresponding to $\lambda_5 = 0$ is

$$\mathbf{a}_5^* = (-.75, -.25, .25, .50, .25).$$

As noted in Section 12.1, $\lambda_5^2 = 0$ implies $\tau_5 = 0$. Thus

$$\tau_5 = \mathbf{a}_5^* \mathbf{g} = -.75g_1 -.25g_2 + .25g_3 + .50g_4 + .25g_5 = 0,$$

$$8g_1 + g_2 = g_3 + 2g_4 + g_5.$$

13.7 Words data of Table 3.2:

	Principal Component Loadings		Variance-Rotated Loadings		Concentration
	$\hat{F}_1$	$\hat{F}_2$	$\hat{F}_1$	$\hat{F}_2$	$\hat{d}^2$
Variables					
Internal words	.805	-.505	.806	.324	.550
Internal words	.836	-.528	.808	.321	.850
Formal words	.883	-.279	.864	.786	.600
Formal words	.714	.658	.185	.866	.640
Variance	1.666	.809	1.884	1.671	5.968
Proportion	.666	.325	.674	.418	.864

The orthogonal matrix $\mathbf{T}$ for the variance rotation is given by (13.45) is

$$\mathbf{T} = \begin{pmatrix} .750 & .661 \\ -.661 & .750 \end{pmatrix}.$$

Thus $\sin \phi = -.661$ and $\phi = -41.4^\circ$. A graphical rotation of -40° would produce results very close to the variance rotation.

13.8. Kaiser-Meyer-Olkin data of Table 3.6.

	Principal Component Loadings		Varimax Rotated Loadings		Communality, R^2	Orthogonal Pattern Loadings	
	F_1	F_2	F_1	F_2		F_1	F_2
Variables							
0 years	.940	-.193	.854	.455	.909	-.118	1.007
R^2 years	.974	-.193	.856	.505	.906	1.108	.960
0 years	.978	.171	.878	.488	.906	.828	.188
R^2 years	.961	.338	.849	.688	.891	1.888	-.121
Variance	3.695	.293	3.008	1.949	3.931		
Proportion	.918	.064	.501	.485	.908		

The Harris-Kaiser orthoblique rotation produced loadings for which the variables have a complexity of 1. These oblique loadings provide a much cleaner simple structure than that given by the varimax loadings. For interpretation, we use F_1 as factor representing variables 1 and 2, and the other factor represents variables 3 and 4. This same clustering of variables can be deduced from the varimax loadings if we simply use the larger of the two loadings for each variable.

The correlation between the two oblique factors is .87. The angle between the oblique axes is $\cos^{-1}(.87) = 29.5^\circ$. With such a small angle between the axes, and a large correlation between the factors, it is clear that a single factor would better represent the variables. This is also borne out by the eigenvalues of the correlation matrix: 3.695, .293, .033, and .017. The first accounts for 92% of the variance and the second for only 6%.

13.9. Rotablock data of Table 6.2.

	Principal Component Loadings		Varimax Rotated Loadings		Communality, R^2
	F_1	F_2	F_1	F_2	
Variables					
Trunk 4 years	.707	.676	.387	.860	.649
Extremities 4 years	.889	.687	.287	.828	.619
Trunk 12 years	.875	-.416	.846	.280	.879
Weight 12 years	.828	-.347	.873	.179	.898
Variance	2.782	1.858	1.931	1.888	3.819
Proportion	.696	.281	.483	.472	.900

The rotation was successful in producing variables with a complexity of 1, that is, partitioning the variables into two groups, each with two variables.

13.10 (a) Fish data of Table 8.17:

	Principal Component Loadings		Rotated Factor Loadings		Communality
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$	
Variables					
y_1	.830	-.408	.874	.294	.831
y_2	.783	-.308	.811	.189	.808
y_3	.808	.432	.278	.879	.831
y_4	.508	.407	.200	.899	.828
Variance	1.177	.858	1.709	1.679	3.888
Proportion	.634	.313	.627	.628	.847

- (b) The loadings for y_1 and y_2 are similar. In $\mathbf{R}$ we see some indication of the reason for this: y_1 and y_2 are more highly correlated than any other pair of variables, and their correlations with y_3 and y_4 are similar:

$$\mathbf{R} = \begin{pmatrix} 1.00 & .71 & .51 & .40 \\ .71 & 1.00 & .38 & .40 \\ .51 & .38 & 1.00 & .87 \\ .40 & .40 & .87 & 1.00 \end{pmatrix}.$$

- (c) By (13.58), the factor score coefficient matrix is

$$\mathbf{B}_0 = \mathbf{R}^{-1} \hat{\mathbf{A}} = \begin{pmatrix} .808 & -.307 \\ .656 & -.207 \\ -.188 & .884 \\ -.194 & .630 \end{pmatrix},$$

where $\hat{\mathbf{A}}$ is the matrix of rotated factor loadings given in part (a). The factor scores are given by (13.59) as follows:

Method 1		Method 2		Method 3	
F_1	F_2	F_1	F_2	F_1	F_2
.344	1.020	-.294	.500	-.1236	2.134
1.250	-.234	-.260	-1.534	-.323	.678
1.047	1.020	-1.468	-1.588	-.473	.627
-.147	-1.083	-.469	-.660	.367	1.130
.219	-.383	.528	-.343	-1.408	-.488
1.007	.678	.818	-.111	.303	.691
1.413	-.386	-.460	-.016	-.454	1.157
-.696	-1.276	-.265	.635	-.363	.360
1.057	-1.676	1.469	-.265	-.236	1.273
.388	-.448	1.371	.265	-1.389	.354
1.326	-.298	1.268	-.027	-1.766	-.111
.684	-.833	-.808	-1.432	-1.636	-.368

- (d) A one-way MANOVA on the two-factor scores comparing the three methods yielded the following values for $\mathbf{E}$ and $\mathbf{H}$:

$$\mathbf{E} = \begin{pmatrix} 11.8005 & 10.3673 \\ 10.3673 & 23.2681 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} 13.1394 & -10.3673 \\ -10.3673 & 9.7819 \end{pmatrix}.$$

The four MANOVA test statistics are $\lambda = .5831$, $V^{(4)} = .6593$, $L^{(4)} = 1.7635$, and $P = .8159$. All are highly significant.

- 13.11 (a) For the first data of Table 5.3, the eigenvalues of $\mathbf{E}$ are 1.273, 1.081, .450, and .195. There is a noticeable gap between 1.081 and .450, and the first two factors account for 83.5% of the variance. Thus $m = 2$ factors seem to be indicated for this set of data.

(b)

	Principal Component Loadings		Varimax Rotated Loadings		Eigenvalues	Orthogonal Pattern Loadings	
	F_1	F_2	F_1	F_2		F_1	F_2
Variance							
%	-.354	.590	-.323	.698	.968	-.383	.698
%	.880	.360	.861	.336	.862	.898	.270
%	.809	-.157	.861	-.176	.823	.867	-.179
%	.827	-.075	.823	-.084	.689	.834	-.087
Varimax	2.273	1.081	1.273	1.081	3.354		
Proportion	.568	.267	.568	.267	.835		

(The variance explained by the varimax rotated factors remains the same as for the initial factors when rounded to three decimal places.)

- (d) In this case, neither of the rotations changes the initial loadings appreciably. The reason for this unusual outcome can be seen in the correlation matrix:

$$\mathbf{R} = \begin{pmatrix} 1.00 & .18 & -.17 & -.03 \\ .18 & 1.00 & .73 & .59 \\ -.17 & .73 & 1.00 & .58 \\ -.03 & .59 & .58 & 1.00 \end{pmatrix}.$$

There are clearly two clusters of variables: $\{y_1\}$ and $\{y_2, y_3, y_4\}$. We would expect two factors corresponding to these groupings to emerge after rotation. That the same pattern surfaces in the initial factor loadings (based on eigenvalues) is due to their affiliation with principal components. As noted in Section 12.8.1, if a variable has small correlations with all other variables, the variable itself will essentially constitute a principal component. In this case, y_1 has this property and makes up most of the second principal component. The first component is comprised of the other three variables.

- 15.62 (a) For the engine data of Table 2.5, the number of eigenvalues greater than 1 is three, but the three account for only 76% of the variance. It requires five eigenvalues to reach 94%. The same plot also indicates four eigenvalues.

(b)

	Principal Component Loadings			Various Rotated Loadings			Communality
	F_1	F_2	F_3	F_1	F_2	F_3	
Variables							
y_1	.838	.461	.478	-.080	.834	.170	.728
y_2	-.328	.870	-.182	-.387	.388	.815	.856
y_3	.304	-.334	-.448	.738	-.328	.688	.328
y_4	.724	-.368	-.110	.798	.283	-.181	.878
y_5	-.408	-.414	.640	-.688	-.813	-.324	.766
y_6	.703	.134	.430	.239	.888	-.099	.762
Variance	1.995	1.334	1.053	1.690	1.403	1.231	4.550
Proportion	.288	.210	.179	.268	.238	.211	.768

- (c) The initial communality estimates for the six variables are given by (13.34) as .215, .215, .313, .285, .181, .348. With these substituted for the diagonal of $\mathbf{R}$, the eigenvalues of $\mathbf{R} - \frac{1}{6}\mathbf{1}\mathbf{1}$ are

Eigenvalue	.364	.368	.215	-.025	-.257	-.330
Proportion	.816	.468	.269	-.028	-.195	-.278
Cumulative	.816	1.284	1.499	1.473	1.278	1.000

The principal factor loadings and varimax rotation are as follows:

	Principal Component Loadings			Varimax Rotated Loadings			Communality, $\hat{d}_i^2$
	f_1	f_2	f_3	f_1	f_2	f_3	
Variables							
y_1	.403	.312	.327	.356	.354	.336	.311
y_2	-.306	.269	-.355	-.288	.361	.269	.345
y_3	.343	-.159	-.397	.403	.060	.337	.376
y_4	.339	-.247	-.090	.364	.255	-.094	.381
y_5	-.386	-.248	.338	-.362	-.088	-.417	.385
y_6	.336	.389	.397	.258	.437	.383	.336

(d) The pattern of loadings is similar in parts (b) and (c), and the interpretation of the three factors would be the same.

13.13. Probe word data of Table 3.5

	Principal Component Loadings		Varimax Rotated Loadings		Communality, $\hat{d}_i^2$	Orthogonal Pattern Loadings	
	f_1	f_2	f_1	f_2		f_1	f_2
Variables							
y_1	.417	-.137	.352	.368	.492	.357	.138
y_2	.438	-.136	.361	.371	.493	.363	-.102
y_3	.474	.388	.414	.376	.647	.348	.364
y_4	.438	-.328	.366	.292	.598	.320	-.357
y_5	.362	.347	.344	.368	.479	-.114	.442
Statistics	3.418	.414	3.294	1.756	4.831		
Proportion	.683	.111	.435	.347	.688		

The loadings for y_2 are similar to those for y_4 in all three sets of loadings. The reason for this can be seen in the correlation matrix.

$$\mathbf{R} = \begin{pmatrix} 1.00 & .54 & .76 & .58 & .41 \\ .54 & 1.00 & .55 & .75 & .35 \\ .76 & .55 & 1.00 & .81 & .69 \\ .58 & .75 & .81 & 1.00 & .52 \\ .41 & .35 & .69 & .52 & 1.00 \end{pmatrix}.$$

The correlations of y_2 with y_1 , y_3 , and y_4 are very similar to the correlations of y_4 with y_1 , y_3 , and y_5 .

CHAPTER 14

14.1 Adding and subtracting X and Y in (14.7) (separately), we obtain

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{y}) &= \sum_{j=1}^p (x_j - X) - (y_j - Y) + (X - Y)d^2 \\ &= \sum_{j=1}^p (x_j - X)^2 + \sum_{j=1}^p (y_j - Y)^2 + p(X - Y)^2 \\ &\quad - 2 \sum_{j=1}^p (x_j - X)(y_j - Y). \end{aligned}$$

The other two terms vanish because $\sum_{j=1}^p (x_j - X) = \sum_{j=1}^p (y_j - Y) = 0$. Substituting $v_1^2 = \sum_{j=1}^p (x_j - X)^2$ and $v_2^2 = \sum_{j=1}^p (y_j - Y)^2$ and adding and subtracting $-2\sqrt{v_1^2 v_2^2} = -2v_1 v_2$, we obtain

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{y}) &= v_1^2 + v_2^2 - 2\sqrt{v_1^2 v_2^2} + p(X - Y)^2 + 2v_1 v_2 \\ &\quad - 2\sqrt{v_1^2 v_2^2} \frac{\sum_{j=1}^p (x_j - X)(y_j - Y)}{\sqrt{v_1^2 v_2^2}} \\ &= (v_1 - v_2)^2 + p(X - Y)^2 + 2v_1 v_2 (1 - \sigma_{xy}). \end{aligned}$$

14.2 (a) Since $\mathbf{Z}_{AB} = \sum_{i=1}^{n_{AB}} \mathbf{z}_i / n_{AB}$, we have by (14.10),

$$\begin{aligned} \text{ESS}_{AB} &= \sum_{i=1}^{n_{AB}} (\mathbf{z}_i - \mathbf{P}_{AB})'(\mathbf{z}_i - \mathbf{P}_{AB}) \\ &= \sum_{i=1}^{n_{AB}} \mathbf{z}_i' \mathbf{z}_i - \sum_{i=1}^{n_{AB}} \mathbf{z}_i' \mathbf{P}_{AB} - \sum_{i=1}^{n_{AB}} \mathbf{P}_{AB}' \mathbf{z}_i \\ &\quad + \sum_{i=1}^{n_{AB}} \mathbf{P}_{AB}' \mathbf{P}_{AB} \\ &= \sum_{i=1}^{n_{AB}} \mathbf{z}_i' \mathbf{z}_i - n_{AB} \mathbf{P}_{AB}' \mathbf{P}_{AB} - n_{AB} \mathbf{P}_{AB}' \mathbf{P}_{AB} \\ &\quad + n_{AB} \mathbf{P}_{AB}' \mathbf{P}_{AB} \\ &= \sum_{i=1}^{n_{AB}} \mathbf{z}_i' \mathbf{z}_i - n_{AB} \mathbf{P}_{AB}' \mathbf{P}_{AB}. \end{aligned}$$

Similarly, $SSD_A = \sum_{i=1}^{n_A} x_i^2 p_i = n_A \overline{x^2} p_A$ and $SSD_B = \sum_{i=1}^{n_B} x_i^2 p_i = n_B \overline{x^2} p_B$. Now

$$\begin{aligned} n_{AB} \overline{x^2}_{AB} \overline{p}_{AB} &= (n_A + n_B) \frac{(n_A \overline{x^2}_A + n_A \overline{x^2}_B) (n_A \overline{p}_A + n_A \overline{p}_B)}{n_A + n_B} \\ &= \frac{n_A^2 \overline{x^2}_A \overline{p}_A + n_A n_B \overline{x^2}_B \overline{p}_B + n_A n_A \overline{x^2}_A \overline{p}_A + n_B^2 \overline{x^2}_B \overline{p}_B}{n_A + n_B}. \end{aligned}$$

Thus

$$\begin{aligned} SSD_{AB} - (SSD_A + SSD_B) &= \sum_{i=1}^{n_A} x_i^2 p_i - \sum_{i=1}^{n_B} x_i^2 p_i - \sum_{i=1}^{n_A} x_i^2 p_i \\ &\quad + n_A \overline{x^2}_B \overline{p}_B + n_B \overline{x^2}_A \overline{p}_A - n_{AB} \overline{x^2}_{AB} \overline{p}_{AB} \\ &= n_A \overline{x^2}_A \overline{p}_A + n_B \overline{x^2}_B \overline{p}_B - n_{AB} \overline{x^2}_{AB} \overline{p}_{AB}. \end{aligned}$$

Show that when the right side of (14.15) is expanded, it reduces to this same expression (see Problem 14.58b).

(b) Multiplying out the right side of (14.15), we have

$$\begin{aligned} n_A \overline{x^2}_A \overline{p}_A &= n_A \overline{x^2}_A \overline{p}_{AB} - n_A \overline{x^2}_{AB} \overline{p}_A + n_A \overline{x^2}_{AB} \overline{p}_{AB} + n_A \overline{x^2}_B \overline{p}_B \\ &\quad - n_B \overline{x^2}_B \overline{p}_{AB} - n_B \overline{x^2}_{AB} \overline{p}_B + n_B \overline{x^2}_{AB} \overline{p}_{AB} \\ &= n_A \overline{x^2}_A \overline{p}_A + n_B \overline{x^2}_B \overline{p}_B - 2n_A \overline{x^2}_A + n_B \overline{x^2}_B \overline{p}_{AB} + (n_A + n_B) \overline{x^2}_{AB} \overline{p}_{AB} \\ &= n_A \overline{x^2}_A \overline{p}_A + n_B \overline{x^2}_B \overline{p}_B - 2n_A + n_B \overline{x^2}_{AB} \overline{p}_{AB} + (n_A + n_B) \overline{x^2}_{AB} \overline{p}_{AB} \\ &= n_A \overline{x^2}_A \overline{p}_A + n_B \overline{x^2}_B \overline{p}_B - (n_A + n_B) \overline{x^2}_{AB} \overline{p}_{AB}. \end{aligned}$$

Substitute $\overline{p}_{AB} = (n_A \overline{p}_A + n_B \overline{p}_B) / (n_A + n_B)$.

14.3 (a) Complete linkage. From Table 14.2, we have

$$D(C, AB) = \frac{1}{2} D(C, A) + \frac{1}{2} D(C, B) + \frac{1}{2} |D(C, A) - D(C, B)| \quad (1)$$

If $D(C, A) \leq D(C, B)$, then $|D(C, A) - D(C, B)| = D(C, A) - D(C, B)$, and equation (1) becomes $D(C, AB) = D(C, A)$. If $D(C, A) > D(C, B)$, then $|D(C, A) - D(C, B)| = D(C, B) - D(C, A)$ and equation (1) becomes $D(C, AB) = D(C, B)$. Thus equation (1) can be written as $D(C, AB) = \max\{D(C, A), D(C, B)\}$, which is equivalent to (14.5), the definition of distance for the complete linkage method.

(b) Average linkage. From Table 14.2, we have

$$D(C, AB) = \frac{n_A}{n_A + n_B} D(C, A) + \frac{n_B}{n_A + n_B} D(C, B). \quad (2)$$

By (14.10), equation (2) can be written as

$$\begin{aligned} D(C, AB) &= \frac{n_A}{n_A + n_B} \cdot \frac{1}{n_C n_A} \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} d(\mathbf{x}_i, \mathbf{z}_j) \\ &\quad + \frac{n_B}{n_A + n_B} \cdot \frac{1}{n_C n_B} \sum_{i=1}^{n_C} \sum_{j=1}^{n_B} d(\mathbf{x}_i, \mathbf{z}_j) \\ &= \frac{1}{n_C(n_A + n_B)} \sum_{i=1}^{n_C} \left[\sum_{j=1}^{n_A} d(\mathbf{x}_i, \mathbf{z}_j) + \sum_{j=1}^{n_B} d(\mathbf{x}_i, \mathbf{z}_j) \right] \\ &= \frac{1}{n_C(n_A + n_B)} \sum_{i=1}^{n_C} \sum_{j=1}^{n_A+n_B} d(\mathbf{x}_i, \mathbf{z}_j), \end{aligned}$$

which, by (14.13), is the definition of distance for the average linkage method.

- (e) Substitute $\bar{y}_{AB} = (n_A \bar{y}_A + n_B \bar{y}_B)/(n_A + n_B)$ in the left side of (14.40) in the statement of Problem 14.3(c) and multiply to obtain

$$\begin{aligned} \bar{y}_C \bar{y}_C &= \frac{2n_A \bar{y}'_A \bar{y}_C}{n_A + n_B} + \frac{2n_B n_B \bar{y}'_B \bar{y}_C}{(n_A + n_B)^2} - \frac{2n_A \bar{y}'_A \bar{y}_C}{n_A + n_B} \\ &\quad + \frac{n_A^2 \bar{y}'_A \bar{y}_A}{(n_A + n_B)^2} + \frac{n_B^2 \bar{y}'_B \bar{y}_B}{(n_A + n_B)^2}. \end{aligned}$$

Similarly, multiply on the right side of (14.40) to obtain the same result.

- (d) Using $n_A = n_B$ in $\bar{y}_{AB} = (n_A \bar{y}_A + n_B \bar{y}_B)/(n_A + n_B)$ in (14.12), we obtain $m_{AB} = \frac{1}{2}(\bar{y}_A + \bar{y}_B)$ in (14.42). Then (14.40) [see part (c)] becomes

$$\begin{aligned} (\bar{y}_C - m_{AB})^2 (\bar{y}_C - m_{AB}) &= \frac{1}{2}(\bar{y}_C - \bar{y}_A)^2 (\bar{y}_C - \bar{y}_A) + \frac{1}{2}(\bar{y}_C - \bar{y}_B)^2 (\bar{y}_C - \bar{y}_B) \\ &\quad - \frac{1}{2}(\bar{y}_A - \bar{y}_B)^2 (\bar{y}_A - \bar{y}_B), \end{aligned}$$

which matches the parameter values for the median method in Table 14.2.

- (e) By (14.10),

$$(\bar{y}_A - \bar{y}_B)^2 (\bar{y}_A - \bar{y}_B) = \frac{n_A + n_B}{n_A n_B} C_{AB},$$

and we have analogous expressions for $(\bar{y}_C - \bar{y}_{AB})^2 (\bar{y}_C - \bar{y}_{AB})$, $(\bar{y}_C - \bar{y}_A)^2 (\bar{y}_C - \bar{y}_A)$, and $(\bar{y}_C - \bar{y}_B)^2 (\bar{y}_C - \bar{y}_B)$. Then (14.40) in part (c) becomes

$$\begin{aligned}
\frac{\partial_C' + \partial_A \partial_B'}{\partial_C \partial_A \partial_B'} k_{C,AB} &= \left(\frac{\partial_A}{\partial_A + \partial_B} \right) \left(\frac{\partial_C' + \partial_A}{\partial_C \partial_A} \right) k_{A,B} \\
&\quad + \left(\frac{\partial_B}{\partial_A + \partial_B} \right) \left(\frac{\partial_C' + \partial_B}{\partial_C \partial_B} \right) k_{B,C} \\
&\quad - \left[\frac{\partial_A \partial_B}{(\partial_A + \partial_B)^2} \right] \left(\frac{\partial_C' + \partial_B}{\partial_A \partial_B} \right) k_{A,B} \\
&= \frac{\partial_A + \partial_C}{\partial_C \partial_A \partial_B} k_{A,C} + \frac{\partial_A + \partial_C}{\partial_C \partial_A \partial_B} k_{B,C} - \frac{1}{\partial_A \partial_B} k_{A,B}.
\end{aligned}$$

Substituting $k_{C,AB}$,

14.4 If $r = 0$, then (14.20) becomes

$$\partial_C(C, AB) = \alpha_A \partial_C(C, A) + \alpha_B \partial_C(C, B) + \beta \partial_C(A, B). \quad (1)$$

By (14.25), we have $\partial_C(A, C) > \partial_C(A, B)$ and $\partial_C(B, C) > \partial_C(A, B)$. Thus, replacing $\partial_C(C, A)$ and $\partial_C(C, B)$ in equation (1) by $\partial_C(A, B)$, we obtain

$$\partial_C(C, AB) > \alpha_A \partial_C(A, B) + \alpha_B \partial_C(A, B) + \beta \partial_C(A, B),$$

which is equivalent to (14.20).

$$\begin{aligned}
14.6 \text{ let } \mathbf{r}_+ &= \frac{1}{\partial^2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{r}_{ij} = \frac{1}{\partial^2} \sum_{i=1}^n \sum_{j=1}^n (\partial \mathbf{r}_{ij} + \mathbf{b}) = \frac{1}{\partial^2} \left(\mathbf{A} \sum_{j=1}^n \mathbf{r}_{ij} + \partial \mathbf{b} \right) \\
&= \mathbf{A} \left(\frac{1}{\partial^2} \sum_{j=1}^n \mathbf{r}_{ij} \right) + \mathbf{b} = \mathbf{A} \mathbf{r}_+ + \mathbf{b}.
\end{aligned}$$

Show similarly that $\mathbf{r}_- = \mathbf{A} \mathbf{r}_- + \mathbf{b}$. Then by (14.6), we have

$$\begin{aligned}
\mathbf{H}_+ &= \mathbf{a} \sum_{i=1}^n (\mathbf{r}_{ii} - \mathbf{r}_+) (\mathbf{r}_{ii} - \mathbf{r}_+)' \\
&= \mathbf{a} \sum_{i=1}^n (\mathbf{A} \mathbf{r}_+ + \mathbf{b} - (\mathbf{A} \mathbf{r}_+ + \mathbf{b})) (\mathbf{A} \mathbf{r}_+ + \mathbf{b} - (\mathbf{A} \mathbf{r}_+ + \mathbf{b}))' \\
&= \mathbf{a} \sum_{i=1}^n (\mathbf{A} \mathbf{r}_+ - \mathbf{A} \mathbf{r}_+) (\mathbf{A} \mathbf{r}_+ - \mathbf{A} \mathbf{r}_+)' \\
&= \mathbf{a} \sum_{i=1}^n \mathbf{A} (\mathbf{r}_{ii} - \mathbf{r}_+) (\mathbf{r}_{ii} - \mathbf{r}_+)' \quad \text{by (2.27)} \\
&= \mathbf{a} \mathbf{A} \left[\sum_{i=1}^n (\mathbf{r}_{ii} - \mathbf{r}_+) (\mathbf{r}_{ii} - \mathbf{r}_+)' \right] \mathbf{A}' \quad \text{by (2.40)} \\
&= \mathbf{A} \mathbf{H}_+ \mathbf{A}'.
\end{aligned}$$

Show similarly that $\mathbf{H}_- = \mathbf{A} \mathbf{H}_- \mathbf{A}'$.

- (d) $w(\mathbf{E}_0) = w(\mathbf{A}\mathbf{E}_0\mathbf{A}') = w(\mathbf{A}'\mathbf{A}\mathbf{E}_0) \neq w(\mathbf{E}_0)$
 (e) $\|\mathbf{E}_0\| = \|\mathbf{A}\mathbf{E}_0\mathbf{A}'\| = \|\mathbf{A}\| \|\mathbf{E}_0\| \|\mathbf{A}'\| = \|\mathbf{A}\|^2 \|\mathbf{E}_0\| = c \|\mathbf{E}_0\|$, where $c > 1$. Thus minimizing $\|\mathbf{E}_0\|$ is equivalent to minimizing $\|\mathbf{E}_0\|_F$.

$$\begin{aligned} \text{(d)} \quad w(\mathbf{E}_0^{-1}\mathbf{H}_0) &= w\{\mathbf{A}\mathbf{E}_0\mathbf{A}'^{-1}(\mathbf{A}\mathbf{E}_0\mathbf{A}')\} \\ &= w\{\mathbf{A}'^{-1}\mathbf{H}_0^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{E}_0\mathbf{A}'\} \\ &= w\{\mathbf{A}'^{-1}\mathbf{H}_0^{-1}\mathbf{H}_0\mathbf{A}'\} \\ &= w\{\mathbf{A}'(\mathbf{A}')^{-1}\mathbf{E}_0^{-1}\mathbf{H}_0\} \\ &= w(\mathbf{E}_0^{-1}\mathbf{H}_0) \end{aligned}$$

- 14.6 There are p parameters in each μ_{ij} , $\frac{1}{2}p(p+1)$ unique parameters in each Σ_{ij} , and $p+1$ unique parameters α_i . Thus the total number is

$$\begin{aligned} xp + g\left[\frac{1}{2}p(p+1)\right] + x + 1 &= g\left[p + \frac{1}{2}p(p+1) + 1\right] + 1 \\ &= \frac{1}{2}g(2p + p^2 + p + 2) + 1 \\ &= \frac{1}{2}g(p^2 + p + 2) + 1 \\ &= \frac{1}{2}g(p+1)(p+2) + 1. \end{aligned}$$

- 14.7 (a) The two-cluster solution from single linkage puts boy No. 20 in one cluster and the other 19 boys in the other cluster.

(b), (c), and (d) Based on the change in distance, average linkage and the other cluster solutions in parts (c) and (d) clearly indicate two clusters. These solutions generally agree and also correspond to a division into two groups seen in the first principal component in Figure 12.5. The separation of the three apparent outliers from the other 17 observations is less pronounced in the cluster analysis than in Figure 12.5. Note that the scale of the second component in Figure 12.5 is much larger than that of the first component, so the separation of points 8, 12, and 20 from the rest is not as large as it appears in the figure. Of the methods in parts (b), (c), and (d), only flexible beta with $\beta = -.50$ and $-.75$ place points 9, 12, and 20 together in one cluster. All others place 9 and 12 in one of the clusters and 20 in the other.

- 14.8 (a) The distance between centroids of the two clusters is $\sqrt{2094.9} = 54.1$.
 (b) From the dendrogram produced by the average linkage method, the largest change in distance corresponds to a two-cluster solution.
 (c) The discriminant function completely separates the two clusters, with no overlap.

- 14.9 (a) Observation 21 seems to be an outlier, because it forms its own cluster in both the single linkage and average linkage methods. The cluster consisting of observations 2, 23, 24, 26, and 30 is the same in all six methods.
 (b) The discriminant function completely separates the two clusters, with no overlap.

- 14.10 (a) The following five clusters were found using as seeds the five observations that are mutually farthest apart.

Cluster	1	2	3	4	5
Observation(s)	9, 15, 16, 18, 19	1, 2, 3, 4, 5, 17	6, 7, 8, 20	10, 11, 12, 13	14

The plot of the first two discriminant functions, observation 14 is relatively far removed from the rest. Clusters 1, 2, and 3 are somewhat closer to each other.

- (b) The following five clusters were found using as seeds the first five observations.

Cluster	1	2	3	4	5
Observation(s)	1, 3, 4	2	5, 17, 18, 19	6, 7, 8, 15, 16, 20	9, 10, 11 12, 13, 14

The plot of the first two discriminant functions shows a pattern different from that in part (a).

- (c) The following five clusters were found using as seeds the centroids of the five-cluster solution resulting from Ward's method.

Cluster	1	2	3	4	5
Observation(s)	6, 7, 8, 15, 16, 20	5, 9, 17, 18, 19	10, 11, 12, 13	1, 2, 3, 4	14

The plot of the first two discriminant functions shows a pattern similar to that found in part (a), with observation 14 isolated. The dendrogram shows that Ward's method gives the same five-cluster solution as the k -means result.

- (d) The following five clusters were found using the k -means method with seeds equal to the centroids of the five clusters from average linkage.

Cluster	1	2	3	4	5
Observation(s)	6, 7, 8, 15, 16, 20	1, 2, 3, 4, 5, 17, 18, 19	10, 11, 12, 13	9	14

The plot of the first two discriminant functions shows a pattern somewhat similar to that in part (a).

In the dendrogram for average linkage, observations 9 and 14 are isolated clusters in the five-cluster solution, which is identical to the five-cluster solution using k -means clustering with these seeds.

- (d) Observation 14 does not appear as an outlier in the plot of the first two principal components, but it does show up as an outlier in the plot of the second and third components. The solutions found in parts (a) and (c) seem to agree most with the principal component plots. This suggests that a different number of initial cluster seeds be used.
- (e) The two clustering solutions are identical. The results are given next.

Cluster	1	2	3
Observations	6, 7, 8, 15, 16, 20	9, 10, 11, 12, 13, 14	1, 2, 3, 4, 5, 17, 18, 19

- (g) The clustering solution is identical to that found in part (f), which indicates that the three-cluster solution is appropriate.

14.11 The number of clusters obtained from the indicated combinations of k and r are shown in the following table. Note that for each pair of values of k and r , the value of r was increased, if necessary, for each point until k points were included in the sphere.

k/r	2	4	6	8	10	12	14	16	18	20
2	10	18	19	19	8	6	4	3	3	2
3	3	3	3	3	3	3	2	2	2	2
4	2	2	2	2	2	2	2	2	2	2
5	1	1	1	1	1	1	1	1	1	1

The maximum value of k that yields a two-cluster solution is 4.

14.12 (a) The number of clusters obtained from the initial combinations of k and r are shown in the following table. The value of r was variable, as noted in Problem 14.11.

k/r	2	4	6	8	10	12	14	16	18	20
1	9	9	9	9	9	9	9	9	9	9
3	3	3	3	2	2	2	2	2	2	2
4	1	1	1	1	1	1	1	1	1	1

- (b) The plot of the first two discriminant functions for $k = 2$ and $r = 1$ shows the three clusters to be well separated.
- (c) The plot of the first two principal components shows the same groupings as in the plot in part (b).

- (d) The plot of the discriminant function shows wide separation of the two clusters, which do not overlap. The three-cluster solution found in part (b) is given next:

Cluster 1	Cluster 2	Cluster 3
Hopson	Rowland	Cambridge
Winkley	Torrington	Corville Park
Wymondleigh	Howley	
Spurlock	Beale-Hopson	
James-Bostington		
Wye		

The two-cluster solution found in part (c) merges clusters 2 and 3 of part (b).

CHAPTER 18

18.1

$$\begin{aligned} \mathbf{B} &= \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \mathbf{A} \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \\ &= \mathbf{A} - \frac{1}{n} \mathbf{A} \mathbf{J} - \frac{1}{n} \mathbf{J} \mathbf{A} + \frac{1}{n^2} \mathbf{J} \mathbf{A} \mathbf{J} \end{aligned} \quad (1)$$

By (2.30),

$$\frac{1}{n} \mathbf{A} \mathbf{J} = \frac{1}{n} \begin{pmatrix} \sum_j a_{1j} \\ \sum_j a_{2j} \\ \vdots \\ \sum_j a_{nj} \end{pmatrix} = \begin{pmatrix} \bar{a}_{1\cdot} \\ \bar{a}_{2\cdot} \\ \vdots \\ \bar{a}_{n\cdot} \end{pmatrix}. \quad (2)$$

Hence,

$$\begin{aligned} \frac{1}{n} \mathbf{A} \mathbf{J} &= \frac{1}{n} [\mathbf{A} \mathbf{J}_1 \mathbf{J}_2 \cdots \mathbf{J}_n] = \left(\frac{1}{n} \mathbf{A} \mathbf{J}_1, \dots, \frac{1}{n} \mathbf{A} \mathbf{J}_n \right) \\ &= \begin{pmatrix} \bar{a}_{1\cdot} & \cdots & \bar{a}_{1n} \\ \bar{a}_{2\cdot} & \cdots & \bar{a}_{2n} \\ \vdots & & \vdots \\ \bar{a}_{n\cdot} & \cdots & \bar{a}_{nn} \end{pmatrix}. \end{aligned}$$

Show that

$$\frac{1}{n}\mathbf{J}\mathbf{A} = \begin{pmatrix} \bar{a}_1 & \bar{a}_1 & \cdots & \bar{a}_n \\ \bar{a}_1 & \bar{a}_1 & \cdots & \bar{a}_n \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_1 & \bar{a}_1 & \cdots & \bar{a}_n \end{pmatrix}.$$

Using equation (2), we obtain

$$\frac{1}{n^2}\mathbf{J}\mathbf{A}\mathbf{J} = \frac{1}{n^2}(1, 1, \dots, 1) \begin{pmatrix} \sum_j a_{1j} \\ \sum_j a_{2j} \\ \vdots \\ \sum_j a_{nj} \end{pmatrix} = \frac{1}{n^2} \sum_j a_{jj} = \bar{a}_1.$$

By (3)-(5),

$$\frac{1}{n^2}\mathbf{J}\mathbf{A}\mathbf{J} = \frac{1}{n^2} \begin{pmatrix} \mathbf{J}\mathbf{A}\mathbf{J} & \cdots & \mathbf{J}\mathbf{A}\mathbf{J} \\ \vdots & & \vdots \\ \mathbf{J}\mathbf{A}\mathbf{J} & \cdots & \mathbf{J}\mathbf{A}\mathbf{J} \end{pmatrix} = \begin{pmatrix} \bar{a}_1 & \cdots & \bar{a}_1 \\ \vdots & & \vdots \\ \bar{a}_1 & \cdots & \bar{a}_1 \end{pmatrix}.$$

Hence the i th element of equation (1) is $b_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}_1$.

- 18.2 (a)** (Sobel 1984, pp. 236-237) The elements of $\mathbf{B} = (b_{ij})$ are defined as $b_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}_1$, where $\bar{a}_j = -\frac{1}{n}\sum_i a_{ij}$. Thus

$$\begin{aligned} -2a_{ij} &= a_{ij}^2 - (a_i - x_j)(a_i + x_j) \\ &= x_i^2 + x_j^2 - 2x_i x_j. \end{aligned}$$

Then

$$\begin{aligned} -2\bar{a}_i &= \frac{1}{n} \sum_{j=1}^n (-2a_{ij}) = \frac{1}{n} \sum_j (x_i^2 + x_j^2 - 2x_i x_j) \\ &= x_i^2 + \frac{1}{n} \sum_j x_j^2 - \frac{2}{n} x_i \left(\sum_j x_j \right) \\ &= x_i^2 + \frac{1}{n} \sum_j x_j^2 - 2\bar{x}x_i. \end{aligned}$$

Similarly, show that

$$\begin{aligned} -2\bar{a}_j &= x_j^2 + \frac{1}{n} \sum_i x_i^2 - 2\bar{x}x_j, \\ -2\bar{x} &= \frac{2}{n} \sum_i x_i^2 - 2\bar{x}\bar{x}. \end{aligned}$$

Solve for $a_1, \bar{a}_1, \bar{a}_2$, and $\bar{a}_3$, and substitute into $b_2 = a_1 - \bar{a}_1 - \bar{a}_2 + \bar{a}_3$ to obtain $b_2 = \mathcal{A}_1 \mathcal{A}_2 - \mathcal{A}_1 \bar{\mathcal{A}}_2 - \bar{\mathcal{A}}_2 \mathcal{A}_2 + \bar{\mathcal{A}}_2 \bar{\mathcal{A}}_2$, which can be factored as $b_2 = (\mathcal{A}_1 - \bar{\mathcal{A}}_2)(\mathcal{A}_2 - \bar{\mathcal{A}}_2)$. Hence

$$\begin{aligned} \mathbf{B} &= \begin{pmatrix} (\mathcal{A}_1 - \bar{\mathcal{A}}_2)(\mathcal{A}_1 - \bar{\mathcal{A}}_1) & \cdots & (\mathcal{A}_1 - \bar{\mathcal{A}}_2)(\mathcal{A}_n - \bar{\mathcal{A}}_n) \\ \vdots & & \vdots \\ (\mathcal{A}_n - \bar{\mathcal{A}}_2)(\mathcal{A}_1 - \bar{\mathcal{A}}_1) & \cdots & (\mathcal{A}_n - \bar{\mathcal{A}}_2)(\mathcal{A}_n - \bar{\mathcal{A}}_n) \end{pmatrix} \\ &= \begin{pmatrix} (\mathcal{A}_1 - \bar{\mathcal{A}}_2) \\ \vdots \\ (\mathcal{A}_n - \bar{\mathcal{A}}_2) \end{pmatrix} (\mathcal{A}_1 - \bar{\mathcal{A}}_1, \dots, \mathcal{A}_n - \bar{\mathcal{A}}_n) \\ &= \mathcal{X}_2 \mathcal{X}_1^T \quad \text{[see (3.1.2)].} \end{aligned}$$

Thus $\mathbf{B}$ is positive semidefinite (see Section 2.7).

- (ii) If $\mathbf{B}$ is positive semidefinite of rank q , then by (2.10.6) and Section 2.13.4, $\mathbf{B}$ can be expressed in the form $\mathbf{B} = \mathbf{V} \mathbf{A} \mathbf{V}^T$, where $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q)$ is an orthogonal matrix of eigenvectors of $\mathbf{B}$, and $\mathbf{A}$ is a diagonal matrix of eigenvalues, q of which are positive, with the rest equal to zero. Letting $\mathbf{A}_1$ be the $q \times q$ upper-left-hand block of $\mathbf{A}$ with positive eigenvalues and $\mathbf{V}_1 = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q)$ be the $n \times q$ matrix with the corresponding eigenvectors, we can write $\mathbf{B} = \mathbf{V}_1 \mathbf{A}_1 \mathbf{V}_1^T$ as

$$\begin{aligned} \mathbf{B} &= (\mathbf{V}_1, \mathbf{V}_2) \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{pmatrix} \\ &= \mathbf{V}_1 \mathbf{A}_1 \mathbf{V}_1^T = \mathbf{V}_1 \mathbf{A}_1^{1/2} \mathbf{A}_1^{1/2} \mathbf{V}_1^T \\ &= \mathbf{Z} \mathbf{Z}^T, \end{aligned} \tag{3}$$

where the $n \times q$ matrix $\mathbf{Z}$ is

$$\begin{aligned} \mathbf{Z} &= \mathbf{V}_1 \mathbf{A}_1^{1/2} = (\sqrt{\lambda_1} \mathbf{v}_1, \sqrt{\lambda_2} \mathbf{v}_2, \dots, \sqrt{\lambda_q} \mathbf{v}_q) \\ &= \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}, \end{aligned}$$

To show that $(\mathbf{z}_1 - \mathbf{z}_2)^T(\mathbf{z}_1 - \mathbf{z}_2)$ is equal to $\mathbf{B}_{22}^2$, we can proceed as follows:

$$(\mathbf{z}_1 - \mathbf{z}_2)^T(\mathbf{z}_1 - \mathbf{z}_2) = \mathbf{z}_1^T \mathbf{z}_1 + \mathbf{z}_2^T \mathbf{z}_2 - 2\mathbf{z}_1^T \mathbf{z}_2. \tag{4}$$

By equation (1), we have

$$\begin{aligned}\mathbf{B} &= \mathbf{Z}\mathbf{Z}^T = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} (z_1, z_2, \dots, z_n) \\ &= \begin{pmatrix} z_1 z_1 & z_1 z_2 & \cdots & z_1 z_n \\ z_2 z_1 & z_2 z_2 & \cdots & z_2 z_n \\ \vdots & \vdots & \ddots & \vdots \\ z_n z_1 & z_n z_2 & \cdots & z_n z_n \end{pmatrix}.\end{aligned}$$

Hence equation (2) becomes

$$\begin{aligned}(a_i - a_j)^T (a_i - a_j) &= a_i^T a_i + a_j^T a_j - 2a_i^T a_j \\ &= a_{ii} + a_{jj} - 2a_{ij}.\end{aligned}\quad (3)$$

Show that substituting $a_{ij} = a_{ji} - \delta_{ij} - \delta_{ji} + \delta_i$ into equation (3) leads to

$$(a_i - a_j)^T (a_i - a_j) = a_{ii} + a_{jj} - 2a_{ij} + \delta_{ii} - \delta_{ji} + \delta_j - \delta_{ji}.$$

Show that the symmetry of $\mathbf{A}$ implies $\delta_{ii} = \delta_{jj}$ and $\delta_{ji} = \delta_{ij}$. Hence,

$$(a_i - a_j)^T (a_i - a_j) = a_{ii} + a_{jj} - 2a_{ij} = -2a_{ij} = -2a_{ji}^T.$$

Since $a_{ij} = -\frac{1}{2}A_{ij}^T = i$ and $-2a_{ij} = A_{ij}^T$,

$$15.3 \text{ (a) } x = \sum_{j=1}^n p_j x_j = \sum_{j=1}^n p_j \left(\frac{p_1}{p_j}, \frac{p_2}{p_j}, \dots, \frac{p_n}{p_j} \right)^T$$

$$= \sum_{j=1}^n (p_{1j}, p_{2j}, \dots, p_{nj})^T \quad \text{by (1.8)(3)}$$

$$= (\sum_{j=1}^n p_{1j}, \sum_{j=1}^n p_{2j}, \dots, \sum_{j=1}^n p_{nj})^T$$

$$= (p_{1*}, p_{2*}, \dots, p_{n*})^T$$

$$\text{(b) } x' = - \sum_{i=1}^n p_i x_i' = - \sum_{i=1}^n p_i \left(\frac{p_1}{p_i}, \frac{p_2}{p_i}, \dots, \frac{p_n}{p_i} \right)^T$$

$$= - \sum_{i=1}^n (p_{i1}, p_{i2}, \dots, p_{in}) \quad \text{by (1.8)(1)}$$

$$= (-\sum_{i=1}^n p_{i1}, -\sum_{i=1}^n p_{i2}, \dots, -\sum_{i=1}^n p_{in})$$

$$= (-p_{*1}, -p_{*2}, \dots, -p_{*n})$$

$$15.4 \text{ (a) } \sum_{j=1}^n p_j = \sum_{j=1}^n \sum_{i=1}^n p_{ij} = \sum_{i,j} a_{ij}/n = n/n = 1,$$

$$x' \mathbf{1} = \sum_{j=1}^n p_j = \sum_{j=1}^n a_{j*}/n = \sum_{j=1}^n \sum_{i=1}^n a_{ij}/n = n/n$$

15.5 By (15.16), (15.24), and (15.18), $p_0 = n_0/n$, $p_i = n_i/n$, and $p_j = n_j/n$. Substituting these into (15.25), we obtain

$$\begin{aligned} x^2 &= \sum_i n \frac{\left(\frac{n_i}{n} - \frac{n_0^2}{n^2}\right)^2}{\frac{n_0^2}{n^2}} = \sum_i n \frac{\left[\frac{1}{n}(n_0 - \frac{n_0^2}{n})\right]^2}{\frac{n_0^2}{n^2}} \\ &= \sum_i n \frac{\frac{1}{n^2}(n_0 - \frac{n_0^2}{n})^2}{\frac{n_0^2}{n^2}} = \sum_i \frac{(n_0 - \frac{n_0^2}{n})^2}{n_0^2}. \end{aligned}$$

15.6 (a) Multiplying numerator and denominator of (15.25) by p_i , we obtain

$$\begin{aligned} x^2 &= \sum_i n \sum_j \frac{p_i}{p_i p_j} (p_0 - p_i p_j)^2 \\ &= \sum_i n p_i \sum_j \frac{1}{p_j} \left[\frac{1}{p_i} (p_0 - p_i p_j) \right]^2 \\ &= \sum_i n p_i \sum_j \left(\frac{p_0}{p_i} - p_j \right)^2 p_j. \end{aligned}$$

15.7 (a) By (15.26), (15.19), (15.12), and (15.18), we obtain

$$\begin{aligned} x^2 &= \sum_i n p_i (b_i - c)(B_i^{-1}(b_i - c)) \\ &= \sum_i n p_i \left(\frac{B_0}{p_i} - p_1, \dots, \frac{B_0}{p_i} - p_k \right) \begin{pmatrix} p_1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & p_k \end{pmatrix}^{-1} \begin{pmatrix} \frac{B_0}{p_i} - p_1 \\ \vdots \\ \frac{B_0}{p_i} - p_k \end{pmatrix} \\ &= \sum_i n p_i \left(\frac{\frac{B_0}{p_i} - p_1}{p_1}, \dots, \frac{\frac{B_0}{p_i} - p_k}{p_k} \right) \begin{pmatrix} \frac{B_0}{p_i} - p_1 \\ \vdots \\ \frac{B_0}{p_i} - p_k \end{pmatrix}. \end{aligned}$$

15.8 (a) By (15.19), $\mathbf{r} = B_0^{-1}\mathbf{r}$. Thus $B_0^{-1}\mathbf{r} = B_0^{-1}B_0^{-1}\mathbf{r} = B_0^{-2}\mathbf{r}$ by (15.15). By (15.13), $c_1^2 = 1$, and therefore $B_0 = \mathbf{1}$. Now

$$B_0^{-1}(\mathbf{p} - \mathbf{w}') = B_0^{-1}\mathbf{p} - B_0^{-1}\mathbf{w}' = \mathbf{R} - B_0\mathbf{p}' = \mathbf{R} - \mathbf{p}'.$$

15.9 By (15.49), $\mathbf{q}_i' = \mathbf{q}_i' A$ (ignoring the centering on p_i). Thus the squared Euclidean distance can be written as

$$\begin{aligned} (\mathbf{x}_i - \mathbf{x}_0)'(\mathbf{x}_i - \mathbf{x}_0) &= (\mathbf{q}_i' - \mathbf{q}_0')(\mathbf{x}_i - \mathbf{x}_0) \\ &= (\mathbf{q}_i' A - \mathbf{q}_0' A)(\mathbf{x}_i - \mathbf{x}_0) \end{aligned}$$

$$\begin{aligned}
 &= (Y_1 - Y_1'BA)(Y_1 - Y_1') \\
 &= (Y_1 - Y_1')^T(Y_1 - Y_1'),
 \end{aligned}$$

since B is orthogonal.

15.10 (a) From $Y_1Y_1' = UA$ in (15.35), we have $Y_1YA^{-1} = U$. Thus

$$\begin{aligned}
 UU' &= Y_1YA^{-1}(A^{-1})'Y_1' \\
 &= Y_1Y(A^{-1})'(A^{-1})'Y_1'.
 \end{aligned} \quad (1)$$

Since $(A^{-1})^T = \text{diag}(1/\lambda_1^2, 1/\lambda_2^2, \dots, 1/\lambda_p^2)$, where the λ_j^2 's are eigenvalues of Y_1Y_1' , the matrix $(A^{-1})^T$ contains eigenvalues of $(Y_1Y_1')^{-1} = [(n-1)S]^{-1} = S^{-1}/(n-1)$ [see (2.115) and (2.116)]. The matrix Y contains eigenvalues of Y_1Y_1' and thereby of $(Y_1Y_1')^{-1}$ [see Section 2.11.8]. Hence we recognize $Y(A^{-1})^T Y'$ as the spectral decomposition of $(Y_1Y_1')^{-1}$ [see (2.108), (2.115), and (2.116)]. Therefore, equation (1) can be written as

$$\begin{aligned}
 UU' &= Y_1Y(A^{-1})^T Y' = Y_1(Y_1Y_1')^{-1}Y_1' \\
 &= Y_1S^{-1}Y_1'/(n-1).
 \end{aligned}$$

(b) If $B = Y_1A$, then $BB' = Y_1AY' = Y_1A^2Y'$. The diagonal matrix A^2 contains the eigenvalues λ_j^2 of the matrix Y_1Y_1' . Thus by (2.115), Y_1A^2Y' is the spectral decomposition of Y_1Y_1' , and

$$BB' = Y_1A^2Y' = Y_1Y_1' = (n-1)S.$$

15.11 By (15.64), (3.63), and (3.64) (ignoring $n-1$ and assuming the y_i 's are centered),

$$\begin{aligned}
 (n_1 - n_1')^T(n_2 - n_2') &= n_1'n_2 + n_2'n_1 - 2n_1'n_2 \\
 &= \sum_1 S^{-1}y_1 + \sum_2 S^{-1}y_2 - 2\sum_1 S^{-1}y_1 \\
 &= (y_1 - y_2)'S^{-1}(y_1 - y_2).
 \end{aligned}$$

15.12 (a) The first ten rows and columns of the matrix B are as follows:

$$\begin{pmatrix}
 12049 & -26851 & -89709 & -20847 & -29135 & -42385 & -12077 & 31751 & 142326 & 82989 \\
 -26851 & 26851 & 17029 & 11438 & 10994 & -14959 & 18149 & -38662 & -14682 & -6188 \\
 -89709 & 17029 & 65970 & 22179 & 21044 & -14085 & 38156 & -26571 & -76891 & -54125 \\
 -20847 & 11438 & 12178 & 27683 & 31808 & -30980 & 80001 & -31181 & -60882 & -52596 \\
 -29135 & 10994 & 54948 & 10108 & 10111 & -29182 & 22565 & -27147 & -13671 & -54687 \\
 -42385 & -14959 & -14085 & -33338 & -34411 & 14791 & -36229 & 27084 & -43871 & 29211 \\
 -12077 & 38156 & 38156 & 10011 & 10011 & -33650 & 36429 & -49582 & -66506 & -58570 \\
 31751 & -38662 & -26571 & -76891 & -54125 & -14682 & -6188 & 12049 & -26851 & -89709 \\
 142326 & -14682 & -6188 & -66506 & -58570 & -54125 & -14682 & 12049 & -26851 & -89709 \\
 82989 & -6188 & -66506 & -58570 & -54125 & -14682 & 12049 & -26851 & -89709 & 82989
 \end{pmatrix}$$

- (d) The first two columns of the matrix $\mathbf{X}$ are given by

City	x_1	x_2	City	x_1	x_2
A	154.1	-19.2	M	381.6	49.5
B	-77.1	35.0	N	21.8	-46.7
C	-298.2	-75.7	O	9.8	30.9
D	-154.9	85.9	P	-375.8	-78.5
E	-189.2	73.2	Q	6.3	17.1
F	126.8	26.9	R	167.8	-88.8
G	-228.8	-149.6	S	-382.3	-178.2
H	223.8	1.5	T	-38.2	-37.3
I	107.9	44.8	U	-355.2	188.8
J	226.7	56.7	V	-225.5	148.2
K	-33.4	22.5	W	-19.3	-34.5
L	1.1	-79.4			

- (a) The matrix multidimensional scaling plot shows the relative positions of the cities.

- 15.13 (a) The multidimensional scaling plot shows two clusters, one for positive values of the first dimension and one for negative values. The two clusters can be interpreted as comfort (positive values) and discomfort (negative values). Hence, the axis of the first dimension can be interpreted as the level of comfort.

- (b) The dendrogram for Ward's method clearly shows two clusters, the same as in part (a).

- 15.14 (a) The initial configuration of points will vary. One example is as follows

x_1	x_2	x_3	x_4	x_5	x_6
1.438	.589	-1.350	.856	-1.463	1.827
-.288	-1.089	-2.687	.658	.819	1.884
-1.777	-.409	.369	.655	-.358	1.179
.091	.561	1.129	-1.546	.180	-.467
-.060	1.361	.745	1.458	.502	-.894
-.737	-.455	-.545	.255	.648	-1.182
-1.971	-.405	-.461	.078	1.448	.809
-1.865	-.173	.687	-.328	1.688	1.838
-.387	.211	-.206	.261	-.355	-1.115
1.449	-.942	.887	-.322	.833	1.146
-1.889	-.039	-1.752	-.509	-1.136	-.226
1.067	.759	.298	.884	-1.389	-.898

- (b) Accuracy will vary. For the seeds given in part (a), STRESS = .0056.
 (c) Accuracy will vary. The plot of STRESS versus k for one solution showed that two dimensions should be retained. The screeplot (MDS plot) showed that France, Mussolini, and Hitler were close together, as well as Churchill and DeGaulle, and Eisenhower and Truman.
 (d) Accuracy will vary. One solution gave results similar to part (c).

- (d) Answers will vary. One solution showed three dimensions. A plot of two dimensions showed Montreal and France together in the center with the others forming a circle around them of almost equally spaced points.

- (e) Answers will vary. One solution was similar to that in part (c).

- 15.12 (a) The correspondence matrix $\mathbf{P}$ is found by dividing each element of Table 15.14 by $n = 1200$ to obtain the following:

Month	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Total
Jan.	.007	.011	.008	.010	.007	.008	.008	.012	.007	.009	.008	.010	.008
Feb.	.010	.007	.010	.006	.007	.014	.013	.014	.005	.009	.014	.010	.009
Mar.	.008	.011	.007	.009	.011	.008	.007	.008	.007	.002	.000	.007	.004
Apr.	.005	.006	.008	.005	.007	.006	.005	.006	.005	.007	.006	.006	.006
May	.006	.008	.008	.005	.000	.008	.007	.007	.006	.008	.007	.008	.075
Jun.	.014	.014	.008	.005	.010	.014	.005	.005	.006	.007	.004	.014	.010
Jul.	.008	.008	.010	.010	.008	.008	.005	.005	.000	.008	.008	.008	.070
Aug.	.008	.014	.008	.010	.008	.007	.005	.006	.006	.006	.006	.008	.006
Sep.	.005	.008	.010	.008	.000	.008	.005	.006	.000	.005	.006	.008	.005
Oct.	.012	.006	.008	.007	.000	.008	.006	.006	.007	.006	.006	.008	.007
Nov.	.005	.005	.012	.008	.000	.008	.005	.008	.005	.006	.005	.008	.007
Dec.	.000	.014	.007	.009	.011	.008	.007	.007	.008	.008	.008	.008	.000
Total	.050	.054	.064	.064	.050	.060	.046	.060	.055	.054	.054	.064	1.000

- (b) The $\mathbf{R}$ matrix is given by

$$\mathbf{R} = \begin{pmatrix} .07 & .11 & .12 & .11 & .07 & .08 & .11 & .10 & .08 & .08 & .08 & .08 & .08 \\ .12 & .08 & .12 & .07 & .07 & .08 & .08 & .11 & .10 & .08 & .08 & .08 & .08 \\ .08 & .08 & .07 & .10 & .08 & .08 & .07 & .08 & .11 & .08 & .08 & .08 & .08 \\ .08 & .06 & .15 & .08 & .15 & .08 & .06 & .07 & .10 & .08 & .12 & .08 & .08 \\ .10 & .11 & .08 & .10 & .07 & .07 & .08 & .08 & .07 & .08 & .08 & .07 & .07 \\ .04 & .06 & .08 & .11 & .10 & .07 & .12 & .14 & .08 & .08 & .11 & .08 & .08 \\ .08 & .04 & .06 & .08 & .14 & .08 & .08 & .08 & .11 & .08 & .10 & .08 & .08 \\ .06 & .08 & .07 & .10 & .10 & .07 & .08 & .07 & .14 & .11 & .08 & .07 & .07 \\ .07 & .08 & .04 & .06 & .08 & .08 & .15 & .11 & .08 & .08 & .08 & .11 & .08 \\ .08 & .08 & .05 & .08 & .06 & .06 & .08 & .14 & .10 & .08 & .08 & .08 & .08 \\ .08 & .07 & .08 & .07 & .14 & .11 & .08 & .10 & .08 & .08 & .07 & .08 & .08 \\ .08 & .08 & .07 & .11 & .07 & .08 & .10 & .10 & .07 & .08 & .08 & .11 & .08 \end{pmatrix},$$

- and the $\mathbf{C}$ matrix is given by

$$\mathbf{C} = \begin{pmatrix} .07 & .11 & .12 & .08 & .06 & .05 & .10 & .08 & .08 & .08 & .08 & .08 & .08 \\ .12 & .08 & .12 & .06 & .04 & .08 & .08 & .08 & .08 & .08 & .08 & .08 & .08 \\ .10 & .08 & .08 & .10 & .05 & .11 & .07 & .07 & .12 & .11 & .08 & .10 & .10 \\ .07 & .08 & .11 & .06 & .11 & .08 & .06 & .08 & .08 & .08 & .11 & .07 & .07 \\ .11 & .10 & .11 & .10 & .08 & .10 & .10 & .10 & .08 & .11 & .11 & .08 & .08 \\ .04 & .06 & .08 & .08 & .10 & .08 & .10 & .11 & .08 & .08 & .10 & .08 & .08 \\ .07 & .04 & .05 & .04 & .12 & .08 & .08 & .08 & .11 & .07 & .08 & .08 & .08 \\ .08 & .10 & .08 & .10 & .10 & .11 & .08 & .07 & .10 & .10 & .08 & .08 & .08 \\ .08 & .08 & .04 & .08 & .07 & .11 & .11 & .08 & .08 & .08 & .10 & .07 & .07 \\ .08 & .08 & .07 & .07 & .14 & .14 & .08 & .08 & .08 & .07 & .08 & .08 & .08 \\ .10 & .10 & .07 & .10 & .06 & .05 & .10 & .08 & .08 & .08 & .08 & .08 & .12 \end{pmatrix}.$$

- (c) The chi-square statistic is 113.7742 with 121 degrees of freedom, which gives a p -value of .000. The two variables appear to be independent.
- (d) In the correspondence plot, the following associations are seen: (November births, June deaths), (March deaths, April deaths, January births), (September births, February deaths), (August births, April births), (May deaths, September deaths, May births).

- 15.36 (a) The correspondence matrix $\mathbf{P}$ is found by dividing each element of Table 15.17 by 8156 to obtain the following:

Part of County	Burglary	Fire	Vandalism	Total
Oslo area	.048	.068	.115	.231
Mid Norway	.038	.038	.112	.188
North Norway	.085	.040	.154	.280
Total	.171	.146	.491	1.008

- (b) The $\mathbf{B}$ matrix is given by

Part of County	Burglary	Fire	Vandalism
Oslo area	.086	.100	.168
Mid Norway	.12	.126	.192
North Norway	.290	.136	.280

and the $\mathbf{C}$ matrix is

Part of County	Burglary	Fire	Vandalism
Oslo area	.028	.037	.037
Mid Norway	.113	.050	.128
North Norway	.561	.111	.105

- (c) The chi-square statistic is 3802.8 with 4 degrees of freedom, which gives a p -value less than .001. The two variables are dependent.
- (d) In the correspondence plot, North Norway is associated with burglary, Oslo is associated with fire, and Mid Norway is associated with vandalism.

15.17 (a) The first matrix is as follows.

No	1014	0	No	1010	1061	1060	1014	1060	1060	1060	1060	1060	1060
Yes	0	1010	1010	0	0	0	0	0	0	0	0	0	0
High debt	No	1010	1010	0	0	1010	1010	0	1010	1010	1010	1010	1010
Low debt	1010	0	0	1010	0	1010	1010	1010	1010	1010	1010	1010	1010
Medium debt	1010	0	0	0	1010	1010	1010	1010	1010	1010	1010	1010	1010
Race— Other	1010	0	1010	1010	1010	1010	0	1010	1010	1010	1010	1010	1010
White	1010	0	1010	1010	1010	0	1010	1010	1010	1010	1010	1010	1010
Female	1010	0	0	1010	1010	1010	1010	0	1010	1010	1010	1010	1010
Male	1010	1010	1010	1010	1010	0	1010	1010	1010	1010	1010	1010	1010
Nonmember	1010	0	1010	1010	1010	1010	1010	0	1010	1010	1010	1010	1010
Member	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010
10, 20	1010	0	0	1010	1010	1010	1010	0	1010	1010	1010	1010	1010
< 10	1010	0	0	1010	1010	1010	1010	0	1010	1010	1010	1010	1010
≥ 10	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010	1010

(b) The column-coordinates for the plot are given by

Variables	$\bar{x}_i$	$\bar{y}_i$
No	-.050	-.080
Yes	1.015	2.761
High debt	1.010	1.660
Low debt	-.050	-.100
Medium debt	.000	-.060
Race—Other	1.000	-.150
White	-.041	.080
Female	.000	-.070
Male	-.000	.070
Nonmember	-.050	-.080
Member	.000	.010
10, 20	-.000	.010
< 10	-.000	-.080
≥ 10	-.071	.010

(c) Some associations seen in the plot are (yes/no—yes, high debt), (female, nonmember), (medium debt), (member, male).

- 15.88 (a) The two-dimensional coordinates of the observation points and variable points are

Observation Points		
Name	Coordinate 1	Coordinate 2
Albania	14.802	-1.302
Armenia	-5.460	1.548
Belgium	-6.877	-1.409
Bulgaria	26.136	3.838
Czech	3.317	-2.892
Denmark	-12.880	1.078
E. Germany	-4.802	-8.468
Finland	-12.282	11.288
France	-6.345	.872
Germany	9.036	3.020
Hungary	10.888	-2.980
Ireland	-10.887	9.882
Italy	8.888	-1.888
Netherlands	-10.888	2.000
Norway	-10.882	-.877
Poland	2.226	2.888
Portugal	.784	-18.750
Romania	19.873	2.791
Spain	3.433	-13.480
Sweden	-14.842	.728
Switzerland	-8.888	4.888
UK	-8.333	.888
USSR	10.888	4.000
W. Germany	-12.534	-3.050
Yugoslavia	23.742	3.348

Variable Points		
Name	Coordinate 1	Coordinate 2
B.ME.W	-.133	.133
W.ME.W	-.129	.045
ROCK	-.087	.028
WILK	-.423	.808
WMA	-.127	-.282
CRASHOLS	.804	.400
STOCKS	-.067	-.085
NOTES	.114	-.050
PROL.VIO	.020	-.149

In the biplot, the arrows for variables are in short in pass through the points for observations.

- (b) The two-dimensional coordinates of the observation points and variable points are given next.

Observation Points

Name	Coordinate 1	Coordinate 2
Austria	.131	-.349
Austria	-.089	.097
Belgium	-.100	-.099
Belgium	.126	.122
Canada	.094	-.077
Denmark	-.127	.054
E. Germany	-.080	-.100
Finland	-.101	.485
France	-.104	.025
Greece	.148	.112
Hungary	.171	-.087
Ireland	-.194	.196
Italy	.103	-.088
Netherlands	-.193	.079
Norway	-.100	-.000
Poland	.041	.110
Portugal	.013	-.017
Romania	.113	.009
Spain	.071	-.190
Sweden	-.143	.027
Switzerland	-.149	.147
UK	-.123	.026
USSR	.173	.160
W. Germany	-.120	-.120
Yugoslavia	.433	.174

Variable Points

Name	Coordinate 1	Coordinate 2
B. MEAT	-.0106	1.002
W. MEAT	-.7.004	1.179
BEER	-.0108	.000
WELK	-20.000	22.000
PIZZA	-7.750	-7.000
CHEESE	52.545	11.005
STARCH	-4.000	-2.000
MILK	0.000	-1.000
FRUIT	1.200	-4.000

In the biplot, the observation points are tightly clustered around the point (0,0), making them difficult to distinguish, whereas variable points are easily discerned. Red meats, white meats, and milk are highly posi-

tively correlated. These three variables are negatively correlated with *size* and *int. reg.*

- (4) The two-dimensional coordinates of the observation points and variable points are as follows.

Observation Points		
Name	Coordinate 1	Coordinate 2
Albania	1.803	-1.204
Armenia	-1.099	.297
Australia	-1.738	-1.284
Austria	1.443	.637
Czech	.423	-1.492
Denmark	-1.734	.364
E. Germany	-1.027	-1.489
Finland	-1.178	2.167
France	-1.832	.129
Germany	1.177	.582
Hungary	1.283	-1.454
Ireland	-1.338	1.828
Italy	.808	-1.272
Netherlands	-1.331	.489
Norway	-1.489	-1.815
Poland	.323	.376
Portugal	.188	-1.218
Romania	2.441	.497
Spain	.348	-1.812
Sweden	-1.888	.129
Switzerland	-1.181	.768
UK	-1.182	.174
US&C	1.111	.838
W. Germany	-1.738	-1.644
Yugoslavia	1.283	.488

Variable Points		
Name	Coordinate 1	Coordinate 2
K.ABOUT	-1.177	.691
W.ABOUT	-1.012	.226
ES&C	-1.505	.189
WILL	-1.555	4.329
PS&I	-1.002	-1.523
CR&P&L	.6726	2.118
ST&ACT&N	-1.022	-1.398
PA&T&S	.380	-1.385
PR&TT.V&C	.138	-1.882

In the biplot, the variable points and observation points are both well spread. Finland scored high on the milk variable. Yugoslavia and Portugal scored high on the cereal variable. Spain and Portugal scored highest on the fish and fish.rvg variables.

- (d) The biplot from part (c) seems better because the scales on the variables and points are more evenly matched.

14.19 (a) The two-dimensional coordinates of the observation points and variable points are as follows.

Observation Points		
Name	Coordinate 1	Coordinate 2
PGM1	-4.111	-4.762
Italy	1.761	.766
PGM2	4.041	-.288
France	4.392	.614
Yugoslavia	-6.768	5.888
USA	1.401	.791
PGM3	1.094	-1.262

Variable Points		
Name	Coordinate 1	Coordinate 2
KGW	.610	-.054
INTBL	.081	.411
HAPPY	.407	-.406
LIKE	.621	-.019
JUST	.354	.785

In the biplot, the arrows for the variables are too short to pass through the points for observations.

- (b) The two-dimensional coordinates of the observation points and variable points are as follows.

Observation Points		
Name	Coordinate 1	Coordinate 2
PGM1	-.621	-.468
Italy	.176	.111
PGM2	.364	-.080
France	.387	.085
Yugoslavia	-.566	.604
USA	.111	.097
PGM3	.341	-.246

Variable Points		
Name	Coordinate 1	Coordinate 2
KIND	0.545	-0.391
INTEL	1.200	2.097
HAPPY	0.255	-3.013
LIKE	0.621	-0.282
JUST	0.094	0.790

In the biplot, the observation points are tightly clustered around the point (0,0), making them difficult to distinguish, whereas variable points are well spaced. Just and intelligent are highly positively correlated, as are kind and likeable.

- (c) The two-dimensional coordinates of the observation points and variable points are as follows.

Observation Points		
Name	Coordinate 1	Coordinate 2
PSM1	-1.433	-1.764
Butter	.091	.200
PSM2	1.083	-0.217
Father	1.122	.220
Teacher	-1.124	1.859
PSM3	.071	.260
PSM3	.043	-0.092

Variable Points		
Name	Coordinate 1	Coordinate 2
KIND	2.507	-1.449
INTEL	.310	1.113
HAPPY	1.000	-1.230
LIKE	24.02	-1.05
JUST	1.000	2.113

In the biplot, the variable points and observation points are both well spaced. Father, sister, PSM2, and PSM3 all scored high on the kind, likeable, and happy variables, whereas teacher and PSM1 scored negatively on those variables.

- (d) The biplot from part (c) seems better because the scales on the variables and points are more evenly matched.

- 15.26 (a) The two-dimensional coordinates of the observation points and variable points are as follows:

Observation Points		
Name	Coordinate 1	Coordinate 2
1	29.113	-5.632
2	29.107	-7.698
3	28.600	-2.080
4	-20.540	-6.087
5	-20.477	-4.277
6	-20.546	2.015
7	-20.179	7.992
8	-20.780	12.655
9	-20.780	6.379
10	-10.688	-3.779
11	-10.529	-1.897
12	20.185	-16.034
13	18.896	-6.688
14	5.66	3.823
15	77.655	1.89
16	12.659	14.284
17	48.040	9.181
18	10.950	14.884
19	10.958	-1.183
20	-10.477	1.148
21	-20.504	-1.355
22	-24.000	-2.424
23	20.600	-5.880
24	-15.675	-8.714
25	-20.000	.768
26	-18.185	-1.960
27	-11.980	-6.917
28	5.255	3.683

Variable Points		
Name	Coordinate 1	Coordinate 2
North	.526	.125
East	.429	.751
South	.579	-.555
West	.492	-.490

In the biplot, the variable points are tightly grouped and the corresponding arrows do not pass through the points for observations.

- (d4) The two-dimensional coordinates of the observation points and variable points are as follows:

Observation Points		
Name	Coordinate 1	Coordinate 2
1	.000	-.145
2	.150	-.181
3	.150	-.098
4	-.140	-.108
5	-.170	-.118
6	-.200	.058
7	-.170	.100
8	-.150	.045
9	-.182	.218
10	-.200	-.094
11	-.200	-.040
12	.170	-.080
13	.000	-.180
14	.000	.070
15	.070	.000
16	.070	.000
17	.257	.127
18	.200	.074
19	.214	-.029
20	-.181	.020
21	-.140	-.020
22	-.200	-.000
23	.127	-.140
24	-.097	-.180
25	-.140	.000
26	-.097	-.040
27	-.070	-.170
28	.000	.000

Variable Points		
Name	Coordinate 1	Coordinate 2
North	89.774	19.026
East	89.898	89.229
South	84.079	-15.215
West	79.082	-19.054

In the figure, the observation points are tightly clustered around the point $(0,0)$, making them difficult to distinguish, whereas variable points are well spaced. All the variables are positively correlated, with north and west showing the closest relationship.

- (d) The two-dimensional coordinates of the observation points and variable points are as follows:

Observation Points		
Name	Coordinate 1	Coordinate 2
1	0.070	-0.900
2	1.000	-1.200
3	1.000	-0.900
4	-1.000	-1.000
5	-2.200	-0.700
6	-0.600	0.000
7	-0.200	1.200
8	-0.000	1.000
9	-0.000	1.000
10	-0.000	-0.000
11	-0.000	-0.200
12	0.000	-0.500
13	0.000	-1.000
14	0.000	0.000
15	0.100	0.200
16	1.000	0.000
17	0.000	0.000
18	0.070	0.000
19	0.700	-0.100
20	-1.000	0.000
21	-1.000	-0.000
22	-0.000	-0.000
23	1.000	-0.000
24	-1.000	-1.000
25	-1.000	0.000
26	-1.000	-0.000
27	-0.000	-1.000
28	0.000	0.000

Variable Points		
Name	Coordinate 1	Coordinate 2
North	0.700	1.000
East	0.070	0.000
South	0.000	-1.000
West	0.000	-0.000

In the biplot, the variable points and observation points are both well spaced. Tree 18 is associated with east, 17 with north, 1 and 3 with south, and 2 and 29 with west.

- (e) The biplot from part (c) seems better because the scales on the variables and points are more evenly matched.

APPENDIX C

Data Sets and SAS Files

Thousands of files are located on the ftp server of John Wiley & Sons (JW) (Scientific, Technical, and Medical) Division.

All data sets in the books.

SAS command files for all numerical examples.

DOWNLOADING DATA SET FILES FROM FTP SERVER

Please read the message in the *multivariate analysis* area before downloading the files. The message has information about the files and describes how to access them.

The files may be accessed through either a standard ftp program or a Web browser using the ftp protocol. To gain ftp access, type the following at your ftp command prompt:

```
ftp>ftp.wiley.com
```

The files are located in the *multivariate analysis* area of the *published*, *tech*, and *med* directory. If you need further information about downloading the files, you can reach Wiley's technical support line at 212-850-6194.

References

- Anderson, M. R. (1971), *Cluster Analysis for Applications*, New York: Academic Press.
- Anderson, B. (1965), "A Semi-Graphical Method for Analysis of Complex Problems," *Psychometrika*, 3, 287-292.
- Andrews, D. F., and Bortolucci, A. M. (1983), *Java*, New York: Springer-Verlag.
- Bailey, R. J. R. (1977), "Tables of the Bootstrap t Statistic," *Journal of the American Statistical Association*, 72, 469-479.
- Bartolucci, R. S. (1981), "Statistical Power with Group Means as the Unit of Analysis," *Journal of Educational Statistics*, 6, 265-283.
- Bernett, F., and Lewis, T. (1979), *Outliers in Statistical Data*, New York: Wiley.
- Bartlett, M. S. (1937), "Properties of Sufficiency and Statistical Tests," *Proceedings of the Royal Society of London, Ser. A*, 160, 268-282.
- Boss, W. D., York, F. L., and Gooder, H. A. (1958), "Testing for Differences Between Methods of Preparing Fish by Use of Discriminant Function," *Industrial Quality Control*, 14, 5-10.
- Bozay, C. G., Demichels, J. J., Ross, V. E., and McCabe, G. D. (1983), "Assessment of Fisher and Logistic Linear and Quadratic Discrimination Models," *Computational Statistics and Data Analysis*, 1, 267-273.
- Brady, R. M. L. (1969), "Threshold Cluster Analysis," *Bulletin of the International Statistical Institute*, 48 (2), 92-6.
- Brill, G. (1945), "Approximate Methods for Calculating Discriminant Functions," *Psychometrika*, 10(3), 303-317.
- Demichels, J. J., and Reed, E. G. (1971), "Some Investigation and Simulation Studies of Canonical Analysis," *Journal of Statistical Computation and Simulation*.
- Goodman, R. J., and Cook, R. D. (1983), "Outlier" (with comments), *Biometrika*, 70, 555-563.
- Gower, J.-P. (1972), *Correspondence Analysis Handbook*, New York: Marcel Dekker.
- Good, R. D. (1983), "Multivariate Analysis of Variance of Repeated Measurements," in: *Proceedings of Measurement Change*, C. W. Hanks (ed.), Madison, Wis.: University of Wisconsin Press, pp. 85-103.
- Good, R. D. (1975), *Multivariate Statistical Methods in Behavioral Research*, New York: McGraw-Hill.
- Good, R. J. (1981), "A Priori Tests in Repeated Measures Designs: Effects of Neurophobic," *Psychometrika*, 46, 241-255.

- Bartlett, C. S. (1946), "B-Statistic of the Association in Groups of Tests," in *Study in Cases del Professor S. G. Corbini Roma*.
- Bell, G. S. P. (1949), "A General Distribution Theory for a Class of Likelihood Criteria," *Biometrika*, 36, 347-365.
- Bell, G. S. P. (1955), "Problems in the Analysis of Growth and Linear Curves," *Biometrika*, 4, 382-389.
- Bell, G. S. P. (1956), "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. II. The Effect of Inequality of Variances and of Correlation between Errors in the Two-Way Classification," *Annals of Mathematical Statistics*, 28, 484-498.
- Bell, G. S. P., and Tsou, P. Y. (1955), "The Representation of Response Functions: An Example of the Link between the Fitted Surface and the Basic Mechanisms of the System," *Biometrika*, 42, 387-393.
- Brown, B. L., Strong, W. J., and Raudenbush, A. C. (1978), "Derivations of Sensitivity from Speeds: Effects of Manipulations of Aerial Parameters," *The Journal of the Aeronautical Society of America*, 34, 28-33.
- Brown, B. L., Williams, B. N., and Bickel, C. D. (1984), "TRFAC: A General Factor Analysis Program," *Journal of Social, Administrative and Models*, 2, 13-24.
- Brown, T. A., and Kapteina, J. (1979), "The Weighted Nearest Neighbor Rule for Class Dependent Sample Sizes," *IEEE Transactions on Information Theory*, IT-25, 817-819.
- Bryce, G. R. (1985), "Some Observations on the Analysis of Growth-Curves," Paper No. SO-825-8, Brigham Young University, Department of Statistics.
- Buck, S. B. A. (1955), "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use With an Electronic Computer," *Journal of the Royal Statistical Society, Series B*, 17, 305-307.
- Burkitt, R. B. (1971), "On the Use of Canonical Analysis to Test MANOVA Hypotheses," Presented at the Annual Meeting of the American Statistical Association, Washington, D.C., August 1971.
- Casselman, T. (1986), "Estimation of a Multivariate Density," *Annals of the Institute of Statistical Mathematics*, 18, 179-189.
- Cameron, E., and Freeling, L. (1978), "Supplemental Asbestos in the Supportive Treatment of Cancer: Reevaluation of Prolongation of Survival Times in Terminal Human Cancer," *Proceedings of the National Academy of Sciences, USA*, 75, 4556-4557.
- Campbell, N. A. (1985), "Robust Procedures in Multivariate Analysis. I. Robust Covariance Estimation," *Applied Statistics*, 34, 251-257.
- Chambers, J. W., and Kleiner, D. (1982), "Graphical Techniques for Multivariate Data and for Clustering," in *Handbook of Statistics*, Vol. 2, P. R. Krishnaiah and L. N. Kanal (eds.), New York: North-Holland, pp. 309-344.
- Chatterjee, H. (1971), "The Use of Pairs to Represent Points in n -dimensional Space Graphically," *Journal of the American Statistical Association*, 66, 361-366.
- Chickman, Gervais, R., and Raudenbush, G. (1979), "The Condensed Nearest Neighbor Rule Using the Concept of Mutual Nearest Neighborhood," *IEEE Transactions on Information Theory*, IT-25, 488-490.
- Clancy, Stan-Gale (1986), *Applied Correspondence Analysis*, Thousand Oaks, Calif.: Sage Publications.
- Cochran, W. G., and Cox, G. M. (1957), *Experimental Design*, 2nd ed., New York: Wiley.
- Cohen, R. G., Bates, P. R., Mandelkern, G. R., and Hayes, T. F. (1987), "Estimates of Test Bias for Several Procedures Based on Conventional Ratios in the Repeated Measures Design," *Psychometrika*, 52, 555-558.

- Crane, E. M., and Nicewander, W. A. (1976), "Some Symmetric, Invariant Measures of Multivariate Association," *Psychometrika*, **41**, 43-54.
- Croquet, H., Kotz, J., Kold, S., and Tork, Y. S. (1985), "Analysis of Incomplete Multivariate Data from Repeated Measurement Experiments," *Biometrics*, **41**, 809-814.
- Critchley, P. (1987), "Influence in Principal-Components Analysis," *Biometrics*, **43**, 557-566.
- Demaris, M. L., and Hand, D. J. (1988), *Analysis of Repeated Measures*, New York: Chapman & Hall.
- D'Agostino, R. B. (1971), "An Omnibus Test of Normality for Moderate and Large Size Samples," *Biometrika*, **58**, 341-349.
- D'Agostino, R. B. (1973), "Small Sample Probability Points for the D-Test of Normality," *Biometrika*, **60**, 219-223.
- D'Agostino, R. B., and Pearson, E. S. (1973), "Tests for Departure from Normality: Empirical Results for the Distributions of b_1 and b_2 ," *Biometrika*, **60**, 613-622; correction, **61**, 547.
- D'Agostino, R. B., and Tieney, D. L. (1975), "Simulated Probability Points of b_1 for Small Samples," *Biometrika*, **62**, 669-672.
- Deviation, M. L. (1971), "Estimating versus Multivariate Tests in Repeated Measures Experiments," *Psychological Bulletin*, **77**, 446-453.
- Deviation, M. L. (1983), *Multivariate Testing*, New York: Wiley.
- Devlin, A. W. (1978a), "Exact Distributions of Hotelling's Generalized T^2 -Test," *Biometrika*, **65**, 187-189.
- Devlin, A. W. (1978b), "Further Applications of a Differential Equation for Hotelling's Generalized T^2 -Test," *Annals of the Institute of Statistical Mathematics*, **22**, 77-87.
- Devlin, A. W. (1980), "Further Tabulation of Hotelling's Generalized T^2 -Test," *Communications in Statistics—Part B Simulation and Computation*, **9**, 331-336.
- Devlin, B. (1989), "Multivariate Analysis of National Travel Records," *American Statistician*, **43**, 110-115.
- Deming, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.
- Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association*, **76**, 754-763.
- Devlin, J. L., and Wertsch, W. D. (1978), "A Mathematical Comparison of the Stability of an Infinite Family of Algorithms in Clustering Algorithms," *Canadian Journal of Statistics*, **5**, 25-38.
- Dunn, R. L., and Clark, P. L. (1974), "Cluster Analysis: a Survey," *Lecture Notes in Economics and Mathematical Systems*, New York: Springer-Verlag.
- Edwards, D., and Karmali, S. (1980), "The Analysis of Contingency Tables by Graphical Methods," *Biometrika*, **67** (2), 353-363.
- Hosmer, A. S. C. (1977), "Radicals of Powerlaw," *Journal of the Royal Statistical Society, Series A*, **140**, 273-283.
- Hosmer, A. S. C., and Gailly, J. S. (1982), "Estimation of Time-response Curves and Their Confidence Bands," *Biometrics*, **38**, 149-155.
- Frank, R. S. (1987), *Introduction to Optimization Methods and Their Application in Statistics*, London: Chapman & Hall.
- Frank, R. S. (1994), *Cluster Analysis*, Jekell, London: Edward Arnold.

- Feare, E. (1973), "A Bayesian approach to Growth Curves," *Biometrika*, **62**, 89-100.
- Feare, E. (1977), "A Two-Stage Model for Growth Curves which Leads to Rao's Covariance Adjustment Estimation," *Biometrika*, **64**, 141-143.
- Feldner, R. E. (1980), "On Planning Repeated Measures Experiments," *Univ. Report 80-004*, M. Carroll University, Bloomington, Ind.
- Feldner, R. E. (1983), "Repeated Measures: The Effect of a Preliminary Test on the Structure of the Covariance Matrix upon the Alpha Levels of the Tests for Experimental Design Factors," M.A. Thesis, Brigham Young University, Department of Statistics.
- Ferguson, T. S. (1963), "On the Rejection of Outliers," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Los Angeles: University of California Press, pp. 255-267.
- Fisher, R. A. (1925), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, **1**, 175-186.
- Fin, E., and Hodges, J. L. (1961), "Discriminatory Analysis, Nonparametric Discrimination: Combinatory Properties," Report No. 3, Project No. 214-0-004, Statistics for Peace Fund, UNAF School of Statistical Methods.
- Flork, V. F., and Chang, F. C. (1967), "Frequency of Selecting Noise Variables in Subset Regression Analysis: A Simulation Study," *American Statistician*, **40**, 84-86.
- Flury, B., and Rousseeu, H. (1981), "Graphical Representation of Multivariate Data by Means of Asymptotic Plans," *Journal of the American Statistical Association*, **76**, 757-765.
- Flury, B., and Rousseeu, H. (1983), "T² Tests, the Linear Two-Class Discriminant Function, and Their Computation by Linear Regression," *American Statistician*, **38**, 20-24.
- Frost, G. B. (1951), "Stability of Head Form in Man," *Genetics*, **3**, 393-394.
- Fukunaga, I. H., and Takyu, J. W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, **23**, 873-880.
- Furnival, G. M., and Wilson, R. W. (1974), "Regularity of Least Squares Estimation," *Biometrics*, **30**, 499-511.
- Gabriel, K. R. (1968), "Simultaneous Test Procedures in Multivariate Analysis of Variance," *Biometrika*, **55**, 455-504.
- Gabriel, K. R. (1969), "Simultaneous Test Procedures—Some Theory of Multiple Comparisons," *Annals of Mathematics and Statistics*, **40**, 229-250.
- Gabriel, K. R. (1971), "The Biplot Graphical Display of Matrices with application to Principal Component Analysis," *Biometrika*, **58**, 453-467.
- Gates, G. W. (1972), "The Reduced Nearest Neighbor Rule," *IEEE Transactions on Information Theory*, **19**, 434-444.
- Gelman, A. (1980), "Growth Curve Analysis," in *Handbook of Statistics*, Vol. 1, P. B. Krishnaiah (ed.), Amsterdam: North Holland, pp. 89-113.
- Gibney, E. S. (1968), "On Discrimination Using Qualitative Variables," *Journal of the American Statistical Association*, **63**, 1355-1412.
- Gnanadesikan, R. (1987), *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd ed., New York: Wiley.
- Gnanadesikan, R., and Kettenring, J. R. (1972), "Subset Estimation, Residuals, and Outlier Detection with Multivariate Data," *Biometrics*, **28**, 91-118.
- Gorden, A. D. (1989), *Classification*, 2nd ed., London: Chapman & Hall/CRC.
- Gower, J. C. (1965), "Some Distance Properties of Least Squares and Vector Methods used in Multivariate Analysis," *Biometrika*, **52**, 1-10.

- Boxer, J. C., and Hoad, D. I. (1966), *Spoken*, London: Chapman & Hall.
- Daybill, P. A. (1969), *Introduction to Matrices with Applications in Statistics, Reliability, and*, Cardiff: Williams.
- Demaree, M. J. (1984), *Theory and applications of Correspondence Analysis*, London: Academic Press.
- Goodman, S. W., and Granger, S. (1955), "On Methods in the Analysis of Profile Data," *Psychometrika*, **20**, 95-112.
- Granger, I. E., and Allen, D. M. (1969), "Analysis of Growth and True Response Curves," *Biometrics*, **25**, 375-381.
- Gutman, I. (1982), *Linear Models: An Introduction*, New York: Wiley.
- Härdle, J. D. F., Härdle, J., and Kerstan, J. (1978), "Variable Kernel Density Estimation in Discriminant Analysis," in *Computer 1978: Proceedings in Computational Statistics*, G. Buehmann (ed.), Vienna: Physica-Verlag, pp. 178-182.
- Härdle, J. D. F., Härdle, J., and von dem Brook, K. (1974), "A Stepwise Discriminant Analysis Program Using Density Estimation," in *Commun. G. Buehmann, P. Pernold, and L. Schumacher (eds.)*, Vienna: Physica-Verlag, pp. 101-118.
- Härdle, J., Smith, D., and Gill, D. S. (1981), "Selection of Adaptive Subsets in Multivariate Multiple Regression," Presented at the Annual Meeting of the American Statistical Association, August 10-13, Detroit, Mich.
- Hoad, D. J., and Gratchkov, G. G. (1976), "An Edited Condensed Nearest Neighbor Rule," *Information Sciences*, **14**, 171-180.
- Hoad, D. J., Day, F., Lunn, A. B., McCowry, R. J., and Geronzi, R. (eds.) (1984), *A Handbook of Small Data Sets*, London: Chapman & Hall.
- Hornik, R. R. (1976), *Modern Factor Analysis*, 3rd rev. ed., Chicago: University of Chicago Press.
- Hu, P. E. (1985), "The Condensed Nearest Neighbor Rule," *IEEE Transactions on Information Theory*, **IT-32**, 545-555.
- Hurlbert, J. A. (1973a), *Clustering Algorithms*, New York: Wiley.
- Hurlbert, J. A. (1973b), "Finite Graphs for Clustering," *Journal of Statistical Computation and Simulation*, **4**, 197-213.
- Hurvich, C. M. (1987), *Matrix algebra from a Geometric Perspective*, New York: Springer-Verlag.
- Hoskins, D. M. (1980), *Classification of Outliers*, London: Chapman & Hall.
- Hu, P. E. (1985), "Rao's Function Concerning Multivariate Tests of Significance," *Biometrics*, **25**, 461-463.
- Heywood, B. B. (1911), "On the Profile Sequence of Real Numbers," *Proceedings of the Royal Society Series A*, **140**, 480-501.
- Hills, D. G. (1982), "A Simulation and Comparison of Three Criterion Functions Used for Subset Selection in Linear Regression," M.A. Thesis, Brigham Young University, Department of Statistics.
- Hocking, R. R. (1975), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, **32**, 1-50.
- Hosking, H. (1971), "The Generalization of Student's Rule," *Annals of Mathematical Statistics*, **2**, 395-398.
- Hosking, H. (1974), "A Generalized F Test and Measures of Multivariate Dispersion," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, **3**, 23-40.

- Box, G. E. P. (1948), "Notes on Box-Cox's Generalized F," *Annals of Mathematical Statistics*, **9**, 236-240.
- Box, G. E. P. (1959), "Projection Pursuit," *Annals of Statistics*, **13**, 405-429.
- Box, G. E. P. (1973), "The Stability of Some Indices of Relative Variable Contribution to Discriminant Analysis," *Journal of Experimental Education*, **44**, 39-64.
- Box, G. E. P., and Jenkins, G. M. (1971), "Empirical Comparison of Univariate and Multivariate Analysis of Variance Procedures," *Psychological Bulletin*, **76**, 46-51.
- Box, G. E. P. (1978), "Some Approximate Tests in Repeated Measures Designs," *Psychometrika*, **43**, 1383-1389.
- Box, G. E. P., and Jenkins, G. M. (1976), "Conditions Under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F Distributions," *Journal of the American Statistical Association*, **71**, 1540-1546.
- Box, G. E. P., and Jenkins, G. M. (1975a), "Estimation of the Item Contribution to Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs," *Journal of Educational Statistics*, **1**(1), 69-82.
- Box, G. E. P., and Jenkins, G. M. (1975b), "A Class of Linear Spectral Models and Analysis for the Study of Longitudinal Data," *Biometrics*, **31**, 831-849.
- Box, G. E. P., and Jenkins, G. M. (1971), "A Univariate-Like Method of Analyzing Growth Curve Data with Nonuniform Variance-Covariance Structures," *Report No. 71-026-R*, Brigham Young University, Department of Statistics.
- Box, G. E. P. (1965), "Principal Components and Factor Analysis: Part I—Principal Components," *Journal of Quality Technology*, **11**, 551-711.
- Box, G. E. P. (1980), *Statistical Graphics for Exploring Multivariate Data*, Thousand Oaks, Calif.: Sage Publications.
- Box, G. E. P., and Jenkins, G. M. (1968), "The Construction of Bivariate and Non-Bivariate Classification," *Computer Journal*, **11**, 117-134.
- Box, G. E. P., and Jenkins, G. M. (1971), *Mathematical Biometry*, London: Wiley.
- Box, G. E. P. (1967), "Two Case Studies in the Application of Principal Component Analysis," *Applied Statistics*, **16**, 225-236.
- Box, G. E. P. (1962), "Efficiency and Robustness in the Use of Repeated Measurements," *Biometrics*, **18**, 611-624.
- Box, G. E. P. (1969), "A Dynamic Programming Algorithm for Cluster Analysis," *Operations Research*, **17**, 1084-1087.
- Box, G. E. P. (1975), *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods*, New York: Springer-Verlag.
- Box, G. E. P. (1977), "Discarding Variables in a Principal Component Analysis. I: Artificial Data," *Applied Statistics*, **26**, 145-173.
- Box, G. E. P. (1977), "Discarding Variables in a Principal Component Analysis. II: Real Data," *Applied Statistics*, **26**, 21-31.
- Box, G. E. P., and Jenkins, G. M. (1967), "What is Projection Pursuit?" *Journal of the Royal Statistical Society, B39*, 1-36.
- Box, G. E. P. (1976), "On Some Multivariate Statistical Methodology with Applications to Statistics, Psychology, and Mathematical Programming," *Journal of the Industrial Mathematics Society*, **26**, 1-18.
- Box, G. E. P. (1976), "A Second Generation Little 379," *Psychometrika*, **41**, 401-415.

- Kaiser, H. F., and Rice, J. (1974), "Little Jiffy, Math PC" *Educational and Psychological Measurement*, **34**, 111-117.
- Kassambara, K., and Makridakis, S. (1998), *Cluster Analysis*, New York: Wiley.
- Kendall, M., Marascuilo, G. P. F., and Magid, A. R. A. (1981), "The Relationship between Real Time and Optimal Leaf Area in Tomatoes, an Example of Superior Multivariate Analysis of Variance," *Statistical Neericulture*, **15**, 231-245.
- Kerns, R., and Nakh, D. N. (1995), *Applied Multivariate Statistics with SAS Software*, 2nd ed, New York: Wiley.
- Kerns, R., and Nakh, D. N. (2000), *Multivariate Data Reduction and Discrimination with SAS Software*, New York: Wiley.
- Khan, A. I., Marascuilo, T., and Nakh, D. G. (1994), "A Test to Determine Clusters of Multivariate Samples with Approximation," *Journal of Multivariate Analysis*, **51**, 281-289.
- Krishnamoorthy, K. G., Koppes, L. L., and Waller, K. E. (1988), *Applied Regression Analysis and Other Multivariate Methods*, Boston: PWS-KENT.
- Kruskal, J. B., and Hartigan, J. A. (1978), "Representing Points in Many Dimensions by Trees and Clusters," *Journal of the American Statistical Association*, **73**, 280-285.
- Kruskal, C. Y. (1972), *A First Course in Multivariate Analysis*, Published by the author, Glendale, VA.
- Kruskal, C. Y., and Aronson, D. R. (1976a), "Fundamentals of Multivariate Analysis, Part I: Inference about Means," *Journal of Quality Technology*, **1**(2), 128-133.
- Kruskal, C. Y., and Aronson, D. R. (1976b), "Fundamentals of Multivariate Analysis, Part II: Inference about Two Treatments," *Journal of Quality Technology*, **1**(3), 191-204.
- Kruskal, C. Y., and Aronson, D. R. (1976), "Fundamentals of Multivariate Analysis, Part IV: Analysis of Variance for Balanced Experiments," *Journal of Quality Technology*, **2**, 35-40.
- Kruskal, J. B. (1964a), "Multidimensional Scaling by Optimizing Criteria of Fit to a Non-metric Hypothesis," *Psychometrika*, **29**, 1-27.
- Kruskal, J. B. (1964b), "Nonmetric Multidimensional Scaling: A Numerical Method," *Psychometrika*, **29**, 115-129.
- Kruskal, J. B., and Moh, M. (1978), *Multidimensional Scaling*, Beverly Hills: Sage Publications.
- Kruskal, W. J. (1971), "Discrimination and Classification Using Both Binary and Continuous Variables," *Journal of the American Statistical Association*, **76**, 782-793.
- Kruskal, W. J. (1976a), "Canonical Representation of the Location Model for Discrimination or Classification," *Journal of the American Statistical Association*, **71**, 840-848.
- Kruskal, W. J. (1977), "The Performance of Fisher's Linear Discriminant Function Under Non-Optimal Conditions," *Biometrika*, **74**, 191-200.
- Kruskal, W. J. (1978), "Using Linear Transformations to Reduce of Binary and Continuous Variables, with Particular Reference to Linear Discriminant Analysis," *Biometrika*, **65**, 33-38.
- Kruskal, W. J. (1980), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," *Biometrika*, **68**, 405-409.
- Lachenbruch, P. A. (1975), *Discriminant Analysis*, New York: Rabin.
- Lachenbruch, P. A., and Graber, M. (1975), "Discriminant Analysis," *Biometrika*, **62**, 69-82.

- Lawes, G. N., and Williams, W. T. (1965), "Computer Programs for Hierarchical Polythetic Classification," *Computer Journal*, 8, 55-64.
- Lawes, G. N., and Williams, W. T. (1967), "A General Theory of Classification Sorting Strategies for Hierarchical Systems," *Computer Journal*, 9, 373-380.
- Levine, R., and Chaffin, D. (1967), "Computer Use in Allocation (Use of Spectroscopic Analysis through Chemical Analysis)," *Analytical Chemistry*, 39 (5), 426-430.
- Lewis, D. N. (1958), "A Generalization of Fisher's Z-Test," *Biometrika*, 45, 185-187.
- Lee, J. C. (1966), "Prediction and Estimation of Growth Curves with Special Covariance Structures," *Journal of the American Statistical Association*, 61, 402-405.
- Lee, J. C., Chang, T. C., and Krishnamo, P. R. (1977), "Approximations to the Distributions of the Unlabeled Ratio Statistics for Testing Certain Structures on the Covariance Matrices of Real Multivariate Normal Populations," in *Mathematical Analysis*, Vol. 4, P. R. Krishnaiah (ed.), Amsterdam: North-Holland, pp. 105-119.
- Lin, C. C., and Mulliken, G. E. (1960), "A Simple Test for Normality Against Asymmetric Alternatives," *Biometrika*, 47, 443-444.
- Limby, H., Willett, J. T., and Halpern, H. (1967), "Canonical Correlation as a Discriminant Test in a Probabilistic Problem," *Biometrika*, 54, 237-244.
- Löffelgren, D. O., and Quisenberry, C. P. (1965), "A Nonparametric Estimate of a Multivariate Density Function," *Annals of Mathematical Statistics*, 36, 1049-1051.
- Lubliner, A. A. (1962), "On the Use of Discriminant Functions in Taxonomy," *Biometrics*, 18, 455-457.
- MacNaughton-Smith, P., Williams, W. T., Cole, M. R., and Mackay, L. G. (1964), "Cluster-Lattice Analysis," *Nature*, 202, 1036-1038.
- Mahalanobis, P. C. (1946), "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Sciences of India*, 11, 46-51.
- Mallows, C. L. (1966), "Choosing Variables in a Linear Regression: A Graphical Aid," (Paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kans.
- Mallows, C. L. (1973), "Some Comments on C_p ," *Technometrics*, 15, 403-423.
- Maritz, R. F. (1970), "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, 57, 519-520.
- Maritz, R. F. (1974), "Applications of Some Measures of Multivariate Skewness and Kurtosis for Testing Normality and Robustness Studies," *Canadian Journal of Statistics*, 12, 113-128.
- Maritz, R. F., Kent, J. T., and Jeffery, J. M. (1975), *Multivariate Analysis*, New York: Academic Press.
- Maritz, R. F., and Rao, P. V. (1975), "A Modified Tukey Outlier Rule for Multiple Outliers," *Communications in Statistics*, 4, 113-126.
- Mather, A. M., and Kappen, K. J. (1975), "Exact Percentage Points for Testing Independence," *Biometrika*, 62, 555-556.
- Maschke, J. W. (1955), "Significance Test for Sphericity of a Normal N -Variate Distribution," *Annals of Mathematical Statistics*, 11, 204-209.
- Maxwell, S. L., and Stewart, R. D. (1962), "Small Sample Profile Analysis with Many Variables," *Psychological Bulletin*, 68, 776-785.
- McCaule, G. P. (1964), "Unlabeled Statistics," *Technometrics*, 6, 137-146.
- McKay, R. J. (1977), "Variable Selection in Multivariate Regression: An Application of Simultaneous Test Procedures," *Journal of the Royal Statistical Society, Series B*, 39, 374-380.

- McKay, R. J. (1975), "The Adequacy of Statistic Tests in Multivariate Regression," *Psychometrika*, **28**, 475-488.
- Milligan, G. W. (1965), "Classifying Validation Results and Implications for Applied Analysis," *Classifying and Classification* (eds P. Ashby, L. J. Hubert and G. de Serres), World Scientific, Singapore, 341-373.
- Milligan, G. W., and Cooper, M. C. (1983), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, **58**, 1-13.
- Moyns, R. (1977), "Hierarchical Grouping Methods and Stopping Rules: An Evaluation," *Computer Journal*, **20**, 338-343.
- Morrissey, D. P. (1973), "The Analysis of a Single Sample of Repeated Measurements," *Biometrics*, **28**, 55-71.
- Morrissey, D. P. (1983), *Applied Linear-Statistical Methods*, Englewood Cliffs, N.J.: Prentice Hall.
- Morrissey, D. P. (1990), *Multivariate Statistical Methods*, 2nd ed., New York: McGraw-Hill.
- Muirhead, R., and Wilks, G. L. (1963), *Applied Regression and Classical Inference: The Case of the Multivariate Papers*, New York: Springer-Verlag.
- Multicolonial, R. P. (1977), "On the Null Distribution of χ^2 for Samples of Size at Most 25, with Tables," *Biometrika*, **64**, 401-409.
- Mullis, K. E. (1982), "Understanding Classical Classification through the General Linear Model and Principal Components," *American Statistician*, **36**, 167-168.
- Mullis, K. E., and Foxman, B. L. (1984), "Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis," *Computational Statistics and Data Analysis*, **3**, 143-158.
- Myers, R. H. (1986), *Classical and Modern Regression with Applications*, 2nd ed., Boston: PWS-Kent.
- Nagengast, B. N., and Pillai, K. C. S. (1973), "The Distribution of the Significance Test Criterion," *Journal of Multivariate Analysis*, **3**, 228-235.
- Nixon, G. (1955), "Three-dimensional Population Percent," *Journal of Applied Statistics*, **68**, 411-415.
- Olson, C. L. (1974), "Comparative Subsystems of Six Tests in Multivariate Analysis of Variance," *Journal of the American Statistical Association*, **69**, 854-858.
- O'Sullivan, J. B., and Willett, C. M. (1986), "Osteoporosis Test: Feasibility in Pregnant and Non-pregnant Women," *American Journal of Clinical Nutrition*, **38**, 345-350.
- Peters, E. (1963), "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, **30**, 1365-1383.
- Pope, H. J. (1985), "Analysis of Repeated Measures Designs with Changing Covariance in Clinical Trials," *Biometrika*, **72**, 767-768.
- Press, S. J., and Bentley, H. G. (eds.) (1975), *Biometrika Tables for Statisticians*, Volume 3, Cambridge: Cambridge University Press.
- Plotnik, R. P. (1983), "Stepwise Statistic Selection and Maximum Likelihood Estimation of Smoothing Factors of Kernel Functions for Nonparametric Regression Functions Estimated by Different Criteria," *Computing and Statistical Research*, **18**, 46-61.
- Pillai, K. C. S. (1964), "On the Distribution of the Largest of Seven Roots of a Matrix in Multivariate Analysis," *Biometrika*, **51**, 275-276.
- Pillai, K. C. S. (1965), "On the Distribution of the Largest Characteristic Root of a Matrix in Multivariate Analysis," *Biometrika*, **52**, 405-414.

- Press, C. (1986), "An Effective Two-dimensional Projection Pursuit Algorithm," *Communications in Statistics—Simulation and Computation*, **15**, 1343–1354.
- Press, H. G. (1982), "Analysis of Variance and Analysis of Regression with More than One Response," *Proceedings of a Symposium on Applications of Statistics and Computers to Plant and Statistical Problems*, 93–109.
- Press, H. P., and Ray, S. N. (1964), "A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems," *Biometrics*, **20**, 513–526.
- Ray, C. R. (1965), "Tests of Significance in Multivariate Analysis," *Biometrics*, **21**, 35–75.
- Ray, C. R. (1974), "Some Problems Involving Linear Hypotheses in Multivariate Analysis," *Biometrics*, **30**, 49–59.
- Ray, C. R. (1980), "Concomitant Adjustment and Related Problems in Multivariate Analysis," in *Multivariate Analysis, II* (Kotz et al., eds), New York: Academic Press, pp. 87–133.
- Ray, C. R. (1973), *Linear Statistical Inference and Its Applications*, 2nd ed., New York: Wiley.
- Ray, C. R. (1983), "Prediction of Future Observations in Polynomial Growth Curve Models," in *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistical Applications and New Directions*, Calcutta: Indian Statistical Institute, pp. 512–528.
- Ray, C. R. (1987), "Prediction of Future Observations in Growth Curve Models," *Statistical Science*, **2**, 434–473.
- Reenan, G. M., and Miller, R. G. (1978), "An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis," *Diabetologia*, **16**, 17–24.
- Schmitt, G. (1962), "Multivariate Regressor-Measurement in Growth Curve Models with Random Effects Covariance Structure," *Journal of the American Statistical Association*, **57**, 146–159.
- Schmitt, J., Hothorn, J. D. E., and Hermann, J. (1993), "A Statistical Comparison of Linear, Quadratic, and Kernel Discrimination," *Journal of Statistical Computation and Simulation*, **35**, 87–139.
- Sencha, A. C. (1982), "On the Use of Correlations to Interpret Canonical Functions," *Biometrika*, **79**, 385–389.
- Sencha, A. C. (1992), "Bias in Approximate Classification Rates in Separate Discriminant Analysis," *Communications in Statistics—Simulation and Computation*, **21**, 573–595.
- Sencha, A. C. (1992b), "Interpretation of Canonical Discriminant Functions, Canonical Variables, and Principal Components," *American Statistician*, **46**, 217–223.
- Sencha, A. C. (1993), "The Contribution of Individual Variables to Hotelling's T^2 , Wilks' Λ , and F ," *Biometrics*, **49**, 479–489.
- Sencha, A. C. (1995), *Multivariate Statistical Inference and Applications*, New York: Wiley.
- Sencha, A. C. (2000), *Linear Models in Statistics*, New York: Wiley.
- Sencha, A. C., and Larson, E. F. (1985), "Bias in Wilks' Λ in Separate Discriminant Analysis," *Communications*, **22**, 345–350.
- Sencha, A. C., and Pua, E. C. (1985), "Inflation of R^2 in Step-Subset Regression," *Psychometrika*, **50**, 49–55.
- Sencha, A. C., and Soth, D. T. (1998), "Assessing the Contribution of Individual Variables Following Rejection of a Multivariate Hypothesis," *Communications in Statistics—Simulation and Computation*, **27** (2), 133–158.
- Sencha, A. C., Williams, R. A., and Young, J. R. (1978), "A Discriminant Analysis of Four Levels of Teacher Competence," *Journal of Experimental Education*, **48**, 46–51.

- Ripley, B. D. (1989), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.
- Roberts, P., and Handcock, C. (1970), "A Unifying Test for Linear Multivariate Analysis: The RSP-Coefficient," *Journal of the Royal Statistical Society, Series C*, **20**, 187-203.
- Rothrock, R. (1961), "Classical Forms and Tests of Hypotheses, Part I: Multivariate Mixed Linear Models," *Statistics Neerlandsche*, **16**, 73-80.
- Roy, J. C., Krishnaiah, H. J., and Mardia, J. I. (1971), "Analysis of Repeated Measurements," *British Journal of Mathematical and Statistical Psychology*, **22**, 205-220.
- Roy, W. G., and Nagao, T. I. (1970), "A Finite Sample Distribution Free Performance Census for Local Discrimination Rules," *Annals of Statistics*, **4**, 505-534.
- Rumsey, A. K., Shepard, R. N., and Nelson, S. B. (eds.) (1977), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, Volume 2: Applications*, New York: Academic Press.
- Rowland, M. (1950), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, **27**, 852-857.
- Royden, J. P. (1963), "Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W," *Applied Statistics*, **12**, 1-10.
- Sargan, J. D. (1961), "A Robust Principal Component Analysis," *Journal of Multivariate Analysis*, **11**, 445-457.
- Satterthwaite, F. E. (1941), "Synthesis of Variances," *Psychometrika*, **6**, 309-316.
- Schwarz, L. I. (1935), "Remarks to Maurice Fisher's article 'Test for definition independence of one class of group distances' which is more generally applicable on Ferguson's theory," *Annals of Mathematics*, **36**, 754-757.
- Schott, J. R., and Fan, J.-C. (1988), "A Multivariate One-Way Classification Model with Random Effects," *Journal of Multivariate Analysis*, **13**, 1-12.
- Schwarzman, F. I., Krikorian, F. S., and Choudhury, A. K. (1973), "Exact Percentage Points of the Distribution of the Trace of a Multivariate Beta Matrix," *Journal of Statistical Computation and Simulation*, **3**, 331-343.
- Schwarz, S. J., and Margolis, B. H. (1962), "Detection of Multivariate Normal Outliers," *Annals of Statistics*, **10**, 945-954.
- Seal, David W. (1992), *Multivariate Density Estimation Theory, Practice and Investigation*, New York: Wiley.
- Seale, S. R. (1982), *Matrix Algebra Useful for Statistics*, New York: Wiley.
- Silks, G. A. F. (1984), *Multivariate Observations*, New York: Wiley.
- Slippy, S. S., and Will, M. B. (1962), "An Analysis of Variance Test for Nonunity-Complete Samples," *Biometrika*, **49**, 391-401.
- Shepard, R. N., Romney, A. K., and Nelson, S. B. (eds.) (1977), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, Volume 1: Theory*, New York: Academic Press.
- Stuart, R. (1964), "Recent Progress and Potential Developments: Some Practical Notes, Multivariate analysis with discussion," *Journal of the Royal Statistical Society, Series A*, **117**, 198-207.
- Gilman, R. W. (1966), *Density Estimation for Statistics and Data Analysis*, New York: Chapman & Hall.
- Sudo, D. K. (1994), "Detection of Multivariate Outliers in Elliptically Symmetric Distributions," *Annals of Statistics*, **12**, 1558-1565.

- Sokal, M., Hapkinow, E., and Pytkowski, T. (1985). *Modern Multivariate Statistical Analysis*. Columbus, Ohio: American Science.
- Sokal, R., Finkbein, K., Kowalski, R., Wojas, K., Kowalska, E., and Sato, M. (1983). "Statistical Research on the Taste Judgement Analysis of the Preliminary Experiment on Sensory and Chemical Character of KIDORI," *Proceedings of the Institute of Statistical Mathematics*, 18, 99-118.
- Smith, N. I. B. (1975). "Plotting Squared Residuals," *Biometrics*, 31, 437-438.
- Smith, D. W., Gil, D. S., and Hammond, J. J. (1965). "Variable Selection in Multivariate Multiple Regression," *Journal of Statistical Computation and Simulation*, 11, 117-127.
- Sned, R. D. (1972). "On the Analysis of Response-Curve Data," *Technometrics*, 14, 47-62.
- Sned, R. D., Auld, S. J., and Wilson, J. R. (1976). "A Unified Method for the Analysis of Growth Studies," *Biometrics*, 32, 835-848.
- Sokal, R. R., and Rohlf, F. J. (1981). *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd ed., San Francisco: W. H. Freeman and Co.
- Sparks, R. S., Christopoulos, G., and Toulkin, L. (1983). "The Multivariate F_p ," *Communications in Statistics—Part A Theory and Methods*, 12 (13), 1775-1793.
- Stevens, J. S., Barden, J. J., and Carpenter, W. T. (1978). "The Use of Clustering Techniques for the Classification of Psychiatric Patients," *British Journal of Psychiatry*, 133, 394-399.
- Thomas, D. R. (1985). "Uniformly Reported Measures Techniques Applied to Multivariate Data," *Psychometrika*, 50, 431-444.
- Thomson, G. H. (1911). *The Statistical Analysis of Human Study*. London: London University Press.
- Tinn, N. H. (1975). *Multivariate Analysis: With Applications in Education and Psychology*. Monterey, Calif.: Brooks/Cole.
- Tinn, N. H. (1980). "Multivariate Analysis of Variance of Reported Measures," in *Handbook of Statistics*, Vol. 1, P. R. Kishinrich (ed.), Amsterdam: North Holland, pp. 41-87.
- Tinn, G. (1986). "Some Applications of Multivariate Analysis to Economic Data," *Journal of the American Statistical Association*, 81, 472-508.
- Timingson, D. M., Murray, G. D., Murray, L. S., Sprigginer, D. L., Stone, A. M., Hoffman, J. D. F., and Gelpie, D. E. (1981). "Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients," *Journal of the Royal Statistical Society Series A*, 144, 141-171.
- Torrey, R. M. (1988). "The Use of a Discriminant Function in the Treatment of Psychological Group Differences," *Psychometrika*, 53 (1), 23-32.
- Ts, C. T., and Hsu, C. F. (1982). "Discriminant Analysis Based on Binary and Continuous Variables," *Journal of the American Statistical Association*, 77, 447-454.
- Wardlaw, W. M. (1976). "Some Implications of the Union-Intersection Principle for Test of Significance," *Journal of Multivariate Analysis*, 6, 183-190.
- Ward, R. E. (1986). "Simple Cases in the Multivariate Analysis of Reported Measurements," *Biometrics*, 42, 691-698.
- Wilks, H. (1981). "Comments on a Paper by Kishor and Bartiga," *Journal of the American Statistical Association*, 76, 765-771.
- Will, P. L. (1987). *The Generalized Variance Ratio or C-Statistic*. Albuquerque, N. M.
- Wong, C. M. (1987). "On the Analysis of Multivariate Reported Measures Data," *Communications in Statistics—Part A Theory and Methods*, 16, 1597-1609.

- Kiefer, J. B. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, **58**, 236-244.
- Koopman, B. J. (1972), "Nonparametric Probability Density Estimation: I. a Summary of Available Methods," *Econometrics*, **44**, 555-585.
- Mardia, B. L. (1967), "Note on Discriminant Functions," *Biometrika*, **54**, 213-220.
- Mardia, B. S. (1972), "Certain Generalizations in the Analysis of Variance," *Biometrika*, **59**, 473-494.
- Mardia, B. S. (1964), "Sample Criteria for Testing Equality of Means, Equality of Variances and Equality of Covariances in a Normal Multivariate Distribution," *Annals of Mathematical Statistics*, **17**, 287-301.
- Mardia, B. S. (1968), "Multivariate Statistical Outline," *Statlib*, Series A, **23**, 407-426.
- Millman, G. J. (1970), "Computing Means of Correlated Variables," *Biometrika*, **57**, 439-441.
- Millman, J. S., and Lerman, A. I. (1981), "A Class of Linear Spectral Models and Analysis for the Study of Longitudinal Data," Technical Report, Colorado State University, Department of Statistics.
- Mohant, D. (1981a), "An Algorithm for Hierarchical Classification," *Biometrika*, **68**, 169-176.
- Mohant, D. (1981b), "Numerical Classification Methods for Deriving Natural Classes," *Nature*, **291**, 97-98.
- Mohant, D. (1971), *A Generalized Approach to Cluster Analysis*. Part of Ph.D. Thesis, University of W. Australia.
- Yang, S. S., and Lee, T. (1987), "Identification of a Multivariate Outlier," Presented at the Annual Meeting of the American Statistical Association, San Francisco, August 1987.
- Narasimhan, I. S. (1966), "Detecting Sequences by Means of Projection Pursuit," *Computr*, **66**, 27-36.
- Young, F. W. (1987), *Multidimensional Scaling: History, Theory and Applications*, R. M. Hammel (ed.), Hillsdale, NJ: Lawrence Erlbaum.
- Young, G., and Bowsholtz, A. S. (1976), "Dispersion of a Set of Points in Terms of Their Mutual Distances," *Psychometrika*, **41**, 19-22.
- Zeyher, G. G. (1979a), "Randomization Analysis of the Completely Randomized Design Restricted to Growth Curves," *Journal of the American Statistical Association*, **74**, 215-223.
- Zeyher, G. G. (1979b), "Randomization Analysis of Randomized Block Design Restricted to Growth and Response Curves," *Communications in Statistics—Part A Theory and Methods*, **8**, 191-205.
- Zhang, L., Hildebrand, M. G., and Drury, C. G. (1990), "Identifying Factors of Cluster and Dispersion in Rating," *Human Factors*, **32** (3), 377-383.

Index

Additional information, test for, 134–139,
204–208

Asymptotic data, 932

Average distance data, 938

algebra, matrix, see *Matrix, algebra*

Analysis of variance

multivariate (MANOVA)

additional information, test for, 204–208

association, measures of, 176–178

assumptions, checking on, 199–200

and canonical correlations, 176–177

contrast, 189–190

discriminant function, 185, 189–193,
194

growth curves, 223–238, *See also* *Growth
curves, Repeated measures
designs*

Hotell T^2 statistic, 189–191

highest order models, 195–198

ind. effed variables, tests on, 183–184,
185–188

discriminant function, 194–195, 191

regression/linear tests, 183–185

posthoc tests, 194

Lewley–Hendley test, 187

table of critical values, 184–188

Multifactor ANOVA, 184

multivariate, 196–198

regional interactions, 196–197

multivariate association, measures of,
173–175

one way, 189–191

contrast, 190–191

colloquial, 181

model, 189

multifactor, 187

Pillai trace, 188

profile analysis, 199–200

repeated measures, 200–220, *See also*
*Repeated measures designs,
Distributions*

Hotell's trace (linear interaction), 188–189

table of critical values, 207–220

stepwise discriminant analysis, 183

stepwise selection of variables, 183

test for additional information, 204–208

test statistics, 189–190

compositional, 189–190, 176–178

and eigenvalues, 188

power of, 176–178

and T^2 , 189

tests on individual variables, 183–184,
183–186, 194

discriminant function, 185, 189–195, 191

regression/linear tests, 183–185

posthoc tests, 194

test on a subspace, 183–184

two way, 189–190

contrast, 190–191

discriminant function, 191

interaction, 189–190

main effects, 189–190

model, 189

tests on individual variables, 191

test statistics, 189

unbalanced one way, 189

linear interaction test, 188

Wilk's λ (Statistical inference), 188–189

chi square approximation, 188

F approximation, 188

power of λ statistic, 207

properties of, 183–184

table of critical values, 200–203

transformations to test F , 183–184

Analysis of variance (cont.)

analysis (ANOVA)

one way, 176–178

variance, 176–180

colleagues, 179–180

DGL, DGL, F-ratios, 178

two way, 180–183

variance, 180

F-test, 180

interaction, 181

main effects, 181–183

model, 180

ANOVA, see Analysis of variance, variance
analysis, discussion of, 171–174, 180–178
bivariate mixed data, 180

Bivariate data, 172

Bivariate data, 178

Bivariate data, 17–19

Bivariate, 171–178

variance of points, 171–174

variance, 174

variance, 174, 177

points for discussion, 174–178

points for variance, 174–178

principal-component approach, 174–178, 177

regular value decomposition, 172–174, 174

Bivariate distribution, 172

Bivariate normal distribution, 16, 14, 16–18, 17

Bivariate data, 172

Bivariate data, 172

Bivariate critical values, 172

table, 172–173

Bivariate data, 172–173

table of critical values, 172–173

Bivariate data, 174

Bivariate data, 174–178

Bivariate data, 172–173

Calculus data, 16

Calculus data, 172

Calculus data, 174, 176, 174–178

variance of points, see Calculus data

definition of, 172–173

and discussion analysis, 172–173

and eigenvalues, 172–173, 172–173

with grouping variables, 174

and MANOVA, 172–173

variance variables, 172–173

and variance of variance, 172, 172–174

and multiple variance, 174–176, 176, 176

propagation of, 174–176

variance analysis, 172–173

and regression, 172–173, 174–176

Calculus data, 172

with test for independence of two variables,

172, 172–173

test of significance, 172–173

all statistical conditions, 172–173

computation of tests, 172–173

and test of independence, 172–173

and test of overall regression, 172–173,

173

effect of variance conditions, 172–173

variance analysis, 172

test of variance in regression, 172–173

Calculus data,

variance analysis, 174

definition of, 172

interpretation, 172–173

by variance conditions coefficients,

172

by variance, 172

by standardized coefficients, 172–173

variance analysis, 172–173

and regression, 172–173

standardized coefficients, 172, 172–173

Calculus data, see Calculus data

Calculus data, see Calculus data

Calculus data,

of variance, 172, 172

of T^2 analysis, 172, 172

Calculus data, see Calculus data

Calculus data, 172

Calculus data, 16, 16–17, 17

Calculus data, 16–17

Calculus data, 16–17

Calculus data, 16

Calculus data, 16–17

assigning a coupling with $n+1$ group, 172

symmetric optimization, 172

variance conditions tests, 172–173

variance data, 172–173 for other data sets

variance data, 172–173

variance data, 172–173

k nearest neighbor rule, 172–173

variance data, 172–173

172–173

variance conditions tests, 172–173

variance conditions tests, 172–173

172–173

variance variables, 172

variance variables, 172–173

k nearest neighbor rule, 172–173

variance data, 172–173

variance conditions tests, 172–173

optimal classification rule (P-value), 172

variance conditions, 172–173

- Commonality, see Factor analysis
- Confidence interval (optimal), 149, 157
- Contingency table
 - graphical analysis of, see Correspondence analysis
 - highway table, 324, 328–329
 - two-way table, 314–316, 318, 321
- Creative plots, 89–92
- Crosstabs
 - control matrices in growth curves, 122–123, 125–126
 - control matrices in repeated measures, 334, 338–339
 - one-sample profile analysis, 143–145
 - one-way ANOVA, 179–180
 - one-way MANOVA, 186–188
 - collinearity, 286
 - one-way MANOVA, 188
 - two-way MANOVA, 189–191
- Covariances, 136
- Covari classification tree, 503–508
- Covariance
 - assumed, see Covariance estimation
 - old roots of single-factor PCA vectors, 19–38
 - into data-covariance, 138–139
 - and loss of vectors, 49–50
 - multiple, see Multiple covariance
 - and orthogonality of two vectors, 38
 - population covariance(Σ), 49
 - sample covariance(s), 49
 - of two linear combinations, 62, 72, 73
- Covariance matrix
 - and covariance matrix, 43
 - factor analysis on, 483–485
 - partitioned, 565
 - population covariance matrix, 43
 - principal-component form, 382–384, 384–387
 - sample covariance matrix, 49–51
 - from covariance matrix, 41
 - from data, 68
- Correspondence analysis, 344–350
 - contingency table
 - highway table, 324, 328–329
 - two-way table, 314–316, 318, 321
 - distance between column points, 325–326
 - distance between row points, 323–324
 - singular value decomposition, 324
 - generalized singular value decomposition, 323
 - correspondence matrix, 315–316
 - additive (graph of contingency table), 344–345
 - independence of rows and columns, testing, 319–321
 - chi square, 323, 325–326
 - inertia of frequencies, 326
 - inertia of inertia, 321
 - inertia of relative frequencies, 324
 - inertia of row and column profiles, 324–325
 - inertia, 343, 344
 - multiple correspondence analysis, 350–354
 - first inertia, 350–351
 - inertia matrix, 354–355
 - profiles of rows and columns, 343–345
 - row and column, 343–345
 - association, 344
 - inertia, 345–346
 - interaction, 346–347
 - point for plotting, 344–345
 - profiles, 343–345
 - single value decomposition, 322, 324
 - generalized singular value decomposition, 322
- Covariance
 - and independence, 46–47
 - and orthogonality, 47–48
 - population covariance (Σ), 46–47
 - sample covariance (s), 46–48
 - expected value of, 47
 - orthogonal relationships, 47
 - of two linear combinations, 62, 68, 72
- Covariance matrix
 - assumed symmetry, 248
 - and covariance matrix, 43
 - of linear combinations of variables, 69–73
 - partitioned, 62–65, 62
 - dependence of y and x and z , 62, 65
 - differences between $\text{cov}\left(\begin{pmatrix} y \\ x \end{pmatrix}\right)$ and $\text{cov}(y, x)$, 63
 - from n observations, 44–46
 - population covariance matrix, 122–123
 - population covariance matrix (Σ), 59–59
 - sample covariance matrix (s), 57–60
 - from data matrix, 58
 - distribution of, 53–55
 - Wishart distribution, 54–55
 - from observations, 59–59
 - profile definition of, 67
 - and sample mean vector, independence of, 62
 - sparsity, 106, 116–117

- test-retest, 283–288. *See also* Tests of hypothesis, correlation matrices
- uniformity of, 28
- uniformity, 286, 292–298
- Yates' correction, 339–341
- Zygodichia, 338
- Estimation/Est** 20
- Estimate**
 - air pollution data, 392
 - airline distances data, 498
 - airline record data, 498
 - bar steel data, 182
 - battle data, 180
 - bills and deaths data, 263
 - blood data, 262
 - blood pressure data, 242
 - breast cancer data, 184
 - by-product data, 242–246
 - calculation data, 38
 - calculation speed data, 240
 - chemical data, 348
 - city values data, 498
 - color-type data, 120
 - correlation, 239
 - crystal data, 124
 - death data, 127
 - distance data, 42
 - drop data, 242–244
 - drugs research data, 428
 - dyslexia data, 120
 - engine data, 181
 - fatality rates data, 128
 - fish data, 264
 - football data, 260–261
 - glucose data, 80–81
 - guinea pig data, 624
 - height-weight data, 42
 - hermaphrodite data, 268, 188
 - iron-oxide data, 267
 - olive data, 241
 - Norway-olive data, 384
 - propyl data, 329
 - propulsion data, 494
 - plasma data, 264
 - plasma drop data, 80, 8
 - pollution data, 242
 - public records data, 78
 - protein data, 488
 - psychological data, 122
 - runner time data, 78
 - suspected data, 128
 - Equilibrium rates data, 88
 - steel distances data, 581
 - statistical data, 171
 - table data, 262
 - temperature data, 128
 - time data, 78
 - test data, 275
 - vertical data, 128–129
 - temperature data, 368
 - test data, 242
 - rating data, 182
 - weight gain data, 240
 - wheat data, 188
 - wire data, 124
- Data, types of, 2–4. *See also* Multivariate data (family functions, 45)
- Death data, 127
- Demographic statistics, 1
- Dependent, 28–29
 - definition of, 26–27
 - of diagonal matrix, 27
 - of inverse, 29
 - of nonsingular matrix, 38
 - of partitioned matrix, 29
 - of positive definite matrix, 38
 - of product, 29
 - as product of eigenvalues, 34
 - of scalar multiple of a matrix, 28
 - of singular matrix, 28
- Distance data, 42
- Diagonal matrix, 8
- Discriminant analysis (descriptive), 176–264
 - and covariance-matrices, 252, 258–259
 - and classification analysis, 176
 - discriminant functions
 - for several groups, 183, 184–185, 194, 271, 276
 - regions of association for, 262
 - for two groups, 184–185, 271–276
 - and distances, 272
 - and eigenvalues, 279–278
 - interpretation of discriminant functions, 280–284
 - correlations (structure coefficients), 281
 - partial F-values, 280
 - variance, 284
 - standardized coefficients, 280
 - properties of, 277
 - vector plots, 271–273
 - selection of variables, 183, 284–286
 - several groups, 271–278
- Standardized discriminant functions, 280–284
- Stepwise discriminant analysis, 183, 284–286
- Tests of significance, 284–288
- Two groups, 171–173
 - and multiple regression, 188–182, 273–278

Exactness analysis (specificity), see

Classification analysis

Exponential matrix, see Covariance matrix

Distance between vectors, 14–17, 83, 118, 119, 123, 271–273

Distribution:

beta, 37

bivariate normal, 46–54, 89–97

chi-square, 81

elliptically symmetric, 108

F , 119, 138, 139, 162–163, 179, 246–249

multivariate normal, see Multivariate normal distribution, Multivariate

normality tests for

multivariate normal, 82–83, 86

test for, see Univariate normality tests for “Wolach”, 91–92

Emp data, 113–114

Euclidean (vector) data, 329

Excess variables, 173–174, 282, 313, 324–327

Ecology data, 152

F matrix, 180–184, 309, 340–348

Regression, 12–15, 163, 167–169, 302–304, 307–309, 403–405, 422–423

Regression, 12–15, 167–169, 302–304, 307–309, 403–405, 422–423

Elliptically symmetric distribution, 108

EM algorithm, 71, 471

Empirical data, 124

Error matrix, 125–127

coldest-cold data, 308

optimal error data, 307

“error”, 308, 309–311

classification variable, 307–308

error validation, 310–311

regression error matrix, 1–2, 128–129, 133–134

lowest method, 103–110, 118

learning error on method, 100–101, 103

partitioning the sample, 1–49

normalization, 107–108

Exponential value

of random matrix, 79

of random matrix $[X^T Y]$, 10–16

of sample covariance matrix $[X^T X]$, 79

of sample mean $[X^T Y]$, 10

of sample mean vector $[X^T Y]$, 16

of sample variance $[X^T Y]$, 1–44

of sum or product of random variables, 40

of univariate random variable $[X^T Y]$, 43

Experimental units, 1

F matrix

ANOVA, 138, 148

between subjects tests in repeated measures, 123, 146, 147

comparing two variances, 274–275

random, 179

equivalent to F^2 , 119, 124, 155–156

in multiple regression, 138, 146–147

partial F test, 127, 138, 143, 163, 164

regression selection, 270–273–276, 148

test for additional information, 107

test for individual variables MANOVA, 153–154

Wolach, 91

exact F transformation for, 142–143

F approximation for, 142–143

Fisher’s test data, 128

Factor analysis, 308–310

assumptions, 403–407

failure of assumptions, consequences of, 411, 444–445

common factors, 409

commensality, 410, 418, 421, 423, 427–428

estimation of, 419, 422, 423, 428

regression, 403–405, 422–423, 427, 432, 440

regression, 410–411, 426, 432

factor scores, 426–443

estimating method, 440

regression method, 439–440

factor, 408–410

common, 408

factorial, 411–412, 428–429

interpretation of, 408–410

number of, 408–409

Hyperbolic, 420–423

loadings

definition of, 408

estimation of, 411–412

estimation of methods, 414

fit of the model, 419

factorial principal factor method,

424–425

Raymond test, 424–425

maximum likelihood method, 423–424

principal component method, 423–424

principal factor method, 423–424

from R or S , 122–123, 127–128

model, 408–414

modeling covariance correlations, 408, 410, 413, 414, 417

number of factor loadings, 426–428

average regression, 427–428

comparison of methods, 428–430

- Linear combination(s) of variables (cont.)
 - mean vector for several linear combinations, 89
 - variance of a single/linear combination, 87, 91–92
- Linear combination of vectors, 39
- Linear hypothesis, 148–149, 199–201, 208–212
- Mahalanobis distance, 76–77, 89
- Maximin rule, 187
- McNemar, 150, 158. *See also* analysis of variance, multivariate
- Mark's inequality
 - applied to, 7–17
 - indicator form, 91–92
 - trial matrix, 328–329
 - Chernoff decomposition, 20–26
 - concentration, 11
 - covariance matrix, 37–39
 - definition, 5–6
 - determinant, 28–29, 34. *See also* Determinant of diagonal matrix, 21
 - of distributions, 29
 - of partitioned matrix, 26
 - of positive-definite matrix, 28
 - of product, 28
 - of order multiple of n matrix, 14
 - of singular matrix, 28
 - orthogonal, 24
- diagonal, 8
- eigenvalues, 32–33. *See also* Eigenvalues
 - characteristic equation, 32
 - and determinant, 34
 - of $I + A$, 34
 - of transposes, 30
 - of positive-definite matrix, 30
 - Perron-Frobenius theorem, 34
 - square root matrix, 36
 - of product, 25
 - singular value decomposition, 30
 - of square matrix, 30
 - of n -matrix matrix, 31
 - split of decomposition, 35
 - and trace, 34
- eigenvectors, 32–37. *See also* Eigenvalues
- equality of, 7
- identity, 8
- indicator matrix, 561–567
- inverse, 20–22
 - of partitioned matrix, 24
 - of product, 24
 - orthogonal, 24
- product, 9
- Q matrix, 9
 - linear combination of, 39
 - lower-triangular matrix, 23
 - relates the matrix and vector, 9–11
- QR-like matrix, 9
- operations with, 9–20
 - distributive law, 11
 - factoring, 12–13, 18
 - product, 10–16, 18–20
 - symmetric, 11
 - orthogonal matrix, 30
 - distributive over addition, 12
 - and eigenvalues, 34–35
 - of matrix and vector, 19
 - of matrix and transpose, 16–19
 - of matrix and vector, 12–13, 16, 18
 - in linear combination, 21
 - non-commutativity of, 11
 - perfect-square form, 17
 - transpose of, 12
 - single product, 13
 - of vector, 14
- rank, 10
 - commutativity of, 30
- reflexive, 16
- relation of, 30, 31–32
- partitioned matrices, 20–22
 - determinant of, 20
 - inverse of, 22
 - product of, 20, 21
 - transpose of, 22
- Perron-Frobenius theorem, 34, 362
- positive definite, 25, 34
- positive semidefinite, 25, 34
- quadratic form, 19
- rule, 22–24
 - full rank, 32
- rules, 9
 - product of vector and matrix, 30
- singular matrix, 24
- singular value decomposition, 30
- size of n matrix, 4
- spectral decomposition, 31
- square root matrix, 36
- sum of products in vector notation, 34
- sum of squares in vector notation, 14
- symmetric, 7, 25
- trace, 30, 34, 39
 - sub-eigenvalues, 34
 - of product, 30
 - of sum, 30
- transpose, 9–7
 - of product, 12
 - of sum, 10

- integers, *q*
- various, see *Varma's*
- (not) not an efficient test vector (N), *q*
- Maximum likelihood estimation, 99–111
 - of covariance matrix, 91
 - of covariance matrix, 99
 - likelihood function, 99
 - of mean vector, 99–111
 - multivariate normal, 99
- Mean
 - generals, 29
 - of linear function, 63, 70
 - population mean (μ), 43
 - of product, 84
 - sample mean ($\bar{y}$), 45–48
 - of sum, 68
- Mean vector, 29–36, 83, 99–102
 - notation, 84
 - population mean vector (μ), 29–36
 - sample mean vector ($\bar{y}$), 45–48
 - from observations, 83
 - distribution of, 84
 - and sample covariance matrix, independence of, 92
- Measurement scale, 1
 - interval scale, 2
 - ordinal scale, 2
 - ratio scale, 2
- Mean data, 241
- Microsimulation data, see *Linear model*
- Missing values, 74–76
- Multidimensionality, 74, 88
- Multidimensional scaling, 204–244
 - classical solution, see *Metric multidimensional scaling*
 - definition, 204–208
 - distance, 208–230
 - metric testings, 204
 - metric multidimensional scaling, 204–208
 - algorithm for finding the points, 208–230
 - and principal-component analysis, 208
 - nonmetric multidimensional scaling, 208, 230–244
 - nonmetric regression, 230–233
 - metric distributions, 230–233
 - STRESS, 233–237
 - principal coordinate analysis, see *Matrix multidimensional scaling*
 - optimal decomposition, 238–240
- Multiple correlation, 94, 94–96, 95 the R^2
 - also R^2
- Multiple correspondence analysis, 126–130
 - test statistics, 126–129
- column correlation, 127, 129–130
- column means, 126–127
- Multiple regression, see *Regression, multiple*
- Multivariate analysis, 1
 - descriptive statistics, 1–2
 - inferential statistics, 2
- Multivariate analysis of variance (MANOVA), see *Analysis of variance, multivariate*
- Multivariate data
 - basic types of, 4
 - plotting of, 58–59
 - specimens of, 57
- Multivariate inference, 1
- Multivariate normal distribution, 82–102
 - applicability of, 82
 - restricted distribution, 99
 - various plots, 84–88
 - density function, 88
 - distribution of $\bar{y}$ and S , 91–92
 - form of, 82
 - independence of $\bar{y}$ and S , 92
 - linear combination of, 84
 - log-likelihood, 97
 - maximum likelihood estimates, 99–104, see also *Maximum likelihood estimation*
 - properties of, 88–90
 - specimens from not all square distributions, 84
 - structural variation, 86
 - test statistics with a simple independence of observations, 87
- Multivariate statistics, both for, 92, 94–95
 - S , 97–98, 102–108
 - and chi square, 98
 - table of critical values, 107
 - dynamic plot, 98
 - water plots, 98, 100
 - distance and location, multivariate, 98–99, 103–104, 108
 - table of critical values, 103–108
- Multivariate regression, see *Regression, multivariate*
- Principal component analysis, 23
 - Normal distribution
 - testing normal, 94, 94–95–95, 100
 - multivariate normal, see *Multivariate normal distribution*
 - multivariate normal, 82–85, 94
- Normality tests for, see *Multivariate normality; Univariate normality*
- Plausibility data, 140
- Principal component analysis, see *Principal component analysis*

- Objectives of this book, 3
 - Observations, 3
 - One-sample *t*-test (see *test*), 177–182
 - Orthogonal matrix, 33
 - Orthogonal polynomial, 222–225
 - table of, 387
 - Orthogonal vectors, 50
 - Outline
 - multivariate
 - cluster, 328–329
 - elliptically symmetric distributions, 333
 - principal components, 385–392
 - diagnostic means, variance, and correlation, 334
 - Wilks' statistic, 352–358
 - outliers
 - accommodation, 382
 - clusters, 338
 - identification, 337
 - masking, 335
 - maximum likelihood method, 336–338
 - distance and Mahalan, 338
 - diagnostic means and variance, 338
 - removing, 338
 - Overall variability, 70–74
 - Partial observation test, 132–134
 - Partial F-tests, 125, 126, 254, 264–266
 - Partitioned matrices (see *Matrix* (matrices)),
 - partitioned matrices
 - Partitioning (see *Cluster* matrices, *partitioning*)
 - Pattern recognition (see *Cluster* analysis)
 - People data, 204
 - Perception data, 439
 - Perceptron (see *Neurons*, 34, 438)
 - Pilot's test statistics
 - definition of, 332
 - table of critical values, 279–280
 - Plotting data, 333
 - Plasma data, 246
 - Plotting multivariate data, 32–33
 - Plumber data, 241
 - Positive definite matrix, 23
 - positive definite sample covariance matrix, 47
 - Propositions for this book, 3
 - Principal components, 385–392
 - algorithm, 385–387
 - and biplots, 333–335
 - and cluster analysis, 364–365, 366, 368–369, 387
 - component scores, 386
 - definition of, 385, 387, 389
 - dimension reduction, 383–384, 385, 387, 389
 - exploratory multivariate test, 382–385, 387–388
 - major axis, 388, 389
 - and factor analysis, 385, 388–389, 392–393
 - properties of, 385–387
 - interpretation, 389–392
 - rotation, 388
 - special patterns in $\mathbf{X}$ or $\mathbf{B}$, 383–385
 - variance/covariance, 382–383
 - large variance of a variable, effect of, 383–384, 385
 - low low principal components, 382, 385, 386
 - maximum variance, 385, 386
 - minimum perpendicular distance to line, 383–385
 - number of components in matrix, 387–388
 - orthogonality of, 385, 383, 384
 - percent of variance, 383, 387
 - and perpendicular regression, 385, 387–389
 - plotting of, 385–387
 - measuring variability, 389–390
 - direction of outliers, 385–386
 - properties of, 383–385
 - proportion of variance, 385
 - robust, 389
 - as rotation of axes, 383–385, 384–385
 - from $\mathbf{X}$ or $\mathbf{B}$, 383–384, 383–387
 - rearrangement of components from $\mathbf{B}$, 387
 - sample specific components, 385
 - scale invariance, both of, 383
 - score graph, 383–385
 - selection of variables, 384–385
 - singular matrix and, 385–386
 - size and shape, 385–386
 - smaller principal components, 382, 384, 385
 - tests of significance for, 387, 389–390
 - variable specific components, 385
 - vectors of, 383, 384
- Protein test data, 78
- Product notation ($\mathbf{X}^T \mathbf{Y}$), 33
- Profile, 139–140
 - profile of observation vectors, 141
- Profile analysis
 - and contrasts, 141–142
 - one sample, 140–141
 - and one-way ANOVA, 140
 - profile, definition of, 139–140
 - and repeated measures, 139
 - normal sample, 140–141
 - two sample, 141–142
 - hypotheses
 - between, 141–142, 140–141
 - within, 144–145, 140–141
 - parallelism, 141–143, 140–141
 - uniformity, 140–141, 143–145
- Proportion points, 43

Pearson data, 493

Psychology, 124

Q - Q plot, 52–54

Quantile-classification functions, 106–107

Quantile-class, 49

Quantiles 52–54, 57

R^2 squared multiple-correlation, 102–103,
104, 106, 108, 109, 114–116, 117, 119,
121–124, 125–126

Random data, 76

Random variable(s):

ignorable, 45

infinite normal distribution, 46, 54,
55–56

in isolation, 49–50

in series, 49–50

in series, 49–50

linear relationships, 47

independence, 46

test for independence, 105–106

table of critical values, 106

unpaired, 49–50

within plot, 50–51

Linear combinations, 104–106

combination(s) of variables

estimation, 45

expected value of, 43

normal, 45

variance of, 44

variance, 43–45

Random variable, 50–56

distance between, 50–51, 55, 116, 121,
121–122

linear functions of, 104–105, for other linear
combination(s) of variables

mean of, 54–56

multivariate variance, 42–45

standardized, 56

variance, 42–45

Tests of statistics, 11–13

See's paradox, 114

Reliability analysis, 171–174

Regression, moments, 100–103

Regression, multiple (see y and normal x):
120–122, 122–127, for other
Regression, multivariate

normal x 's, 120–121

estimation of β :

normal x 's, 121–124

variances, 120–121

least squares, 121–124

estimation of σ^2 , 120–121

test x 's, 120–121

model, 121–124

assumptions, 121–124

corrected for mean-correlation, 121

multiple correlation, 121

R^2 squared multiple correlation, 102–103,
104, 106, 108, for other R^2

normal x 's, 121–124, 121

regression coefficients, 120

102, 104–106, 106–109, 109–109, 116

108, 109–109

subset selection, 120–121

all possible subsets, 120–121

criteria for selection (R^2 , r^2 , C_p),
121–124

comparison of criteria, 121

stepwise selection, 120–121

tests of hypotheses, 120–121

full and reduced model, 100–101

partial F test, 104–105

overall regression, 100–101

value of the y 's, 120–121

variables:

dependent x 's, 100

independent x 's, 101

regression x 's, 100

response x 's, 101

Regression, multivariate (normal x 's and normal
 y 's), 101–104, 101–104

correlation, moments of, 100–101

normal x 's, 101–101

estimation of β (matrix of regression
coefficients)

corrected x 's, 142–143

variances, 141

least squares, 139–141

properties of estimation, 141–142

estimation of β , 141

test x 's, 139–140

Class: Markov chains, 116

model, 117–119

assumptions, 119

corrected for mean-correlation, 142–143

normal x 's, 100

regression coefficients, matrix of (β , β),
100

subset selection, 120–121

all possible subsets, 120–121

criteria for selection (R^2 , R^2 , C_p),
121–124

stepwise procedure, 120–121

partial F-test, 104–105

value of the x 's, 120–121

value of the y 's, 120–121

tests of hypotheses, 140–141

F matrix, 139, 142–144

- Regression, multivariate (cont.)**
 full and reduced model
 on the x 's, 385–389
 on the y 's, 389–393
 with correlated covariates, 375–376
- Residuals**, 343–344
 overall regression, 343–347
 with correlated covariates, 375
 comparison of two models, 349
 Levene's testing test, 343
 Pillai's test, 343
 root-of- R , 345
 Roy's trace (linear interaction), 364–365
 "Wilks'" λ test (likelihood ratio), 364
 subset-of-the- x 's, 351–353
 with correlated covariates, 375, 376
 subset-of-the- y 's, 353–355
- Repeated measures**, 338
Repeated measures designs, 294–323, *see also*
Greenhouse-Geisser
 assumptions, 304–307
 comparison of observations, 297–312
 contrast matrices, 308, 309, 326
 doubly multivariate tests, 323
 higher order designs, 314–323
 multivariate approach, 307–309, *see also*
 316–317
 one sample, 296–311
 Blocked tests test, 298–300
 and randomized/block designs, 308
 old profile analysis, 320
 overall samples, 311–312
 covariance approach, 304–307
- Replication statistics**, 32
Researcher's role, 1
Rival observations, 394
Romantic data, 171
Rotation, *see* Factor analysis
Roy's maximum
 definition of, 364–365
 table of critical values, 370–377
- Sampling units**, 1
Scaling, 4
Scale of measurement, 1
 Scatterplot, 30–33, 35, 328
Seismic data, 263
Selection of variables, 263, 303–305, 336–338
Singular value decomposition, 36, 361, 364,
 362–363
 generalized singular value decomposition,
 362
Size and shape, 352–353
Skewness, 34–35, 38–39, 304
- Stepwise data**, 136
Stem data, 70
Specific estimates, *see* Factor analysis
Spectral decomposition, 33, 362, 363–364,
 365–366
Squared multiple correlation, *see* R^2
Standard deviation, 34
Standard errors, 36
Statistics, 375
Subject selection of variables, 234, 303–305,
 363–364
SYNOPSIS, 541–542
Subspaces, 32–34
 conditional distribution of, 36
 covariance matrix of, 32–34
 distribution of sum of SS
 independence of, 35, 37
 manipulation, 32–34, 36
 test of, 136–139, 154–156, 347–348,
 353–354
Summation statistics (Σ), 3
Survival data, 139–141
- T-tests**
 characteristic form, 137, 139
 contrasts, 179
 applied levels in profile analysis, 140
 growth curves, 134, 136
 matched pairs, 132–133
 one sample, 137
 paired observations, 132–133
 repeated measures, 139–141
 two samples, 133–135, 137
- T^2 statistics**
 additional information, *see* for, 154–156
 assumptions for, 122
 characteristic form, 138, 139
 c/s square approximation for, 120
 comparison of, 138–139
 by MANOVA, 136
 by regression, 130–132
 and F^2 distribution, 139, 139, 154–156
 full and reduced models, 127
 likelihood ratio test, 126
 matched pairs, 134–136
 one sample, 137–139
 paired observations, 134–136
 and profile analysis, 139–140
 one sample, 136–138
 two samples, 141–143
 properties of, 135–136, 133–134
 for contrasts, 136–139
 table of critical values for T^2 , 158–161
 two samples, 135–139

Time-series, *see* statistical, *see* Cluster analysis
 Temperature data, 266

Type of hypothesis

- accepting H_0 , 118
- for additional information, 134–139, 214–218, 347–348, 353–355
- paired F tests, 171, 176, 182
- rejection statistics, 248–249
 - one covariance matrix, 248–249
 - independence:
 - individual variables, 241–243
 - table of exact critical values, 241
 - normal subspaces, 241–244
 - two subspaces, 250–254
 - multivariate correlations, 241
 - a quadratic matrix B_0 , 146, 149
 - sphericity, 250–252
 - uniformly, compound symmetry, 246, 253–254
- rejection covariance matrices, 214–218
 - Hot's M test, 217–218
 - table of exact critical values, 218–220
- on individual variables, 126–130
 - exact critical values for, 127
 - table, 342–343
 - discreetness theories, 128–132
 - experimental error var, 134–135
 - paired F tests, 171, 182
 - paired test, 134–135
- Multivariate test, 126, *see* also Latent class analysis
- Two factor correlations:
 - one sample ($H_0: C_p = 0$), 117, 149–149, 186–187
 - two samples ($H_0: C_p = C_p(1)$), 141–145
- mean vector:
 - Method moments, 136
 - one sample, Σ known, 114–117
 - one sample, Σ unknown, 117–121
 - second samples, 158–171
 - two sample F^2 test, 122–126
- multivariate vs. univariate testing, 1–3, 112–113, 115–117, 127–130
- paired observations (matched pairs), 122–124
 - sub-variate, 114–116
 - univariate, 112–113
- paired F tests, 171, 181, 182
- power of a test, 113
- protected tests, 128–129
- on regression coefficients, 124–125, 183–189

one subspace, 139–146, 218–219, 347–349, 353–355

variate tests

- ANOVA, F test, 158–159, 169–169
 - one sample t -test t -test, σ known, 113
 - one sample t -test t -test, σ unknown, 117
 - paired observations, 152–153
 - test of variance, 224–225
 - two sample t -test, 124–125, 127
 - variance, equality of, 134–135
- Total sample variance, 76, 80, 105, 110–111, 117

Type of α matrix, 81, 84, 85

Type data, 262

Two sample test for equal mean vectors, 122–123

Union intersection test, 144–145

Unit:

- experimental, 1
- research, 1
- sampling, 1

Univariate normal distribution, 10–11, 81

Univariate normality tests for, 92–94

- F -hypothesis's D -matrix, 94
- table of critical values, 93
- procedure of the test, 94–97
- normal probability paper, 94
- D - D' plot, 95–94
- quantile, 93–94, 97
- skewness and kurtosis, 94–95
- table of critical values, 94–94
- transformation of correlation, 94

Variables, 1, *see* also Random variables

- recommends, 1
- dummy variables, 125, 126, 182, 183, 184–187
- linear combinations of, 61–71

Variance:

- generalized sample variance, 71
- paired variance, 121
- population variance σ^2 , 44
- sample variance (s^2), 61
- total sample variance, 79
- Variance of variance matrix, *see* Covariance matrix

Variance matrix, *see* Covariance matrix

Variance matrix, $\sigma^2 A - \sigma^2 B$

Variance:

- σ matrix, 8
- additive σ^2 , 8

Regression model**assumes**

- multicollinearity, 56–57
- homoscedastic error point, 58
- heteroscedastic regression, 56–57

generating *cf.* 8 **β -values** *cf.* 9**length** *cf.* 14**linear combination** *cf.* 19**linear independence and dependence** *cf.* 20**normalised** *cf.* 9**notation for matrix** 8**observational values** 55–56**collapsing** 33, 50**perpendicular** 48**product** *cf.* 14–15**the product** 14**error sub-column of a matrix** 15–16**standardised** *cf.* 9**submatrix** 52–53**sum of products** 14**sum of squares** 14**transpose** *cf.* 9–11**unit matrix** 9**Using lists** 33**Weight** *cf.* 14–15, 23**W least squares** 166**Wilder's χ^2 test statistic****definition** *cf.* 163–164**partial χ^2 statistic** 153**table of critical values** 164–171**Wisher distribution** 51–52**Wishart test** 124

WILEY SERIES IN PROBABILITY AND STATISTICS
ESTABLISHED BY WALTER A. SHEPPARD AND SAMUEL S. WILKS

*Editors: David J. Balding, Peter Dumbold, Neil A. C. Cressie,
Richard J. Fisher, Dale M. Johnson, J. B. Kadane, Louise M. Ryan,
David W. Scott, Adrian F. M. Smith, Josef L. Teugels*
Editors Emeriti: Ho Barnett, J. Stuart Hunter, David G. Kendall

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and probabilists, the titles span both state-of-the-art developments in the field and classical methods.

Covering the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computational practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

- ADRIAN F. M. SMITH and JOSEF L. TEUGELS · *Statistical Methods for Forecasting*
AGRESTI · *Analysis of Ordinal/Categorical Data*
AGRESTI · *An Introduction to Categorical Data Analysis*
AGRESTI · *Categorical Data Analysis, Second Edition*
ANGEL · *Mathematics of Chance*
ANDERSON · *An Introduction to Mathematical Statistical Analysis, Second Edition*
*ANDERSON · *The Statistical Analysis of Time Series*
ANDERSON, ALBERT, BIRCK, CARRIS, VANDERLIL, and WUNDERLICH ·
Statistical Methods for Comparative Studies
ANDERSON and LOYNER · *The Teaching of Practical Statistics*
ARMSTRONG and DAVIS (Editors) · *Advances in Biometry*
ARNOLD, BALAKRISHNAN, and MATHIAJAN · *Records*
*ASCHENBACH and BERGE · *Mathematical Programming in Statistics*
*BAILEY · *The Elements of Stochastic Processes with Applications to the Natural Sciences*
BALAKRISHNAN and KOUTRAS · *Risks and Ruins with Applications*
BARNETT · *Comparative Statistical Inference, Third Edition*
BARNETT and LEWIS · *Quality in Statistical Data, Third Edition*
BARTHOLOMEW and PERMALOMARCA (Editors) · *Probability and Statistical Inference*
BASILEVSKY · *Statistical Factor Analysis and Related Methods: Theory and Applications*
BASU and BISHOP · *Statistical Methods for the Reliability of Repaired Systems*
BATES and RATTI · *Nonlinear Regression Analysis and Its Applications*
BETHGEBER, SANTNER, and GOLDMAN · *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*
BILAL · *Combining Diagnostic Collinearity and Weak Data in Regression*
BILLEY, ELLI, and WELCH · *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*
BENDAT and PERSOL · *Random Data: Analysis and Measurement Procedures, Third Edition*

*Also available in a lower priced paperback edition in the Wiley Classics Library.

- HEBERT, CHALLONER, and CHURCH - Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
- HERNANDEZ and SMITH - Bayesian Theory
- HIRST and MILLER - Elements of Applied Stochastic Processes, Third Edition
- HORTINACHARYA and JOHNSON - Statistical Concepts and Methods
- HORTINACHARYA and RAYBURN - Stochastic Processes with Applications
- HULLGROVE - Convergence of Probability Measures, Second Edition
- HULLGROVE - Probability and Measures, Third Edition
- JURKES and JOGGI - Alternative Methods of Regression
- KLEINBERG and MERTZT - Reliability - Case Studies in Reliability and Maintenance
- KLEINBERG and MERTZT - Reliability: Modeling, Prediction, and Optimization
- KLOPFELD - Fourier Analysis of Time Series: An Introduction, Second Edition
- KOLLER - Discrete Equations with Latent Variables
- KOROVKOV - Existence and Stability of Stochastic Processes
- KOZLOV - Numerical Methods for Stochastic Processes
- KOZ - Bayesian Inference in Statistical Analysis
- KOZ - R. A. Fisher, the Life of a Scientist
- KOZ and GRAPPEL - Empirical Model-Building and Response Surfaces
- *KOZ and GRAPPEL - Evolutionary Operations: A Statistical Method for Process Improvement
- KOZ, HUNTER, and HUNTER - Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
- KOZ and LUCERO - Statistical Control by Monitoring and Feedback Adjustment
- KRANENHARTS - Numerical Methods in Finance: A MATLAB-Based Introduction
- KRYWEN and HILLANDER - Statistics: A Biomedical Introduction
- KRUMER, DOMBOI, and LAJOS - Nonparametric Analysis of Longitudinal Data in Factorial Experiments
- KUCCLEN - Large Deviation Techniques in Decision, Simulation, and Estimation
- KUCCLEN and DELLAND - Sequential Stochastic Optimization
- KUHN - Time Series: Applications to Finance
- KUTTERBERG and BLAKE - Reliability Analysis and Linear Regression
- KUTTERBERG and PRICE - Regression Analysis by Example, Third Edition
- KYRNICK - Bootstrap Methods: A Practitioner's Guide
- KYRNICK and PREE - Introductory Biostatistics for the Health Sciences
- KYRNICK and DELUTNER - Geostatistics: Modeling Spatial Uncertainty
- KYRW and LEE - Design and Analysis of Clinical Trials: Concepts and Methodologies
- CLARK and DIBNEY - Probability and Random Processes: A First Course with Applications, Second Edition
- *LACOMBE and COE - Experimental Design, Second Edition
- LACOMBE - Bayesian Financial Modelling
- LACOMBE - Practical Nonparametric Statistics, Second Edition
- COE - Regression Graphics
- COE and WEISBERG - Applied Regression Including Computing and Graphics
- COE and WEISBERG - An Introduction to Regression Graphics
- COOPER - Experiments with Machines, Designs, Models, and the Analysis of Machine Data, Third Edition
- COOPER and THOMAS - Concepts of Information Theory
- COE - A Handbook of Introductory Statistical Methods
- *COE - Planning of Experiments
- COOPER - Statistics for Spatial Data, Revised Edition
- COOPER and PURVATH - Limit Theorems in Change Point Analysis
- DANIEL - Applications of Statistics in Industrial Experimentation
- DANIEL - Biostatistics: A Foundation for Analysis in the Health Sciences, Sixth Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

- DANIEL** - *Fitting Equations to Data: Computer Analysis of Multifactor Data, Second Edition*
DAVIS - *Order Statistics, Second Edition*
DEBOWY, FERNBERG, and KAJANE - *Statistics and the Law*
DEI, CAMPILLO - *Statistical Process Adjustment for Quality Control*
DETTI and STEINEN - *The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis*
DEU and MEKELBOE - *Fractional Factorial Plans*
DELLON and GORDEN - *Multivariate Analysis: Methods and Applications*
DEDER - *Alternative Methods of Regression*
DEWE and SCABO - *Sampling Inspection Tables, Second Edition*
DOOR - *Stochastic Processes*
DOVON and WILKINSON - *Statistics for Research, Second Edition*
DRAPER and SMITH - *Applied Regression Analysis, Third Edition*
DRYDEN and MARDAL - *Statistical Shape Analysis*
DUDWICZ and NIKOLSK - *Modern Mathematical Statistics*
DUNN and CLARKE - *Applied Statistics: Analysis of Variance and Regression, Second Edition*
DUNN and CLARKE - *Basic Statistics: A Primer for the Biomedical Sciences, Third Edition*
DUPUIS and ELLIS - *A Weak Convergence Approach to the Theory of Large Deviations*
ELANEC-KERBERN and KERNON - *Survival Models and Data Analysis*
ETHER and KURTZ - *Markov Processes: Characterization and Convergence*
EVANS, RANTNER, and PRINCECE - *Statistical Distributions, Third Edition*
FELLER - *An Introduction to Probability Theory and Its Applications, Volume I, Third Edition, Revised, Volume II, Second Edition*
FRISCH and VAN BELLE - *Biostatistics: A Methodology for the Health Sciences*
FLEISS - *The Design and Analysis of Clinical Experiments*
FLEISS - *Statistical Methods for Rates and Proportions, Second Edition*
FLEMING and HARRINGTON - *Coupling Processes and Survival Analysis*
FELLER - *Introduction to Statistical Time Series, Second Edition*
FELLER - *Measurement Error Models*
GAUJANI - *Nonlinear Statistical Models*
GHOSH, GHOSHAPYAK, and GH - *Sequential Estimation*
GH - *Nonlinear Multivariate Analysis*
GLASSBERMAN and YAO - *Markov Structures in Discrete-Event Systems*
GHANADEKIAN - *Methods for Statistical Data Analysis of Multivariate Observations, Second Edition*
GH, GORDON and LEWIS - *Asymptotic Problems, Development, and Statistical Issues*
GREENWOOD and NIELSEN - *A Guide to Chi-Square Testing*
GRONL and HARRIS - *Fundamentals of Queueing Theory, Third Edition*
HAJANI and HAJI PIRI - *Statistical Models in Engineering*
HAJANI and MICHEL - *Statistical Inference: A Guide for Practitioners*
HALL - *A History of Probability and Statistics and Their Applications Before 1750*
HALL - *A History of Mathematical Statistics from 1750 to 1950*
HAMPTON - *Reliant Statistics: The Approach Based on Influence Functions*
HANNAN and DEUTLER - *The Statistical Theory of Linear Systems*
HERRMETER - *Computations for the Analysis of Designed Experiments*
HENDRY and SINGH - *Design and Inference in Finite Population Sampling*
HILLER - *MACPYMA for Statistics*
HINGULIAN and KAPTEHINE - *Design and Analysis of Experiments, Volume I: Introduction to Experimental Design*
HOAGLIN, MOSTELLER, and TUKEY - *Exploratory Approaches to Analysis of Variance*

- HEALING, MCSTELLER, and TUCKY · *Exploring Data Tables, Trends and Shapes*
- HEALING, MCSTELLER, and TUCKY · *Understanding Robust and Exploratory Data Analysis*
- HECHBERG and TAMRANE · *Multiple Comparison Procedures*
- HOCKING · *Methods and Applications of Linear Models: Regression and the Analysis of Variance, Second Edition*
- HOLL · *Introductory Mathematical Statistics, Fifth Edition*
- HODGE and KLEINMAN · *Loss Distributions*
- HOLLANDER and WOLFE · *Nonparametric Statistical Methods, Second Edition*
- HOSMER and LEMESHOW · *Applied Logistic Regression, Second Edition*
- HOSMER and LEMESHOW · *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*
- HEYLAND and RABRAND · *System Reliability Theory: Models and Statistical Methods*
- FRIBER · *Robust Statistics*
- HUBERTY · *Applied Discriminant Analysis*
- HUNT and KENNEDY · *Financial Derivatives in Theory and Practice*
- HUNGUEN, BURAN, and DUBUC · *Collected Works of Jacques Huguen—*
with Commentary
- IRWIN and CONOVER · *A Modern Approach to Statistics*
- JACKSON · *A User's Guide to Principal Components*
- JOHN · *Statistical Methods in Engineering and Quality Assurance*
- JOHNSON · *Multivariate Statistical Distributions*
- JOHNSON and BALAKRISHNAN · *Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz*
- JUDGE, GRIFITHS, HILL, LUTENFORD, and LEE · *The Theory and Practice of Econometrics, Second Edition*
- JOHNSON and KOTZ · *Distributions in Statistics*
- JOHNSON and KOTZ (editors) · *Leading Personalities in Statistical Science: Profiles from Scientific Careers in the Present*
- JOHNSON, KOTZ, and BALAKRISHNAN · *Continuous Univariate Distributions, Volume 1, Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · *Continuous Univariate Distributions, Volume 2, Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · *Discrete Multivariate Distributions*
- JOHNSON, KOTZ, and GOMEZ · *Exponential Distributions, Second Edition*
- JURK, KOVA, and BEN · *Robust Statistical Procedures: Asymptotics and Interactions*
- JUREK and MIKOSIN · *Operator Limit Distributions in Probability Theory*
- KADANE · *Bayesian Methods and Ethics in a Clinical Trial Design*
- KADANE AND SCHUM · *A Probabilistic Analysis of the Ozone and Cancer Evidence*
- KALBFLEISCH and PRENTICE · *The Statistical Analysis of Failure Time Data, Second Edition*
- KASS and VOSS · *Geometrical Foundations of Asymptotic Inference*
- KALFMAN and BOCKREITW · *Finding Groups in Data: An Introduction to Cluster Analysis*
- KRUM and FORMAN · *Regression Models for Time Series Analysis*
- KUPPALA, RABHON, CHAKRI, and LI · *Shape and Shape Theory*
- KLEIN · *Advanced Calculus with Applications in Statistics, Second Edition*
- KLEIN, MATHER, and KIRK · *Statistical Tools for Mixed Linear Models*
- KLEINMAN, FINKEL, and WILLARD · *Loss Models: From Data to Decisions*
- KLEINMAN, FINKEL, and WILLARD · *Solutions Manual for Accompany Loss Models From Data to Decisions*
- KOTZ, BALAKRISHNAN, and JOHNSON · *Continuous Multivariate Distributions, Volume 1, Second Edition*
- KOTZ and JOHNSON (editors) · *Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index*

KOTZ and RIBBEN (editors) - *Encyclopedia of Statistical Sciences: Supplement Volume*

KOTZ, READ, and RANKS (editors) - *Encyclopedia of Statistical Sciences: Update Volume 1*

KOTZ, READ, and RANKS (editors) - *Encyclopedia of Statistical Sciences: Update Volume 2*

KOVALENKO, KUZNETSOV, and PRIG - *Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications*

LACHEN - *Biostatistical Methods: The Assessment of Relative Risks*

LAD - *Operational/Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*

LAMPART - *Probability: A Survey of the Mathematical Theory, Second Edition*

LANCH, STAN, HILLARD, HILLINGER, CONQUEST, and ORSHANICH - *Case Studies in Biometry*

LAERCH - *Introduction to Probability Theory and Statistical Inference, Third Edition*

LARLEY - *Statistical Models and Methods for Lifetime Data, Second Edition*

LARSON - *Statistical Methods in System Engineering*

LE - *Applied Categorical Data Analysis*

LE - *Applied Survival Analysis*

LEE and WANG - *Statistical Methods for Survival Data Analysis, Third Edition*

LEPAGE and BILLARD - *Exploring the Limits of Bootstrap*

LITTLE and MCLEOD (editors) - *Multilevel Modelling of Health Statistics*

LIAN - *Statistical Group Comparison*

LIPOWITZ - *Lectures on the Grouping Method*

LINSART and LICHEN - *Model Selection*

LITTLE and RUPP - *Statistical Analysis with Missing Data, Second Edition*

LLOYD - *The Statistical Analysis of Categorical Data*

MASON and McBRID - *Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition*

MALLER and SOOTHY - *Survival Analysis with Long-Term Survivors*

MALLORY - *Design, Data, and Analysis for Some Friends of Childhood Death*

MANN, SCHAPIRO, and SCHNEPPENALL - *Methods for Statistical Analysis of Reliability and Life Data*

MANTON, WOODBURY, and TOLLEY - *Statistical Applications Using Fuzzy Sets*

MARINA and RUPP - *Directional Statistics*

MARON, DETROT, and RICH - *Statistical Design and Analysis of Experiments with Applications to Engineering and Science, Second Edition*

McCOLLON and MURIEL - *Generalized, Linear, and Mixed Models*

McFADDEN - *Management of Data in Clinical Trials*

McLACHLAN - *Discriminant Analysis and Statistical Pattern Recognition*

McLACHLAN and KIRKPATRICK - *The EM Algorithm and Extensions*

McLACHLAN and PEE - *Finite Mixture Models*

McNEIL - *Epidemiological Research Methods*

MEYER and SCHAPIRO - *Statistical Methods for Reliability Data*

MERCHANDISE and SCHAPIRO - *Local Distributions for Tests of Independent Random Vectors: Theory, Tests in Theory and Practice*

*MILLER - *Survival Analysis, Second Edition*

MONTGOMERY, PECK, and VINCE - *Introduction to Linear Regression Analysis, Third Edition*

RECHENTHALER and TUCKER - *Configured Polysampling: A Guide to Practical Reliability*

McBRIDE - *Aspects of Mathematical Statistical Theory*

McBRIDE - *Statistical 2.0 Statistical Implementation, Design Data Analysis, and Nonlinear Optimization*

*Now available in a lower priced paperback edition in the Wiley-Interscience Library.

- MYERS and MOENCHMIST - *Regression Surface Methodology: Process and Profile Optimization Using Designed Experiments, Second Edition*
- MYERS, MOENCHMIST, and YINHO - *Generalized Linear Models: With Applications in Engineering and the Sciences*
- NELSON - *Accelerated Testing, Statistical Models, Test Plans, and Data Analysis*
- NELSON - *Applied Life Data Analysis*
- NEWMAN - *Statistical Methods in Epidemiology*
- OCBO - *Applied Probability and Stochastic Processes in Engineering and Physical Sciences*
- ORABI, BOOTS, RUSHARA, and CHU - *Spatial Transactions: Concepts and Applications of General Diagrams, Second Edition*
- OLIVER and SMITH - *Influence Diagrams, Belief Nets and Decision Analysis*
- PANAGIOTIS - *Forecasting with Dynamic Regression Models*
- PANAGIOTIS - *Forecasting with Univariate Box-Jenkins Models: Concepts and Case Studies*
- PARZEN - *Modern Probability Theory and Its Applications*
- PENK, TIAO, and TAY - *A Course in Time-Series Analysis*
- PLANTADON - *Clinical Trials: A Methodologic Perspective*
- PORT - *Theoretical Probability for Applications*
- POURAHMADI - *Foundations of Time Series Analysis and Prediction Theory*
- PREER - *Bayesian Statistics: Principles, Models, and Applications*
- PREIS - *Subjective and Objective Bayesian Statistics, Second Edition*
- PREER and TAYLOR - *The Subjectivity of Scientists and the Bayesian Approach*
- PURCELLERUM - *Optimal Experimental Design*
- PURI, YILAPLANA, and WERTZ - *New Perspectives in Theoretical and Applied Statistics*
- PUTERMAN - *Markov Decision Processes: Discrete Stochastic Dynamic Programming*
- RAO - *Linear Statistical Inference and Its Applications, Second Edition*
- REINER - *Linear Models in Statistics*
- REINER - *Methods of Multivariate Analysis, Second Edition*
- REINER - *Multivariate Statistical Inference with Applications*
- REPLEY - *Spatial Statistics*
- REPLEY - *Stochastic Simulation*
- ROBINSON - *Practical Strategies for Experimenting*
- ROBERTS and SALLER - *An Introduction to Probability and Statistics, Second Edition*
- ROLD, SCHROEDER, SCHROEDER, and TUGGLES - *Stochastic Processes for Insurance and Finance*
- ROSENBERGER and LACHEN - *Randomization in Clinical Trials: Theory and Practice*
- ROSS - *Introduction to Probability and Statistics for Engineers and Scientists*
- ROUSSELEW and LEROY - *Robust Regression and Outlier Detection*
- RUBIN - *Multiple Imputation for Nonresponse in Surveys*
- RUBINSTEIN - *Simulation and the Monte Carlo Method*
- RUBINSTEIN and MELAMED - *Modern Simulation and Modeling*
- RYAN - *Modern Regression Methods*
- RYAN - *Statistical Methods for Quality Improvement, Second Edition*
- SAMUEL, CHAN, and SCOTT (Editors) - *Sensitivity Analysis*
- SCHIFFER - *The Analysis of Variance*
- SCHWENK - *Smoothing and Regression: Approaches, Comparison, and Application*
- SCHWERT - *Matrix Analysis for Statistics*
- SCHWISS - *Theory and Applications of Stochastic Differential Equations*
- SCOTT - *Multivariate Density Estimation: Theory, Practice, and Visualization*
- SEARLE - *Linear Models*
- SEARLE - *Linear Models for Unbalanced Data*
- SEARLE - *Matrix Algebra Useful for Statistics*
- SEARLE, CASE, LA, and MCELLOCH - *Variance Components*
- SEARLE and WELLET - *Matrix Algebra for Applied Economists*

ZERBER - Linear Regression Analysis
 ZERBER - Multivariate Observations
 ZERBER and WILD - Nonlinear Regression
 ZIMMERT - Stochastic Dynamic Programming and the Control of Queueing Systems
 *ZIRBLING - Approximation Theorems of Mathematical Statistics
 ZLOTNIK and YOFFE - Probability and Finance: It's Only a Game!
 SMALL and McLEISH - Hilbert Space Methods in Probability and Statistical Inference
 ZIEGLER - Multilevel Mathematics Statistics
 STAPLETON - Linear Statistical Models
 STAUDTE and ZIEGLER - Robust Statistics and Testing
 STOVAK, KENDALL, and MITCHELL - Stochastic Geometry and Its Applications, Second Edition
 STOVAK and STOVAK - Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
 STYAN - The Collected Papers of T. W. Anderson, 1943-1993
 SUTTON, ABRAMS, KONIS, SHELDON, and SONO - Methods for Meta-Analysis in Medical Research
 TAKENAKA - Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
 THOMPSON - Empirical Model Building
 THOMPSON - Sampling, Second Edition
 THOMPSON - Simulation: A Modeler's Approach
 THOMPSON and DEBER - Adaptive Sampling
 THOMPSON, WILLIAMS, and FENDLEY - Models for Investors in Real World Markets
 TIAN, BINGHARIL, HILL, PINK, and STOLER (editors) - Sex, or Quality and Quantity with Design, Control, and Substrates
 TERNET - LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
 TNAV - Analysis of Financial Time Series
 OTTON and FREDERSON - Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
 TANIGUCHI - Statistical Factor Analysis
 TIEDAUF - Statistical Modeling by Examples
 WEINBERG - Applied Linear Regression, Second Edition
 WELSH - Aspects of Statistical Inference
 WESTFALL and YOUNG - Resampling-Based Multiple Testing: Examples and Methods for p -Value Adjustment
 WHITTAKER - Graphical Models in Applied Multivariate Statistics
 WINGO - Optimization Methods in Economics: Applications of Threshold Accepting
 WOODHURST and WOODHURST - Econometrics, Second Edition
 WOODING - Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
 WOODSON and CLARKE - Statistical Methods for the Analysis of Biomedical Data, Second Edition
 WU and HAMADA - Experiments: Planning, Analysis, and Parameter Design Optimization
 TANO - The Construction Theory of Dimensionable Markov Processes
 *ZELLNER - An Introduction to Bayesian Inference in Econometrics
 ZHOU, CHEN, ZHANG, and MCCLELL - Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.