An Introduction to Multivariate Statistical Analysis

Third Edition

T. W. ANDERSON

Stanford University
Department of Statistics
Stanford, CA

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2003 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc. Hoboken, New Jersey
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 Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permreq@wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data

Anderson, T. W. (Theodore Wilbur), 1918-
An introduction to multivariate statistical analysis / Theodore W. Anderson.--3rd ed.
p. cm.-- (Wiley series in probability and mathematical statistics)
Includes bibliographical references and index.
ISBN 0-471-36091-0 (cloth : acid-free paper)
1. Multivariate analysis. I. Title. II. Series.

QA278.A516 2003
519.5'35--dc21 2002034317

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To

DOROTHY

Contents

Preface to the Third Edition

For some forty years the first and second editions of this book have been used by students to acquire a basic knowledge of the theory and methods of multivariate statistical analysis. The book has also served a wider community of statisticians in furthering their understanding and proficiency in this field. Since the second edition was published, multivariate analysis has been developed and extended in many directions. Rather than attempting to cover, or even survey, the enlarged scope, I have elected to elucidate several aspects that are particularly interesting and useful for methodology and comprehension.

Earlier editions included some methods that could be carried out on an adding machine! In the twenty-first century, however, computational techniques have become so highly developed and improvements come so rapidly that it is impossible to include all of the relevant methods in a volume on the general mathematical theory. Some aspects of statistics exploit computational power such as the resampling technologies; these are not covered here.

The definition of multivariate statistics implies the treatment of variables that are interrelated. Several chapters are devoted to measures of correlation and tests of independence. A new chapter, "Patterns of Dependence; Graphical Models" has been added. A so-called graphical model is a set of vertices or nodes identifying observed variables together with a new set of edges suggesting dependences between variables. The algebra of such graphs is an outgrowth and development of path analysis and the study of causal chains. A graph may represent a sequence in time or logic and may suggest causation of one set of variables by another set.

Another new topic systematically presented in the third edition is that of elliptically contoured distributions. The multivariate normal distribution, which is characterized by the mean vector and covariance matrix, has a limitation that the fourth-order moments of the variables are determined by the first- and second-order moments. The class of elliptically contoured distribution relaxes this restriction. A density in this class has contours of equal density which are ellipsoids as does a normal density, but the set of fourth-order moments has one further degree of freedom. This topic is expounded by the addition of sections to appropriate chapters.

Reduced rank regression developed in Chapters 12 and 13 provides a method of reducing the number of regression coefficients to be estimated in the regression of one set of variables to another. This approach includes the limited-information maximum-likelihood estimator of an equation in a simultaneous equations model.

The preparation of the third edition has been benefited by advice and comments of readers of the first and second editions as well as by reviewers of the current revision. In addition to readers of the earlier editions listed in those prefaces I want to thank Michael Perlman and Kathy Richards for their assistance in getting this manuscript ready.

T. W. ANDERSON

Stanford, California
February 2003

Preface to the Second Edition

Twenty-six years have passed since the first edition of this book was published. During that time great advances have been made in multivariate statistical analysis—particularly in the areas treated in that volume. This new edition purports to bring the original edition up to date by substantial revision, rewriting, and additions. The basic approach has been maintained, namely, a mathematically rigorous development of statistical methods for observations consisting of several measurements or characteristics of each subject and a study of their properties. The general outline of topics has been retained.

The method of maximum likelihood has been augmented by other considerations. In point estimation of the mean vector and covariance matrix alternatives to the maximum likelihood estimators that are better with respect to certain loss functions, such as Stein and Bayes estimators, have been introduced. In testing hypotheses likelihood ratio tests have been supplemented by other invariant procedures. New results on distributions and asymptotic distributions are given; some significant points are tabulated. Properties of these procedures, such as power functions, admissibility, unbiasedness, and monotonicity of power functions, are studied. Simultaneous confidence intervals for means and covariances are developed. A chapter on factor analysis replaces the chapter sketching miscellaneous results in the first edition. Some new topics, including simultaneous equations models and linear functional relationships, are introduced. Additional problems present further results.

It is impossible to cover all relevant material in this book; what seems most important has been included. For a comprehensive listing of papers until 1966 and books until 1970 the reader is referred to *A Bibliography of Multivariate Statistical Analysis* by Anderson, Das Gupta, and Styan (1972). Further references can be found in *Multivariate Analysis: A Selected and*

Abstracted Bibliography, 1957–1972 by Subrahmaniam and Subrahmaniam (1973).

I am in debt to many students, colleagues, and friends for their suggestions and assistance; they include Yasuo Amemiya, James Berger, Byoung-Seon Choi, Arthur Cohen, Margery Cruise, Somesh Das Gupta, Kai-Tai Fang, Gene Golub, Aaron Han, Takeshi Hayakawa, Jogi Henna, Huang Hsu, Fred Huffer, Mituaki Huzii, Jack Kiefer, Mark Knowles, Sue Leurgans, Alex McMillan, Masashi No, Ingram Olkin, Kartik Patel, Michael Perlman, Allen Sampson, Ashis Sen Gupta, Andrew Siegel, Charles Stein, Patrick Strout, Akimichi Takemura, Joe Verducci, Marlos Viana, and Y. Yajima. I was helped in preparing the manuscript by Dorothy Anderson, Alice Lundin, Amy Schwartz, and Pat Struse. Special thanks go to Johanne Thiffault and George P. H. Styan for their precise attention. Support was contributed by the Army Research Office, the National Science Foundation, the Office of Naval Research, and IBM Systems Research Institute.

Seven tables of significance points are given in Appendix B to facilitate carrying out test procedures. Tables 1, 5, and 7 are Tables 47, 50, and 53, respectively, of *Biometrika Tables for Statisticians*, Vol. 2, by E. S. Pearson and H. O. Hartley; permission of the Biometrika Trustees is hereby acknowledged. Table 2 is made up from three tables prepared by A. W. Davis and published in *Biometrika* (1970a), *Annals of the Institute of Statistical Mathematics* (1970b) and *Communications in Statistics, B. Simulation and Computation* (1980). Tables 3 and 4 are Tables 6.3 and 6.4, respectively, of *Concise Statistical Tables*, edited by Ziro Yamauti (1977) and published by the Japanese Standards Association; this book is a concise version of *Statistical Tables and Formulas with Computer Applications, JSA-1972*. Table 6 is Table 3 of *The Distribution of the Sphericity Test Criterion*, ARL 72-0154, by B. N. Nagarsenker and K. C. S. Pillai, Aerospace Research Laboratories (1972). The author is indebted to the authors and publishers listed above for permission to reproduce these tables.

T. W. ANDERSON

Stanford, California
June 1984

Preface to the First Edition

This book has been designed primarily as a text for a two-semester course in multivariate statistics. It is hoped that the book will also serve as an introduction to many topics in this area to statisticians who are not students and will be used as a reference by other statisticians.

For several years the book in the form of dittoed notes has been used in a two-semester sequence of graduate courses at Columbia University; the first six chapters constituted the text for the first semester, emphasizing correlation theory. It is assumed that the reader is familiar with the usual theory of univariate statistics, particularly methods based on the univariate normal distribution. A knowledge of matrix algebra is also a prerequisite; however, an appendix on this topic has been included.

It is hoped that the more basic and important topics are treated here, though to some extent the coverage is a matter of taste. Some of the more recent and advanced developments are only briefly touched on in the late chapter.

The method of maximum likelihood is used to a large extent. This leads to reasonable procedures; in some cases it can be proved that they are optimal. In many situations, however, the theory of desirable or optimum procedures is lacking.

Over the years this manuscript has been developed, a number of students and colleagues have been of considerable assistance. Allan Birnbaum, Harold Hotelling, Jacob Horowitz, Howard Levene, Ingram Olkin, Gobind Seth, Charles Stein, and Henry Teicher are to be mentioned particularly. Acknowledgements are also due to other members of the Graduate Mathematical

Statistics Society at Columbia University for aid in the preparation of the manuscript in dittoed form. The preparation of this manuscript was supported in part by the Office of Naval Research.

T. W. ANDERSON

Center for Advanced Study
in the Behavioral Sciences
Stanford, California
December 1957

CHAPTER 1

Introduction

1.1. MULTIVARIATE STATISTICAL ANALYSIS

Multivariate statistical analysis is concerned with data that consist of sets of measurements on a number of individuals or objects. The sample data may be heights and weights of some individuals drawn randomly from a population of school children in a given city, or the statistical treatment may be made on a collection of measurements, such as lengths and widths of petals and lengths and widths of sepals of iris plants taken from two species, or one may study the scores on batteries of mental tests administered to a number of students.

The measurements made on a single individual can be assembled into a column vector. We think of the entire vector as an observation from a multivariate population or distribution. When the individual is drawn randomly, we consider the vector as a random vector with a distribution or probability law describing that population. The set of observations on all individuals in a sample constitutes a sample of vectors, and the vectors set side by side make up the matrix of observations.† The data to be analyzed then are thought of as displayed in a matrix or in several matrices.

We shall see that it is helpful in visualizing the data and understanding the methods to think of each observation vector as constituting a point in a Euclidean space, each coordinate corresponding to a measurement or variable. Indeed, an early step in the statistical analysis is plotting the data; since

†When data are listed on paper by individual, it is natural to print the measurements on one individual as a row of the table; then one individual corresponds to a *row* vector. Since we prefer to operate algebraically with column vectors, we have chosen to treat observations in terms of *column* vectors. (In practice, the basic data set may well be on cards, tapes, or disks.)

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

most statisticians are limited to two-dimensional plots, two coordinates of the observation are plotted in turn.

Characteristics of a univariate distribution of essential interest are the mean as a measure of location and the standard deviation as a measure of variability; similarly the mean and standard deviation of a univariate sample are important summary measures. In multivariate analysis, the means and variances of the separate measurements—for distributions and for samples—have corresponding relevance. An essential aspect, however, of multivariate analysis is the dependence between the different variables. The dependence between two variables may involve the covariance between them, that is, the average products of their deviations from their respective means. The covariance standardized by the corresponding standard deviations is the correlation coefficient; it serves as a measure of degree of dependence. A set of summary statistics is the mean vector (consisting of the univariate means) and the covariance matrix (consisting of the univariate variances and bivariate covariances). An alternative set of summary statistics with the same information is the mean vector, the set of standard deviations, and the correlation matrix. Similar parameter quantities describe location, variability, and dependence in the population or for a probability distribution. The multivariate *normal* distribution is completely determined by its mean vector and covariance matrix, and the sample mean vector and covariance matrix constitute a sufficient set of statistics.

The measurement and analysis of dependence between variables, between sets of variables, and between variables and sets of variables are fundamental to multivariate analysis. The multiple correlation coefficient is an extension of the notion of correlation to the relationship of one variable to a set of variables. The partial correlation coefficient is a measure of dependence between two variables when the effects of other correlated variables have been removed. The various correlation coefficients computed from samples are used to estimate corresponding correlation coefficients of distributions. In this book tests of hypotheses of independence are developed. The properties of the estimators and test procedures are studied for sampling from the multivariate normal distribution.

A number of statistical problems arising in multivariate populations are straightforward analogs of problems arising in univariate populations; the suitable methods for handling these problems are similarly related. For example, in the univariate case we may wish to test the hypothesis that the mean of a variable is zero; in the multivariate case we may wish to test the hypothesis that the vector of the means of several variables is the zero vector. The analog of the Student t-test for the first hypothesis is the generalized T^2-test. The analysis of variance of a single variable is adapted to vector observations; in regression analysis, the dependent quantity may be a vector variable. A comparison of variances is generalized into a comparison of covariance matrices.

The test procedures of univariate statistics are generalized to the multivariate case in such ways that the dependence between variables is taken into account. These methods may not depend on the coordinate system; that is, the procedures may be invariant with respect to linear transformations that leave the null hypothesis invariant. In some problems there may be families of tests that are invariant; then choices must be made. Optimal properties of the tests are considered.

For some other purposes, however, it may be important to select a coordinate system so that the variates have desired statistical properties. One might say that they involve characterizations of inherent properties of normal distributions and of samples. These are closely related to the algebraic problems of canonical forms of matrices. An example is finding the normalized linear combination of variables with maximum or minimum variance (finding principal components); this amounts to finding a rotation of axes that carries the covariance matrix to diagonal form. Another example is characterizing the dependence between two sets of variates (finding canonical correlations). These problems involve the characteristic roots and vectors of various matrices. The statistical properties of the corresponding sample quantities are treated.

Some statistical problems arise in models in which means and covariances are restricted. Factor analysis may be based on a model with a (population) covariance matrix that is the sum of a positive definite diagonal matrix and a positive semidefinite matrix of low rank; linear structural relationships may have a similar formulation. The simultaneous equations system of econometrics is another example of a special model.

1.2. THE MULTIVARIATE NORMAL DISTRIBUTION

The statistical methods treated in this book can be developed and evaluated in the context of the multivariate normal distribution, though many of the procedures are useful and effective when the distribution sampled is not normal. A major reason for basing statistical analysis on the normal distribution is that this probabilistic model approximates well the distribution of continuous measurements in many sampled populations. In fact, most of the methods and theory have been developed to serve statistical analysis of data. Mathematicians such as Adrian (1808), Laplace (1811), Plana (1813), Gauss

(1823), and Bravais (1846) studied the bivariate normal density. Francis Galton, the geneticist, introduced the ideas of correlation, regression, and homoscedasticity in the study of pairs of measurements, one made on a parent and one in an offspring. [See, e.g., Galton (1889).] He enunciated the theory of the multivariate normal distribution as a generalization of observed properties of samples.

Karl Pearson and others carried on the development of the theory and use of different kinds of correlation coefficients† for studying problems in genetics, biology, and other fields. R. A. Fisher further developed methods for agriculture, botany, and anthropology, including the discriminant function for classification problems. In another direction, analysis of scores on mental tests led to a theory, including *factor analysis*, the sampling theory of which is based on the normal distribution. In these cases, as well as in agricultural experiments, in engineering problems, in certain economic problems, and in other fields, the multivariate normal distributions have been found to be sufficiently close approximations to the populations so that statistical analyses based on these models are justified.

The univariate normal distribution arises frequently because the effect studied is the sum of many independent random effects. Similarly, the multivariate normal distribution often occurs because the multiple measurements are sums of small independent effects. Just as the central limit theorem leads to the univariate normal distribution for single variables, so does the general central limit theorem for several variables lead to the multivariate normal distribution.

Statistical theory based on the normal distribution has the advantage that the multivariate methods based on it are extensively developed and can be studied in an organized and systematic way. This is due not only to the need for such methods because they are of practical use, but also to the fact that normal theory is amenable to exact mathematical treatment. The suitable methods of analysis are mainly based on standard operations of matrix algebra; the distributions of many statistics involved can be obtained exactly or at least characterized; and in many cases optimum properties of procedures can be deduced.

The point of view in this book is to state problems of inference in terms of the multivariate normal distributions, develop efficient and often optimum methods in this context, and evaluate significance and confidence levels in these terms. This approach gives coherence and rigor to the exposition, but, by its very nature, cannot exhaust consideration of multivariate statistical analysis. The procedures are appropriate to many nonnormal distributions,

†For a detailed study of the development of the ideas of correlation, see Walker (1931).

but their adequacy may be open to question. Roughly speaking, inferences about means are robust because of the operation of the central limit theorem, but inferences about covariances are sensitive to normality, the variability of sample covariances depending on fourth-order moments.

This inflexibility of normal methods with respect to moments of order greater than two can be reduced by including a larger class of elliptically contoured distributions. In the univariate case the normal distribution is determined by the mean and variance; higher-order moments and properties such as peakedness and long tails are functions of the mean and variance. Similarly, in the multivariate case the means and covariances or the means, variances, and correlations determine all of the properties of the distribution. That limitation is alleviated in one respect by consideration of a broad class of elliptically contoured distributions. That class maintains the dependence structure, but permits more general peakedness and long tails. This study leads to more robust methods.

The development of computer technology has revolutionized multivariate statistics in several respects. As in univariate statistics, modern computers permit the evaluation of observed variability and significance of results by resampling methods, such as the bootstrap and cross-validation. Such methodology reduces the reliance on tables of significance points as well as eliminates some restrictions of the normal distribution.

Nonparametric techniques are available when nothing is known about the underlying distributions. Space does not permit inclusion of these topics as well as other considerations of data analysis, such as treatment of outliers and transformations of variables to approximate normality and homoscedasticity.

The availability of modern computer facilities makes possible the analysis of large data sets and that ability permits the application of multivariate methods to new areas, such as image analysis, and more effective analysis of data, such as meteorological. Moreover, new problems of statistical analysis arise, such as sparseness of parameter or data matrices. Because hardware and software development is so explosive and programs require specialized knowledge, we are content to make a few remarks here and there about computation. Packages of statistical programs are available for most of the methods.

CHAPTER 2

The Multivariate Normal Distribution

2.1. INTRODUCTION

In this chapter we discuss the multivariate normal distribution and some of its properties. In Section 2.2 are considered the fundamental notions of multivariate distributions: the definition by means of multivariate density functions, marginal distributions, conditional distributions, expected values, and moments. In Section 2.3 the multivariate normal distribution is defined; the parameters are shown to be the means, variances, and covariances or the means, variances, and correlations of the components of the random vector. In Section 2.4 it is shown that linear combinations of normal variables are normally distributed and hence that marginal distributions are normal. In Section 2.5 we see that conditional distributions are also normal with means that are linear functions of the conditioning variables; the coefficients are regression coefficients. The variances, covariances, and correlations—called partial correlations—are constants. The multiple correlation coefficient is the maximum correlation between a scalar random variable and linear combination of other random variables; it is a measure of association between one variable and a set of others. The fact that marginal and conditional distributions of normal distributions are normal makes the treatment of this family of distributions coherent. In Section 2.6 the characteristic function, moments, and cumulants are discussed. In Section 2.7 elliptically contoured distributions are defined; the properties of the normal distribution are extended to this larger class of distributions.

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

2.2. NOTIONS OF MULTIVARIATE DISTRIBUTIONS

2.2.1. Joint Distributions

In this section we shall consider the notions of joint distributions of several variables, derived marginal distributions of subsets of variables, and derived conditional distributions. First consider the case of two (real) random variables† X and Y. Probabilities of events defined in terms of these variables can be obtained by operations involving the *cumulative distribution function* (abbreviated as cdf),

$$F(x,y) = \Pr\{X \le x,\ Y \le y\}, \tag{1}$$

defined for every pair of real numbers (x, y). We are interested in cases where $F(x, y)$ is absolutely continuous; this means that the following partial derivative exists almost everywhere:

$$\frac{\partial^2 F(x,y)}{\partial x\, \partial y} = f(x,y), \tag{2}$$

and

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)\, du\, dv. \tag{3}$$

The nonnegative function $f(x, y)$ is called the *density* of X and Y. The pair of random variables (X, Y) defines a random point in a plane. The probability that (X, Y) falls in a rectangle is

$$\begin{aligned} (4)\qquad & \Pr\{x \le X \le x+\Delta x,\ y \le Y \le y+\Delta y\} \\ &= F(x+\Delta x, y+\Delta y) - F(x+\Delta x, y) - F(x, y+\Delta y) + F(x,y) \\ &= \int_{y}^{y+\Delta y}\int_{x}^{x+\Delta x} f(u,v)\, du\, dv \end{aligned}$$

$(\Delta x > 0,\ \Delta y > 0)$. The probability of the random point (X, Y) falling in any set E for which the following integral is defined (that is, any measurable set E) is

$$\Pr\{(X,Y) \in E\} = \int\int_E f(x,y)\, dx\, dy. \tag{5}$$

†In Chapter 2 we shall distinguish between random variables and running variables by use of capital and lowercase letters, respectively. In later chapters we may be unable to hold to this convention because of other complications of notation.

This follows from the definition of the integral [as the limit of sums of the sort (4)]. If $f(x, y)$ is continuous in both variables, the *probability element* $f(x, y)\,\Delta y\,\Delta x$ is approximately the probability that X falls between x and $x + \Delta x$ and Y falls between y and $y + \Delta y$ since

$$(6) \qquad \Pr\{x \le X \le x + \Delta x,\, y \le Y \le y + \Delta y\} = \int_y^{y+\Delta y} \int_x^{x+\Delta x} f(u, v)\, du\, dv$$
$$= f(x_0, y_0)\,\Delta x\,\Delta y$$

for some x_0, y_0 $(x \le x_0 \le x + \Delta x,\ y \le y_0 \le y + \Delta y)$ by the mean value theorem of calculus. Since $f(u, v)$ is continuous, (6) is approximately $f(x, y)\,\Delta x\,\Delta y$. In fact,

$$(7) \qquad \lim_{\substack{\Delta y \to 0 \\ \Delta y \to 0}} \frac{1}{\Delta x\,\Delta y} \left| \Pr\{x \le X \le x + \Delta x,\ y \le Y \le y + \Delta y\} - f(x, y)\,\Delta x\,\Delta y \right| = 0.$$

Now we consider the case of p random variables $X_1, X_2, \dots, X_p$. The cdf is

$$(8) \qquad F(x_1, \dots, x_p) = \Pr\{X_1 \le x_1, \dots, X_p \le x_p\}$$

defined for every set of real numbers $x_1, \dots, x_p$. The density function, if $F(x_1, \dots, x_p)$ is absolutely continuous, is

$$(9) \qquad \frac{\partial^p F(x_1, \dots, x_p)}{\partial x_1 \cdots \partial x_p} = f(x_1, \dots, x_p)$$

(almost everywhere), and

$$(10) \qquad F(x_1, \dots, x_p) = \int_{-\infty}^{x_p} \cdots \int_{-\infty}^{x_1} f(u_1, \dots, u_p)\, du_1 \cdots du_p.$$

The probability of falling in any (measurable) set R in the p-dimensional Euclidean space is

$$(11) \qquad \Pr\{(X_1, \dots, X_p) \in R\} = \int \underset{R}{\cdots} \int f(x_1, \dots, x_p)\, dx_1 \cdots dx_p.$$

The probability element $f(x_1, \dots, x_p)\,\Delta x_1 \cdots \Delta x_p$ is approximately the probability $\Pr\{x_1 \le X_1 \le x_1 + \Delta x_1, \dots, x_p \le X_p \le x_p + \Delta x_p\}$ if $f(x_1, \dots, x_p)$ is continuous. The joint moments are defined as†

$$(12)\qquad \mathscr{E}X_1^{h_1}\cdots X_p^{h_p}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}x_1^{h_1}\cdots x_p^{h_p}f(x_1,\dots,x_p)\,dx_1\cdots dx_p.$$

2.2.2. Marginal Distributions

Given the cdf of two random variables X, Y as being $F(x, y)$, the marginal cdf of X is

$$(13)\qquad \begin{aligned}\Pr\{X\le x\}&=\Pr\{X\le x, Y\le\infty\}\\&=F(x,\infty).\end{aligned}$$

Let this be $F(x)$. Clearly

$$(14)\qquad F(x)=\int_{-\infty}^{x}\int_{-\infty}^{\infty}f(u,v)\,dv\,du.$$

We call

$$(15)\qquad \int_{-\infty}^{\infty}f(u,v)\,dv=f(u),$$

say, the *marginal density* of X. Then (14) is

$$(16)\qquad F(x)=\int_{-\infty}^{x}f(u)\,du.$$

In a similar fashion we define $G(y)$, the marginal cdf of Y, and $g(y)$, the marginal density of Y.

Now we turn to the general case. Given $F(x_1,\dots,x_p)$ as the cdf of $X_1,\dots,X_p$, we wish to find the marginal cdf of some of $X_1,\dots,X_p$, say, of $X_1,\dots,X_r$ $(r<p)$. It is

$$(17)\qquad \begin{aligned}&\Pr\{X_1\le x_1,\dots,X_r\le x_r\}\\&\quad=\Pr\{X_1\le x_1,\dots,X_r\le x_r, X_{r+1}\le\infty,\dots,X_p\le\infty\}\\&\quad=F(x_1,\dots,x_r,\infty,\dots,\infty).\end{aligned}$$

The marginal density of $X_1,\dots,X_r$ is

$$(18)\qquad \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,\dots,x_r,u_{r+1},\dots,u_p)\,du_{r+1}\cdots du_p.$$

† $\mathscr{E}$ will be used to denote *mathematical expectation.*

The marginal distribution and density of any other subset of $X_1, \ldots, X_p$ are obtained in the obviously similar fashion.

The joint moments of a subset of variates can be computed from the marginal distribution; for example,

$$\begin{aligned}(19)\quad \mathscr{E}X_1^{h_1} \cdots X_r^{h_r} &= \mathscr{E}X_1^{h_1} \cdots X_r^{h_r} X_{r+1}^0 \cdots X_p^0 \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{h_1} \cdots x_r^{h_r} f(x_1, \ldots, x_p)\, dx_1 \cdots dx_p \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{h_1} \cdots x_r^{h_r} \\ &\quad \cdot \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p)\, dx_{r+1} \cdots dx_p \right] dx_1 \cdots dx_r.\end{aligned}$$

2.2.3. Statistical Independence

Two random variables X, Y with cdf $F(x, y)$ are said to be *independent* if

$$(20)\qquad F(x,y) = F(x)G(y),$$

where $F(x)$ is the marginal cdf of X and $G(y)$ is the marginal cdf of Y. This implies that the density of X, Y is

$$\begin{aligned}(21)\qquad f(x,y) &= \frac{\partial^2 F(x,y)}{\partial x\, \partial y} = \frac{\partial^2 F(x)G(y)}{\partial x\, \partial y} \\ &= \frac{dF(x)}{dx} \frac{dG(y)}{dy} \\ &= f(x)g(y).\end{aligned}$$

Conversely, if $f(x, y) = f(x)g(y)$, then

$$\begin{aligned}(22)\qquad F(x,y) &= \int_{-\infty}^{y} \int_{-\infty}^{x} f(u,v)\, du\, dv = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u)g(v)\, du\, dv \\ &= \int_{-\infty}^{x} f(u)\, du \int_{-\infty}^{y} g(v)\, dv = F(x)G(y).\end{aligned}$$

Thus an equivalent definition of independence in the case of densities existing is that $f(x, y) = f(x)g(y)$. To see the implications of statistical independence, given any $x_1 < x_2$, $y_1 < y_2$, we consider the probability

$$\begin{aligned}(23)\qquad &\Pr\{x_1 \le X \le x_2,\ y_1 \le Y \le y_2\} \\ &\qquad = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(u,v)\, du\, dv = \int_{x_1}^{x_2} f(u)\, du \int_{y_1}^{y_2} g(v)\, dv \\ &\qquad = \Pr\{x_1 \le X \le x_2\} \Pr\{y_1 \le Y \le y_2\}.\end{aligned}$$

The probability of X falling in a given interval and Y falling in a given interval is the product of the probability of X falling in the interval and the probability of Y falling in the other interval.

If the cdf of $X_1, \dots, X_p$ is $F(x_1, \dots, x_p)$, the set of random variables is said to be *mutually independent* if

$$F(x_1, \dots, x_p) = F_1(x_1) \cdots F_p(x_p), \tag{24}$$

where $F_i(x_i)$ is the marginal cdf of X_i, $i = 1, \dots, p$. The set $X_1, \dots, X_r$ is said to be independent of the set $X_{r+1}, \dots, X_p$ if

$$F(x_1, \dots, x_p) = F(x_1, \dots, x_r, \infty, \dots, \infty) \cdot F(\infty, \dots, \infty, x_{r+1}, \dots, x_r). \tag{25}$$

One result of independence is that joint moments factor. For example, if $X_1, \dots, X_p$ are mutually independent, then

$$\begin{aligned} \mathscr{E} X_1^{h_1} \cdots X_p^{h_p} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{h_1} \cdots x_p^{h_p} f_1(x_1) \cdots f_p(x_p)\, dx_1 \cdots dx_p \\ &= \prod_{i=1}^{p} \int_{-\infty}^{\infty} x_i^{h_i} f_i(x_i)\, dx_i \\ &= \prod_{i=1}^{p} \{\mathscr{E} X_i^{h_i}\}. \end{aligned} \tag{26}$$

2.2.4. Conditional Distributions

If A and B are two events such that the probability of A and B occurring simultaneously is $P(AB)$ and the probability of B occurring is $P(B) > 0$, then the conditional probability of A occurring given that B has occurred is $P(AB)/P(B)$. Suppose the event A is X falling in the interval $[x_1, x_2]$ and the event B is Y falling in $[y_1, y_2]$. Then the conditional probability that X falls in $[x_1, x_2]$, given that Y falls in $[y_1, y_2]$, is

$$\begin{aligned} \Pr\{x_1 \le X \le x_2 | y_1 \le Y \le y_2\} &= \frac{\Pr\{x_1 \le X \le x_2,\ y_1 \le Y \le y_2\}}{\Pr\{y_1 \le Y \le y_2\}} \\ &= \frac{\int_{x_1}^{x_2} \int_{y_1}^{y_2} f(u, v)\, dv\, du}{\int_{y_1}^{y_2} g(v)\, dv}. \end{aligned} \tag{27}$$

Now let $y_1 = y$, $y_2 = y + \Delta y$. Then for a continuous density,

$$\int_{y}^{y+\Delta y} g(v)\, dv = g(y^*)\, \Delta y, \tag{28}$$

where $y \le y^* \le y + \Delta y$. Also

$$\int_y^{y+\Delta y} f(u,v)\,dv = f[u, y^*(u)]\,\Delta y, \tag{29}$$

where $y \le y^*(u) \le y + \Delta y$. Therefore,

$$\Pr\{x_1 \le X \le x_2 | y \le Y \le y + \Delta y\} = \int_{x_1}^{x_2} \frac{f[u, y^*(u)]}{g(y^*)}\,du. \tag{30}$$

It will be noticed that for fixed y and Δy (> 0), the integrand of (30) behaves as a univariate density function. Now for y such that $g(y) > 0$, we define $\Pr\{x_1 \le X \le x_2 | Y = y\}$, the probability that X lies between x_1 and x_2, given that Y is y, as the limit of (30) as $\Delta y \to 0$. Thus

$$\Pr\{x_1 \le X \le x_2 | Y = y\} = \int_{x_1}^{x_2} f(u|y)\,du, \tag{31}$$

where $f(u|y) = f(u,y)/g(y)$. For given y, $f(u|y)$ is a density function and is called the *conditional density* of X given y. We note that if X and Y are independent, $f(x|y) = f(x)$.

In the general case of $X_1, \dots, X_p$ with cdf $F(x_1, \dots, x_p)$, the conditional density of $X_1, \dots, X_r$, given $X_{r+1} = x_{r+1}, \dots, X_p = x_p$, is

$$\frac{f(x_1, \dots, x_p)}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(u_1, \dots, u_r, x_{r+1}, \dots, x_p)\,du_1 \cdots du_r}. \tag{32}$$

For a more general discussion of conditional probabilities, the reader is referred to Chung (1974), Kolmogorov (1950), Loève (1977), (1978), and Neveu (1965).

2.2.5. Transformation of Variables

Let the density of $X_1, \dots, X_p$ be $f(x_1, \dots, x_p)$. Consider the p real-valued functions

$$y_i = y_i(x_1, \dots, x_p), \qquad i = 1, \dots, p. \tag{33}$$

We assume that the transformation from the x-space to the y-space is one-to-one;[†] the inverse transformation is

$$x_i = x_i(y_1, \dots, y_p), \qquad i = 1, \dots, p. \tag{34}$$

[†] More precisely, we assume this is true for the part of the x-space for which $f(x_1, \dots, x_p)$ is positive.

Let the random variables $Y_1,\ldots,Y_p$ be defined by

$$Y_i = y_i(X_1,\ldots,X_p), \qquad i=1,\ldots,p. \tag{35}$$

Then the density of $Y_1,\ldots,Y_p$ is

$$g(y_1,\ldots,y_p) = f\left[x_1(y_1,\ldots,y_p),\ldots,x_p(y_1,\ldots,y_p)\right]J(y_1,\ldots,y_p), \tag{36}$$

where $J(y_1,\ldots,y_p)$ is the Jacobian

$$J(y_1,\ldots,y_p) = \operatorname{mod}\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \end{vmatrix}. \tag{37}$$

We assume the derivatives exist, and "mod" means modulus or absolute value of the expression following it. The probability that $(X_1,\ldots,X_p)$ falls in a region R is given by (11); the probability that $(Y_1,\ldots,Y_p)$ falls in a region S is

$$\Pr\{(Y_1,\ldots,Y_p)\in S\} = \int \underset{S}{\cdots}\int g(y_1,\ldots,y_p)\,dy_1\cdots dy_p. \tag{38}$$

If S is the transform of R, that is, if each point of R transforms by (33) into a point of S and if each point of S transforms into R by (34), then (11) is equal to (38) by the usual theory of transformation of multiple integrals. From this follows the assertion that (36) is the density of $Y_1,\ldots,Y_p$.

2.3. THE MULTIVARIATE NORMAL DISTRIBUTION

The univariate normal density function can be written

$$ke^{-\frac{1}{2}\alpha(x-\beta)^2} = ke^{-\frac{1}{2}(x-\beta)\alpha(x-\beta)}, \tag{1}$$

where α is positive and k is chosen so that the integral of (1) over the entire x-axis is unity. The density function of a multivariate normal distribution of $X_1,\ldots,X_p$ has an analogous form. The scalar variable x is replaced by a vector

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; \tag{2}$$

the scalar constant β is replaced by a vector

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}; \tag{3}$$

and the positive constant α is replaced by a positive definite (symmetric) matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}. \tag{4}$$

The square $\alpha(x-\beta)^2 = (x-\beta)\alpha(x-\beta)$ is replaced by the quadratic form

$$(x-b)'A(x-b) = \sum_{i,j=1}^{p} a_{ij}(x_i - b_i)(x_j - b_j). \tag{5}$$

Thus the density function of a p-variate normal distribution is

$$f(x_1, \ldots, x_p) = Ke^{-\frac{1}{2}(x-b)'A(x-b)}, \tag{6}$$

where K (>0) is chosen so that the integral over the entire p-dimensional Euclidean space of $x_1, \ldots, x_p$ is unity.

Written in matrix notation, the similarity of the multivariate normal density (6) to the univariate density (1) is clear. Throughout this book we shall use matrix notation and operations. The reader is referred to the Appendix for a review of matrix theory and for definitions of our notation for matrix operations.

We observe that $f(x_1, \ldots, x_p)$ is nonnegative. Since A is positive definite,

$$(x-b)'A(x-b) \geq 0, \tag{7}$$

and therefore the density is bounded; that is,

$$f(x_1, \ldots, x_p) \leq K. \tag{8}$$

Now let us determine K so that the integral of (6) over the p-dimensional space is one. We shall evaluate

$$K^* = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-b)'A(x-b)}\, dx_p \cdots dx_1. \tag{9}$$

We use the fact (see Corollary A.1.6 in the Appendix) that if A is positive definite, there exists a nonsingular matrix C such that

$$C'AC = I, \tag{10}$$

where I denotes the identity and C' the transpose of C. Let

$$x - b = Cy, \tag{11}$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}. \tag{12}$$

Then

$$(x-b)'A(x-b) = y'C'ACy = y'y. \tag{13}$$

The Jacobian of the transformation is

$$J = \operatorname{mod}|C|, \tag{14}$$

where mod$|C|$ indicates the absolute value of the determinant of C. Thus (9) becomes

$$K^* = \operatorname{mod}|C| \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}y'y}\, dy_p \cdots dy_1. \tag{15}$$

We have

$$e^{-\frac{1}{2}y'y} = \exp\left(-\frac{1}{2}\sum_{i=1}^{p} y_i^2\right) = \prod_{i=1}^{p} e^{-\frac{1}{2}y_i^2}, \tag{16}$$

where $\exp(z) = e^z$. We can write (15) as

$$\begin{aligned} K^* &= \operatorname{mod}|C| \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_1^2} \cdots e^{-\frac{1}{2}y_p^2}\, dy_p \cdots dy_1 \\ &= \operatorname{mod}|C| \prod_{i=1}^{p} \left\{ \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_i^2}\, dy_i \right\} \\ &= \operatorname{mod}|C| \prod_{i=1}^{p} \{\sqrt{2\pi}\} \\ &= \operatorname{mod}|C| (2\pi)^{\frac{1}{2}p} \end{aligned} \tag{17}$$

by virtue of

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2}\,dt = 1. \tag{18}$$

Corresponding to (10) is the determinantal equation

$$|C'|\cdot|A|\cdot|C| = |I|. \tag{19}$$

Since

$$|C'| = |C|, \tag{20}$$

and since $|I| = 1$, we deduce from (19) that

$$\operatorname{mod}|C| = 1/\sqrt{|A|}. \tag{21}$$

Thus

$$K = 1/K^* = \sqrt{|A|}\,(2\pi)^{-\frac{1}{2}p}. \tag{22}$$

The normal density function is

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{1}{2}p}} e^{-\frac{1}{2}(x-b)'A(x-b)}. \tag{23}$$

We shall now show the significance of b and A by finding the first and second moments of $X_1,\ldots,X_p$. It will be convenient to consider these random variables as constituting a random vector

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}. \tag{24}$$

We shall define generally a random matrix and the expected value of a random matrix; a random vector is considered as a special case of a random matrix with one column.

Definition 2.3.1. *A random matrix Z is a matrix*

$$Z = (Z_{gh}), \qquad g = 1,\ldots,m, \quad h = 1,\ldots,n, \tag{25}$$

of random variables $Z_{11},\ldots,Z_{mn}$.

If the random variables $Z_{11}, \ldots, Z_{mn}$ can take on only a finite number of values, the random matrix Z can be one of a finite number of matrices, say $Z(1), \ldots, Z(q)$. If the probability of $Z = Z(i)$ is p_i, then we should like to define $\mathscr{E}Z$ as $\sum_{i=1}^{q} Z(i)p_i$. Then $\mathscr{E}Z = (\mathscr{E}Z_{gh})$. If the random variables $Z_{11}, \ldots, Z_{mn}$ have a joint density, then by operating with Riemann sums we can define $\mathscr{E}Z$ as the limit (if the limit exists) of approximating sums of the kind occurring in the discrete case; then again $\mathscr{E}Z = (\mathscr{E}Z_{gh})$. Therefore, in general we shall use the following definition:

Definition 2.3.2. *The* expected value of a random *matrix* Z *is*

$$\mathscr{E}Z = (\mathscr{E}Z_{gh}), \qquad g = 1, \ldots, m, \quad h = 1, \ldots, n. \tag{26}$$

In particular if Z is X defined by (24), the expected value

$$\mathscr{E}X = \begin{pmatrix} \mathscr{E}X_1 \\ \vdots \\ \mathscr{E}X_p \end{pmatrix} \tag{27}$$

is the *mean* or *mean vector* of X. We shall usually denote this mean vector by μ. If Z is $(X-\mu)(X-\mu)'$, the expected value is

$$\mathscr{C}(X) = \mathscr{E}(X-\mu)(X-\mu)' = \left[\mathscr{E}(X_i-\mu_i)(X_j-\mu_j)\right], \tag{28}$$

the *covariance* or *covariance matrix* of X. The ith diagonal element of this matrix, $\mathscr{E}(X_i-\mu_i)^2$, is the *variance* of X_i, and the i,jth off-diagonal element, $\mathscr{E}(X_i-\mu_i)(X_j-\mu_j)$, is the *covariance* of X_i and X_j, $i \neq j$. We shall usually denote the covariance matrix by Σ. Note that

$$\mathscr{C}(X) = \mathscr{E}(XX' - \mu X' - X\mu' + \mu\mu') = \mathscr{E}XX' - \mu\mu'. \tag{29}$$

The operation of taking the expected value of a random matrix (or vector) satisfies certain rules which we can summarize in the following lemma:

Lemma 2.3.1. *If Z is an $m \times n$ random matrix, D is an $l \times m$ real matrix, E is an $n \times q$ real matrix, and F is an $l \times q$ real matrix, then*

$$\mathscr{E}(DZE + F) = D(\mathscr{E}Z)E + F. \tag{30}$$

Proof. The element in the ith row and jth column of $\mathscr{E}(\boldsymbol{DZE}+\boldsymbol{F})$ is

$$\mathscr{E}\Big(\sum_{h,g} d_{ih} Z_{hg} e_{gj} + f_{ij}\Big) = \sum_{h,g} d_{ih}(\mathscr{E} Z_{hg}) e_{gj} + f_{ij}, \tag{31}$$

which is the element in the ith row and jth column of $\boldsymbol{D}(\mathscr{E}\boldsymbol{Z})\boldsymbol{E}+\boldsymbol{F}$. ∎

Lemma 2.3.2. *If* $\boldsymbol{Y} = \boldsymbol{DX}+\boldsymbol{f}$, *where* $\boldsymbol{X}$ *is a random vector, then*

$$\mathscr{E}\boldsymbol{Y} = \boldsymbol{D}\,\mathscr{E}\boldsymbol{X}+\boldsymbol{f}, \tag{32}$$

$$\mathscr{C}(\boldsymbol{Y}) = \boldsymbol{D}\mathscr{C}(\boldsymbol{X})\boldsymbol{D}'. \tag{33}$$

Proof. The first assertion follows directly from Lemma 2.3.1, and the second from

$$\begin{aligned}\mathscr{C}(\boldsymbol{Y}) &= \mathscr{E}(\boldsymbol{Y}-\mathscr{E}\boldsymbol{Y})(\boldsymbol{Y}-\mathscr{E}\boldsymbol{Y})' \\ &= \mathscr{E}[\boldsymbol{DX}+\boldsymbol{f}-(\boldsymbol{D}\,\mathscr{E}\boldsymbol{X}+\boldsymbol{f})][\boldsymbol{DX}+\boldsymbol{f}-(\boldsymbol{D}\,\mathscr{E}\boldsymbol{X}+\boldsymbol{f})]' \\ &= \mathscr{E}[\boldsymbol{D}(\boldsymbol{X}-\mathscr{E}\boldsymbol{X})][\boldsymbol{D}(\boldsymbol{X}-\mathscr{E}\boldsymbol{X})]' \\ &= \mathscr{E}[\boldsymbol{D}(\boldsymbol{X}-\mathscr{E}\boldsymbol{X})(\boldsymbol{X}-\mathscr{E}\boldsymbol{X})'\boldsymbol{D}'],\end{aligned} \tag{34}$$

which yields the right-hand side of (33) by Lemma 2.3.1. ∎

When the transformation corresponds to (11), that is, $\boldsymbol{X} = \boldsymbol{CY}+\boldsymbol{b}$, then $\mathscr{E}\boldsymbol{X} = \boldsymbol{C}\,\mathscr{E}\boldsymbol{Y}+\boldsymbol{b}$. By the transformation theory given in Section 2.2, the density of $\boldsymbol{Y}$ is proportional to (16); that is, it is

$$\frac{1}{(2\pi)^{\frac{1}{2}p}} e^{-\frac{1}{2}\boldsymbol{y}'\boldsymbol{y}} = \prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_j^2}\right\}. \tag{35}$$

The expected value of the ith component of $\boldsymbol{Y}$ is

$$\begin{aligned}\mathscr{E}Y_i &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} y_i \prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_j^2}\right\} dy_1\cdots dy_p \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y_i e^{-\frac{1}{2}y_i^2}\,dy_i \prod_{\substack{j=1\\ j\neq i}}^{p}\left\{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_j^2}\,dy_j\right\} \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y_i e^{-\frac{1}{2}y_i^2}\,dy_i \\ &= 0.\end{aligned} \tag{36}$$

The last equality follows because† $y_i e^{-\frac{1}{2}y_i^2}$ is an odd function of y_i. Thus $\mathscr{E}Y = \mathbf{0}$. Therefore, the mean of X, denoted by $\boldsymbol{\mu}$, is

$$\boldsymbol{\mu} = \mathscr{E}X = \boldsymbol{b}. \tag{37}$$

From (33) we see that $\mathscr{C}(X) = C(\mathscr{E}YY')C'$. The i,jth element of $\mathscr{E}YY'$ is

$$\mathscr{E}Y_iY_j = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} y_i y_j \prod_{h=1}^{p} \left\{ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_h^2} \right\} dy_1 \cdots dy_p \tag{38}$$

because the density of Y is (35). If $i = j$, we have

$$\begin{aligned} \mathscr{E}Y_i^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i^2 e^{-\frac{1}{2}y_i^2}\, dy_i \prod_{\substack{h=1 \\ h \neq i}}^{p} \left\{ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_h^2}\, dy_h \right\} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i^2 e^{-\frac{1}{2}y_i^2}\, dy_i \\ &= 1. \end{aligned} \tag{39}$$

The last equality follows because the next to last expression is the expected value of the square of a variable normally distributed with mean 0 and variance 1. If $i \neq j$, (38) becomes

$$\begin{aligned} \mathscr{E}Y_iY_j &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i e^{-\frac{1}{2}y_i^2}\, dy_i \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_j e^{-\frac{1}{2}y_j^2}\, dy_j \\ &\qquad \cdot \prod_{\substack{h=1 \\ h \neq i,j}}^{p} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_h^2}\, dy_h \right\} \\ &= 0, \qquad\qquad i \neq j, \end{aligned} \tag{40}$$

since the first integration gives 0. We can summarize (39) and (40) as

$$\mathscr{E}YY' = I. \tag{41}$$

Thus

$$\mathscr{E}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})' = CIC' = CC'. \tag{42}$$

From (10) we obtain $A = (C')^{-1}C^{-1}$ by multiplication by $(C')^{-1}$ on the left and by C^{-1} on the right. Taking inverses on both sides of the equality

†Alternatively, the last equality follows because the next to last expression is the expected value of a normally distributed variable with mean 0.

gives us

$$CC' = A^{-1}. \tag{43}$$

Thus, the covariance matrix of X is

$$\Sigma = \mathscr{E}(X-\mu)(X-\mu)' = A^{-1}. \tag{44}$$

From (43) we see that Σ is positive definite. Let us summarize these results.

Theorem 2.3.1. *If the density of a p-dimensional random vector X is* (23), *then the expected value of X is b and the covariance matrix is A^{-1}. Conversely, given a vector μ and a positive definite matrix Σ, there is a multivariate normal density*

$$(2\pi)^{-\frac{1}{2}p}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} \tag{45}$$

such that the expected value of the vector with this density is μ and the covariance matrix is Σ.

We shall denote the density (45) as $n(x|\mu,\Sigma)$ and the distribution law as $N(\mu,\Sigma)$.

The ith diagonal element of the covariance matrix, σ_{ii}, is the variance of the ith component of X; we may sometimes denote this by σ_i^2. The *correlation coefficient* between X_i and X_j is defined as

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}. \tag{46}$$

This measure of association is symmetric in X_i and X_j: $\rho_{ij} = \rho_{ji}$. Since

$$\begin{pmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ji} & \sigma_{jj} \end{pmatrix} = \begin{pmatrix} \sigma_i^2 & \sigma_i\sigma_j\rho_{ij} \\ \sigma_i\sigma_j\rho_{ij} & \sigma_j^2 \end{pmatrix} \tag{47}$$

is positive definite (Corollary A.1.3 of the Appendix), the determinant

$$\begin{vmatrix} \sigma_i^2 & \sigma_i\sigma_j\rho_{ij} \\ \sigma_i\sigma_j\rho_{ij} & \sigma_j^2 \end{vmatrix} = \sigma_i^2\sigma_j^2\left(1-\rho_{ij}^2\right) \tag{48}$$

is positive. Therefore, $-1 < \rho_{ij} < 1$. (For singular distributions, see Section 2.4.) The multivariate normal density can be parametrized by the means μ_i, $i = 1,\dots,p$, the variances σ_i^2, $i = 1,\dots,p$, and the correlations ρ_{ij}, $i<j$, $i,j = 1,\dots,p$.

As a special case of the preceding theory, we consider the *bivariate* normal distribution. The mean vector is

$$\mathscr{E}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \tag{49}$$

the covariance matrix may be written

$$\Sigma = \mathscr{E}\begin{pmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 \end{pmatrix} \tag{50}$$

$$= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix},$$

where σ_1^2 is the variance of X_1, σ_2^2 the variance of X_2, and ρ the correlation between X_1 and X_2. The inverse of (50) is

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \dfrac{1}{\sigma_1^2} & -\dfrac{\rho}{\sigma_1\sigma_2} \\ -\dfrac{\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} \end{pmatrix}. \tag{51}$$

The density function of X_1 and X_2 is

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}. \tag{52}$$

Theorem 2.3.2. *The correlation coefficient ρ of any bivariate distribution is invariant with respect to transformations $X_i^* = b_iX_i + c_i$, $b_i > 0$, $i = 1, 2$. Every function of the parameters of a bivariate normal distribution that is invariant with respect to such transformations is a function of ρ.*

Proof. The variance of X_i^* is $b_i^2\sigma_i^2$, $i = 1, 2$, and the covariance of X_1^* and X_2^* is $b_1b_2\sigma_1\sigma_2\rho$ by Lemma 2.3.2. Insertion of these values into the definition of the correlation between X_1^* and X_2^* shows that it is ρ. If $f(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ is invariant with respect to such transformations, it must be $f(0, 0, 1, 1, \rho)$ by choice of $b_i = 1/\sigma_i$ and $c_i = -\mu_i/\sigma_i$, $i = 1, 2$. ■

The correlation coefficient ρ is the natural *measure of association* between X_1 and X_2. Any function of the parameters of the bivariate normal distribution that is independent of the scale and location parameters is a function of ρ. The *standardized* variable (or standard score) is $Y_i = (X_i - \mu_i)/\sigma_i$. The mean squared difference between the two standardized variables is

$$\mathscr{E}(Y_1 - Y_2)^2 = 2(1 - \rho). \tag{53}$$

The smaller (53) is (that is, the larger ρ is), the more similar Y_1 and Y_2 are. If $\rho > 0$, X_1 and X_2 tend to be positively related, and if $\rho < 0$, they tend to be negatively related. If $\rho = 0$, the density (52) is the product o' the marginal densities of X_1 and X_2; hence X_1 and X_2 are independent.

It will be noticed that the density function (45) is constant on ellipsoids

$$(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c \tag{54}$$

for every positive value of c in a p-dimensional Euclidean space. The center of each ellipsoid is at the point $\boldsymbol{\mu}$. The shape and orientation of the ellipsoid are determined by $\boldsymbol{\Sigma}$, and the size (given $\boldsymbol{\Sigma}$) is determined by c. Because (54) is a sphere if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, $n(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ is known as a *spherical* normal density.

Let us consider in detail the bivariate case of the density (52). We transform coordinates by $(x_i - \mu_i)/\sigma_i = y_i$, $i = 1, 2$, so that the centers of the loci of constant density are at the origin. These loci are defined by

$$\frac{1}{1 - \rho^2}(y_1^2 - 2\rho y_1 y_2 + y_2^2) = c. \tag{55}$$

The intercepts on the y_1-axis and y_2-axis are equal. If $\rho > 0$, the major axis of the ellipse is along the 45° line with a length of $2\sqrt{c(1 + \rho)}$, and the minor axis has a length of $2\sqrt{c(1 - \rho)}$. If $\rho < 0$, the major axis is along the 135° line with a length of $2\sqrt{c(1 - \rho)}$, and the minor axis has a length of $2\sqrt{c(1 + \rho)}$. The value of ρ determines the ratio of these lengths. In this bivariate case we can think of the density function as a surface above the plane. The contours of equal density are contours of equal altitude on a topographical map; they indicate the shape of the hill (or probability surface). If $\rho > 0$, the hill will tend to run along a line with a positive slope; most of the hill will be in the first and third quadrants. When we transform back to $x_i = \sigma_i y_i + \mu_i$, we expand each contour by a factor of σ_i in the direction of the ith axis and shift the center to (μ_1, μ_2).

The numerical values of the cdf of the univariate normal variable are obtained from tables found in most statistical texts. The numerical values of

$$F(x_1, x_2) = \Pr\{X_1 \le x_1, X_2 \le x_2\} \tag{56}$$
$$= \Pr\left\{\frac{X_1 - \mu_1}{\sigma_1} \le y_1, \frac{X_2 - \mu_2}{\sigma_2} \le y_2\right\},$$

where $y_1 = (x_1 - \mu_1)/\sigma_1$ and $y_2 = (x_2 - \mu_2)/\sigma_2$, can be found in Pearson (1931). An extensive table has been given by the National Bureau of Standards (1959). A bibliography of such tables has been given by Gupta (1963). Pearson has also shown that

$$F(x_1, x_2) = \sum_{j=0}^{\infty} \rho^j \tau_j(y_1)\tau_j(y_2), \tag{57}$$

where the so-called *tetrachoric functions* $\tau_j(y)$ are tabulated in Pearson (1930) up to $\tau_{19}(y)$. Harris and Soms (1980) have studied generalizations of (57).

2.4. THE DISTRIBUTION OF LINEAR COMBINATIONS OF NORMALLY DISTRIBUTED VARIATES; INDEPENDENCE OF VARIATES; MARGINAL DISTRIBUTIONS

One of the reasons that the study of normal multivariate distributions is so useful is that marginal distributions and conditional distributions derived from multivariate normal distributions are also normal distributions. Moreover, linear combinations of multivariate normal variates are again normally distributed. First we shall show that if we make a nonsingular linear transformation of a vector whose components have a joint distribution with a normal density, we obtain a vector whose components are jointly distributed with a normal density.

Theorem 2.4.1. *Let X (with p components) be distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then*

$$Y = CX \tag{1}$$

is distributed according to $N(C\boldsymbol{\mu}, C\boldsymbol{\Sigma}C')$ for C nonsingular.

Proof. The density of Y is obtained from the density of X, $n(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, by replacing $\boldsymbol{x}$ by

$$\boldsymbol{x} = C^{-1}\boldsymbol{y}, \tag{2}$$

and multiplying by the Jacobian of the transformation (2),

$$(3)\quad \operatorname{mod}|C^{-1}| = \frac{1}{\operatorname{mod}|C|} = \sqrt{\frac{1}{|C|^2}} = \sqrt{\frac{|\Sigma|}{|C|\cdot|\Sigma|\cdot|C'|}} = \frac{|\Sigma|^{\frac{1}{2}}}{|C\Sigma C'|^{\frac{1}{2}}}.$$

The quadratic form in the exponent of $n(x|\mu,\Sigma)$ is

$$(4)\qquad Q = (x-\mu)'\Sigma^{-1}(x-\mu).$$

The transformation (2) carries Q into

$$\begin{aligned}(5)\qquad Q &= (C^{-1}y-\mu)'\Sigma^{-1}(C^{-1}y-\mu)\\ &= (C^{-1}y-C^{-1}C\mu)'\Sigma^{-1}(C^{-1}y-C^{-1}C\mu)\\ &= [C^{-1}(y-C\mu)]'\Sigma^{-1}[C^{-1}(y-C\mu)]\\ &= (y-C\mu)'(C^{-1})'\Sigma^{-1}C^{-1}(y-C\mu)\\ &= (y-C\mu)'(C\Sigma C')^{-1}(y-C\mu)\end{aligned}$$

since $(C^{-1})' = (C')^{-1}$ by virtue of transposition of $CC^{-1} = I$. Thus the density of Y is

$$\begin{aligned}(6)\quad & n(C^{-1}y|\mu,\Sigma)\operatorname{mod}|C|^{-1}\\ &\qquad = (2\pi)^{-\frac{1}{2}p}|C\Sigma C'|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}(y-C\mu)'(C\Sigma C')^{-1}(y-C\mu)\right]\\ &\qquad = n(y|C\mu,C\Sigma C'). \qquad \blacksquare\end{aligned}$$

Now let us consider two sets of random variables $X_1,\dots,X_q$ and $X_{q+1},\dots,X_p$ forming the vectors

$$(7)\qquad X^{(1)} = \begin{pmatrix} X_1\\ \vdots \\ X_q\end{pmatrix}, \qquad X^{(2)} = \begin{pmatrix} X_{q+1}\\ \vdots \\ X_p\end{pmatrix}.$$

These variables form the random vector

$$(8)\qquad X = \begin{pmatrix} X^{(1)}\\ X^{(2)}\end{pmatrix} = \begin{pmatrix} X_1\\ \vdots \\ X_p\end{pmatrix}.$$

Now let us assume that the p variates have a joint normal distribution with mean vectors

$$(9)\qquad \mathscr{E}X^{(1)} = \mu^{(1)}, \qquad \mathscr{E}X^{(2)} = \mu^{(2)},$$

and covariance matrices

$$\mathscr{E}(X^{(1)}-\mu^{(1)})(X^{(1)}-\mu^{(1)})' = \Sigma_{11}, \tag{10}$$

$$\mathscr{E}(X^{(2)}-\mu^{(2)})(X^{(2)}-\mu^{(2)})' = \Sigma_{22}, \tag{11}$$

$$\mathscr{E}(X^{(1)}-\mu^{(1)})(X^{(2)}-\mu^{(2)})' = \Sigma_{12}. \tag{12}$$

We say that the random vector X has been partitioned in (8) into subvectors, that

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} \tag{13}$$

has been partitioned similarly into subvectors, and that

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{14}$$

has been partitioned similarly into submatrices. Here $\Sigma_{21} = \Sigma_{12}'$. (See Appendix, Section A.3.)

We shall show that $X^{(1)}$ and $X^{(2)}$ are independently normally distributed if $\Sigma_{12} = \Sigma_{21}' = 0$. Then

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}. \tag{15}$$

Its inverse is

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix}. \tag{16}$$

Thus the quadratic form in the exponent of $n(x|\mu, \Sigma)$ is

$$\begin{aligned} Q &= (x-\mu)'\Sigma^{-1}(x-\mu) \\ &= \left[(x^{(1)}-\mu^{(1)})', (x^{(2)}-\mu^{(2)})'\right] \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} x^{(1)}-\mu^{(1)} \\ x^{(2)}-\mu^{(2)} \end{pmatrix} \\ &= \left[(x^{(1)}-\mu^{(1)})'\Sigma_{11}^{-1}, (x^{(2)}-\mu^{(2)})'\Sigma_{22}^{-1}\right] \begin{pmatrix} x^{(1)}-\mu^{(1)} \\ x^{(2)}-\mu^{(2)} \end{pmatrix} \\ &= (x^{(1)}-\mu^{(1)})'\Sigma_{11}^{-1}(x^{(1)}-\mu^{(1)}) + (x^{(2)}-\mu^{(2)})'\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)}) \\ &= Q_1 + Q_2, \end{aligned} \tag{17}$$

say, where

$$
\begin{aligned}
Q_1 &= (x^{(1)} - \mu^{(1)})' \Sigma_{11}^{-1} (x^{(1)} - \mu^{(1)}), \\
Q_2 &= (x^{(2)} - \mu^{(2)})' \Sigma_{22}^{-1} (x^{(2)} - \mu^{(2)}).
\end{aligned} \tag{18}
$$

Also we note that $|\Sigma| = |\Sigma_{11}| \cdot |\Sigma_{22}|$. The density of X can be written

$$
\begin{aligned}
n(x|\mu, \Sigma) &= \frac{1}{(2\pi)^{\frac{1}{2}p} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}Q} \\
&= \frac{1}{(2\pi)^{\frac{1}{2}q} |\Sigma_{11}|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_1} \cdot \frac{1}{(2\pi)^{\frac{1}{2}(p-q)} |\Sigma_{22}|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_2} \\
&= n(x^{(1)}|\mu^{(1)}, \Sigma_{11}) n(x^{(2)}|\mu^{(2)}, \Sigma_{22}).
\end{aligned} \tag{19}
$$

The marginal density of $X^{(1)}$ is given by the integral

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} n(x|\mu, \Sigma)\, dx_{q+1} \cdots dx_p \\
&= n(x^{(1)}|\mu^{(1)}, \Sigma_{11}) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} n(x^{(2)}|\mu^{(2)}, \Sigma_{22})\, dx_{q+1} \cdots dx_p \\
&= n(x^{(1)}|\mu^{(1)}, \Sigma_{11}).
\end{aligned} \tag{20}
$$

Thus the marginal distribution of $X^{(1)}$ is $N(\mu^{(1)}, \Sigma_{11})$; similarly the marginal distribution of $X^{(2)}$ is $N(\mu^{(2)}, \Sigma_{22})$. Thus the joint density of $X_1, \dots, X_p$ is the product of the marginal density of $X_1, \dots, X_q$ and the marginal density of $X_{q+1}, \dots, X_p$, and therefore the two sets of variates are independent. Since the numbering of variates can always be done so that $X^{(1)}$ consists of any subset of the variates, we have proved the sufficiency in the following theorem:

Theorem 2.4.2. *If $X_1, \dots, X_p$ have a joint normal distribution, a necessary and sufficient condition for one subset of the random variables and the subset consisting of the remaining variables to be independent is that each covariance of a variable from one set and a variable from the other set is* 0.

The necessity follows from the fact that if X_i is from one set and X_j from the other, then for any density (see Section 2.2.3)

$$
\begin{aligned}
(21)\qquad \sigma_{ij} &= \mathscr{E}(X_i-\mu_i)(X_j-\mu_j)\\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_i-\mu_i)(x_j-\mu_j)f(x_1,\ldots,x_q)\\
&\quad\cdot f(x_{q+1},\ldots,x_p)\,dx_1\cdots dx_p\\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_i-\mu_i)f(x_1,\ldots,x_q)\,dx_1\cdots dx_q\\
&\quad\cdot\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_j-\mu_j)f(x_{q+1},\ldots,x_p)\,dx_{q+1}\cdots dx_p\\
&= 0.
\end{aligned}
$$

Since $\sigma_{ij}=\sigma_i\sigma_j\rho_{ij}$, and $\sigma_i, \sigma_j \neq 0$ (we tacitly assume that Σ is nonsingular), the condition $\sigma_{ij}=0$ is equivalent to $\rho_{ij}=0$. Thus if one set of variates is uncorrelated with the remaining variates, the two sets are independent. It should be emphasized that the implication of independence by lack of correlation depends on the assumption of normality, but the converse is always true.

Let us consider the special case of the bivariate normal distribution. Then $X^{(1)}=X_1$, $X^{(2)}=X_2$, $\mu^{(1)}=\mu_1$, $\mu^{(2)}=\mu_2$, $\Sigma_{11}=\sigma_{11}=\sigma_1^2$, $\Sigma_{22}=\sigma_{22}=\sigma_2^2$, and $\Sigma_{12}=\Sigma_{21}=\sigma_{12}=\sigma_1\sigma_2\rho_{12}$. Thus if X_1 and X_2 have a bivariate normal distribution, they are independent if and only if they are uncorrelated. If they are uncorrelated, the marginal distribution of X_i is normal with mean μ_i and variance σ_i^2. The above discussion also proves the following corollary:

Corollary 2.4.1. *If X is distributed according to $N(\mu,\Sigma)$ and if a set of components of X is uncorrelated with the other components, the marginal distribution of the set is multivariate normal with means, variances, and covariances obtained by taking the corresponding components of μ and Σ, respectively.*

Now let us show that the corollary holds even if the two sets are not independent. We partition X, μ, and Σ as before. We shall make a nonsingular linear transformation to subvectors

$$
(22)\qquad Y^{(1)}=X^{(1)}+BX^{(2)},
$$

$$
(23)\qquad Y^{(2)}=X^{(2)},
$$

choosing B so that the components of $Y^{(1)}$ are uncorrelated with the components of $Y^{(2)} = X^{(2)}$. The matrix $\boldsymbol{B}$ must satisfy the equation

$$
\begin{aligned}
(24) \qquad \mathbf{0} &= \mathscr{E}(Y^{(1)} - \mathscr{E}Y^{(1)})(Y^{(2)} - \mathscr{E}Y^{(2)})' \\
&= \mathscr{E}(X^{(1)} + EX^{(2)} - \mathscr{E}X^{(1)} - \boldsymbol{B}\,\mathscr{E}X^{(2)})(X^{(2)} - \mathscr{E}X^{(2)})' \\
&= \mathscr{E}\big[(X^{(1)} - \mathscr{E}X^{(1)}) + \boldsymbol{B}(X^{(2)} - \mathscr{E}X^{(2)})\big](X^{(2)} - \mathscr{E}X^{(2)})' \\
&= \boldsymbol{\Sigma}_{12} + \boldsymbol{B}\boldsymbol{\Sigma}_{22}.
\end{aligned}
$$

Thus $\boldsymbol{B} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ and

$$
(25) \qquad Y^{(1)} = X^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}X^{(2)}.
$$

The vector

$$
(26) \qquad \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} = Y = \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} X
$$

is a nonsingular transform of X, and therefore has a normal distribution with

$$
\begin{aligned}
(27) \qquad \mathscr{E}\begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} &= \mathscr{E}\begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} X \\
&= \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu}^{(1)} \\ \boldsymbol{\nu}^{(2)} \end{pmatrix} \\
&= \boldsymbol{\nu},
\end{aligned}
$$

say, and

$$
\begin{aligned}
(28) \quad \mathscr{C}(Y) &= \mathscr{E}(Y - \boldsymbol{\nu})(Y - \boldsymbol{\nu})' \\
&= \begin{pmatrix} \mathscr{E}(Y^{(1)} - \boldsymbol{\nu}^{(1)})(Y^{(1)} - \boldsymbol{\nu}^{(1)})' & \mathscr{E}(Y^{(1)} - \boldsymbol{\nu}^{(1)})(Y^{(2)} - \boldsymbol{\nu}^{(2)})' \\ \mathscr{E}(Y^{(2)} - \boldsymbol{\nu}^{(2)})(Y^{(1)} - \boldsymbol{\nu}^{(1)})' & \mathscr{E}(Y^{(2)} - \boldsymbol{\nu}^{(2)})(Y^{(2)} - \boldsymbol{\nu}^{(2)})' \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix}
\end{aligned}
$$

since

$$
\begin{aligned}
(29)\quad \mathscr{E}(Y^{(1)}-\boldsymbol{\nu}^{(1)})(Y^{(1)}-\boldsymbol{\nu}^{(1)})' &= \mathscr{E}\big[(X^{(1)}-\boldsymbol{\mu}^{(1)})-\Sigma_{12}\Sigma_{22}^{-1}(X^{(2)}-\boldsymbol{\mu}^{(2)})\big] \\
&\quad \cdot\big[(X^{(1)}-\boldsymbol{\mu}^{(1)})-\Sigma_{12}\Sigma_{22}^{-1}(X^{(2)}-\boldsymbol{\mu}^{(2)})\big]' \\
&= \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}+\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.
\end{aligned}
$$

Thus $Y^{(1)}$ and $Y^{(2)}$ are independent, and by Corollary 2.4.1 $X^{(2)}=Y^{(2)}$ has the marginal distribution $N(\boldsymbol{\mu}^{(2)}, \Sigma_{22})$. Because the numbering of the components of X is arbitrary, we can state the following theorem:

Theorem 2.4.3. *If X is distributed according to $N(\boldsymbol{\mu}, \Sigma)$, the marginal distribution of any set of components of X is multivariate normal with means, variances, and covariances obtained by taking the corresponding components of $\boldsymbol{\mu}$ and Σ, respectively.*

Now consider any transformation

$$
(30)\qquad Z = DX,
$$

where Z has q components and D is a $q \times p$ real matrix. The expected value of Z is

$$
(31)\qquad \mathscr{E}Z = D\boldsymbol{\mu},
$$

and the covariance matrix is

$$
(32)\qquad \mathscr{E}(Z-D\boldsymbol{\mu})(Z-D\boldsymbol{\mu})' = D\Sigma D'.
$$

The case $q=p$ and D nonsingular has been treated above. If $q \le p$ and D is of rank q, we can find a $(p-q)\times p$ matrix E such that

$$
(33)\qquad \begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} D \\ E \end{pmatrix} X
$$

is a nonsingular transformation. (See Appendix, Section A.3.) Then Z and W have a joint normal distribution, and Z has a marginal normal distribution by Theorem 2.4.3. Thus for D of rank q (and X having a nonsingular distribution, that is, a density) we have proved the following theorem:

Theorem 2.4.4. *If X is distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{X}$ is distributed according to $N(\boldsymbol{D}\boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}')$, where $\boldsymbol{D}$ is a $q \times p$ matrix of rank $q \le p$.*

The remainder of this section is devoted to the *singular* or *degenerate* normal distribution and the extension of Theorem 2.4.4 to the case of any matrix D. A singular distribution is a distribution in p-space that is concentrated on a lower dimensional set; that is, the probability associated with any set not intersecting the given set is 0. In the case of the singular normal distribution the mass is concentrated on a given linear set [that is, the intersection of a number of $(p-1)$-dimensional hyperplanes]. Let y be a set of coordinates in the linear set (the number of coordinates equaling the dimensionality of the linear set); then the *parametric* definition of the linear set can be given as $x = Ay + \lambda$, where A is a $p \times q$ matrix and λ is a p-vector. Suppose that Y is normally distributed in the q-dimensional linear set; then we say that

$$X = AY + \lambda \tag{34}$$

has a singular or degenerate normal distribution in p-space. If $\mathscr{E}Y = \boldsymbol{\nu}$, then $\mathscr{E}X = A\boldsymbol{\nu} + \lambda = \boldsymbol{\mu}$, say. If $\mathscr{E}(Y-\boldsymbol{\nu})(Y-\boldsymbol{\nu})' = T$, then

$$\mathscr{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})' = \mathscr{E}A(Y-\boldsymbol{\nu})(Y-\boldsymbol{\nu})'A' = ATA' = \Sigma, \tag{35}$$

say. It should be noticed that if $p > q$, then Σ is singular and therefore has no inverse, and thus we cannot write the normal density for X. In fact, X cannot have a density at all, because the fact that the probability of any set not intersecting the q-set is 0 would imply that the density is 0 almost everywhere.

Now, conversely, let us see that if X has mean $\boldsymbol{\mu}$ and covariance matrix Σ of rank r, it can be written as (34) (except for 0 probabilities), where X has an arbitrary distribution, and Y of r $(\le p)$ components has a suitable distribution. If Σ is of rank r, there is a $p \times p$ nonsingular matrix B such that

$$\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}' = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}, \tag{36}$$

where the identity is of order r. (See Theorem A.4.1 of the Appendix.) The transformation

$$BX = V = \begin{pmatrix} V^{(1)} \\ V^{(2)} \end{pmatrix} \tag{37}$$

defines a random vector V with covariance matrix (36) and a mean vector

$$\mathscr{E}V = B\boldsymbol{\mu} = \boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}^{(1)} \\ \boldsymbol{\nu}^{(2)} \end{pmatrix}, \tag{38}$$

say. Since the variances of the elements of $V^{(2)}$ are zero, $V^{(2)} = \boldsymbol{\nu}^{(2)}$ with probability 1. Now partition

$$B^{-1} = (C \quad D), \tag{39}$$

where C consists of r columns. Then (37) is equivalent to

$$X = B^{-1}V = (C \quad D)\begin{pmatrix} V^{(1)} \\ V^{(2)} \end{pmatrix} = CV^{(1)} + DV^{(2)}. \tag{40}$$

Thus with probability 1

$$X = CV^{(1)} + D\boldsymbol{\nu}^{(2)}, \tag{41}$$

which is of the form of (34) with C as A, $V^{(1)}$ as Y, and $D\boldsymbol{\nu}^{(2)}$ as $\boldsymbol{\lambda}$.

Now we give a formal definition of a normal distribution that includes the singular distribution.

Definition 2.4.1. *A random vector X of p components with $\mathscr{E}X = \boldsymbol{\mu}$ and $\mathscr{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})' = \Sigma$ is said to be* normally distributed [*or is said to be distributed according to $N(\boldsymbol{\mu}, \Sigma)$*] *if there is a transformation* (34), *where the number of rows of A is p and the number of columns is the rank of Σ, say r, and Y (of r components) has a nonsingular normal distribution, that is, has a density*

$$ke^{-\frac{1}{2}(y-\boldsymbol{\nu})'T^{-1}(y-\boldsymbol{\nu})}. \tag{42}$$

It is clear that if Σ has rank p, then A can be taken to be I and $\boldsymbol{\lambda}$ to be $\mathbf{0}$; then $X = Y$ and Definition 2.4.1 agrees with Section 2.3. To avoid redundancy in Definition 2.4.1 we could take $T = I$ and $\boldsymbol{\nu} = \mathbf{0}$.

Theorem 2.4.5. *If X is distributed according to $N(\boldsymbol{\mu}, \Sigma)$, then $Z = DX$ is distributed according to $N(D\boldsymbol{\mu}, D\Sigma D')$.*

This theorem includes the cases where X may have a nonsingular or a singular distribution and D may be nonsingular or of rank less than q. Since X can be represented by (34), where Y has a nonsingular distribution

$N(\boldsymbol{\nu}, \boldsymbol{T})$, we can write

$$Z = DAY + D\boldsymbol{\lambda}, \tag{43}$$

where DA is $q \times r$. If the rank of DA is r, the theorem is proved. If the rank is less than r, say s, then the covariance matrix of $\boldsymbol{Z}$,

$$DATA'D' = E, \tag{44}$$

say, is of rank s. By Theorem A.4.1 of the Appendix, there is a nonsingular matrix

$$F = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \tag{45}$$

such that

$$\begin{aligned} FEF' &= \begin{pmatrix} F_1EF_1' & F_1EF_2' \\ F_2EF_1' & F_2EF_2' \end{pmatrix} \\ &= \begin{pmatrix} (F_1DA)T(F_1DA)' & (F_1DA)T(F_2DA)' \\ (F_2DA)T(F_1DA)' & (F_2DA)T(F_2DA)' \end{pmatrix} = \begin{pmatrix} I_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \end{aligned} \tag{46}$$

Thus F_1DA is of rank s (by the converse of Theorem A.1.1 of the Appendix), and $F_2DA = \mathbf{0}$ because each diagonal element of $(F_2DA)T(F_2DA)'$ is a quadratic form in a row of F_2DA with positive definite matrix $\boldsymbol{T}$. Thus the covariance matrix of FZ is (46), and

$$FZ = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} DAY + FD\boldsymbol{\lambda} = \begin{pmatrix} F_1DAY \\ \mathbf{0} \end{pmatrix} + FD\boldsymbol{\lambda} = \begin{pmatrix} U_1 \\ \mathbf{0} \end{pmatrix} + FD\boldsymbol{\lambda}, \tag{47}$$

say. Clearly U_1 has a nonsingular normal distribution. Let $F^{-1} = (G_1 \quad G_2)$. Then

$$Z = G_1U_1 + D\boldsymbol{\lambda}, \tag{48}$$

which is of the form (34). ∎

The developments in this section can be illuminated by considering the geometric interpretation put forward in the previous section. The density of X is constant on the ellipsoids (54) of Section 2.3. Since the transformation (2) is a linear transformation (i.e., a change of coordinate axes), the density of

Y is constant on ellipsoids

$$(y - C\mu)'(C\Sigma C')^{-1}(y - C\mu) = k. \tag{49}$$

The marginal distribution of $X^{(1)}$ is the projection of the mass of the distribution of X onto the q-dimensional space of the first q coordinate axes. The surfaces of constant density are again ellipsoids. The projection of mass on any line is normal.

2.5. CONDITIONAL DISTRIBUTIONS AND MULTIPLE CORRELATION COEFFICIENT

2.5.1. Conditional Distributions

In this section we find that conditional distributions derived from joint normal distribution are normal. The conditional distributions are of a particularly simple nature because the means depend only linearly on the variates held fixed, and the variances and covariances do not depend at all on the values of the fixed variates. The theory of partial and multiple correlation discussed in this section was originally developed by Karl Pearson (1896) for three variables and extended by Yule (1897a, 1897b).

Let X be distributed according to $N(\mu, \Sigma)$ (with Σ nonsingular). Let us partition

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \tag{1}$$

as before into q- and $(p-q)$-component subvectors, respectively. We shall use the algebra developed in Section 2.4 here. The joint density of $Y^{(1)} = X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}X^{(2)}$ and $Y^{(2)} = X^{(2)}$ is

$$n\big(y^{(1)} \mid \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big)n\big(y^{(2)} \mid \mu^{(2)}, \Sigma_{22}\big).$$

The density of $X^{(1)}$ and $X^{(2)}$ then can be obtained from this expression by substituting $x^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)}$ for $y^{(1)}$ and $x^{(2)}$ for $y^{(2)}$ (the Jacobian of this transformation being 1); the resulting density of $X^{(1)}$ and $X^{(2)}$ is

(2)
$$\begin{aligned} f(x^{(1)}, x^{(2)}) = {} & \frac{1}{(2\pi)^{\frac{1}{2}q}\sqrt{|\Sigma_{11\cdot2}|}} \exp\Big\{-\tfrac{1}{2}\big[(x^{(1)} - \mu^{(1)}) - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})\big]' \\ & \cdot \Sigma_{11\cdot2}^{-1}\big[(x^{(1)} - \mu^{(1)}) - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})\big]\Big\} \\ & \cdot \frac{1}{(2\pi)^{\frac{1}{2}(p-q)}\sqrt{|\Sigma_{22}|}} \exp\big[-\tfrac{1}{2}(x^{(2)} - \mu^{(2)})'\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})\big], \end{aligned}$$

where

(3) $$\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

This density must be $n(x|\mu, \Sigma)$. The conditional density of $X^{(1)}$ given that $X^{(2)} = x^{(2)}$ is the quotient of (2) and the marginal density of $X^{(2)}$ at the point $x^{(2)}$, which is $n(x^{(2)}|\mu^{(2)}, \Sigma_{22})$, the second factor of (2). The quotient is

(4)
$$f(x^{(1)}|x^{(2)}) = \frac{1}{(2\pi)^{\frac{1}{2}q}\sqrt{|\Sigma_{11\cdot 2}|}}\exp\left\{-\tfrac{1}{2}\left[(x^{(1)} - \mu^{(1)}) - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})\right]'\right.$$
$$\left.\cdot\Sigma_{11\cdot 2}^{-1}\left[(x^{(1)} - \mu^{(1)}) - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})\right]\right\}.$$

It is understood that $x^{(2)}$ consists of $p - q$ numbers. The density $f(x^{(1)}|x^{(2)})$ is a q-variate normal density with mean

(5) $$\mathscr{E}(X^{(1)}|x^{(2)}) = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)}) = \nu(x^{(2)}),$$

say, and covariance matrix

(6) $$\mathscr{E}\left\{\left[X^{(1)} - \nu(x^{(2)})\right]\left[X^{(1)} - \nu(x^{(2)})\right]'\middle|x^{(2)}\right\} = \Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

It should be noted that the mean of $X^{(1)}$ given $x^{(2)}$ is simply a linear function of $x^{(2)}$, and the covariance matrix of $X^{(1)}$ given $x^{(2)}$ does not depend on $x^{(2)}$ at all.

Definition 2.5.1. *The matrix* $\mathbf{B} = \Sigma_{12}\Sigma_{22}^{-1}$ *is the* matrix of regression coefficients *of* $X^{(1)}$ *on* $x^{(2)}$.

The element in the ith row and $(k-q)$th column of $\mathbf{B} = \Sigma_{12}\Sigma_{22}^{-1}$ is often denoted by

(7) $$\beta_{ik\cdot q+1,\dots,k-1,k+1,\dots,p}, \qquad i = 1,\dots,q, \quad k = q+1,\dots,p.$$

The vector $\mu^{(1)} + \mathbf{B}(x^{(2)} - \mu^{(2)})$ is called the *regression function*.

Let $\sigma_{ij\cdot q+1,\dots,p}$ be the i,jth element of $\Sigma_{11\cdot 2}$. We call these *partial covariances*; $\sigma_{ii\cdot q+1,\dots,p}$ is a *partial variance*.

Definition 2.5.2

(8) $$\rho_{ij\cdot q+1,\dots,p} = \frac{\sigma_{ij\cdot q+1,\dots,p}}{\sqrt{\sigma_{ii\cdot q+1,\dots,p}}\sqrt{\sigma_{jj\cdot q+1,\dots,p}}}, \qquad i,j = 1,\dots,q,$$

is the partial correlation *between* X_i *and* X_j *holding* $X_{q+1},\dots,X_p$ *fixed*.

The numbering of the components of X is arbitrary and q is arbitrary. Hence, the above serves to define the conditional distribution of any q components of X given any other $p-q$ components. In the case of partial covariances and correlations the conditioning variables are indicated by the subscripts after the dot, and in the case of regression coefficients the dependent variable is indicated by the first subscript, the relevant conditioning variable by the second subscript, and the other conditioning variables by the subscripts after the dot. Further, the notation accommodates the conditional distribution of any q variables conditional on any other $r-q$ variables $(q \leq r \leq p)$.

Theorem 2.5.1. *Let the components of X be divided into two groups composing the subvectors $X^{(1)}$ and $X^{(2)}$. Suppose the mean $\boldsymbol{\mu}$ is similarly divided into $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$, and suppose the covariance matrix $\boldsymbol{\Sigma}$ of X is divided into $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{22}$, the covariance matrices of $X^{(1)}$, of $X^{(1)}$ and $X^{(2)}$, and of $X^{(2)}$, respectively. Then if the distribution of X is normal, the conditional distribution of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ is normal with mean $\boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x^{(2)} - \boldsymbol{\mu}^{(2)})$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.*

As an example of the above considerations let us consider the bivariate normal distribution and find the conditional distribution of X_1 given $X_2 = x_2$. In this case $\boldsymbol{\mu}^{(1)} = \mu_1$, $\boldsymbol{\mu}^{(2)} = \mu_2$, $\boldsymbol{\Sigma}_{11} = \sigma_1^2$, $\boldsymbol{\Sigma}_{12} = \sigma_1\sigma_2\rho$, and $\boldsymbol{\Sigma}_{22} = \sigma_2^2$. Thus the 1×1 matrix of regression coefficients is $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} = \sigma_1\rho/\sigma_2$, and the 1×1 matrix of partial covariances is

$$\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \sigma_1^2 - \sigma_1^2\sigma_2^2\rho^2/\sigma_2^2 = \sigma_1^2(1-\rho^2). \tag{9}$$

The density of X_1 given x_2 is $n[x_1|\mu_1 + (\sigma_1\rho/\sigma_2)(x_2 - \mu_2), \sigma_1^2(1-\rho^2)]$. The mean of this conditional distribution increases with x_2 when ρ is positive and decreases with increasing x_2 when ρ is negative. It may be noted that when $\sigma_1 = \sigma_2$, for example, the mean of the conditional distribution of x_1 does not increase relative to μ_1 as much as x_2 increases relative to μ_2. [Galton (1889) observed that the average heights of sons whose fathers' heights were above average tended to be less than the fathers' heights; he called this effect "regression towards mediocrity."] The larger $|\rho|$ is, the smaller the variance of the conditional distribution, that is, the more information x_2 gives about x_1. This is another reason for considering ρ a measure of association between X_1 and X_2.

A geometrical interpretation of the theory is enlightening. The density $f(x_1, x_2)$ can be thought of as a surface $z = f(x_1, x_2)$ over the x_1, x_2-plane. If we intersect this surface with the plane $x_2 = c$, we obtain a curve $z = f(x_1, c)$ over the line $x_2 = c$ in the x_1, x_2-plane. The ordinate of this curve is proportional to the conditional density of X_1 given $x_2 = c$; that is, it is proportional to the ordinate of the curve of a univariate normal distribution. In the more general case it is convenient to consider the ellipsoids of constant density in the p-dimensional space. Then the surfaces of constant density of $f(x_1, \dots, x_q | c_{q+1}, \dots, c_p)$ are the intersections of the surfaces of constant density of $f(x_1, \dots, x_p)$ and the hyperplanes $x_{q+1} = c_{q+1}, \dots, x_p = c_p$; these are again ellipsoids.

Further clarification of these ideas may be had by consideration of an actual population which is idealized by a normal distribution. Consider, for example, a population of father–son pairs. If the population is reasonably homogeneous, the heights of fathers and the heights of corresponding sons have approximately a normal distribution (over a certain range). A conditional distribution may be obtained by considering sons of all fathers whose height is, say, 5 feet, 9 inches (to the accuracy of measurement); the heights of these sons will have an approximate univariate normal distribution. The mean of this normal distribution will differ from the mean of the heights of sons whose fathers' heights are 5 feet, 4 inches, say, but the variances will be about the same.

We could also consider triplets of observations, the height of a father, height of the oldest son, and height of the next oldest son. The collection of heights of two sons given that the fathers' heights are 5 feet, 9 inches is a conditional distribution of two variables; the correlation between the heights of oldest and next oldest sons is a partial correlation coefficient. Holding the fathers' heights constant eliminates the effect of heredity from fathers; however, one would expect that the partial correlation coefficient would be positive, since the effect of mothers' heredity and environmental factors would tend to cause brothers' heights to vary similarly.

As we have remarked above, any conditional distribution obtained from a normal distribution is normal with the mean a linear function of the variables held fixed and the covariance matrix constant. In the case of nonnormal distributions the conditional distribution of one set of variates on another does not usually have these properties. However, one can construct nonnormal distributions such that some conditional distributions have these properties. This can be done by taking as the density of X the product $n[\boldsymbol{x}^{(1)} | \boldsymbol{\mu}^{(1)} + \boldsymbol{\beta}(x^{(2)} - \boldsymbol{\mu}^{(2)}), \boldsymbol{\Sigma}_{11\cdot 2}] f(x^{(2)})$, where $f(x^{(2)})$ is an arbitrary density.

2.5.2. The Multiple Correlation Coefficient

We again consider X partitioned into $X^{(1)}$ and $X^{(2)}$. We shall study some properties of $\boldsymbol{\beta} X^{(2)}$.

Definition 2.5.3. *The vector* $X^{(1\cdot 2)} = X^{(1)} - \mu^{(1)} - \beta(X^{(2)} - \mu^{(2)})$ *is the vector of* residuals *of* $X^{(1)}$ *from its regression on* $X^{(2)}$.

Theorem 2.5.2. *The components of* $X^{(1\cdot 2)}$ *are uncorrelated with the components of* $X^{(2)}$.

Proof. The vector $X^{(1\cdot 2)}$ is $Y^{(1)} - \mathscr{E}Y^{(1)}$ in (25) of Section 2.4. ∎

Let $\sigma'_{(i)}$ be the ith row of Σ_{12}, and $\beta'_{(i)}$ the ith row of β (i.e., $\beta'_{(i)} = \sigma'_{(i)}\Sigma_{22}^{-1}$). Let $\mathscr{V}(Z)$ be the variance of Z.

Theorem 2.5.3. *For every vector* α

$$\mathscr{V}\big(X_i^{(1\cdot 2)}\big) \le \mathscr{V}\big(X_i - \alpha' X^{(2)}\big). \tag{10}$$

Proof. By Theorem 2.5.2

$$\begin{aligned}
\mathscr{V}\big(X_i - \alpha' X^{(2)}\big) &= \mathscr{E}\big[X_i - \mu_i - \alpha'(X^{(2)} - \mu^{(2)})\big]^2 \\
&= \mathscr{E}\big[X_i^{(1\cdot 2)} - \mathscr{E}X_i^{(1\cdot 2)} + (\beta_{(i)} - \alpha)'(X^{(2)} - \mu^{(2)})\big]^2 \\
&= \mathscr{V}\big[X_i^{(1\cdot 2)}\big] + (\beta_{(i)} - \alpha)'\mathscr{E}(X^{(2)} - \mu^{(2)})(X^{(2)} - \mu^{(2)})'(\beta_{(i)} - \alpha) \\
&= \mathscr{V}\big(X_i^{(1\cdot 2)}\big) + (\beta_{(i)} - \alpha)'\Sigma_{22}(\beta_{(i)} - \alpha).
\end{aligned} \tag{11}$$

Since Σ_{22} is positive definite, the quadratic form in $\beta_{(i)} - \alpha$ is nonnegative and attains its minimum of 0 at $\alpha = \beta_{(i)}$. ∎

Since $\mathscr{E}X_i^{(1\cdot 2)} = 0$, $\mathscr{V}(X_i^{(1\cdot 2)}) = \mathscr{E}(X_i^{(1\cdot 2)})^2$. Thus $\mu_i + \beta'_{(i)}(X^{(2)} - \mu^{(2)})$ is the *best linear predictor* of X_i in the sense that of all functions of $X^{(2)}$ of the form $a'X^{(2)} + c$, the mean squared error of the above is a minimum.

Theorem 2.5.4. *For every vector* α

$$\operatorname{Corr}\big(X_i, \beta'_{(i)}X^{(2)}\big) \ge \operatorname{Corr}\big(X_i, \alpha' X^{(2)}\big). \tag{12}$$

Proof. Since the correlation between two variables is unchanged when either or both is multiplied by a positive constant, we can assume that

$\mathscr{E}[\alpha'(X^{(2)} - \mu^{(2)})]^2 = \mathscr{E}[\beta'_{(i)}(X^{(2)} - \mu^{(2)})]^2$. Then the expansion of (10) is

$$\sigma_{ii} - 2\,\mathscr{E}(X_i - \mu_i)\beta'_{(i)}(X^{(2)} - \mu^{(2)}) + \mathscr{V}\big(\beta'_{(i)}X^{(2)}\big) \tag{13}$$
$$\leq \sigma_{ii} - 2\,\mathscr{E}(X_i - \mu_i)\alpha'(X^{(2)} - \mu^{(2)}) + \mathscr{V}(\alpha' X^{(2)}).$$

This leads to

$$\frac{\mathscr{E}(X_i - \mu_i)\beta'_{(i)}(X^{(2)} - \mu^{(2)})}{\sqrt{\sigma_{ii}\mathscr{V}\big(\beta'_{(i)}X^{(2)}\big)}} \geq \frac{\mathscr{E}(X_i - \mu_i)\alpha'(X^{(2)} - \mu^{(2)})}{\sqrt{\sigma_{ii}\mathscr{V}(\alpha' X^{(2)})}}. \qquad \blacksquare \tag{14}$$

Definition 2.5.4. *The maximum correlation between X_i and the linear combination $\alpha' X^{(2)}$ is called the multiple correlation coefficient between X_i and $X^{(2)}$.*

It follows that this is

$$\bar{R}_{i\cdot q+1,\dots,p} = \frac{\mathscr{E}\beta'_{(i)}(X^{(2)} - \mu^{(2)})(X_i - \mu_i)}{\sqrt{\sigma_{ii}}\sqrt{\mathscr{E}\beta'_{(i)}(X^{(2)} - \mu^{(2)})(X^{(2)} - \mu^{(2)})'\beta_{(i)}}} \tag{15}$$
$$= \frac{\sigma'_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}}{\sqrt{\sigma_{ii}}\sqrt{\sigma'_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}}} = \frac{\sqrt{\sigma'_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}}}{\sqrt{\sigma_{ii}}}.$$

A useful formula is

$$1 - \bar{R}^2_{i\cdot q+1,\dots,p} = \frac{\sigma_{ii} - \sigma'_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}}{\sigma_{ii}} = \frac{|\Sigma_i|}{\sigma_{ii}|\Sigma_{22}|}, \tag{16}$$

where Theorem A.3.2 of the Appendix has been applied to

$$\Sigma_i = \begin{pmatrix} \sigma_{ii} & \sigma'_{(i)} \\ \sigma_{(i)} & \Sigma_{22} \end{pmatrix}. \tag{17}$$

Since

$$\sigma_{ii\cdot q+1,\dots,p} = \sigma_{ii} - \sigma'_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}, \tag{18}$$

it follows that

$$\sigma_{ii\cdot q+1,\dots,p} = \big(1 - \bar{R}^2_{i\cdot q+1,\dots,p}\big)\sigma_{ii}. \tag{19}$$

This shows incidentally that any partial variance of a component of X cannot be greater than the variance. In fact, the larger $\bar{R}_{i\cdot q+1,\dots,p}$ is, the greater the reduction in variance on going to the conditional distribution. This fact is another reason for considering the multiple correlation coefficient a measure of association between X_i and $X^{(2)}$.

That $\boldsymbol{\beta}'_{(i)} X^{(2)}$ is the best linear predictor of X_i and has the maximum correlation between X_i and linear functions of $X^{(2)}$ depends only on the covariance structure, without regard to normality. Even if X does not have a normal distribution, the regression of $X^{(1)}$ on $X^{(2)}$ can be defined by $\boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(X^{(2)} - \boldsymbol{\mu}^{(2)})$; the residuals can be defined by Definition 2.5.3; and partial covariances and correlations can be defined as the covariances and correlations of residuals yielding (3) and (8). Then these quantities do not necessarily have interpretations in terms of conditional distributions. In the case of normality $\mu_i + \boldsymbol{\beta}'_{(i)}(x^{(2)} - \boldsymbol{\mu}^{(2)})$ is the conditional expectation of X_i given $X^{(2)} = x^{(2)}$. Without regard to normality, $X_i - \mathscr{E}X_i|X^{(2)}$ is uncorrelated with any function of $X^{(2)}$, $\mathscr{E}X_i|X^{(2)}$ minimizes $\mathscr{E}[X_i - h(X^{(2)})]^2$ with respect to functions $h(X^{(2)})$ of $X^{(2)}$, and $\mathscr{E}X_i|X^{(2)}$ maximizes the correlation between X_i and functions of $X^{(2)}$. (See Problems 2.48 to 2.51.)

2.5.3. Some Formulas for Partial Correlations

We now consider relations between several conditional distributions obtained by holding several different sets of variates fixed. These relations are useful because they enable us to compute one set of conditional parameters from another set. A very special case is

$$\rho_{12\cdot 3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}; \tag{20}$$

this follows from (8) when $p = 3$ and $q = 2$. We shall now find a generalization of this result. The derivation is tedious, but is given here for completeness.

Let

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ X^{(3)} \end{pmatrix}, \tag{21}$$

where $X^{(1)}$ is of p_1 components, $X^{(2)}$ of p_2 components, and $X^{(3)}$ of p_3 components. Suppose we have the conditional distribution of $X^{(1)}$ and $X^{(2)}$ given $X^{(3)} = x^{(3)}$; how do we find the conditional distribution of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ and $X^{(3)} = x^{(3)}$? We use the fact that the conditional density of $X^{(1)}$

given $X^{(2)} = x^{(2)}$ and $X^{(3)} = x^{(3)}$ is

$$(22) \qquad f(x^{(1)}|x^{(2)}, x^{(3)}) = \frac{f(x^{(1)}, x^{(2)}, x^{(3)})}{f(x^{(2)}, x^{(3)})}$$
$$= \frac{f(x^{(1)}, x^{(2)}, x^{(3)})/f(x^{(3)})}{f(x^{(2)}, x^{(3)})/f(x^{(3)})}$$
$$= \frac{f(x^{(1)}, x^{(2)}|x^{(3)})}{f(x^{(2)}|x^{(3)})}.$$

In the case of normality the conditional covariance matrix of $X^{(1)}$ and $X^{(2)}$ given $X^{(3)} = x^{(3)}$ is

$$(23) \qquad \mathscr{C}\left[\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}\middle| x^{(3)}\right] = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1} (\Sigma_{31} \quad \Sigma_{32})$$
$$= \begin{pmatrix} \Sigma_{11\cdot 3} & \Sigma_{12\cdot 3} \\ \Sigma_{21\cdot 3} & \Sigma_{22\cdot 3} \end{pmatrix},$$

say, where

$$(24) \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix}.$$

The conditional covariance of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ and $X^{(3)} = x^{(3)}$ is calculated from the conditional covariances of $X^{(1)}$ and $X^{(2)}$ given $X^{(3)} = x^{(3)}$ as

$$(25) \qquad \mathscr{C}[X^{(1)}|x^{(2)}, x^{(3)}] = \Sigma_{11\cdot 3} - \Sigma_{12\cdot 3}(\Sigma_{22\cdot 3})^{-1}\Sigma_{21\cdot 3}.$$

This result permits the calculation of $\sigma_{ij\cdot p_1+1,\dots,p}$, $i,j = 1,\dots,p_1$, from $\sigma_{ij\cdot p_1+p_2,\dots,p}$, $i,j = 1,\dots,p_1+p_2$.

In particular, for $p_1 = q$, $p_2 = 1$, and $p_3 = p - q - 1$, we obtain

$$(26) \qquad \sigma_{ij\cdot q+1,\dots,p} = \sigma_{ij\cdot q+2,\dots,p} - \frac{\sigma_{i,q+1\cdot q+2,\dots,p}\,\sigma_{j,q+1\cdot q+2,\dots,p}}{\sigma_{q+1,q+1\cdot q+2,\dots,p}}, \qquad i,j = 1,\dots,q.$$

Since

$$(27) \qquad \sigma_{ii\cdot q+1,\dots,p} = \sigma_{ii\cdot q+2,\dots,p}\left(1 - \rho^2_{i,q+1\cdot q+2,\dots,p}\right),$$

we obtain

$$\rho_{ij\cdot q+1,\ldots,p} = \frac{\rho_{ij\cdot q+2,\ldots,p} - \rho_{i,q+1\cdot q+2,\ldots,p}\,\rho_{j,q+1\cdot q+2,\ldots,p}}{\sqrt{1-\rho^2_{i,q+1\cdot q+2,\ldots,p}}\sqrt{1-\rho^2_{j,q+1\cdot q+2,\ldots,p}}}. \tag{28}$$

This is a useful recursion formula to compute from $\{\rho_{ij}\}$ in succession $\{\rho_{ij\cdot p}\}, \{\rho_{ij\cdot p-1,p}\}, \ldots, \rho_{12\cdot 3,\ldots,p}$.

2.6. THE CHARACTERISTIC FUNCTION; MOMENTS

2.6.1. The Characteristic Function

The characteristic function of a multivariate normal distribution has a form similar to the density function. From the characteristic function, moments and cumulants can be found easily.

Definition 2.6.1. *The* characteristic function *of a random vector X is*

$$\phi(t) = \mathscr{E}e^{it'X} \tag{1}$$

defined for every real vector t.

To make this definition meaningful we need to define the expected value of a complex-valued function of a random vector.

Definition 2.6.2. *Let the complex-valued function $g(x)$ be written as $g(x) = g_1(x) + ig_2(x)$, where $g_1(x)$ and $g_2(x)$ are real-valued. Then the expected value of $g(X)$ is*

$$\mathscr{E}g(X) = \mathscr{E}g_1(X) + i\,\mathscr{E}g_2(X). \tag{2}$$

In particular, since $e^{i\theta} = \cos\theta + i\sin\theta$,

$$\mathscr{E}e^{it'X} = \mathscr{E}\cos t'X + i\,\mathscr{E}\sin t'X. \tag{3}$$

To evaluate the characteristic function of a vector X, it is often convenient to use the following lemma:

Lemma 2.6.1. *Let $X' = (X^{(1)\prime}X^{(2)\prime})$. If $X^{(1)}$ and $X^{(2)}$ are independent and $g(x) = g^{(1)}(x^{(1)})g^{(2)}(x^{(2)})$, then*

$$\mathscr{E}g(X) = \mathscr{E}g^{(1)}(X^{(1)})\,\mathscr{E}g^{(2)}(X^{(2)}). \tag{4}$$

Proof. If $g(x)$ is real-valued and X has a density,

$$
\begin{aligned}
(5)\quad \mathscr{E}g(X) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x)f(x)\,dx_1\cdots dx_p \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g^{(1)}(x^{(1)})g^{(2)}(x^{(2)})f^{(1)}(x^{(1)})f^{(2)}(x^{(2)})\,dx_1\cdots dx_p \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g^{(1)}(x^{(1)})f^{(1)}(x^{(1)})\,dx_1\cdots dx_q \\
&\quad \cdot\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g^{(2)}(x^{(2)})f^{(2)}(x^{(2)})\,dx_{q+1}\cdots dx_p \\
&= \mathscr{E}g^{(1)}(X^{(1)})\,\mathscr{E}g^{(2)}(X^{(2)}).
\end{aligned}
$$

If $g(x)$ is complex-valued,

$$
\begin{aligned}
(6)\quad g(x) &= \left[g_1^{(1)}(x^{(1)}) + ig_2^{(1)}(x^{(1)})\right]\left[g_1^{(2)}(x^{(2)}) + ig_2^{(2)}(x^{(2)})\right] \\
&= g_1^{(1)}(x^{(1)})g_1^{(2)}(x^{(2)}) - g_2^{(1)}(x^{(1)})g_2^{(2)}(x^{(2)}) \\
&\quad + i\left[g_2^{(1)}(x^{(1)})g_1^{(2)}(x^{(2)}) + g_1^{(1)}(x^{(1)})g_2^{(2)}(x^{(2)})\right].
\end{aligned}
$$

Then

$$
\begin{aligned}
(7)\quad \mathscr{E}g(X) &= \mathscr{E}\left[g_1^{(1)}(X^{(1)})g_1^{(2)}(X^{(2)}) - g_2^{(1)}(X^{(1)})g_2^{(2)}(X^{(2)})\right] \\
&\quad + i\,\mathscr{E}\left[g_2^{(1)}(X^{(1)})g_1^{(2)}(X^{(2)}) + g_1^{(1)}(X^{(1)})g_2^{(2)}(X^{(2)})\right] \\
&= \mathscr{E}g_1^{(1)}(X^{(1)})\,\mathscr{E}g_1^{(2)}(X^{(2)}) - \mathscr{E}g_2^{(1)}(X^{(1)})\,\mathscr{E}g_2^{(2)}(X^{(2)}) \\
&\quad + i\left[\mathscr{E}g_2^{(1)}(X^{(1)})\,\mathscr{E}g_1^{(2)}(X^{(2)}) + \mathscr{E}g_1^{(1)}(X^{(1)})\,\mathscr{E}g_2^{(2)}(X^{(2)})\right] \\
&= \left[\mathscr{E}g_1^{(1)}(X^{(1)}) + i\,\mathscr{E}g_2^{(1)}(X^{(1)})\right]\left[\mathscr{E}g_1^{(2)}(X^{(2)}) + i\,\mathscr{E}g_2^{(2)}(X^{(2)})\right] \\
&= \mathscr{E}g^{(1)}(X^{(1)})\,\mathscr{E}g^{(2)}(X^{(2)}). \qquad \blacksquare
\end{aligned}
$$

By applying Lemma 2.6.1 successively to $g(X) = e^{it'X}$, we derive

Lemma 2.6.2. *If the components of X are mutually independent,*

$$
(8)\qquad \mathscr{E}e^{it'X} = \prod_{j=1}^{p}\mathscr{E}e^{it_jX_j}.
$$

We now find the characteristic function of a random vector with a normal distribution.

Theorem 2.6.1. *The characteristic function of X distributed according to $N(\mu, \Sigma)$ is*

$$\phi(t) = \mathscr{E}e^{it'X} = e^{it'\mu - \frac{1}{2}t'\Sigma t} \tag{9}$$

for every real vector t.

Proof. From Corollary A.1.6 of the Appendix we know there is a nonsingular matrix C such that

$$C'\Sigma^{-1}C = I. \tag{10}$$

Thus

$$\Sigma^{-1} = C'^{-1}C^{-1} = (CC')^{-1}. \tag{11}$$

Let

$$X - \mu = CY. \tag{12}$$

Then Y is distributed according to $N(0, I)$.

Now the characteristic function of Y is

$$\psi(u) = \mathscr{E}e^{iu'Y} = \prod_{j=1}^{p} \mathscr{E}e^{iu_jY_j}. \tag{13}$$

Since Y_j is distributed according to $N(0,1)$,

$$\psi(u) = \prod_{j=1}^{p} e^{-\frac{1}{2}u_j^2} = e^{-\frac{1}{2}u'u}. \tag{14}$$

Thus

$$\begin{aligned} \phi(t) &= \mathscr{E}e^{it'X} = \mathscr{E}e^{it'(CY+\mu)} \\ &= e^{it'\mu}\,\mathscr{E}e^{it'CY} \\ &= e^{it'\mu}e^{-\frac{1}{2}(t'C)(t'C)'} \end{aligned} \tag{15}$$

for $t'C = u'$; the third equality is verified by writing both sides of it as integrals. But this is

$$\begin{aligned} \phi(t) &= e^{it'\mu}e^{-\frac{1}{2}t'CC't} \\ &= e^{it'\mu - \frac{1}{2}t'\Sigma t} \end{aligned} \tag{16}$$

by (11). This proves the theorem. ■

The characteristic function of the normal distribution is very useful. For example, we can use this method of proof to demonstrate the results of Section 2.4. If $Z = DX$, then the characteristic function of Z is

$$\mathscr{E} e^{it'Z} = \mathscr{E} e^{it'DX} = \mathscr{E} e^{i(D't)'X} \tag{17}$$
$$= e^{i(D't)'\mu - \frac{1}{2}(D't)'\Sigma(D't)}$$
$$= e^{it'(D\mu) - \frac{1}{2}t'(D\Sigma D')t},$$

which is the characteristic function of $N(D\mu, D\Sigma D')$ (by Theorem 2.6.1).

It is interesting to use the characteristic function to show that it is only the multivariate normal distribution that has the property that every linear combination of variates is normally distributed. Consider a vector Y of p components with density $f(y)$ and characteristic function

$$\psi(u) = \mathscr{E} e^{iu'Y} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{iu'y} f(y)\, dy_1 \cdots dy_p, \tag{18}$$

and suppose the mean of Y is μ and the covariance matrix is Σ. Suppose $u'Y$ is normally distributed for every u. Then the characteristic function of such linear combination is

$$\mathscr{E} e^{itu'Y} = e^{itu'\mu - \frac{1}{2}t^2 u'\Sigma u}. \tag{19}$$

Now set $t = 1$. Since the right-hand side is then the characteristic function of $N(\mu, \Sigma)$, the result is proved (by Theorem 2.6.1 above and 2.6.3 below).

Theorem 2.6.2. *If every linear combination of the components of a vector Y is normally distributed, then Y is normally distributed.*

It might be pointed out in passing that it is essential that *every* linear combination be normally distributed for Theorem 2.6.2 to hold. For instance, if $Y = (Y_1, Y_2)'$ and Y_1 and Y_2 are not independent, then Y_1 and Y_2 can each have a marginal normal distribution. An example is most easily given geometrically. Let X_1, X_2 have a joint normal distribution with means 0. Move the same mass in Figure 2.1 from rectangle A to C and from B to D. It will be seen that the resulting distribution of Y is such that the marginal distributions of Y_1 and Y_2 are the same as X_1 and X_2, respectively, which are normal, and yet the joint distribution of Y_1 and Y_2 is not normal.

This example can be used also to demonstrate that two variables, Y_1 and Y_2, can be uncorrelated and the marginal distribution of each may be normal,

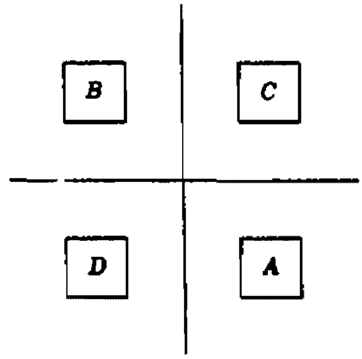

Figure 2.1

but the pair need not have a joint normal distribution and need not be independent. This is done by choosing the rectangles so that for the resultant distribution the expected value of $Y_1 Y_2$ is zero. It is clear geometrically that this can be done.

For future reference we state two useful theorems concerning characteristic functions.

Theorem 2.6.3. *If the random vector X has the density $f(x)$ and the characteristic function $\phi(t)$, then*

$$f(x) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-it'x} \phi(t)\, dt_1 \cdots dt_p. \tag{20}$$

This shows that the characteristic function determines the density function uniquely. If X does not have a density, the characteristic function uniquely defines the probability of any *continuity interval.* In the univariate case a continuity interval is an interval such that the cdf does not have a discontinuity at an endpoint of the interval.

Theorem 2.6.4. *Let $\{F_j(x)\}$ be a sequence of* cdfs, *and let $\{\phi_j(t)\}$ be the sequence of corresponding characteristic functions. A necessary and sufficient condition for $F_j(x)$ to converge to a* cdf *$F(x)$ is that, for every t, $\phi_j(t)$ converges to a limit $\phi(t)$ that is continuous at $t = 0$. When this condition is satisfied, the limit $\phi(t)$ is identical with the characteristic function of the limiting distribution $F(x)$.*

For the proofs of these two theorems, the reader is referred to Cramér (1946), Sections 10.6 and 10.7.

2.6.2. The Moments and Cumulants

The moments of $X_1, \dots, X_p$ with a joint normal distribution can be obtained from the characteristic function (9). The mean is

$$\mathscr{E}X_h = \frac{1}{i}\frac{\partial \phi}{\partial t_h}\bigg|_{t=0} \tag{21}$$

$$= \frac{1}{i}\Big\{-\sum_j \sigma_{hj}t_j + i\mu_h\Big\}\phi(t)\bigg|_{t=0}$$

$$= \mu_h.$$

The second moment is

$$\mathscr{E}X_h X_j = \frac{1}{i^2}\frac{\partial^2 \phi}{\partial t_h\, \partial t_j}\bigg|_{t=0} \tag{22}$$

$$= \frac{1}{i^2}\Big\{\Big(-\sum_k \sigma_{hk}t_k + i\mu_h\Big)\Big(-\sum_k \sigma_{kj}t_k + i\mu_j\Big) - \sigma_{hj}\Big\}\phi(t)\bigg|_{t=0}$$

$$= \sigma_{hj} + \mu_h\mu_j.$$

Thus

$$\operatorname{Variance}(X_i) = \mathscr{E}(X_i - \mu_i)^2 = \sigma_{ii}, \tag{23}$$

$$\operatorname{Covariance}(X_i, X_j) = \mathscr{E}(X_i - \mu_i)(X_j - \mu_j) = \sigma_{ij}. \tag{24}$$

Any third moment about the mean is

$$\mathscr{E}(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k) = 0. \tag{25}$$

The fourth moment about the mean is

$$\mathscr{E}(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)(X_l - \mu_l) = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}. \tag{26}$$

Every moment of odd order is 0.

Definition 2.6.3. *If all the moments of a distribution exist, then the* cumulants *are the coefficients* κ *in*

$$\log \phi(t) = \sum_{s_1, \dots, s_p = 0}^{\infty} \kappa_{s_1 \cdots s_p} \frac{(it_1)^{s_1} \cdots (it_p)^{s_p}}{s_1! \cdots s_p!}. \tag{27}$$

In the case of the multivariate normal distribution $\kappa_{10\cdots 0}=\mu_1,\dots,\kappa_{0\cdots 01}=\mu_p$, $\kappa_{20\cdots 0}=\sigma_{11},\dots,\kappa_{0\cdots 02}=\sigma_{pp}$, $\kappa_{110\cdots 0}=\sigma_{12},\dots$. The cumulants for which $\Sigma s_i>2$ are 0.

2.7. ELLIPTICALLY CONTOURED DISTRIBUTIONS

2.7.1. Spherically and Elliptically Contoured Distributions

It was noted at the end of Section 2.3 that the density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is constant on concentric ellipsoids

$$(1)\qquad (x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})=k.$$

A general class of distributions with this property is the class of *elliptically contoured distributions* with density

$$(2)\qquad |\boldsymbol{\Lambda}|^{-\frac{1}{2}}g\left[(x-\boldsymbol{\nu})'\boldsymbol{\Lambda}^{-1}(x-\boldsymbol{\nu})\right],$$

where $\boldsymbol{\Lambda}$ is a positive definite matrix, $g(\cdot)\geq 0$, and

$$(3)\qquad \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}g(y'y)\,dy_1\cdots dy_p=1.$$

If C is a nonsingular matrix such that $C'\boldsymbol{\Lambda}^{-1}C=I$, the transformation $x-\boldsymbol{\nu}=Cy$ carries the density (2) to the density $g(y'y)$. The contours of constant density of $g(y'y)$ are spheres centered at the origin. The class of such densities is known as the *spherically contoured distributions*. Elliptically contoured distributions do not necessarily have densities, but in this exposition only distributions with densities will be treated for statistical inference.

A spherically contoured density can be expressed in polar coordinates by the transformation

$$(4)\qquad \begin{aligned}
y_1&=r\sin\theta_1,\\
y_2&=r\cos\theta_1\sin\theta_2,\\
y_3&=r\cos\theta_1\cos\theta_2\sin\theta_3,\\
&\ \vdots\\
y_{p-1}&=r\cos\theta_1\cos\theta_2\cdots\cos\theta_{p-2}\sin\theta_{p-2},\\
y_p&=r\cos\theta_1\cos\theta_2\cdots\cos\theta_{p-2}\cos\theta_{p-1},
\end{aligned}$$

where $-\frac{1}{2}\pi < \theta_i \le \frac{1}{2}\pi$, $i = 1, \ldots, p-2$, $-\pi < \theta_{p-1} \le \pi$, and $0 \le r < \infty$. Note that $y'y = r^2$. The Jacobian of the transformation (4) is $r^{p-1} \cos^{p-2}\theta_1 \cos^{p-3}\theta_2 \cdots \cos\theta_{p-2}$. See Problem 7.1. If $g(y'y)$ is the density of Y, then the density of $R, \Theta_1, \ldots, \Theta_{p-1}$ is

$$r^{p-1} \cos^{p-2}\theta_1 \cos^{p-3}\theta_2 \cdots \cos\theta_{p-2}\, g(r^2). \tag{5}$$

Note that $R, \Theta_1, \ldots, \Theta_{p-1}$ are independently distributed. Since

$$\int_{-\pi/2}^{\pi/2} \cos^{h-1}\theta \, d\theta = \frac{\Gamma(\frac{1}{2}h)\Gamma(\frac{1}{2})}{\Gamma[\frac{1}{2}(h+1)]} \tag{6}$$

(Problem 7.2), the marginal density of R is

$$C(p)g(r^2)r^{p-1}, \tag{7}$$

where

$$\begin{aligned} C(p) &= \frac{2\pi^{\frac{1}{2}p}}{\Gamma(\frac{1}{2}p)} \\ &= \int_{-\pi}^{\pi}\int_{-\pi/2}^{\pi/2}\cdots\int_{-\pi/2}^{\pi/2} \cos^{p-2}\theta_1 \cos^{p-3}\theta_2 \cdots \cos\theta_{p-2}\, d\theta_1 \cdots d\theta_{p-2}\, d\theta_{p-1}. \end{aligned} \tag{8}$$

The marginal density of Θ_i is $\Gamma[\frac{1}{2}(p-i)]\cos^{p-i-1}\theta/\{\Gamma(\frac{1}{2})\Gamma[\frac{1}{2}(p-i-1)]\}$, $i = 1, \ldots, p-2$, and of θ_{p-1} is $1/(2\pi)$.

In the normal case of $N(\mathbf{0}, I)$ the density of Y is

$$g(y'y) = (2\pi)^{-\frac{1}{2}p} \exp(-\tfrac{1}{2}y'y),$$

and the density of $R = (Y'Y)^{\frac{1}{2}}$ is $r^{p-1}\exp(-\frac{1}{2}r^2)/[2^{\frac{1}{2}p-1}\Gamma(\frac{1}{2}p)]$. The density of $r^2 = v$ is $v^{\frac{1}{2}p-1}e^{-\frac{1}{2}v}/[2^{\frac{1}{2}p}\Gamma(\frac{1}{2}p)]$. This is the χ^2-density with p degrees of freedom.

The constant $C(p)$ is the surface area of a sphere of unit radius in p dimensions. The random vector U with coordinates $\sin\Theta_1, \cos\Theta_1 \sin\Theta_2, \ldots, \cos\Theta_1 \cos\Theta_2 \cdots \cos\Theta_{p-1}$, where $\Theta_1, \ldots, \Theta_{p-1}$ are independently distributed each with the uniform distribution over $(-\pi/2, \pi/2)$ except for Θ_{p-1} having the uniform distribution over $(-\pi, \pi)$, is said to be *uniformly distributed on the unit sphere*. (This is the simplest example of a spherically contoured distribution not having a density.) A stochastic representation of Y with the density $g(y'y)$ is

$$Y \overset{d}{=} RU, \tag{9}$$

where R has the density (7).

Since each of the densities of $\Theta_1, \dots, \Theta_{p-1}$ are even,

$$\mathscr{E}U = \mathbf{0}. \tag{10}$$

Because R and U are independent,

$$\mathscr{E}Y = \mathbf{0} \tag{11}$$

if $\mathscr{E}R < \infty$. Further,

$$\mathscr{E}YY' = \mathscr{E}R^2 \mathscr{E}UU' \tag{12}$$

if $\mathscr{E}R^2 < \infty$. By symmetry $\mathscr{E}U_1^2 = \cdots = \mathscr{E}U_p^2 = 1/p$ because $\Sigma_{i=1}^p U_i^2 = 1$. Again by symmetry $\mathscr{E}U_1U_2 = \mathscr{E}U_1U_3 = \cdots = \mathscr{E}U_{p-1}U_p$. In particular $\mathscr{E}U_1U_2 = \mathscr{E} \sin\Theta_1 \cos\Theta_1 \sin\Theta_2$, the integrand of which is an odd function of θ_1 and of θ_2. Hence, $\mathscr{E}U_iU_j = 0$, $i \neq j$. To summarize,

$$\mathscr{E}UU' = (1/p)\mathbf{I}_p \tag{13}$$

and

$$\mathscr{E}YY' = (1/p)\mathscr{E}R^2\mathbf{I}_p \tag{14}$$

(if $\mathscr{E}R^2 < \infty$).

The distinguishing characteristic of the class of spherically contoured distributions is that $OY \overset{d}{=} Y$ for every orthogonal matrix O.

Theorem 2.7.1. *If Y has the density $g(y'y)$, then $Z = OY$, where $O'O = I$, has the density $g(z'z)$.*

Proof. The transformation $z = Oy$ has Jacobian 1. ∎

We shall extend the definition of Y being spherically contoured to any distribution with the property $OY \overset{d}{=} Y$.

Corollary 2.7.1. *If Y is spherically contoured with stochastic representation $Y \overset{d}{=} RU$ with $R^2 = Y'Y$, then U is spherically contoured.*

Proof. If $Z = OY$ and hence $Z \overset{d}{=} Y$, and Z has the stochastic representation $Z = SV$, where $S^2 = Z'Z$, then $S = R$ and $V = OU \overset{d}{=} U$. ∎

The density of $X = \nu + CY$ is (2). From (11) and (14) we derive the following theorem:

Theorem 2.7.2. *If X has the density* (2) *and* $\mathscr{E}R^2 < \infty$,

$$\mathscr{E}X = \mu = \nu, \qquad \mathscr{C}(X) = \mathscr{E}(X-\mu)(X-\mu)' = \Sigma = (1/p)\,\mathscr{E}R^2\Lambda. \tag{15}$$

In fact if $\mathscr{E}R^m < \infty$, a moment of X of order h $(\leq m)$ is $\mathscr{E}(X_1 - \mu_1)^{h_1} \cdots (X_p - \mu_p)^{h_p} = \mathscr{E}Z_1^{h_1} \cdots Z_p^{h_p}\mathscr{E}R^h/\mathscr{E}(\chi_p^2)^{\frac{1}{2}h}$, where Z has the distribution $N(\mathbf{0}, \Sigma)$ and $h = h_1 + \cdots + h_p$.

Theorem 2.7.3. *If X has the density* (2), $\mathscr{E}R^2 < \infty$, *and* $f[c\mathscr{C}(X)] = f[\mathscr{C}(X)]$ *for all* $c > 0$, *then* $f[\mathscr{C}(X)] = f(\Sigma)$.

In particular $\rho_{ij}(X) = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} = \lambda_{ij}/\sqrt{\lambda_{ii}\lambda_{jj}}$, *where* $\Sigma = (\sigma_{ij})$ *and* $\Lambda = (\lambda_{ij})$.

2.7.2. Distributions of Linear Combinations; Marginal Distributions

First we consider a spherically contoured distribution with density $g(y'y)$. Let $y' = (y_1', y_2')$, where y_1 and y_2 have q and $p - q$ components, respectively. The marginal density of y_2 is

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1'y_1 + y_2'y_2)\, dy_1 \cdots dy_q. \tag{16}$$

Express y_1 in polar coordinates (4) with r replaced by r_1 and p replaced by q. Then the marginal density of y_2 is

$$g_2(y_2'y_2) = C(q)\int_0^{\infty} g\left(r_1^2 + y_2'y_2\right) r_1^{q-1}\, dr_1. \tag{17}$$

This expression shows that the marginal distribution of y_2 has a density which is spherically contoured.

Now consider a vector $X' = (X^{(1)\prime}, X^{(2)\prime})$ with density (2). If $\mathscr{E}R^2 < \infty$, the covariance matrix of X is (15) partitioned as (14) of Section 2.4. Let $Z^{(1)} = X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}X^{(2)} = X^{(1)} - \Lambda_{12}\Lambda_{22}^{-1}X^{(2)}$, $Z^{(2)} = X^{(2)}$, $\tau^{(1)} = \nu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\nu^{(2)} = \nu^{(1)} - \Lambda_{12}\Lambda_{22}^{-1}\nu^{(2)}$, $\tau^{(2)} = \nu^{(2)}$. Then the density of $Z' = (Z^{(1)\prime}, Z^{(2)\prime})$ is

$$\begin{aligned} &|\Lambda_{11\cdot 2}|^{-\frac{1}{2}}|\Lambda_{22}|^{-\frac{1}{2}} g\big[(z^{(1)} - \tau^{(1)})'\Lambda_{11\cdot 2}(z^{(1)} - \tau^{(1)}) \\ &\qquad + (z^{(2)} - \nu^{(2)})'\Lambda_{22}'(z^{(2)} - \nu^{(2)})\big]. \end{aligned} \tag{18}$$

Note that $Z^{(1)}$ and $Z^{(2)}$ are uncorrelated even though possibly dependent. Let C_1 and C_2 be $q \times q$ and $(p-q)\times(p-q)$ matrices satisfying $C_1\Lambda_{11\cdot 2}^{-1}C_1' = I_q$ and $C_2\Lambda_{22}^{-1}C_2' = I_{p-q}$. Define $y^{(1)}$ and $y^{(2)}$ by $z^{(1)} - \tau^{(1)} = C_1 y^{(1)}$ and $z^{(2)} - \nu^{(2)} = C_2 y^{(2)}$. Then $Y^{(1)}$ and $Y^{(2)}$ have the density $g(y^{(1)\prime}y^{(1)} + y^{(2)\prime}y^{(2)})$. The marginal density of $Y^{(2)}$ is (17), and the marginal density of $X^{(2)} = Z^{(2)}$ is

$$
(19) \qquad |\Lambda_{22}|^{-\frac{1}{2}} g_2\left[(x^{(2)} - \nu^{(2)})'\Lambda_{22}^{-1}(x^{(2)} - \nu^{(2)})\right]
$$
$$
= C(q)\int_0^\infty g\left[r_1^2 + (x^{(2)} - \nu^{(2)})'\Lambda_{22}^{-1}(x^{(2)} - \nu^{(2)}\right] r_1^{q-1}\, dr_1.
$$

The moments of Y_2 can be calculated from the moments of Y.

The generalization of Theorem 2.4.1 to elliptically contoured distributions is the following: Let X with p components have the density (2). Then $Y = CX$ has the density $|C\Lambda C'|^{-\frac{1}{2}} g[(x - C\nu)'(C\Lambda C')^{-1}(x - C\nu)]$ for C nonsingular.

The generalization of Theorem 2.4.4 is the following: If X has the density (2), then $Z = DX$ has the density

$$
(20) \qquad |D\Lambda D'|^{-\frac{1}{2}} g_2\left[(z - D\nu)'(D\Lambda D')^{-1}(z - D\nu)\right],
$$

where D is a $q \times p$ matrix of rank $q \le p$ and g_2 is given by (17).

We can also characterize marginal distributions in terms of the representation (9). Consider

$$
(21) \qquad Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} \overset{d}{=} RU = R\begin{pmatrix} U^{(1)} \\ U^{(2)} \end{pmatrix},
$$

where $Y^{(1)}$ and $U^{(1)}$ have q components and $Y^{(2)}$ and $U^{(2)}$ have $p - q$ components. Then $R_2^2 = Y^{(2)\prime}Y^{(2)}$ has the distribution of $R^2 U^{(2)\prime}U^{(2)}$, and

$$
(22) \qquad U^{(2)\prime}U^{(2)} = \frac{U^{(2)\prime}U^{(2)}}{U'U} \overset{d}{=} \frac{Y^{(2)\prime}Y^{(2)}}{Y'Y}.
$$

In the case $Y \sim N(0, I_p)$, (22) has the beta distribution, say $B(p-q, q)$, with density

$$
(23) \qquad \frac{\Gamma(p/2)}{\Gamma(q/2)\Gamma[(p-q)/2]} z^{\frac{1}{2}(p-q)-1}(1-z)^{\frac{1}{2}q-1}, \qquad 0 \le z \le 1.
$$

Hence, in general,

$$
(24) \qquad Y^{(2)} \overset{d}{=} R_2 V,
$$

where $R_2^2 \overset{d}{=} R^2 b$, $b \sim B(p-q, q)$, V has the uniform distribution of $\mathbf{v}'\mathbf{v} = 1$ in p_2 dimensions, and R^2, b, and V are independent. All marginal distributions are elliptically contoured.

2.7.3. Conditional Distributions and Multiple Correlation Coefficient

The density of the conditional distribution of $\mathbf{y}_1$ given $\mathbf{y}_2$ when $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2')'$ has the spherical density $g(\mathbf{y}'\mathbf{y})$ is

$$\frac{g(\mathbf{y}_1'\mathbf{y}_1 + \mathbf{y}_2'\mathbf{y}_2)}{g_2(\mathbf{y}_2'\mathbf{y}_2)} = \frac{g(\mathbf{y}_1'\mathbf{y}_1 + r_2^2)}{g_2(r_2^2)}, \tag{25}$$

where the marginal density $g_2(\mathbf{y}_2'\mathbf{y}_2)$ is given by (17) and $r_2^2 = \mathbf{y}_2'\mathbf{y}_2$. In terms of $\mathbf{y}_1$, (25) is a spherically contoured distribution (depending on r_2^2).

Now consider $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ with density (2). The conditional density of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ is

(26)

$$\begin{aligned}
&|\mathbf{\Lambda}_{11\cdot2}|^{-\frac{1}{2}} g\{[(\mathbf{x}^{(1)} - \mathbf{v}^{(1)})' - (\mathbf{x}^{(2)} - \mathbf{v}^{(2)})'\mathbf{B}']\mathbf{\Lambda}_{11\cdot2}^{-1}[\mathbf{x}^{(1)} - \mathbf{v}^{(1)} - \mathbf{B}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})] \\
&\qquad\qquad + (\mathbf{x}^{(2)} - \mathbf{v}^{(2)})'\mathbf{\Lambda}_{22}^{-1}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})\} \\
&\quad \div g_2[(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})'\mathbf{\Lambda}_{22}^{-1}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})] \\
&= |\mathbf{\Lambda}_{11\cdot2}|^{-\frac{1}{2}} g\{[\mathbf{x}^{(1)} - \mathbf{v}^{(1)} - \mathbf{B}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})]'\mathbf{\Lambda}_{11\cdot2}^{-1}[\mathbf{x}^{(1)} - \mathbf{v}^{(1)} - \mathbf{B}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})] + r_2^2\} \\
&\quad \div g_2(r_2^2),
\end{aligned}$$

where $r_2^2 = (\mathbf{x}^{(2)} - \mathbf{v}^{(2)})'\mathbf{\Lambda}_{22}^{-1}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})$ and $\mathbf{B} = \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}$. The density (26) is elliptically contoured in $\mathbf{x}^{(1)} - \mathbf{v}^{(1)} - \mathbf{B}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)})$ as a function of $\mathbf{x}^{(1)}$. The conditional mean of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ is

$$\mathscr{E}(\mathbf{X}^{(1)}|\mathbf{x}^{(2)}) = \mathbf{v}^{(1)} + \mathbf{B}(\mathbf{x}^{(2)} - \mathbf{v}^{(2)}) \tag{27}$$

if $\mathscr{E}(R_1^2|\mathbf{y}_2'\mathbf{y}_2 = r_2^2) < \infty$ in (25), where $R_1^2 = \mathbf{Y}_1'\mathbf{Y}_1$. Also the conditional covariance matrix is $(\mathscr{E}r_2^2/q)\mathbf{\Lambda}_{11\cdot2}$. It follows that Definition 2.5.2 of the partial correlation coefficient holds when $(\sigma_{ij\cdot q+1,\dots,p}) = \mathbf{\Sigma}_{11\cdot2} = \mathbf{\Sigma}_{11} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ and $\mathbf{\Sigma}$ is the parameter matrix given above.

Theorems 2.5.2, 2.5.3, and 2.5.4 are true for any elliptically contoured distribution for which $\mathscr{E}R^2 < \infty$.

2.7.4. The Characteristic Function; Moments

The characteristic function of a random vector $\mathbf{Y}$ with a spherically contoured distribution $\mathscr{E}e^{i\mathbf{t}'\mathbf{Y}}$ has the property of invariance over orthogonal transformations, that is,

$$\mathscr{E}e^{it'OY} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{it'Oy} g(y'y)\, dy_1 \cdots dy_p \tag{28}$$
$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{it'z} g(z'z)\, dz_1 \cdots dz_p$$
$$= \mathscr{E}e^{it'Z},$$

where $Z = OY$ also has the density $g(y'y)$. The equality (28) for all orthogonal O implies $\mathscr{E}e^{it'Z}$ is a function of $t't$. We write

$$\mathscr{E}e^{it'Y} = \phi(t't). \tag{29}$$

Then for $X = \mu + CY$

$$\mathscr{E}e^{it'X} = e^{it'\mu}\,\mathscr{E}e^{it'CY} \tag{30}$$
$$= e^{it'\mu}\phi(t'CC't)$$
$$= e^{it'\mu}\phi(t'\Lambda t)$$

when $\Lambda = CC'$. Conversely, any characteristic function of the form $e^{it'\mu}\phi(t'\Lambda t)$ corresponding to a density corresponds to a random vector X with the density (2).

The moments of X with an elliptically contoured distribution can be found from the characteristic function $e^{it'\mu}\phi(t'\Sigma t)$ or from the representation $X = \mu + RCU$, where $C'\Lambda^{-1}C = I$. Note that

$$\mathscr{E}R^2 = C(p)\int_0^{\infty} r^{p+1} g(r^2)\, dr = -2p\phi'(0), \tag{31}$$

$$\mathscr{E}R^4 = C(p)\int_0^{\infty} r^{p+3} g(r^2)\, dr = 4p(p+2)\phi''(0). \tag{32}$$

Consider the higher-order moments of $Y = RU$. The odd-order moments of R are 0, and hence the odd-order moments of Y are 0.

We have

$$\mathscr{E}(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k) = 0. \tag{33}$$

In fact, all moments of $X - \mu$ of odd order are 0.

Consider $\mathscr{E}U_iU_jU_kU_l$. Because $U'U = 1$,

$$1 = \sum_{i,j=1}^{p} \mathscr{E}U_i^2U_j^2 = p\,\mathscr{E}U_1^4 + p(p-1)\,\mathscr{E}U_1^2U_2^2. \tag{34}$$

Integration of $\mathscr{E}\sin^4\Theta_1$ gives $\mathscr{E}U_1^4 = 3/[p(p+2)]$; then (34) implies $\mathscr{E}U_1^2U_2^2 = 1/[p(p+2)]$. Hence $\mathscr{E}Y_i^4 = 3\mathscr{E}R^4/[p(p+2)]$ and $\mathscr{E}Y_1^2Y_2^2 = \mathscr{E}R^4/[p(p+2)]$. Unless $i=j=k=l$ or $i=j\neq k=l$ or $i=k\neq j=l$ or $i=l\neq j=k$, we have $\mathscr{E}U_iU_jU_kU_l = 0$. To summarize $\mathscr{E}U_iU_jU_kU_l = (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})/[p(p+2)]$. The fourth-order moments of X are

$$
\begin{aligned}
(35)\qquad \mathscr{E}(X_i-\mu_i)(X_j-\mu_j)(X_k-\mu_k)(X_l-\mu_l)
&= \frac{\mathscr{E}R^4}{p(p+2)}(\lambda_{ij}\lambda_{kl} + \lambda_{ik}\lambda_{jl} + \lambda_{il}\lambda_{jk}) \\
&= \frac{\mathscr{E}R^4}{(\mathscr{E}R^2)^2}\frac{p}{p+2}(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}).
\end{aligned}
$$

The fourth cumulant of the ith component of X standardized by its standard deviation is

$$
\begin{aligned}
(36)\qquad \frac{\mathscr{E}(X_i-\mu_i)^4}{\left[\mathscr{E}(X_i-\mu_i)^2\right]^2} - 3
&= \frac{\dfrac{3\mathscr{E}R^4}{p(p+2)} - 3\left(\dfrac{\mathscr{E}R^2}{p}\right)^2}{\left(\dfrac{\mathscr{E}R^2}{p}\right)^2} \\
&= 3\left[\frac{\mathscr{E}R^4}{(\mathscr{E}R^2)^2}\frac{p}{p+2} - 1\right] = 3\left[\frac{\phi''(0)}{[\phi'(0)]^2} - 1\right] \\
&= 3\kappa,
\end{aligned}
$$

say. This is known as the *kurtosis*. (Note that κ is $\frac{1}{3}\mathscr{E}\{(X_i-\mu_i)^4/[\mathscr{E}(X_i-\mu_i)^2]^2\} - 1$.) The standardized fourth cumulant is 3κ for every component of X. The fourth cumulant of X_i, X_j, X_k, and X_l is

(37)

$$
\begin{aligned}
\kappa_{ijkl} &= \mathscr{E}(X_i-\mu_i)(X_j-\mu_j)(X_k-\mu_k)(X_l-\mu_l) - (\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) \\
&= \kappa(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}).
\end{aligned}
$$

For the normal distribution $\kappa = 0$. The fourth-order moments can be written

$$
\begin{aligned}
(38)\qquad \mathscr{E}(X_i-\mu_i)(X_j-\mu_j)(X_k-\mu_k)(X_l-\mu_l)
&= (1+\kappa)(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}).
\end{aligned}
$$

More detail about elliptically contoured distributions can be found in Fang and Zhang (1990).

The class of elliptically contoured distributions generalizes the normal distribution, introducing more flexibility; the kurtosis is not required to be 0. The typical "bell-shaped surface" of $|\Lambda|^{-\frac{1}{2}}g[(x-\nu)'\Lambda^{-1}(x-\nu)]$ can be more or less peaked than in the case of the normal distribution. In the next subsection some examples are given.

2.7.5. Examples

(1) *The multivariate t-distribution.* Suppose $Z \sim N(\mathbf{0}, I_p)$, $ms^2 \stackrel{d}{=} \chi_m^2$, and Z and s^2 are independent. Define $Y=(1/s)Z$. Then the density of Y is

$$\frac{\Gamma\left(\frac{m+p}{2}\right)}{\Gamma\left(\frac{m}{2}\right)m^{p/2}\pi^{p/2}}\left(1+\frac{y'y}{m}\right)^{-\frac{m+p}{2}}, \tag{39}$$

and

$$\frac{R^2}{p}=\frac{\|Y\|^2}{p}\sim F_{p,m}=\frac{m}{p}\frac{\chi_p^2}{\chi_m^2}. \tag{40}$$

If $X=\mu+CY$, the density of X is

$$\frac{\Gamma\left(\frac{m+p}{2}\right)}{\Gamma\left(\frac{m}{2}\right)m^{p/2}\pi^{p/2}}|\Lambda|^{-\frac{1}{2}}\left[1+\frac{(x-\mu)'\Lambda^{-1}(x-\mu)}{m}\right]^{-\frac{1}{2}(m+p)}. \tag{41}$$

(2) *Contaminated normal.* The contaminated normal distribution is a mixture of two normal distributions with proportional covariance matrices and the same mean vector. The density can be written

$$\begin{aligned}(1-\varepsilon)&\frac{1}{(2\pi)^{p/2}|\Lambda|^{\frac{1}{2}}}e^{-\frac{1}{2}(x-\mu)'\Lambda^{-1}(x-\mu)}\\&+\varepsilon\frac{1}{(2\pi)^{p/2}|c\Lambda|^{\frac{1}{2}}}e^{-(1/2c)(x-\mu)'\Lambda^{-1}(x-\mu)},\end{aligned} \tag{42}$$

where $c>0$ and $0\le\varepsilon\le1$. Usually ε is rather small and c rather large.

(3) *Mixtures of normal distributions.* Let $w(v)$ be a cumulative distribution function over $0\le v\le\infty$. Then a mixture of normal densities is defined by

$$\int_0^\infty n\left(x|\mu,\frac{1}{v^2}\Sigma\right)dw(v), \tag{43}$$

which is an elliptically contoured density. The random vector X with this density has a representation $X = wZ$, where $Z \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $w \sim w(w)$ are independent.

Fang, Kotz, and Ng (1990) have discussed (43) and have given other examples of elliptically contoured distributions.

PROBLEMS

2.1. (Sec. 2.2) Let
$$f(x,y) = 1, \quad 0 \le x \le 1,\ 0 \le y \le 1,$$
$$= 0, \quad \text{otherwise.}$$

Find:

(a) $F(x, y)$.

(b) $F(x)$.

(c) $f(x)$.

(d) $f(x|y)$. [Note: $f(x_0|y_0) = 0$ if $f(x_0, y_0) = 0$.]

(e) $\mathscr{E}X^nY^m$.

(f) Prove X and Y are independent.

2.2. (Sec. 2.2) Let
$$f(x,y) = 2, \quad 0 \le y \le x \le 1,$$
$$= 0, \quad \text{otherwise.}$$

Find:

(a) $F(x, y)$.

(b) $F(x)$.

(c) $f(x)$.

(d) $G(y)$.

(e) $g(y)$.

(f) $f(x|y)$.

(g) $f(y|x)$.

(h) $\mathscr{E}X^nY^m$.

(i) Are X and Y independent?

2.3. (Sec. 2.2) Let $f(x, y) = C$ for $x^2 + y^2 \le k^2$ and 0 elsewhere. Prove $C = 1/(\pi k^2)$, $\mathscr{E}X = \mathscr{E}Y = 0$, $\mathscr{E}X^2 = \mathscr{E}Y^2 = k^2/4$, and $\mathscr{E}XY = 0$. Are X and Y independent?

2.4. (Sec. 2.2) Let $F(x_1, x_2)$ be the joint cdf of X_1, X_2, and let $F_i(x_i)$ be the marginal cdf of X_i, $i = 1, 2$. Prove that if $F_i(x_i)$ is continuous, $i = 1, 2$, then $F(x_1, x_2)$ is continuous.

2.5. (Sec. 2.2) Show that if the set $X_1, \dots, X_r$ is independent of the set $X_{r+1}, \dots, X_p$, then

$$\mathscr{E}g(X_1, \dots, X_r)h(X_{r+1}, \dots, X_p) = \mathscr{E}g(X_1, \dots, X_r)\,\mathscr{E}h(X_{r+1}, \dots, X_p).$$

2.6. (Sec. 2.3) Sketch the ellipses $f(x, y) = 0.06$, where $f(x, y)$ is the bivariate normal density with

(a) $\mu_x = 1$, $\mu_y = 2$, $\sigma_x^2 = 1$, $\sigma_y^2 = 1$, $\rho_{xy} = 0$.
(b) $\mu_x = 0$, $\mu_y = 0$, $\sigma_x^2 = 1$, $\sigma_y^2 = 1$, $\rho_{xy} = 0$.
(c) $\mu_x = 0$, $\mu_y = 0$, $\sigma_x^2 = 1$, $\sigma_y^2 = 1$, $\rho_{xy} = 0.2$.
(d) $\mu_x = 0$, $\mu_y = 0$, $\sigma_x^2 = 1$, $\sigma_y^2 = 1$, $\rho_{xy} = 0.8$.
(e) $\mu_x = 0$, $\mu_y = 0$, $\sigma_x^2 = 4$, $\sigma_y^2 = 1$, $\rho_{xy} = 0.8$.

2.7. (Sec. 2.3) Find $\boldsymbol{b}$ and A so that the following densities can be written in the form of (23). Also find μ_x, μ_y, σ_x, σ_y and ρ_{xy}.

(a) $\dfrac{1}{2\pi}\exp\{-\tfrac{1}{2}[(x-1)^2 + (y-2)^2]\}$.

(b) $\dfrac{1}{2.4\pi}\exp\left(-\dfrac{x^2/4 - 1.6xy/2 + y^2}{0.72}\right)$.

(c) $\dfrac{1}{2\pi}\exp[-\tfrac{1}{2}(x^2 + y^2 + 4x - 6y + 13)]$.

(d) $\dfrac{1}{2\pi}\exp[-\tfrac{1}{2}(2x^2 + y^2 + 2xy - 22x - 14y + 65)]$.

2.8. (Sec. 2.3) For each matrix A in Problem 2.7 find C so that $C'AC = I$.

2.9. (Sec. 2.3) Let $\boldsymbol{b} = \boldsymbol{0}$.

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

(a) Write the density (23).
(b) Find $\boldsymbol{\Sigma}$.

2.10. (Sec. 2.3) Prove that the principal axes of (55) of Section 2.3 are along the 45° and 135° lines with lengths $2\sqrt{c(1+\rho)}$ and $2\sqrt{c(1-\rho)}$, respectively, by transforming according to $y_1 = (z_1 + z_2)/\sqrt{2}$, $y_2 = (z_1 - z_2)/\sqrt{2}$.

2.11. (Sec. 2.3) Suppose the scalar random variables $X_1, \dots, X_n$ are independent and have a density which is a function only of $x_1^2 + \cdots + x_n^2$. Prove that the X_i are normally distributed with mean 0 and common variance. Indicate the mildest conditions on the density for your proof.

2.12. (Sec. 2.3) Show that if $\Pr\{X \ge 0, Y \ge 0\} = \alpha$ for the distribution

$$N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right],$$

then $\rho = \cos(1 - 2\alpha)\pi$. [*Hint:* Let $X = U, Y = \rho U + \sqrt{1-\rho^2}\,V$ and verify $\rho = \cos 2\pi(\frac{1}{2} - \alpha)$ geometrically.]

2.13. (Sec. 2.3) Prove that if $\rho_{ij} = \rho$, $i \ne j$, $i, j = 1, \dots, p$, then $\rho \ge -1/(p-1)$.

2.14. (Sec. 2.3) *Concentration ellipsoid.* Let the density of the p-component $\boldsymbol{Y}$ be $f(y) = \Gamma(\frac{1}{2}p + 1)/[(p+2)\pi]^{\frac{1}{2}p}$ for $y'y \le p + 2$ and 0 elsewhere. Then $\mathscr{E}\boldsymbol{Y} = \mathbf{0}$ and $\mathscr{E}\boldsymbol{Y}\boldsymbol{Y}' = \boldsymbol{I}$ (Problem 7.4). From this result prove that if the density of $\boldsymbol{X}$ is $g(\boldsymbol{x}) = \sqrt{|A|}\,\Gamma(\frac{1}{2}p + 1)/[(p+2)\pi]^{\frac{1}{2}p}$ for $(\boldsymbol{x} - \boldsymbol{\mu})'A(\boldsymbol{x} - \boldsymbol{\mu}) \le p + 2$ and 0 elsewhere, then $\mathscr{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\mathscr{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = A^{-1}$.

2.15. (Sec. 2.4) Show that when $\boldsymbol{X}$ is normally distributed the components are mutually independent if and only if the covariance matrix is diagonal.

2.16. (Sec. 2.4) Find necessary and sufficient conditions on A so that $A\boldsymbol{Y} + \boldsymbol{\lambda}$ has a continuous cdf.

2.17. (Sec. 2.4) Which densities in Problem 2.7 define distributions in which X and Y are independent?

2.18. (Sec. 2.4)

(a) Write the marginal density of X for each case in Problem 2.6.
(b) Indicate the marginal distribution of X for each case in Problem 2.7 by the notation $N(a, b)$.
(c) Write the marginal density of X_1 and X_2 in Problem 2.9.

2.19. (Sec. 2.4) What is the distribution of $Z = X - Y$ when X and Y have each of the densities in Problem 2.6?

2.20. (Sec. 2.4) What is the distribution of $X_1 + 2X_2 - 3X_3$ when X_1, X_2, X_3 have the distribution defined in Problem 2.9?

2.21. (Sec. 2.4) Let $\boldsymbol{X} = (X_1, X_2)'$, where $X_1 = X$ and $X_2 = aX + b$ and X has the distribution $N(0, 1)$. Find the cdf of $\boldsymbol{X}$.

2.22. (Sec. 2.4) Let $X_1, \dots, X_N$ be independently distributed, each according to $N(\mu, \sigma^2)$.

(a) What is the distribution of $\boldsymbol{X} = (X_1, \dots, X_N)'$? Find the vector of means and the covariance matrix.
(b) Using Theorem 2.4.4, find the marginal distribution of $\bar{X} = \Sigma X_i/N$.

2.23. (Sec. 2.4) Let $X_1, \dots, X_N$ be independently distributed with X_i having distribution $N(\beta + \gamma z_i, \sigma^2)$, where z_i is a given number, $i = 1, \dots, N$, and $\Sigma_i z_i = 0$.

(a) Find the distribution of $(X_1, \dots, X_N)'$.
(b) Find the distribution of $\bar{X}$ and $g = \Sigma X_i z_i / \Sigma z_i^2$ for $\Sigma z_i^2 > 0$.

2.24. (Sec. 2.4) Let $(X_1, Y_1)', (X_2, Y_2)', (X_3, Y_3)'$ be independently distributed, $(X_i, Y_i)'$ according to

$$N\left[\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}\right], \qquad i = 1, 2, 3.$$

(a) Find the distribution of the six variables.
(b) Find the distribution of $(\bar{X}, \bar{Y})'$.

2.25. (Sec. 2.4) Let X have a (singular) normal distribution with mean $\mathbf{0}$ and covariance matrix

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

(a) Prove $\mathbf{\Sigma}$ is of rank 1.
(b) Find $\boldsymbol{a}$ so $\boldsymbol{X} = \boldsymbol{a}'Y$ and Y has a nonsingular normal distribution, and give the density of Y.

2.26. (Sec. 2.4) Let

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

(a) Find a vector $\boldsymbol{u} \neq \mathbf{0}$ so that $\mathbf{\Sigma}\boldsymbol{u} = \mathbf{0}$. [*Hint:* Take cofactors of any column.]
(b) Show that any matrix of the form $G = (H \;\; \boldsymbol{u})$, where H is 3×2, has the property

$$G'\mathbf{\Sigma}G = \begin{pmatrix} H'\mathbf{\Sigma}H & 0 \\ \mathbf{0} & 0 \end{pmatrix}.$$

(c) Using (a) and (b), find $\boldsymbol{B}$ to satisfy (36).
(d) Find $\boldsymbol{B}^{-1}$ and partition according to (39).
(e) Verify that $CC' = \mathbf{\Sigma}$.

2.27. (Sec. 2.4) Prove that if the joint (marginal) distribution of X_1 and X_2 is singular (that is, degenerate), then the joint distribution of X_1, X_2, and X_3 is singular.

2.28. (Sec. 2.5) In each part of Problem 2.6, find the conditional distribution of X given $Y=y$, find the conditional distribution of Y given $X=x$, and plot each regression line on the appropriate graph in Problem 2.6.

2.29. (Sec. 2.5) Let $\boldsymbol{\mu}=\mathbf{0}$ and

$$\boldsymbol{\Sigma}=\begin{pmatrix} 1. & 0.80 & -0.40 \\ 0.80 & 1. & -0.56 \\ -0.40 & -0.56 & 1. \end{pmatrix}.$$

(a) Find the conditional distribution of X_1 and X_3, given $X_2=x_2$.

(b) What is the partial correlation between X_1 and X_3 given X_2?

2.30. (Sec. 2.5) In Problem 2.9, find the conditional distribution of X_1 and X_2 given $X_3=x_3$.

2.31. (Sec. 2.5) Verify (20) directly from Theorem 2.5.1.

2.32. (Sec. 2.5)

(a) Show that finding $\boldsymbol{\alpha}$ to maximize the absolute value of the correlation between X_i and $\boldsymbol{\alpha}'\boldsymbol{X}^{(2)}$ is equivalent to maximizing $(\boldsymbol{\sigma}'_{(i)}\boldsymbol{\alpha})^2$ subject to $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{22}\boldsymbol{\alpha}$ constant.

(b) Find $\boldsymbol{\alpha}$ by maximizing $(\boldsymbol{\sigma}'_{(i)}\boldsymbol{\alpha})^2-\lambda(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{22}\boldsymbol{\alpha}-c)$, where c is a constant and λ is a Lagrange multiplier.

2.33. (Sec. 2.5) *Invariance of the multiple correlation coefficient.* Prove that $\bar{R}_{i\cdot q+1,\dots,p}$ is an invariant characteristic of the multivariate normal distribution of X_i and $\boldsymbol{X}^{(2)}$ under the transformation $x_i^*=b_ix_i+c_i$ for $b_i\neq 0$ and $\boldsymbol{X}^{(2)*}=\boldsymbol{H}\boldsymbol{X}^{(2)}+\boldsymbol{k}$ for $\boldsymbol{H}$ nonsingular and that every function of μ_i, σ_{ii}, $\boldsymbol{\sigma}_{(i)}$, $\boldsymbol{\mu}^{(2)}$, and $\boldsymbol{\Sigma}_{22}$ that is invariant is a function of $\bar{R}_{i\cdot q+1,\dots,p}$.

2.34. (Sec. 2.5) Prove that

$$1-\bar{R}^2_{i\cdot q+1,\dots,p}=\frac{1}{|\rho_{kj}|}\begin{vmatrix} 1 & \rho_{ij} \\ \rho_{ki} & \rho_{kj} \end{vmatrix}, \qquad k,j=q+1,\dots,p.$$

2.35. (Sec. 2.5) Find the multiple correlation coefficient between X_1 and (X_2, X_3) in Problem 2.29.

2.36. (Sec. 2.5) Prove explicitly that if $\boldsymbol{\Sigma}$ is positive definite,

$$|\boldsymbol{\Sigma}|=|\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|\cdot|\boldsymbol{\Sigma}_{22}|.$$

2.37. (Sec. 2.5) Prove *Hadamard's inequality*

$$|\boldsymbol{\Sigma}| \le \prod_{i=1}^{p} \sigma_{ii}.$$

[*Hint:* Using Problem 2.36, prove $|\boldsymbol{\Sigma}| \le \sigma_{11}|\boldsymbol{\Sigma}_{22}|$, where $\boldsymbol{\Sigma}_{22}$ is $(p-1)\times(p-1)$, and apply induction.]

2.38. (Sec. 2.5) Prove equality holds in Problem 2.37 if and only if $\boldsymbol{\Sigma}$ is diagonal.

2.39. (Sec. 2.5) Prove $\beta_{12\cdot3} = \sigma_{12\cdot3}/\sigma_{22\cdot3} = \rho_{13\cdot2}\sigma_{1\cdot2}/\sigma_{3\cdot2}$ and $\beta_{13\cdot2} = \sigma_{13\cdot2}/\sigma_{33\cdot2} = \rho_{13\cdot2}\sigma_{1\cdot2}/\sigma_{3\cdot2}$, where $\sigma_{i\cdot k}^2 = \sigma_{ii\cdot k}$.

2.40. (Sec. 2.5) Let (X_1, X_2) have the density $n(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{\Sigma}) = f(x_1, x_2)$. Let the density of X_2 given $X_1 = x_1$ be $f(x_2|x_1)$. Let the joint density of X_1, X_2, X_3 be $f(x_1, x_2)f(x_3|x_1)$. Find the covariance matrix of X_1, X_2, X_3 and the partial correlation between X_2 and X_3 for given X_1.

2.41. (Sec. 2.5) Prove $1 - \overline{R}_{1\cdot23}^2 = (1-\rho_{13}^2)(1-\rho_{12\cdot3}^2)$. [*Hint:* Use the fact that the variance of X_1 in the conditional distribution given x_2 and x_3 is $(1-\overline{R}_{1\cdot23}^2)\sigma_{11}$.]

2.42. (Sec. 2.5) If $p = 2$, can there be a difference between the simple correlation between X_1 and x_2 and the multiple correlation between X_1 and $X^{(2)} = X_2$? Explain.

2.43. (Sec. 2.5) Prove

$$\beta_{ik\cdot q+1,\dots,k-1,k+1,\dots,p} = \frac{\sigma_{ik\cdot q-1,\dots,k-1,k+1,\dots,p}}{\sigma_{kk\cdot q+1,\dots,k-1,k+1,\dots,p}}$$

$$= \rho_{ik\cdot q-1,\dots,k-1,k+1,\dots,p}\frac{\sigma_{i\cdot q+1,\dots,k-1,k+1,\dots,p}}{\sigma_{k\cdot q+1,\dots,k-1,k+1,\dots,p}},$$

$i = 1,\dots,q$, $k = q+1,\dots,p$, where $\sigma_{j\cdot q+1,\dots,k-1,k+1,\dots,p}^2 = \sigma_{jj\cdot q+1,\dots,k-1,k+1,\dots,p}$, $j = i, k$. [*Hint:* Prove this for the special case $k = q+1$ by using Problem 2.56 with $p_1 = q$, $p_2 = 1$, $p_3 = p - q - 1$.]

2.44. (Sec. 2.5) Give a necessary and sufficient condition for $\overline{R}_{i\cdot q+1,\dots,p} = 0$ in terms of $\sigma_{i,q+1},\dots,\sigma_{ip}$.

2.45. (Sec. 2.5) Show

$$1 - \overline{R}_{i\cdot q+1,\dots,p}^2 = \left(1-\rho_{ip}^2\right)\left(1-\rho_{i,p-1\cdot p}^2\right)\cdots\left(1-\rho_{i,q-1\cdot q+2,\dots,p}^2\right).$$

[*Hint:* Use (19) and (27) successively.]

2.46. (Sec. 2.5) Show

$$\rho^2_{ij\cdot q+1,\dots,p} = \beta_{ij\cdot q+1,\dots,p}\,\beta_{ji\cdot q+1,\dots,p}.$$

2.47. (Sec. 2.5) Prove

$$\rho_{12\cdot 3\cdots p} = \frac{-\sigma^{12}}{\sqrt{\sigma^{11}\sigma^{22}}}.$$

[*Hint:* Apply Theorem A.3.2 of the Appendix to the cofactors used to calculate σ^{ij}.]

2.48. (Sec. 2.5) Show that for any joint distribution for which the expectations exist and any function $h(x^{(2)})$ that

$$\mathscr{E}\big(X_i - \mathscr{E}X_i|X^{(2)}\big)h(X^{(2)}) = 0.$$

[*Hint:* In the above take the expectation first with respect to X_i conditional on $X^{(2)}$.]

2.49. (Sec. 2.5) Show that for any function $h(x^{(2)})$ and any joint distribution of X_i and $X^{(2)}$ for which the relevant expectations exist, $\mathscr{E}[X_i - h(X^{(2)})]^2 = \mathscr{E}[X_i - g(X^{(2)})]^2 + \mathscr{E}[g(X^{(2)}) - h(X^{(2)})]^2$, where $g(x^{(2)}) = \mathscr{E}X_i|x^{(2)}$ is the conditional expectation of X_i given $X^{(2)} = x^{(2)}$. Hence $g(X^{(2)})$ minimizes the mean squared error of prediction. [*Hint:* Use Problem 2.48.]

2.50. (Sec. 2.5) Show that for any function $h(x^{(2)})$ and any joint distribution of X_i and $X^{(2)}$ for which the relevant expectations exist, the correlation between X_i and $h(X^{(2)})$ is not greater than the correlation between X_i and $g(X^{(2)})$, where $g(x^{(2)}) = \mathscr{E}X_i|x^{(2)}$.

2.51. (Sec. 2.5) Show that for any vector function $h(x^{(2)})$

$$\mathscr{E}\big[X^{(1)} - h(X^{(2)})\big]\big[X^{(1)} - h(X^{(2)})\big]' - \mathscr{E}[X^{(1)} - \mathscr{E}X^{(1)}|X^{(2)}][X^{(1)} - \mathscr{E}X^{(1)}|X^{(2)}]'$$

is positive semidefinite. Note this generalizes Theorem 2.5.3 and Problem 2.49.

2.52. (Sec. 2.5) Verify that $\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} = -\mathbf{\Psi}_{11}^{-1}\mathbf{\Psi}_{12}$, where $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}$ is partitioned similarly to $\mathbf{\Sigma}$.

2.53. (Sec. 2.5) Show

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{\Sigma}_{11\cdot 2}^{-1} & -\mathbf{\Sigma}_{11\cdot 2}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ -\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11\cdot 2}^{-1} & \mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11\cdot 2}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} + \mathbf{\Sigma}_{22}^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22}^{-1} \end{pmatrix} + \begin{pmatrix} I \\ -\mathbf{\beta}' \end{pmatrix} \mathbf{\Sigma}_{11\cdot 2}^{-1}(I \quad -\mathbf{\beta}),$$

where $\mathbf{\beta} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$. [*Hint:* Use Theorem A.3.3 of the Appendix and the fact that $\mathbf{\Sigma}^{-1}$ is symmetric.]

2.54. (Sec. 2.5) Use Problem 2.53 to show that

$$x'\Sigma^{-1}x = \left(x^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)}\right)'\Sigma_{11\cdot 2}^{-1}\left(x^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)}\right) + x^{(2)\prime}\Sigma_{22}^{-1}x^{(2)}.$$

2.55. (Sec. 2.5) Show

$$\begin{aligned}\mathscr{E}\left(X^{(1)}|x^{(2)}, x^{(3)}\right) = \mu^{(1)} &+ \Sigma_{13}\Sigma_{33}^{-1}\left(x^{(3)} - \mu^{(3)}\right)\\ &+ \left(\Sigma_{12} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{32}\right)\left(\Sigma_{22} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{32}\right)^{-1}\\ &\cdot\left[x^{(2)} - \mu^{(2)} - \Sigma_{23}\Sigma_{33}^{-1}\left(x^{(3)} - \mu^{(3)}\right)\right].\end{aligned}$$

2.56. (Sec. 2.5) Prove by matrix algebra that

$$\begin{aligned}\Sigma_{11} - (\Sigma_{12}\Sigma_{13})\begin{pmatrix}\Sigma_{22} & \Sigma_{23}\\ \Sigma_{32} & \Sigma_{33}\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{21}\\ \Sigma_{31}\end{pmatrix} &= \Sigma_{11} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{31}\\ &-\left(\Sigma_{12} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{32}\right)\left(\Sigma_{22} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{32}\right)^{-1}\left(\Sigma_{21} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{31}\right).\end{aligned}$$

2.57. (Sec. 2.5) *Invariance of the partial correlation coefficient.* Prove that $\rho_{12\cdot 3,\dots,p}$ is invariant under the transformations $x_i^* = a_i x_i + b_i' x^{(3)} + c_i$, $a_i > 0$, $i = 1, 2$, $x^{(3)*} = Cx^{(3)} + d$, where $x^{(3)} = (x_3, \dots, x_p)'$, and that any function of μ and Σ that is invariant under these transformations is a function of $\rho_{12\cdot 3,\dots,p}$.

2.58. (Sec. 2.5) Suppose $X^{(1)}$ and $X^{(2)}$ of q and $p-q$ components, respectively, have the density

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}p}}e^{-\frac{1}{2}Q},$$

where

$$\begin{aligned}Q = \left(x^{(1)} - \mu^{(1)}\right)'A_{11}\left(x^{(1)} - \mu^{(1)}\right) &+ \left(x^{(1)} - \mu^{(1)}\right)'A_{12}\left(x^{(2)} - \mu^{(2)}\right)\\ + \left(x^{(2)} - \mu^{(2)}\right)'A_{21}\left(x^{(1)} - \mu^{(1)}\right) &+ \left(x^{(2)} - \mu^{(2)}\right)'A_{22}\left(x^{(2)} - \mu^{(2)}\right).\end{aligned}$$

Show that Q can be written as $Q_1 + Q_2$, where

$$Q_1 = \left[\left(x^{(1)} - \mu^{(1)}\right) + A_{11}^{-1}A_{12}\left(x^{(2)} - \mu^{(2)}\right)\right]'A_{11}\left[\left(x^{(1)} - \mu^{(1)}\right) + A_{11}^{-1}A_{12}\left(x^{(2)} - \mu^{(2)}\right)\right]$$

$$Q_2 = \left(x^{(2)} - \mu^{(2)}\right)'\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)\left(x^{(2)} - \mu^{(2)}\right).$$

Show that the marginal density of $X^{(2)}$ is

$$\frac{|A_{22} - A_{21}A_{11}^{-1}A_{12}|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}(p-q)}}e^{-\frac{1}{2}Q_2}.$$

Show that the conditional density of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ is

$$\frac{|A_{11}|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}q}}e^{-\frac{1}{2}Q_1}$$

(without using the Appendix). This problem is meant to furnish an *alternative* proof of Theorems 2.4.3 and 2.5.1.

2.59. (Sec. 2.6) Prove Lemma 2.6.2 in detail.

2.60. (Sec. 2.6) Let Y be distributed according to $N(\mathbf{0}, \Sigma)$. Differentiating the characteristic function, verify (25) and (26).

2.61. (Sec. 2.6) Verify (25) and (26) by using the transformation $X - \mu = CY$, where $\Sigma = CC'$, and integrating the density of Y.

2.62. (Sec. 2.6) Let the density of (X, Y) be

$$\begin{array}{ll} 2n(x|0,1)n(y|0,1), & 0 \le y \le x < \infty, \quad 0 \le -x \le y < \infty, \\ & 0 \le -y \le -x < \infty, \quad 0 \le x \le -y < \infty, \\ 0 & \text{otherwise}. \end{array}$$

Show that $X, Y, X+Y, X-Y$ each have a marginal normal distribution.

2.63. (Sec. 2.6) Suppose X is distributed according to $N(\mathbf{0}, \Sigma)$. Let $\Sigma = (\sigma_1, \dots, \sigma_p)$. Prove

$$\begin{aligned} \mathscr{E}(XX' \otimes XX') &= \Sigma \otimes \Sigma + \operatorname{vec} \Sigma \, (\operatorname{vec} \Sigma)' + \begin{bmatrix} \sigma_1 \sigma_1' & \cdots & \sigma_p \sigma_1' \\ \vdots & & \vdots \\ \sigma_1 \sigma_p' & \cdots & \sigma_p \sigma_p' \end{bmatrix} \\ &= (I + K)(\Sigma \otimes \Sigma) + \operatorname{vec} \Sigma \, (\operatorname{vec} \Sigma)', \end{aligned}$$

where

$$\operatorname{vec} \Sigma = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_p \end{bmatrix}, \qquad K = \begin{bmatrix} \varepsilon_1 \varepsilon_1' & \cdots & \varepsilon_p \varepsilon_1' \\ \vdots & & \vdots \\ \varepsilon_1 \varepsilon_p' & \cdots & \varepsilon_p \varepsilon_p' \end{bmatrix},$$

and ε_i is a column vector with 1 in the ith position and 0's elsewhere.

2.64. *Complex normal distribution.* Let $(X', Y')'$ have a normal distribution with mean vector $(\mu_X', \mu_Y')'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \Gamma & -\Phi \\ \Phi & \Gamma \end{pmatrix},$$

where Γ is positive definite and $\Phi = -\Phi'$ (skew symmetric). Then $Z = X + iY$ is said to have a complex normal distribution with mean $\theta = \mu_X + i\mu_Y$ and covariance matrix $\mathscr{E}(Z - \theta)(Z - \theta)^* = P = Q + iR$, where $Z^* = X' - iY'$. Note that P is Hermitian and positive definite.

(a) Show $Q = 2\Gamma$ and $R = 2\Phi$.
(b) Show $|P|^2 = |2\Sigma|$. [*Hint:* $|\Gamma + i\Phi| = |\Gamma - i\Phi|$.]

(c) Show

$$P^{-1} = (Q + RQ^{-1}R)^{-1} - iQ^{-1}R(Q + RQ^{-1}R)^{-1}.$$

Note that the inverse of a Hermitian matrix is Hermitian.

(d) Show that the density of X and Y can be written

$$\pi^{-p}|P|^{-1}e^{-(z-\theta)^* P^{-1}(z-\theta)}.$$

2.65. *Complex normal (continued).* If Z has the complex normal distribution of Problem 2.64, show that $W = AZ$, where A is a nonsingular complex matrix, has the complex normal distribution with mean $A\theta$ and covariance matrix $\mathscr{C}(W) = APA^*$.

2.66. Show that the characteristic function of Z defined in Problem 2.64 is

$$\mathscr{E}e^{i\mathscr{R}(u^*Z)} = e^{i\mathscr{R}u^*\theta - u^*Pu},$$

where $\mathscr{R}(x+iy) = x$.

2.67. (Sec. 2.2) Show that $\int_{-a}^{a} e^{-x^2/2}\,dx/\sqrt{2\pi}$ is approximately $(1 - e^{-2a^2/\pi})^{1/2}$. [*Hint:* The probability that (X, Y) falls in a square is approximately the probability that (X, Y) falls in an approximating circle [Pólya (1949)].]

2.68. (Sec. 2.7) For the multivariate t-distribution with density (41) show that $\mathscr{E}X = \mu$ and $\mathscr{C}(X) = [m/(m-2)]\Lambda$.

CHAPTER 3

Estimation of the Mean Vector and the Covariance Matrix

3.1. INTRODUCTION

The multivariate normal distribution is specified completely by the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The first statistical problem is how to estimate these parameters on the basis of a sample of observations. In Section 3.2 it is shown that the maximum likelihood estimator of $\boldsymbol{\mu}$ is the sample mean; the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is proportional to the matrix of sample variances and covariances. A sample variance is a sum of squares of deviations of observations from the sample mean divided by one less than the number of observations in the sample; a sample covariance is similarly defined in terms of cross products. The sample covariance matrix is an unbiased estimator of $\boldsymbol{\Sigma}$.

The distribution of the sample mean vector is given in Section 3.3, and it is shown how one can test the hypothesis that $\boldsymbol{\mu}$ is a given vector when $\boldsymbol{\Sigma}$ is known. The case of $\boldsymbol{\Sigma}$ unknown will be treated in Chapter 5.

Some theoretical properties of the sample mean are given in Section 3.4, and the Bayes estimator of the population mean is derived for a normal a priori distribution. In Section 3.5 the James–Stein estimator is introduced; improvements over the sample mean for the mean squared error loss function are discussed.

In Section 3.6 estimators of the mean vector and covariance matrix of elliptically contoured distributions and the distributions of the estimators are treated.

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

3.2. THE MAXIMUM LIKELIHOOD ESTIMATORS OF THE MEAN VECTOR AND THE COVARIANCE MATRIX

Given a sample of (vector) observations from a p-variate (nondegenerate) normal distribution, we ask for estimators of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the distribution. We shall deduce the maximum likelihood estimators.

It turns out that the method of maximum likelihood is very useful in various estimation and hypothesis testing problems concerning the multivariate normal distribution. The maximum likelihood estimators or modifications of them often have some optimum properties. In the particular case studied here, the estimators are asymptotically efficient [Cramér (1946), Sec. 33.3].

Suppose our sample of N observations on $\boldsymbol{X}$ distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, where $N > p$. The likelihood function is

$$(1) \qquad \begin{aligned} L &= \prod_{\alpha=1}^{N} n(\boldsymbol{x}_\alpha | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{(2\pi)^{\frac{1}{2}pN} |\boldsymbol{\Sigma}|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2} \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) \right]. \end{aligned}$$

In the likelihood function the vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ are fixed at the sample values and L is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. To emphasize that these quantities are variables (and not parameters) we shall denote them by $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$. Then the logarithm of the likelihood function is

$$(2) \qquad \begin{aligned} \log L = &-\tfrac{1}{2}pN \log 2\pi - \tfrac{1}{2}N \log |\boldsymbol{\Sigma}^*| \\ &- \tfrac{1}{2} \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}^*)' \boldsymbol{\Sigma}^{*-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}^*). \end{aligned}$$

Since $\log L$ is an increasing function of L, its maximum is at the same point in the space of $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ as the maximum of L. The maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the vector $\boldsymbol{\mu}^*$ and the positive definite matrix $\boldsymbol{\Sigma}^*$ that maximize $\log L$. (It remains to be seen that the supremum of $\log L$ is attained for a positive definite matrix $\boldsymbol{\Sigma}^*$.)

Let the *sample mean vector* be

$$(3) \qquad \bar{\boldsymbol{x}} = \frac{1}{N} \sum_{\alpha=1}^{N} \boldsymbol{x}_\alpha = \begin{pmatrix} \frac{1}{N} \sum_{\alpha=1}^{N} x_{1\alpha} \\ \vdots \\ \frac{1}{N} \sum_{\alpha=1}^{N} x_{p\alpha} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix},$$

where $x_\alpha = (x_{1\alpha}, \dots, x_{p\alpha})'$ and $\bar{x}_i = \sum_{\alpha=1}^N x_{i\alpha}/N$, and let the matrix of sums of squares and cross products of deviations about the mean be

$$A = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' \tag{4}$$
$$= \left[\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) \right], \qquad i, j = 1, \dots, p.$$

It will be convenient to use the following lemma:

Lemma 3.2.1. *Let $x_1, \dots, x_N$ be N (p-component) vectors, and let $\bar{x}$ be defined by* (3). *Then for any vector b*

$$\sum_{\alpha=1}^{N} (x_\alpha - b)(x_\alpha - b)' = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' + N(\bar{x} - b)(\bar{x} - b)'. \tag{5}$$

Proof

(6)

$$\sum_{\alpha=1}^{N} (x_\alpha - b)(x_\alpha - b)' = \sum_{\alpha=1}^{N} \left[(x_\alpha - \bar{x}) + (\bar{x} - b) \right] \left[(x_\alpha - \bar{x}) + (\bar{x} - b) \right]'$$
$$= \sum_{\alpha=1}^{N} \left[(x_\alpha - \bar{x})(x_\alpha - \bar{x})' + (x_\alpha - \bar{x})(\bar{x} - b)' \right.$$
$$\left. + (\bar{x} - b)(x_\alpha - \bar{x})' + (x - xb)(\bar{x} - b)' \right]$$
$$= \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' + \left[\sum_{\alpha=1}^{N} (x_\alpha - \bar{x}) \right] (\bar{x} - b)'$$
$$+ (\bar{x} - b) \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})' + N(\bar{x} - b)(\bar{x} - b)'.$$

The second and third terms on the right-hand side are 0 because $\sum(x_\alpha - \bar{x}) = \sum x_\alpha - N\bar{x} = 0$ by (3). ■

When we let $b = \mu^*$, we have

(7)

$$\sum_{\alpha=1}^{N} (x_\alpha - \mu^*)(x_\alpha - \mu^*)' = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' + N(\bar{x} - \mu^*)(\bar{x} - \mu^*)'$$
$$= A + N(\bar{x} - \mu^*)(\bar{x} - \mu^*)'.$$

Using this result and the properties of the trace of a matrix ($\operatorname{tr} CD = \sum c_{ij} d_{ji} = \operatorname{tr} DC$), we have

(8)
$$\begin{aligned}\sum_{\alpha=1}^{N} (x_\alpha - \mu^*)'\Sigma^{*-1}(x_\alpha - \mu^*) &= \operatorname{tr} \sum_{\alpha=1}^{N} (x_\alpha - \mu^*)'\Sigma^{*-1}(x_\alpha - \mu^*) \\ &= \operatorname{tr} \sum_{\alpha=1}^{N} \Sigma^{*-1}(x_\alpha - \mu^*)(x_\alpha - \mu^*)' \\ &= \operatorname{tr} \Sigma^{*-1}A + \operatorname{tr} \Sigma^{*-1}N(\bar{x} - \mu^*)(\bar{x} - \mu^*)' \\ &= \operatorname{tr} \Sigma^{*-1}A + N(\bar{x} - \mu^*)'\Sigma^{*-1}(\bar{x} - \mu^*).\end{aligned}$$

Thus we can write (2) as

(9)
$$\begin{aligned}\log L = &-\tfrac{1}{2}pN \log(2\pi) - \tfrac{1}{2}N \log|\Sigma^*| \\ &-\tfrac{1}{2}\operatorname{tr} \Sigma^{*-1}A - \tfrac{1}{2}N(\bar{x} - \mu^*)'\Sigma^{*-1}(\bar{x} - \mu^*).\end{aligned}$$

Since Σ^* is positive definite, Σ^{*-1} is positive definite, and $N(\bar{x} - \mu^*)'\Sigma^{*-1}(\bar{x} - \mu^*) \geq 0$ and is 0 if and only if $\mu^* = \bar{x}$. To maximize the second and third terms of (9) we use the following lemma (which is also used in later chapters):

Lemma 3.2.2. *If D is positive definite of order p, the maximum of*

(10)
$$f(G) = N \log|G| - \operatorname{tr} G^{-1}D$$

with respect to positive definite matrices G exists, occurs at $G = (1/N)D$, and has the value

(11)
$$f[(1/N)D] = pN \log N - N \log|D| - pN.$$

Proof. Let $D = EE'$ and $E'G^{-1}E = H$. Then $G = EH^{-1}E'$, and $|G| = |E| \cdot |H^{-1}| \cdot |E'| = |H^{-1}| \cdot |EE'| = |D|/|H|$, and $\operatorname{tr} G^{-1}D = \operatorname{tr} G^{-1}EE' = \operatorname{tr} E'G^{-1}E = \operatorname{tr} H$. Then the function to be maximized (with respect to positive definite H) is

(12)
$$f = -N \log|D| + N \log|H| - \operatorname{tr} H.$$

Let $H = TT'$, where T is lower triangular (Corollary A.1.7). Then the maximum of

(13)
$$\begin{aligned}f &= -N \log|D| + N \log|T|^2 - \operatorname{tr} TT' \\ &= -N \log|D| + \sum_{i=1}^{p} (N \log t_{ii}^2 - t_{ii}^2) - \sum_{i>j} t_{ij}^2\end{aligned}$$

occurs at $t_{ii}^2 = N$, $t_{ij} = 0$, $i \neq j$; that is, at $\boldsymbol{H} = N\boldsymbol{I}$. Then $\boldsymbol{G} = (1/N)\boldsymbol{E}\boldsymbol{E}' = (1/N)\boldsymbol{D}$. ■

Theorem 3.2.1. *If $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ constitute a sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $p < N$, the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = (1/N)\sum_{\alpha=1}^N \boldsymbol{x}_\alpha$ and $\hat{\boldsymbol{\Sigma}} = (1/N)\sum_{\alpha=1}^N (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$, respectively.*

Other methods of deriving the maximum likelihood estimators have been discussed by Anderson and Olkin (1985). See Problems 3.4, 3.8, and 3.12.

Computation of the estimate $\hat{\boldsymbol{\Sigma}}$ is made easier by the specialization of Lemma 3.2.1 ($\boldsymbol{b} = \boldsymbol{0}$)

$$\sum_{\alpha=1}^N (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})' = \sum_{\alpha=1}^N \boldsymbol{x}_\alpha \boldsymbol{x}_\alpha' - N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'. \tag{14}$$

An element of $\sum_{\alpha=1}^N \boldsymbol{x}_\alpha \boldsymbol{x}_\alpha'$ is computed as $\sum_{\alpha=1}^N x_{i\alpha} x_{j\alpha}$, and an element of $N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'$ is computed as $N\bar{x}_i \bar{x}_j$ or $(\sum_{\alpha=1}^N x_{i\alpha})(\sum_{\alpha=1}^N x_{j\alpha})/N$. It should be noted that if $N > p$. the probability is 1 of drawing a sample so that (14) is positive definite; see Problem 3.17.

The covariance matrix can be written in terms of the variances or standard deviations and correlation coefficients. These are uniquely defined by the variances and covariances. We assert that the maximum likelihood estimators of functions of the parameters are those functions of the maximum likelihood estimators of the parameters.

Lemma 3.2.3. *Let $f(\theta)$ be a real-valued function defined on a set S, and let ϕ be a single-valued function, with a single-valued inverse, on S to a set S^*; that is, to each $\theta \in S$ there corresponds a unique $\theta^* \in S^*$, and, conversely, to each $\theta^* \in S^*$ there corresponds a unique $\theta \in S$. Let*

$$g(\theta^*) = f[\phi^{-1}(\theta^*)]. \tag{15}$$

Then if $f(\theta)$ attains a maximum at $\theta = \theta_0$, $g(\theta^*)$ attains a maximum at $\theta^* = \theta_0^* = \phi(\theta_0)$. If the maximum of $f(\theta)$ at θ_0 is unique, so is the maximum of $g(\theta^*)$ at θ_0^*.

Proof. By hypothesis $f(\theta_0) \geq f(\theta)$ for all $\theta \in S$. Then for any $\theta^* \in S^*$

$$g(\theta^*) = f[\phi^{-1}(\theta^*)] = f(\theta) \leq f(\theta_0) = g[\phi(\theta_0)] = g(\theta_0^*). \tag{16}$$

Thus $g(\theta^*)$ attains a maximum at θ_0^*. If the maximum of $f(\theta)$ at θ_0 is unique, there is strict inequality above for $\theta \neq \theta_0$, and the maximum of $g(\theta^*)$ is unique. ■

We have the following corollary:

Corollary 3.2.1. *If on the basis of a given sample $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are maximum likelihood estimators of the parameters $\theta_1, \ldots, \theta_m$ of a distribution, then $\phi_1(\hat{\theta}_1, \ldots, \hat{\theta}_m), \ldots, \phi_m(\hat{\theta}_1, \ldots, \hat{\theta}_m)$ are maximum likelihood estimators of $\phi_1(\theta_1, \ldots, \theta_m), \ldots, \phi_m(\theta_1, \ldots, \theta_m)$ if the transformation from $\theta_1, \ldots, \theta_m$ to $\phi_1, \ldots, \phi_m$ is one-to-one.*† *If the estimators of $\theta_1, \ldots, \theta_m$ are unique, then the estimators of $\phi_1, \ldots, \phi_m$ are unique.*

Corollary 3.2.2. *If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ constitutes a sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$ ($\rho_{ii} = 1$), then the maximum likelihood estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = (1/N)\sum_\alpha \boldsymbol{x}_\alpha$; the maximum likelihood estimator of σ_i^2 is $\hat{\sigma}_i^2 = (1/N)\sum_\alpha (x_{i\alpha} - \bar{x}_i)^2 = (1/N)(\sum_\alpha x_{i\alpha}^2 - N\bar{x}_i^2)$, where $x_{i\alpha}$ is the ith component of $\boldsymbol{x}_\alpha$ and $\bar{x}_i$ is the ith component of $\bar{\boldsymbol{x}}$; and the maximum likelihood estimator of ρ_{ij} is*

$$(17) \qquad \hat{\rho}_{ij} = \frac{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^N (x_{i\alpha} - \bar{x}_i)^2} \sqrt{\sum_{\alpha=1}^N (x_{j\alpha} - \bar{x}_j)^2}}$$

$$= \frac{\sum_{\alpha=1}^N x_{i\alpha} x_{j\alpha} - N\bar{x}_i \bar{x}_j}{\sqrt{\sum_{\alpha=1}^N x_{i\alpha}^2 - N\bar{x}_i^2} \sqrt{\sum_{\alpha=1}^N x_{j\alpha}^2 - N\bar{x}_j^2}}.$$

Proof. The set of parameters $\mu_i = \mu_i$, $\sigma_i^2 = \sigma_{ii}$, and $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ is a one-to-one transform of the set of parameters μ_i and σ_{ij}. Therefore, by Corollary 3.2.1 the estimator of μ_i is $\hat{\mu}_i$, of σ_i^2 is $\hat{\sigma}_{ii}$, and of ρ_{ij} is

$$(18) \qquad \hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}. \qquad \blacksquare$$

Pearson (1896) gave a justification for this estimator of ρ_{ij}, and (17) is sometimes called the *Pearson correlation coefficient*. It is also called the *simple correlation coefficient*. It is usually denoted by r_{ij}.

†The assumption that the transformation is one-to-one is made so that the set $\phi_1, \ldots, \phi_m$ uniquely defines the likelihood. An alternative in case $\theta^* = \phi(\theta)$ does not have a unique inverse is to define $s(\theta^*) = \{\theta: \phi(\theta) = \theta^*\}$ and $g(\theta^*) = \sup f(\theta) | \theta \in S(\theta^*)$, which is considered the "induced likelihood" when $f(\theta)$ is the likelihood function. Then $\hat{\theta}^* = \phi(\hat{\theta})$ maximizes $g(\theta^*)$, for $g(\theta^*) = \sup f(\theta) | \theta \in S(\theta^*) \geq \sup f(\theta) | \theta \in S = f(\hat{\theta}) = g(\hat{\theta}^*)$ for all $\theta^* \in S^*$. [See, e.g., Zehna (1966).]

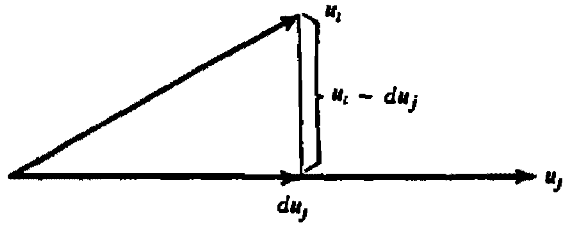

Figure 3.1

A convenient geometrical interpretation of this sample $(x_1, x_2, \dots, x_N) = X$ is in terms of the rows of X. Let

$$(19) \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pN} \end{pmatrix} = \begin{pmatrix} u_1' \\ \vdots \\ u_p' \end{pmatrix};$$

that is, u_i' is the ith row of X. The vector u_i can be considered as a vector in an N-dimensional space with the αth coordinate of one endpoint being $x_{i\alpha}$ and the other endpoint at the origin. Thus the sample is represented by p vectors in N-dimensional Euclidean space. By definition of the Euclidean metric, the squared length of u_i (that is, the squared distance of one endpoint from the other) is $u_i'u_i = \sum_{\alpha=1}^N x_{i\alpha}^2$.

Now let us show that the cosine of the angle between u_i and u_j is $u_i'u_j/\sqrt{u_i'u_i u_j'u_j} = \sum_{\alpha=1}^N x_{i\alpha}x_{j\alpha}/\sqrt{\sum_{\alpha=1}^N x_{i\alpha}^2 \sum_{\alpha=1}^N x_{j\alpha}^2}$. Choose the scalar d so the vector du_j is orthogonal to $u_i - du_j$; that is, $0 = du_j'(u_i - du_j) = d(u_j'u_i - du_j'u_j)$. Therefore, $d = u_j'u_i/u_j'u_j$. We decompose u_i into $u_i - du_j$ and du_j $[u_i = (u_i - du_j) + du_j]$ as indicated in Figure 3.1. The absolute value of the cosine of the angle between u_i and u_j is the length of du_j divided by the length of u_i; that is, it is $\sqrt{du_j'(du_j)/u_i'u_i} = \sqrt{du_j'u_j d/u_i'u_i}$; the cosine is $u_i'u_j/\sqrt{u_i'u_i u_j'u_j}$. This proves the desired result.

To give a geometric interpretation of a_{ii} and $a_{ij}/\sqrt{a_{ii}a_{jj}}$, we introduce the equiangular line, which is the line going through the origin and the point $(1, 1, \dots, 1)$. See Figure 3.2. The projection of u_i on the vector $\varepsilon = (1, 1, \dots, 1)'$ is $(\varepsilon'u_i/\varepsilon'\varepsilon)\varepsilon = (\sum_\alpha x_{i\alpha}/\sum_\alpha 1)\varepsilon = \bar{x}_i\varepsilon = (\bar{x}_i, \bar{x}_i, \dots, \bar{x}_i)'$. Then we decompose u_i into $\bar{x}_i\varepsilon$, the projection on the equiangular line, and $u_i - \bar{x}_i\varepsilon$, the projection of u_i on the plane perpendicular to the equiangular line. The squared length of $u_i - \bar{x}_i\varepsilon$ is $(u_i - \bar{x}_i\varepsilon)'(u_i - \bar{x}_i\varepsilon) = \sum_\alpha(x_{i\alpha} - \bar{x}_i)^2$; this is $N\hat{\sigma}_{ii} = a_{ii}$. Translate $u_i - \bar{x}_i\varepsilon$ and $u_j - \bar{x}_j\varepsilon$, so that each vector has an endpoint at the origin; the αth coordinate of the first vector is $x_{i\alpha} - \bar{x}_i$, and of

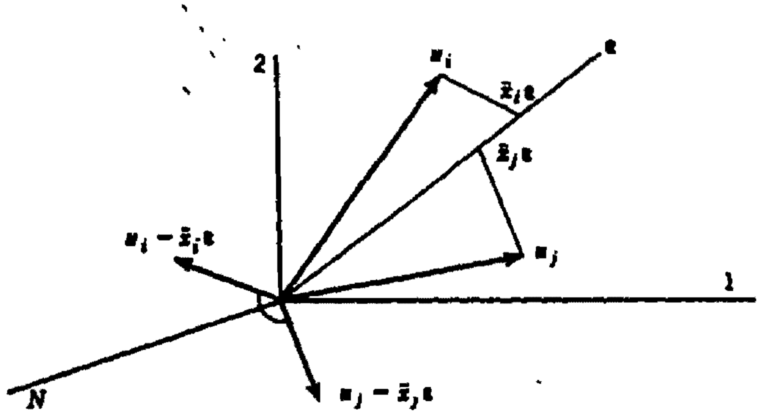

Figure 3.2

the second is $x_{j\alpha} - \bar{x}_j$. The cosine of the angle between these two vectors is

$$r_{ij} = \frac{(u_i - \bar{x}_i\varepsilon)'(u_j - \bar{x}_j\varepsilon)}{\sqrt{(u_i - \bar{x}_i\varepsilon)'(u_i - \bar{x}_i\varepsilon)(u_j - \bar{x}_j\varepsilon)'(u_j - \bar{x}_j\varepsilon)}} \tag{20}$$

$$= \frac{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)^2 \sum_{\alpha=1}^{N}(x_{j\alpha} - \bar{x}_j)^2}}.$$

As an example of the calculations consider the data in Table 3.1 and graphed in Figure 3.3, taken from Student (1908). The measurement $x_{11} = 1.9$ on the first patient is the increase in the number of hours of sleep due to the use of the sedative A, $x_{21} = 0.7$ is the increase in the number of hours due to

Table 3.1. Increase in Sleep

Patient	Drug A x_1	Drug B x_2
1	1.9	0.7
2	0.8	−1.6
3	1.1	−0.2
4	0.1	−1.2
5	−0.1	−0.1
6	4.4	3.4
7	5.5	3.7
8	1.6	0.8
9	4.6	0.0
10	3.4	2.0

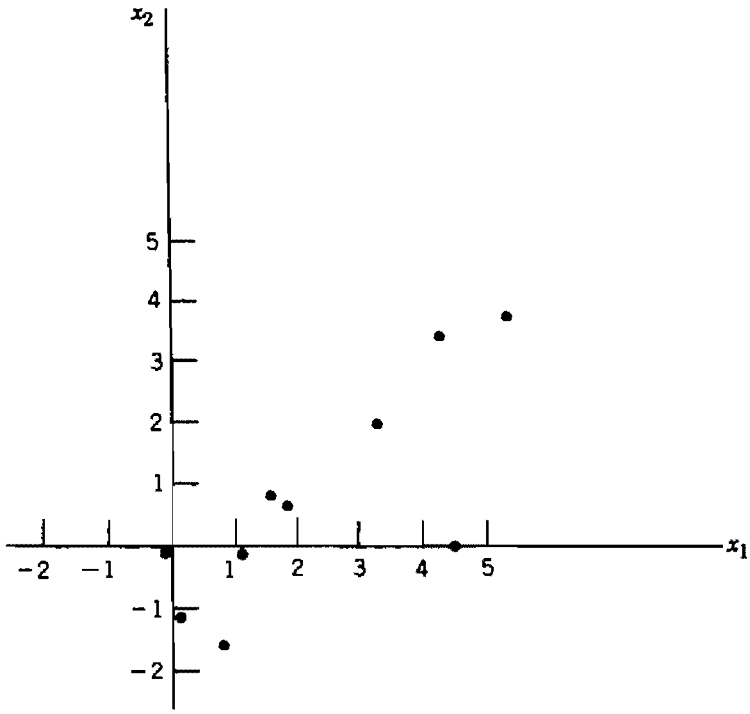

Figure 3.3. Increase in sleep.

sedative B, and so on. Assuming that each pair (i.e., each row in the table) is an observation from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we find that

$$\hat{\boldsymbol{\mu}} = \bar{x} = \begin{pmatrix} 2.33 \\ 0.75 \end{pmatrix},$$

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 3.61 & 2.56 \\ 2.56 & 2.88 \end{pmatrix}, \tag{21}$$

$$S = \begin{pmatrix} 4.01 & 2.85 \\ 2.85 & 3.20 \end{pmatrix},$$

and $\hat{\rho}_{12} = r_{12} = 0.7952$. ($S$ will be defined later.)

3.3. THE DISTRIBUTION OF THE SAMPLE MEAN VECTOR; INFERENCE CONCERNING THE MEAN WHEN THE COVARIANCE MATRIX IS KNOWN

3.3.1. Distribution Theory

In the univariate case the mean of a sample is distributed normally and independently of the sample variance. Similarly, the sample mean $\bar{X}$ defined in Section 3.2 is distributed normally and independently of $\hat{\boldsymbol{\Sigma}}$.

To prove this result we shall make a transformation of the set of observation vectors. Because this kind of transformation is used several times in this book, we first prove a more general theorem.

Theorem 3.3.1. *Suppose $X_1, \dots, X_N$ are independent, where X_α is distributed according to $N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma})$. Let $C = (c_{\alpha\beta})$ be an $N \times N$ orthogonal matrix. Then $Y_\alpha = \sum_{\beta=1}^N c_{\alpha\beta} X_\beta$ is distributed according to $N(\boldsymbol{\nu}_\alpha, \boldsymbol{\Sigma})$, where $\boldsymbol{\nu}_\alpha = \sum_{\beta=1}^N c_{\alpha\beta} \boldsymbol{\mu}_\beta$, $\alpha = 1, \dots, N$, and $Y_1, \dots, Y_N$ are independent.*

Proof. The set of vectors $Y_1, \dots, Y_N$ have a joint normal distribution, because the entire set of components is a set of linear combinations of the components of $X_1, \dots, X_N$, which have a joint normal distribution. The expected value of Y_α is

$$\text{(1)} \qquad \begin{aligned} \mathscr{E} Y_\alpha &= \mathscr{E} \sum_{\beta=1}^N c_{\alpha\beta} X_\beta = \sum_{\beta=1}^N c_{\alpha\beta} \mathscr{E} X_\beta \\ &= \sum_{\beta=1}^N c_{\alpha\beta} \boldsymbol{\mu}_\beta = \boldsymbol{\nu}_\alpha. \end{aligned}$$

The covariance matrix between Y_α and Y_γ is

$$\text{(2)} \qquad \begin{aligned} \mathscr{C}(Y_\alpha, Y_\gamma') &= \mathscr{E}(Y_\alpha - \boldsymbol{\nu}_\alpha)(Y_\gamma - \boldsymbol{\nu}_\gamma)' \\ &= \mathscr{E}\left[\sum_{\beta=1}^N c_{\alpha\beta}(X_\beta - \boldsymbol{\mu}_\beta)\right]\left[\sum_{\varepsilon=1}^N c_{\gamma\varepsilon}(X_\varepsilon - \boldsymbol{\mu}_\varepsilon)'\right] \\ &= \sum_{\beta, \varepsilon=1}^N c_{\alpha\beta} c_{\gamma\varepsilon} \mathscr{E}(X_\beta - \boldsymbol{\mu}_\beta)(X_\varepsilon - \boldsymbol{\mu}_\varepsilon)' \\ &= \sum_{\beta, \varepsilon=1}^N c_{\alpha\beta} c_{\gamma\varepsilon} \delta_{\beta\varepsilon} \boldsymbol{\Sigma} \\ &= \sum_{\beta=1}^N c_{\alpha\beta} c_{\gamma\beta} \boldsymbol{\Sigma} \\ &= \delta_{\alpha\gamma} \boldsymbol{\Sigma}, \end{aligned}$$

where $\delta_{\alpha\gamma}$ is the Kronecker delta ($=1$ if $\alpha = \gamma$ and $=0$ if $\alpha \neq \gamma$). This shows that Y_α is independent of Y_γ, $\alpha \neq \gamma$, and Y_α has the covariance matrix $\boldsymbol{\Sigma}$. ■

We also use the following general lemma:

Lemma 3.3.1. *If $C = (c_{\alpha\beta})$ is orthogonal, then $\sum_{\alpha=1}^N x_\alpha x'_\alpha = \sum_{\alpha=1}^N y_\alpha y'_\alpha$, where $y_\alpha = \sum_{\beta=1}^N c_{\alpha\beta} x_\beta$, $\alpha = 1, \dots, N$.*

Proof

$$\sum_{\alpha=1}^N y_\alpha y'_\alpha = \sum_\alpha \sum_\beta c_{\alpha\beta} x_\beta \sum_\gamma c_{\alpha\gamma} x'_\gamma \tag{3}$$

$$= \sum_{\beta,\gamma} \left(\sum_\alpha c_{\alpha\beta} c_{\alpha\gamma} \right) x_\beta x'_\gamma$$

$$= \sum_{\beta,\gamma} \delta_{\beta\gamma} x_\beta x'_\gamma$$

$$= \sum_\beta x_\beta x'_\beta. \qquad \blacksquare$$

Let $X_1, \dots, X_N$ be independent, each distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. There exists an $N \times N$ orthogonal matrix $B = (b_{\alpha\beta})$ with the last row

$$(1/\sqrt{N}, \dots, 1/\sqrt{N}). \tag{4}$$

(See Lemma A.4.2.) This transformation is a rotation in the N-dimensional space described in Section 3.2 with the equiangular line going into the Nth coordinate axis. Let $A = N\hat{\boldsymbol{\Sigma}}$, defined in Section 3.2, and let

$$Z_\alpha = \sum_{\beta=1}^N b_{\alpha\beta} X_\beta. \tag{5}$$

Then

$$Z_N = \sum_{\beta=1}^N b_{N\beta} X_\beta = \sum_{\beta=1}^N \frac{1}{\sqrt{N}} X_\beta = \sqrt{N}\,\bar{X}. \tag{6}$$

By Lemma 3.3.1 we have

$$A = \sum_{\alpha=1}^N X_\alpha X'_\alpha - N\bar{X}\bar{X}' \tag{7}$$

$$= \sum_{\alpha=1}^N Z_\alpha Z'_\alpha - Z_N Z'_N$$

$$= \sum_{\alpha=1}^{N-1} Z_\alpha Z'_\alpha.$$

Since Z_N is independent of $Z_1, \dots, Z_{N-1}$, the mean vector $\bar{X}$ is independent of A. Since

$$\mathscr{E} Z_N = \sum_{\beta=1}^{N} b_{N\beta} \mathscr{E} X_\beta = \sum_{\beta=1}^{N} \frac{1}{\sqrt{N}} \boldsymbol{\mu} = \sqrt{N}\, \boldsymbol{\mu}, \tag{8}$$

Z_N is distributed according to $N(\sqrt{N}\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\bar{X} = (1/\sqrt{N}) Z_N$ is distributed according to $N[\boldsymbol{\mu}, (1/N)\boldsymbol{\Sigma}]$. We note

$$\begin{aligned} \mathscr{E} Z_\alpha &= \sum_{\beta=1}^{N} b_{\alpha\beta} \mathscr{E} X_\beta = \sum_{\beta=1}^{N} b_{\alpha\beta} \boldsymbol{\mu} \\ &= \sum_{\beta=1}^{N} b_{\alpha\beta} b_{N\beta} \sqrt{N}\, \boldsymbol{\mu} \\ &= 0, \qquad \alpha \neq N. \end{aligned} \tag{9}$$

Theorem 3.3.2. *The mean of a sample of size N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is distributed according to $N[\boldsymbol{\mu}, (1/N)\boldsymbol{\Sigma}]$ and independently of $\hat{\boldsymbol{\Sigma}}$, the maximum likelihood estimator of $\boldsymbol{\Sigma}$. $N\hat{\boldsymbol{\Sigma}}$ is distributed as $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$, where Z_α is distributed according to $N(\mathbf{0}, \boldsymbol{\Sigma})$, $\alpha = 1, \dots, N-1$, and $Z_1, \dots, Z_{N-1}$ are independent.*

Definition 3.3.1. *An estimator t of a parameter vector $\boldsymbol{\theta}$ is* unbiased *if and only if $\mathscr{E}_{\boldsymbol{\theta}} t = \boldsymbol{\theta}$.*

Since $\mathscr{E}\bar{X} = (1/N)\mathscr{E}\sum_{\alpha=1}^{N} X_\alpha = \boldsymbol{\mu}$, the sample mean is an unbiased estimator of the population mean. However,

$$\mathscr{E}\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\mathscr{E}\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha' = \frac{N-1}{N}\boldsymbol{\Sigma}. \tag{10}$$

Thus $\hat{\boldsymbol{\Sigma}}$ is a biased estimator of $\boldsymbol{\Sigma}$. We shall therefore define

$$S = \frac{1}{N-1}A = \frac{1}{N-1}\sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' \tag{11}$$

as the *sample covariance matrix*. It is an unbiased estimator of $\boldsymbol{\Sigma}$ and the diagonal elements are the usual (unbiased) sample variances of the components of X.

3.3.2. Tests and Confidence Regions for the Mean Vector When the Covariance Matrix Is Known

A statistical problem of considerable importance is that of testing the hypothesis that the mean vector of a normal distribution is a given vector.

and a related problem is that of giving a confidence region for the unknown vector of means. We now go on to study these problems under the assumption that the covariance matrix $\mathbf{\Sigma}$ is known. In Chapter 5 we consider these problems when the covariance matrix is unknown.

In the univariate case one bases a test or a confidence interval on the fact that the difference between the sample mean and the population mean is normally distributed with mean zero and known variance; then tables of the normal distribution can be used to set up significance points or to compute confidence intervals. In the multivariate case one uses the fact that the difference between the sample mean vector and the population mean vector is normally distributed with mean vector zero and known covariance matrix. One could set up limits for each component on the basis of the distribution, but this procedure has the disadvantages that the choice of limits is somewhat arbitrary and in the case of tests leads to tests that may be very poor against some alternatives, and, moreover, such limits are difficult to compute because tables are available only for the bivariate case. The procedures given below, however, are easily computed and furthermore can be given general intuitive and theoretical justifications.

The procedures and evaluation of their properties are based on the following theorem:

Theorem 3.3.3. *If the m-component vector Y is distributed according to $N(\boldsymbol{\nu}, T)$ (nonsingular), then $Y'T^{-1}Y$ is distributed according to the noncentral χ^2-distribution with m degrees of freedom and noncentrality parameter $\boldsymbol{\nu}'T^{-1}\boldsymbol{\nu}$. If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is the central χ^2-distribution.*

Proof. Let C be a nonsingular matrix such that $CTC' = I$, and define $Z = CY$. Then Z is normally distributed with mean $\mathscr{E}Z = C\mathscr{E}Y = C\boldsymbol{\nu} = \boldsymbol{\lambda}$, say, and covariance matrix $\mathscr{E}(Z-\boldsymbol{\lambda})(Z-\boldsymbol{\lambda})' = \mathscr{E}C(Y-\boldsymbol{\nu})(Y-\boldsymbol{\nu})'C' = CTC' = I$. Then $Y'T^{-1}Y = Z'(C')^{-1}T^{-1}C^{-1}Z = Z'(CTC')^{-1}Z = Z'Z$, which is the sum of squares of the components of Z. Similarly $\boldsymbol{\nu}'T^{-1}\boldsymbol{\nu} = \boldsymbol{\lambda}'\boldsymbol{\lambda}$. Thus $Y'T^{-1}Y$ is distributed as $\sum_{i=1}^m Z_i^2$, where $Z_1, \dots, Z_m$ are independently normally distributed with means $\lambda_1, \dots, \lambda_m$, respectively, and variances 1. By definition this distribution is the noncentral χ^2-distribution with noncentrality parameter $\sum_{i=1}^m \lambda_i^2$. See Section 3.3.3. If $\lambda_1 = \cdots = \lambda_m = 0$, the distribution is central. (See Problem 7.5.) ■

Since $\sqrt{N}(\bar{X} - \boldsymbol{\mu})$ is distributed according to $N(\mathbf{0}, \mathbf{\Sigma})$, it follows from the theorem that

$$N(\bar{X} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\bar{X} - \boldsymbol{\mu}) \tag{12}$$

has a (central) χ^2-distribution with p degrees of freedom. This is the fundamental fact we use in setting up tests and confidence regions concerning $\boldsymbol{\mu}$.

Let $\chi_p^2(\alpha)$ be the number such that

$$\Pr\{\chi_p^2 > \chi_p^2(\alpha)\} = \alpha. \tag{13}$$

Thus

$$\Pr\left\{N(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) > \chi_p^2(\alpha)\right\} = \alpha. \tag{14}$$

To test the hypothesis that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0$ is a specified vector, we use as our critical region

$$N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \chi_p^2(\alpha). \tag{15}$$

If we obtain a sample such that (15) is satisfied, we reject the null hypothesis. It can be seen intuitively that the probability is greater than α of rejecting the hypothesis if $\boldsymbol{\mu}$ is very different from $\boldsymbol{\mu}_0$, since in the space of $\bar{\mathbf{x}}$ (15) defines an ellipsoid with center at $\boldsymbol{\mu}_0$, and when $\boldsymbol{\mu}$ is far from $\boldsymbol{\mu}_0$ the density of $\bar{\mathbf{x}}$ will be concentrated at a point near the edge or outside of the ellipsoid. The quantity $N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$ is distributed as a noncentral χ^2 with p degrees of freedom and noncentrality parameter $N(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ when $\bar{\mathbf{X}}$ is the mean of a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [given by Bose (1936a), (1936b)]. Pearson (1900) first proved Theorem 3.3.3 for $\boldsymbol{\nu} = \mathbf{0}$.

Now consider the following statement made on the basis of a sample with mean $\bar{\mathbf{x}}$: "The mean of the distribution satisfies

$$N(\bar{\mathbf{x}} - \boldsymbol{\mu}^*)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}^*) \leq \chi_p^2(\alpha) \tag{16}$$

as an inequality on $\boldsymbol{\mu}^*$." We see from (14) that the probability that a sample will be drawn such that the above statement is true is $1 - \alpha$ because the event in (14) is equivalent to the statement being false. Thus, the set of $\boldsymbol{\mu}^*$ satisfying (16) is a confidence region for $\boldsymbol{\mu}$ with confidence $1 - \alpha$.

In the p-dimensional space of $\bar{\mathbf{x}}$, (15) is the surface and exterior of an ellipsoid with center $\boldsymbol{\mu}_0$, the shape of the ellipsoid depending on $\boldsymbol{\Sigma}^{-1}$ and the size on $(1/N)\chi_p^2(\alpha)$ for given $\boldsymbol{\Sigma}^{-1}$. In the p-dimensional space of $\boldsymbol{\mu}^*$ (16) is the surface and interior of an ellipsoid with its center at $\bar{\mathbf{x}}$. If $\boldsymbol{\Sigma}^{-1} = \mathbf{I}$, then (14) says that the probability is α that the distance between $\bar{\mathbf{x}}$ and $\boldsymbol{\mu}$ is greater than $\sqrt{\chi_p^2(\alpha)/N}$.

Theorem 3.3.4. *If $\bar{x}$ is the mean of a sample of N drawn from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is known, then* (15) *gives a critical region of size α for testing the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, and* (16) *gives a confidence region for $\boldsymbol{\mu}$ of confidence $1-\alpha$. Here $\chi_p^2(\alpha)$ is chosen to satisfy* (13).

The same technique can be used for the corresponding two-sample problems. Suppose we have a sample $\{x_\alpha^{(1)}\}$, $\alpha = 1, \dots, N_1$, from the distribution $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$, and a sample $\{x_\alpha^{(2)}\}$, $\alpha = 1, \dots, N_2$, from a second normal population $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ with the same covariance matrix. Then the two sample means

$$\bar{x}^{(1)} = \frac{1}{N_1} \sum_{\alpha=1}^{N_1} x_\alpha^{(1)}, \qquad \bar{x}^{(2)} = \frac{1}{N_2} \sum_{\alpha=1}^{N_2} x_\alpha^{(2)} \tag{17}$$

are distributed independently according to $N[\boldsymbol{\mu}^{(1)}, (1/N_1)\boldsymbol{\Sigma}]$ and $N[\boldsymbol{\mu}^{(2)}, (1/N_2)\boldsymbol{\Sigma}]$, respectively. The difference of the two sample means, $y = \bar{x}^{(1)} - \bar{x}^{(2)}$, is distributed according to $N\{\boldsymbol{\nu}, [(1/N_1) + (1/N_2)]\boldsymbol{\Sigma}\}$, where $\boldsymbol{\nu} = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$. Thus

$$\frac{N_1 N_2}{N_1 + N_2} (y - \boldsymbol{\nu})' \boldsymbol{\Sigma}^{-1} (y - \boldsymbol{\nu}) \le \chi_p^2(\alpha) \tag{18}$$

is a confidence region for the difference $\boldsymbol{\nu}$ of the two mean vectors, and a critical region for testing the hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ is given by

$$\frac{N_1 N_2}{N_1 + N_2} (\bar{x}^{(1)} - \bar{x}^{(2)})' \boldsymbol{\Sigma}^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) > \chi_p^2(\alpha). \tag{19}$$

Mahalanobis (1930) suggested $(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$ as a measure of the *distance* squared between two populations. Let C be a matrix such that $\boldsymbol{\Sigma} = CC'$ and let $\boldsymbol{\nu}^{(i)} = C^{-1}\boldsymbol{\mu}^{(i)}$, $i = 1, 2$. Then the distance squared is $(\boldsymbol{\nu}^{(1)} - \boldsymbol{\nu}^{(2)})'(\boldsymbol{\nu}^{(1)} - \boldsymbol{\nu}^{(2)})$, which is the Euclidean distance squared.

3.3.3. The Noncentral χ^2-Distribution; the Power Function

The power function of the test (15) of the null hypothesis that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ can be evaluated from the noncentral χ^2-distribution. The central χ^2-distribution is the distribution of the sum of squares of independent (scalar) normal variables with means 0 and variances 1; the noncentral χ^2-distribution is the generalization of this when the means may be different from 0. Let Y (of p components) be distributed according to $N(\boldsymbol{\lambda}, I)$. Let Q be an orthogonal matrix with elements of the first row being

$$q_{1i} = \frac{\lambda_i}{\sqrt{\boldsymbol{\lambda}'\boldsymbol{\lambda}}}, \qquad i = 1, \dots, p. \tag{20}$$

Then $\boldsymbol{Z} = \boldsymbol{Q}\boldsymbol{Y}$ is distributed according to $N(\boldsymbol{\tau}, \boldsymbol{I})$, where

$$\boldsymbol{\tau} = \begin{pmatrix} \tau \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{21}$$

and $\tau = \sqrt{\boldsymbol{\lambda}'\boldsymbol{\lambda}}$. Let $V = \boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{Z}'\boldsymbol{Z} = \sum_{i=1}^p Z_i^2$. Then $W = \sum_{i=2}^p Z_i^2$ has a χ^2-distribution with $p-1$ degrees of freedom (Problem 7.5), and Z_1 and W have as joint density

$$\begin{aligned} &\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_1-\tau)^2} \frac{1}{2^{\frac{1}{2}(p-1)}\Gamma\left[\frac{1}{2}(p-1)\right]} w^{\frac{1}{2}(p-1)-1} e^{-\frac{1}{2}w} \\ &\qquad = C e^{-\frac{1}{2}(\tau^2+z_1^2+w)} w^{\frac{1}{2}(p-3)} e^{\tau z_1} \\ &\qquad = C e^{-\frac{1}{2}(\tau^2+z_1^2+w)} w^{\frac{1}{2}(p-3)} \sum_{\alpha=0}^{\infty} \frac{\tau^\alpha z_1^\alpha}{\alpha!}, \end{aligned} \tag{22}$$

where $C^{-1} = 2^{\frac{1}{2}p}\sqrt{\pi}\,\Gamma[\frac{1}{2}(p-1)]$. The joint density of $V = W + Z_1^2$ and Z_1 is obtained by substituting $w = v - z_1^2$ (the Jacobian being 1):

$$C e^{-\frac{1}{2}(\tau^2+v)} \left(v - z_1^2\right)^{\frac{1}{2}(p-3)} \sum_{\alpha=0}^{\infty} \frac{\tau^\alpha z_1^\alpha}{\alpha!}. \tag{23}$$

The joint density of V and $U = Z_1/\sqrt{V}$ is $(dz_1 = \sqrt{v}\,du)$

$$C e^{-\frac{1}{2}(\tau^2+v)} v^{\frac{1}{2}(p-2)} \left(1 - u^2\right)^{\frac{1}{2}(p-3)} \sum_{\alpha=0}^{\infty} \frac{\tau^\alpha v^{\frac{1}{2}\alpha} u^\alpha}{\alpha!}. \tag{24}$$

The admissible range of z_1 given v is $-\sqrt{v}$ to $\sqrt{v}$, and the admissible range of u is -1 to 1. When we integrate (24) with respect to u term by term, the terms for α odd integrate to 0, since such a term is an odd function of u. In the other integrations we substitute $u = \sqrt{s}$ $(du = \frac{1}{2}ds/\sqrt{s})$ to obtain

$$
\begin{aligned}
(25)\qquad \int_{-1}^{1}(1-u^2)^{\frac{1}{2}(p-3)}u^{2\beta}\,du &= 2\int_0^1 (1-u^2)^{\frac{1}{2}(p-3)}u^{2\beta}\,du \\
&= \int_0^1 (1-s)^{\frac{1}{2}(p-3)}s^{\beta-\frac{1}{2}}\,ds \\
&= B\left[\tfrac{1}{2}(p-1), \beta+\tfrac{1}{2}\right] \\
&= \frac{\Gamma\left[\frac{1}{2}(p-1)\right]\Gamma(\beta+\frac{1}{2})}{\Gamma(\frac{1}{2}p+\beta)},
\end{aligned}
$$

by the usual properties of the beta and gamma functions. Thus the density of V is

$$
(26)\qquad \frac{1}{2^{\frac{1}{2}p}\sqrt{\pi}}e^{-\frac{1}{2}(\tau^2+v)}v^{\frac{1}{2}p-1}\sum_{\beta=0}^{\infty}\frac{(\tau^2)^{\beta}v^{\beta}}{(2\beta)!}\frac{\Gamma(\beta+\frac{1}{2})}{\Gamma(\frac{1}{2}p+\beta)}.
$$

We can use the *duplication formula for the gamma function* $\Gamma(2\beta+1) = (2\beta)!$ (Problem 7.37),

$$
(27)\qquad \Gamma(2\beta+1) = \Gamma(\beta+\tfrac{1}{2})\Gamma(\beta+1)2^{2\beta}/\sqrt{\pi},
$$

to rewrite (26) as

$$
(28)\qquad \frac{1}{2^{\frac{1}{2}p}}e^{-\frac{1}{2}(\tau^2+v)}v^{\frac{1}{2}p-1}\sum_{\beta=0}^{\infty}\left(\frac{\tau^2}{4}\right)^{\beta}\frac{1}{\beta!\,\Gamma(\frac{1}{2}p+\beta)}v^{\beta}.
$$

This is the density of the *noncentral* χ^2*-distribution* with p degrees of freedom and noncentrality parameter τ^2.

Theorem 3.3.5. *If Y of p components is distributed according to $N(\boldsymbol{\lambda}, I)$, then $V = Y'Y$ has the density* (28), *where* $\tau^2 = \boldsymbol{\lambda}'\boldsymbol{\lambda}$.

To obtain the power function of the test (15), we note that $\sqrt{N}(\bar{X} - \boldsymbol{\mu}_0)$ has the distribution $N[\sqrt{N}(\boldsymbol{\mu} - \boldsymbol{\mu}_0), \boldsymbol{\Sigma}]$. From Theorem 3.3.3 we obtain the following corollary:

Corollary 3.3.1. *If $\bar{X}$ is the mean of a random sample of N drawn from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $N(\bar{X} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{X} - \boldsymbol{\mu}_0)$ has a noncentral χ^2-distribution with p degrees of freedom and noncentrality parameter $N(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$.*

3.4. THEORETICAL PROPERTIES OF ESTIMATORS OF THE MEAN VECTOR

3.4.1. Properties of Maximum Likelihood Estimators

It was shown in Section 3.3.1 that $\bar{x}$ and S are unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. In this subsection we shall show that $\bar{x}$ and S are sufficient statistics and are complete.

Sufficiency

A statistic $\boldsymbol{T}$ is *sufficient* for a family of distributions of X or for a parameter $\boldsymbol{\theta}$ if the conditional distribution of X given $\boldsymbol{T}=\boldsymbol{t}$ does not depend on $\boldsymbol{\theta}$ [e.g., Cramér (1946), Section 32.4]. In this sense the statistic $\boldsymbol{T}$ gives as much information about $\boldsymbol{\theta}$ as the entire sample X. (Of course, this idea depends strictly on the assumed family of distributions.)

Factorization Theorem. *A statistic $t(y)$ is sufficient for $\boldsymbol{\theta}$ if and only if the density $f(y|\boldsymbol{\theta})$ can be factored as*

$$f(y|\boldsymbol{\theta})=g[t(y),\boldsymbol{\theta}]h(y), \tag{1}$$

where $g[t(y),\boldsymbol{\theta}]$ and $h(y)$ are nonnegative and $h(y)$ does not depend on $\boldsymbol{\theta}$.

Theorem 3.4.1. *If $x_1,\dots,x_N$ are observations from $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\bar{x}$ and S are sufficient for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. If $\boldsymbol{\mu}$ is given, $\sum_{\alpha=1}^N(x_\alpha-\boldsymbol{\mu})(x_\alpha-\boldsymbol{\mu})'$ is sufficient for $\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}$ is given, $\bar{x}$ is sufficient for $\boldsymbol{\mu}$.*

Proof. The density of $X_1,\dots,X_N$ is

$$\begin{aligned}(2)\qquad &\prod_{\alpha=1}^N n(x_\alpha|\boldsymbol{\mu},\boldsymbol{\Sigma})\\ &=(2\pi)^{-\frac12Np}|\boldsymbol{\Sigma}|^{-\frac12N}\exp\left[-\tfrac12\operatorname{tr}\boldsymbol{\Sigma}^{-1}\sum_{\alpha=1}^N(x_\alpha-\boldsymbol{\mu})(x_\alpha-\boldsymbol{\mu})'\right]\\ &=(2\pi)^{-\frac12Np}|\boldsymbol{\Sigma}|^{-\frac12N}\exp\left\{-\tfrac12\left[N(\bar{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{x}-\boldsymbol{\mu})+(N-1)\operatorname{tr}\boldsymbol{\Sigma}^{-1}S\right]\right\}.\end{aligned}$$

The right-hand side of (2) is in the form of (1) for $\bar{x}$, S, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and the middle is in the form of (1) for $\sum_{\alpha=1}^N(x_\alpha-\boldsymbol{\mu})(x_\alpha-\boldsymbol{\mu})'$, $\boldsymbol{\Sigma}$; in each case $h(x_1,\dots,x_N)=1$. The right-hand side is in the form of (1) for $\bar{x},\boldsymbol{\mu}$ with $h(x_1,\dots,x_N)=\exp\{-\frac12(N-1)\operatorname{tr}\boldsymbol{\Sigma}^{-1}S\}$. ■

Note that if $\boldsymbol{\Sigma}$ is given, $\bar{x}$ is sufficient for $\boldsymbol{\mu}$, but if $\boldsymbol{\mu}$ is given, S is not sufficient for $\boldsymbol{\Sigma}$.

Completeness

To prove an optimality property of the T^2-test (Section 5.5), we need the result that $(\bar{x}, S)$ is a complete sufficient set of statistics for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Definition 3.4.1. *A family of distributions of y indexed by $\boldsymbol{\theta}$ is* complete *if for every real-valued function $g(y)$,*

$$\mathscr{E}_{\boldsymbol{\theta}} g(y) \equiv 0 \tag{3}$$

identically in $\boldsymbol{\theta}$ implies $g(y) = 0$ except for a set of y of probability 0 for every $\boldsymbol{\theta}$.

If the family of distributions of a sufficient set of statistics is complete, the set is called a complete sufficient set.

Theorem 3.4.2. *The sufficient set of statistics $\bar{x}$, S is complete for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ when the sample is drawn from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Proof. We can define the sample in terms of $\bar{x}$ and $z_1, \dots, z_n$ as in Section 3.3 with $n = N - 1$. We assume for any function $g(\bar{x}, A) = g(\bar{x}, nS)$ that

$$\begin{aligned}\int \cdots \int & K|\boldsymbol{\Sigma}|^{-\frac{1}{2}N} g\left(\bar{x}, \sum_{\alpha=1}^{n} z_\alpha z_\alpha'\right) \\ & \cdot \exp\left\{-\tfrac{1}{2}\left[N(\bar{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{x}-\boldsymbol{\mu}) + \sum_{\alpha=1}^{n} z_\alpha' \boldsymbol{\Sigma}^{-1} z_\alpha\right]\right\} \\ & \cdot d\bar{x} \prod_{\alpha=1}^{n} dz_\alpha \equiv 0, \qquad \forall \boldsymbol{\mu}, \boldsymbol{\Sigma},\end{aligned} \tag{4}$$

where $K = \sqrt{N}(2\pi)^{-\frac{1}{2}pN}$, $d\bar{x} = \prod_{i=1}^{p} d\bar{x}_i$, and $dz_\alpha = \prod_{i=1}^{p} dz_{i\alpha}$. If we let $\boldsymbol{\Sigma}^{-1} = \boldsymbol{I} - 2\boldsymbol{\Theta}$, where $\boldsymbol{\Theta} = \boldsymbol{\Theta}'$ and $\boldsymbol{I} - 2\boldsymbol{\Theta}$ is positive definite, and let $\boldsymbol{\mu} = (\boldsymbol{I} - 2\boldsymbol{\Theta})^{-1} \boldsymbol{t}$, then (4) is

$$\begin{aligned} 0 \equiv \int \cdots \int & K|\boldsymbol{I} - 2\boldsymbol{\Theta}|^{\frac{1}{2}N} g\left(\bar{x}, \sum_{\alpha=1}^{n} z_\alpha z_\alpha'\right) \\ & \cdot \exp\left\{-\tfrac{1}{2}\left[\operatorname{tr}(\boldsymbol{I} - 2\boldsymbol{\Theta})\left(\sum_{\alpha=1}^{n} z_\alpha z_\alpha' + N\bar{x}\bar{x}'\right)\right.\right. \\ & \qquad \left.\left. -2N\boldsymbol{t}'\bar{x} + N\boldsymbol{t}'(\boldsymbol{I} - 2\boldsymbol{\Theta})^{-1}\boldsymbol{t}\right]\right\} d\bar{x} \prod_{\alpha=1}^{n} dz_\alpha \\ = |\boldsymbol{I} - 2\boldsymbol{\Theta}|^{\frac{1}{2}N} & \exp\left\{-\tfrac{1}{2}N\boldsymbol{t}'(\boldsymbol{I} - 2\boldsymbol{\Theta})^{-1}\boldsymbol{t}\right\} \int \cdots \int g(\bar{x}, \boldsymbol{B} - N\bar{x}\bar{x}') \\ & \cdot \exp\left[\operatorname{tr} \boldsymbol{\Theta}\boldsymbol{B} + \boldsymbol{t}'(N\bar{x})\right] n\left[\bar{x} \mid \mathbf{0}, (1/N)\boldsymbol{I}\right] \prod_{\alpha=1}^{n} n(z_\alpha \mid \mathbf{0}, \boldsymbol{I})\, d\bar{x} \prod_{\alpha=1}^{n} dz_\alpha,\end{aligned} \tag{5}$$

where $B = \sum_{\alpha=1}^{n} z_\alpha z_\alpha' + N\bar{x}\bar{x}'$. Thus

$$(6) \qquad 0 \equiv \mathscr{E} g(\bar{x}, B - N\bar{x}\bar{x}') \exp[\operatorname{tr} \Theta B + t'(N\bar{x})]$$
$$= \int \cdots \int g(\bar{x}, B - N\bar{x}\bar{x}') \exp[\operatorname{tr} \Theta B + t'(N\bar{x})] h(\bar{x}, B)\, d\bar{x}\, dB,$$

where $h(\bar{x}, B)$ is the joint density of $\bar{x}$ and B and $dB = \prod_{i \le j} db_{ij}$. The right-hand side of (6) is the Laplace transform of $g(\bar{x}, B - N\bar{x}\bar{x}')h(\bar{x}, B)$. Since this is 0, $g(\bar{x}, A) = 0$ except for a set of measure 0. ∎

Efficiency

If a q-component random vector Y has mean vector $\mathscr{E}Y = \nu$ and covariance matrix $\mathscr{E}(Y - \nu)(Y - \nu)' = \Psi$, then

$$(7) \qquad (y - \nu)'\Psi^{-1}(y - \nu) = q + 2$$

is called the *concentration ellipsoid* of Y. [See Cramér (1946), p. 300.] The density defined by a uniform distribution over the interior of this ellipsoid has the same mean vector and covariance matrix as Y. (See Problem 2.14.) Let θ be a vector of q parameters in a distribution, and let t be a vector of unbiased estimators (that is, $\mathscr{E}t = \theta$) based on N observations from that distribution with covariance matrix Ψ. Then the ellipsoid

$$(8) \qquad N(t - \theta)' \mathscr{E}\left(\frac{\partial \log f}{\partial \theta}\right)\left(\frac{\partial \log f}{\partial \theta}\right)'(t - \theta) = q + 2$$

lies entirely within the ellipsoid of concentration of t; $\partial \log f / \partial \theta$ denotes the column vector of derivatives of the density of the distribution (or probability function) with respect to the components of θ. The discussion by Cramér (1946, p. 495) is in terms of scalar observations, but it is clear that it holds true for vector observations. If (8) is the ellipsoid of concentration of t, then t is said to be efficient. In general, the ratio of the volume of (8) to that of the ellipsoid of concentration defines the efficiency of t. In the case of the multivariate normal distribution, if $\theta = \mu$, then $\bar{x}$ is efficient. If θ includes both μ and Σ, then $\bar{x}$ and S have efficiency $[(N-1)/N]^{p(p+1)/2}$. Under suitable regularity conditions, which are satisfied by the multivariate normal distribution,

$$(9) \qquad \mathscr{E}\left(\frac{\partial \log f}{\partial \theta}\right)\left(\frac{\partial \log f}{\partial \theta}\right)' = -\mathscr{E}\frac{\partial^2 \log f}{\partial \theta\, \partial \theta'}.$$

This is the *information matrix* for one observation. The Cramér–Rao lower bound is that for any unbiased estimator t the matrix

$$(10) \qquad N\mathscr{E}(t-\theta)(t-\theta)' - \left[-\mathscr{E}\frac{\partial^2 \log f}{\partial\theta\,\partial\theta'}\right]^{-1}$$

is positive semidefinite. (Other lower bounds can also be given.)

Consistency

Definition 3.4.2. *A sequence of vectors $t_n = (t_{1n}, \dots, t_{mn})'$, $n = 1, 2, \dots$, is a* consistent *estimator of $\theta = (\theta_1, \dots, \theta_m)'$ if $\operatorname{plim}_{n\to\infty} t_{in} = \theta_i$, $i = 1, \dots, m$.*

By the law of large numbers each component of the sample mean $\bar{x}$ is a consistent estimator of that component of the vector of expected values μ if the observation vectors are independently and identically distributed with mean μ, and hence $\bar{x}$ is a consistent estimator of μ. Normality is not involved.

An element of the sample covariance matrix is

$$(11) \quad s_{ij} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(x_{i\alpha}-\mu_i)(x_{j\alpha}-\mu_j) - \frac{N}{N-1}(\bar{x}_i-\mu_i)(\bar{x}_j-\mu_j)$$

by Lemma 3.2.1 with $b = \mu$. The probability limit of the second term is 0. The probability limit of the first term is σ_{ij} if $x_1, x_2, \dots$ are independently and identically distributed with mean μ and covariance matrix Σ. Then S is a consistent estimator of Σ.

Asymptotic Normality

First we prove a multivariate central limit theorem.

Theorem 3.4.3. *Let the m-component vectors $Y_1, Y_2, \dots$ be independently and identically distributed with means $\mathscr{E}Y_\alpha = \nu$ and covariance matrices $\mathscr{E}(Y_\alpha - \nu)(Y_\alpha - \nu)' = T$. Then the limiting distribution of $(1/\sqrt{n})\sum_{\alpha=1}^{n}(Y_\alpha - \nu)$ as $n \to \infty$ is $N(0, T)$.*

Proof. Let

$$(12) \qquad \phi_n(t, u) = \mathscr{E}\exp\left[iut'\frac{1}{\sqrt{n}}\sum_{\alpha=1}^{n}(Y_\alpha - \nu)\right],$$

where u is a scalar and t an m-component vector. For fixed t, $\phi_n(t, u)$ can be considered as the characteristic function of $(1/\sqrt{n})\sum_{\alpha=1}^{n}(t'Y_\alpha - \mathscr{E}t'Y_\alpha)$. By the univariate central limit theorem [Cramér (1946), p. 215], the limiting distribution is $N(0, \boldsymbol{t}'\boldsymbol{T}\boldsymbol{t})$. Therefore (Theorem 2.6.4),

$$\lim_{n\to\infty} \phi_n(\boldsymbol{t}, u) = e^{-\frac{1}{2}u^2 \boldsymbol{t}'\boldsymbol{T}\boldsymbol{t}} \tag{13}$$

for every u and $\boldsymbol{t}$. (For $\boldsymbol{t} = \mathbf{0}$ a special and obvious argument is used.) Let $u = 1$ to obtain

$$\lim_{n\to\infty} \mathscr{E} \exp\left[i\boldsymbol{t}' \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\boldsymbol{Y}_\alpha - \boldsymbol{\nu}) \right] = e^{-\frac{1}{2}\boldsymbol{t}'\boldsymbol{T}\boldsymbol{t}} \tag{14}$$

for every $\boldsymbol{t}$. Since $e^{-\frac{1}{2}\boldsymbol{t}'\boldsymbol{T}\boldsymbol{t}}$ is continuous at $\boldsymbol{t} = \mathbf{0}$, the convergence is uniform in some neighborhood of $\boldsymbol{t} = \mathbf{0}$. The theorem follows. ■

Now we wish to show that the sample covariance matrix is asymptotically normally distributed as the sample size increases.

Theorem 3.4.4. *Let $A(n) = \sum_{\alpha=1}^{N} (\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}}_N)(\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}}_N)'$, where $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ are independently distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $n = N - 1$. Then the limiting distribution of $\boldsymbol{B}(n) = (1/\sqrt{n})[A(n) - n\boldsymbol{\Sigma}]$ is normal with mean $\mathbf{0}$ and covariances*

$$\mathscr{E} b_{ij}(n) b_{kl}(n) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}. \tag{15}$$

Proof. As shown earlier, $A(n)$ is distributed as $A(n) = \sum_{\alpha=1}^{n} \boldsymbol{Z}_\alpha \boldsymbol{Z}_\alpha'$, where $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \dots$ are distributed independently according to $N(\mathbf{0}, \boldsymbol{\Sigma})$. We arrange the elements of $\boldsymbol{Z}_\alpha \boldsymbol{Z}_\alpha'$ in a vector such as

$$\boldsymbol{Y}_\alpha = \begin{pmatrix} Z_{1\alpha}^2 \\ Z_{1\alpha}Z_{2\alpha} \\ \vdots \\ Z_{2\alpha}^2 \\ \vdots \\ Z_{p\alpha}^2 \end{pmatrix}, \tag{16}$$

the moments of $\boldsymbol{Y}_\alpha$ can be deduced from the moments of $\boldsymbol{Z}_\alpha$ as given in Section 2.6. We have $\mathscr{E} Z_{i\alpha} Z_{j\alpha} = \sigma_{ij}$, $\mathscr{E} Z_{i\alpha} Z_{j\alpha} Z_{k\alpha} Z_{l\alpha} = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$, $\mathscr{E}(Z_{i\alpha}Z_{j\alpha} - \sigma_{ij})(Z_{k\alpha}Z_{l\alpha} - \sigma_{kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$. Thus the vectors $\boldsymbol{Y}_\alpha$ defined by (16) satisfy the conditions of Theorem 3.4.3 with the elements of $\boldsymbol{\nu}$ being the elements of $\boldsymbol{\Sigma}$ arranged in vector form similar to (16) and the elements of T being given above. If the elements of $A(n)$ are arranged in vector form similar to (16), say the vector $W(n)$, then $W(n) - n\boldsymbol{\nu} = \sum_{\alpha=1}^{n}(Y_\alpha - \boldsymbol{\nu})$. By Theorem 3.4.3, $(1/\sqrt{n})[W(n) - n\boldsymbol{\nu}]$ has a limiting normal distribution with mean $\mathbf{0}$ and the covariance matrix of Y_α. ∎

The elements of $B(n)$ will have a limiting normal distribution with mean $\mathbf{0}$ if $x_1, x_2, \ldots$ are independently and identically distributed with finite fourth-order moments, but the covariance structure of $B(n)$ will depend on the fourth-order moments.

3.4.2. Decision Theory

It may be enlightening to consider estimation in terms of decision theory. We review some of the concepts. An observation x is made on a random variable X (which may be a vector) whose distribution P_θ depends on a parameter θ which is an element of a set Θ. The statistician is to make a decision d in a set D. A *decision procedure* is a function $\delta(x)$ whose domain is the set of values of X and whose range is D. The *loss* in making decision d when the distribution is P_θ is a nonnegative function $L(\theta, d)$. The evaluation of a procedure $\delta(x)$ is on the basis of the *risk function*

$$R(\theta, \delta) = \mathscr{E}_\theta L[\theta, \delta(X)]. \tag{17}$$

For example, if d and θ are univariate, the loss may be squared error, $L(\theta, d) = (\theta - d)^2$, and the risk is the mean squared error $\mathscr{E}_\theta[\delta(X) - \theta]^2$.

A decision procedure $\delta(x)$ is *as good as* a procedure $\delta^*(x)$ if

$$R(\theta, \delta) \le R(\theta, \delta^*), \qquad \forall \theta; \tag{18}$$

$\delta(x)$ is *better than* $\delta^*(x)$ if (18) holds with a strict inequality for at least one value of θ. A procedure $\delta^*(x)$ is *inadmissible* if there exists another procedure $\delta(x)$ that is better than $\delta^*(x)$. A procedure is *admissible* if it is not inadmissible (i.e., if there is no procedure better than it) in terms of the given loss function. A class of procedures is *complete* if for any procedure not in the class there is a better procedure in the class. The class is *minimal complete* if it does not contain a proper complete subclass. If a minimal complete class exists, it is identical to the class of admissible procedures. When such a class is available, there is no (mathematical) need to use a procedure outside the minimal complete class. Sometimes it is convenient to refer to an *essentially complete* class, which is a class of procedures such that for every procedure outside the class there is one in the class that is just as good.

For a given procedure the risk function is a function of the parameter. If the parameter can be assigned an a priori distribution, say, with density $\rho(\theta)$, then the average loss from use of a decision procedure $\delta(x)$ is

$$r(\rho, \delta) = \mathscr{E}_\rho R(\theta, \delta) = \mathscr{E}_\rho \mathscr{E}_\theta L[\theta, \delta(X)]. \tag{19}$$

Given the a priori density ρ, the decision procedure $\delta(x)$ that minimizes $r(\rho, \delta)$ is the *Bayes procedure*, and the resulting minimum of $r(\rho, \delta)$ is the *Bayes risk*. Under general conditions Bayes procedures are admissible and admissible procedures are Bayes or limits of Bayes procedures. If the density of X given θ is $f(x|\theta)$, the joint density of X and θ is $f(x|\theta)\rho(\theta)$ and the average risk of a procedure $\delta(x)$ is

$$\begin{aligned} r(\rho, \delta) &= \int_\Theta \int_X L[\theta, \delta(x)] f(x|\theta) \rho(\theta)\, dx\, d\theta \\ &= \int_X \left\{ \int_\Theta L[\theta, \delta(x)] g(\theta|x)\, d\theta \right\} f(x)\, dx; \end{aligned} \tag{20}$$

here

$$f(x) = \int_\Theta f(x|\theta) \rho(\theta)\, d\theta, \qquad g(\theta|x) = \frac{f(x|\theta)\rho(\theta)}{f(x)} \tag{21}$$

are the marginal density of X and the a posteriori density of θ given x. The procedure that minimizes $r(\rho, \delta)$ is one that for each x minimizes the expression in braces on the right-hand side of (20), that is, the expectation of $L[\theta, \delta(x)]$ with respect to the a posteriori distribution. If θ and d are vectors ($\boldsymbol{\theta}$ and $\boldsymbol{d}$) and $L(\boldsymbol{\theta}, \boldsymbol{d}) = (\boldsymbol{\theta} - \boldsymbol{d})'Q(\boldsymbol{\theta} - \boldsymbol{d})$, where Q is positive definite, then

$$\begin{aligned} \mathscr{E}_{\boldsymbol{\theta}|x} L[\boldsymbol{\theta}, \boldsymbol{d}(x)] = {}& \mathscr{E}_{\boldsymbol{\theta}|x}[\boldsymbol{\theta} - \mathscr{E}(\boldsymbol{\theta}|x)]'Q[\boldsymbol{\theta} - \mathscr{E}(\boldsymbol{\theta}|x)] \\ & + [\mathscr{E}(\boldsymbol{\theta}|x) - \boldsymbol{d}(x)]'Q[\mathscr{E}(\boldsymbol{\theta}|x) - \boldsymbol{d}(x)]. \end{aligned} \tag{22}$$

The minimum occurs at $\boldsymbol{d}(x) = \mathscr{E}(\boldsymbol{\theta}|x)$, the mean of the a posteriori distribution.

Theorem 3.4.5. *If $x_1, \dots, x_N$ are independently distributed, each x_α according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{\mu}$ has an a priori distribution $N(\boldsymbol{\nu}, \boldsymbol{\Phi})$, then the a posteriori distribution of $\boldsymbol{\mu}$ given $x_1, \dots, x_N$ is normal with mean*

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} \bar{x} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\nu} \tag{23}$$

and covariance matrix

$$\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Phi}. \tag{24}$$

Proof. Since $\bar{x}$ is sufficient for $\boldsymbol{\mu}$, we need only consider $\bar{x}$, which has the distribution of $\boldsymbol{\mu}+\boldsymbol{v}$, where $\boldsymbol{v}$ has the distribution $N[\mathbf{0},(1/N)\boldsymbol{\Sigma}]$ and is independent of $\boldsymbol{\mu}$. Then the joint distribution of $\boldsymbol{\mu}$ and $\bar{x}$ is

$$N\left[\begin{pmatrix} \boldsymbol{\nu} \\ \boldsymbol{\nu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Phi} & \boldsymbol{\Phi} \\ \boldsymbol{\Phi} & \boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma} \end{pmatrix}\right]. \tag{25}$$

The mean of the conditional distribution of $\boldsymbol{\mu}$ given $\bar{x}$ is (by Theorem 2.5.1)

$$\boldsymbol{\nu}+\boldsymbol{\Phi}\left(\boldsymbol{\Phi}+\frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}(\bar{x}-\boldsymbol{\nu}), \tag{26}$$

which reduces to (23). ■

Corollary 3.4.1. *If $x_1,\dots,x_N$ are independently distributed, each x_α according to $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ has an a priori distribution $N(\boldsymbol{\nu},\boldsymbol{\Phi})$, and the loss function is $(d-\boldsymbol{\mu})'Q(d-\boldsymbol{\mu})$, then the Bayes estimator of $\boldsymbol{\mu}$ is* (23).

The Bayes estimator of $\boldsymbol{\mu}$ is a kind of weighted average of $\bar{x}$ and $\boldsymbol{\nu}$, the prior mean of $\boldsymbol{\mu}$. If $(1/N)\boldsymbol{\Sigma}$ is small compared to $\boldsymbol{\Phi}$ (e.g., if N is large), $\boldsymbol{\nu}$ is given little weight. Put another way, if $\boldsymbol{\Phi}$ is large, that is, the prior is relatively uninformative, a large weight is put on $\bar{x}$. In fact, as $\boldsymbol{\Phi}$ tends to ∞ in the sense that $\boldsymbol{\Phi}^{-1}\to\mathbf{0}$, the estimator approaches $\bar{x}$.

A decision procedure $\delta_0(x)$ is *minimax* if

$$\sup_\theta R(\theta,\delta_0)=\inf_\delta \sup_\theta R(\theta,\delta). \tag{27}$$

Theorem 3.4.6. *If $x_1,\dots,x_N$ are independently distributed each according to $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and the loss function is $(d-\boldsymbol{\mu})'Q(d-\boldsymbol{\mu})$, then $\bar{x}$ is a minimax estimator.*

Proof. This follows from a theorem in statistical decision theory that if a procedure δ_0 is *extended Bayes* [i.e., if for arbitrary ε, $r(\rho,\delta_0)\le r(\rho,\delta_\rho)+\varepsilon$ for suitable ρ, where δ_ρ is the corresponding Bayes procedure] and if $R(\theta,\delta_0)$ is constant, then δ_0 is minimax. [See, e.g., Ferguson (1967), Theorem 3 of Section 2.11.] We find

$$\begin{aligned} R(\boldsymbol{\mu},\bar{x}) &= \mathscr{E}(\bar{x}-\boldsymbol{\mu})'Q(\bar{x}-\boldsymbol{\mu}) \\ &= \mathscr{E}\operatorname{tr} Q(\bar{x}-\boldsymbol{\mu})(\bar{x}-\boldsymbol{\mu})' \\ &= \frac{1}{N}\operatorname{tr} Q\boldsymbol{\Sigma}. \end{aligned} \tag{28}$$

Let (23) be $d(\bar{x})$. Its average risk is

$$(29) \qquad \mathscr{E}_{\bar{x}}\mathscr{E}_{\mu}\{\operatorname{tr} Q[d(\bar{x})-\mu][d(\bar{x})-\mu]'|\bar{x}\}$$

$$= \mathscr{E}_{\bar{x}} \operatorname{tr} Q\left[\Phi - \Phi\left(\Phi + \frac{1}{N}\Sigma\right)^{-1}\Phi\right] = \operatorname{tr} Q\Phi\left(\Phi+\frac{1}{N}\Sigma\right)^{-1}\frac{1}{N}\Sigma$$

$$= \operatorname{tr} Q\left(I + \frac{1}{N}\Sigma\Phi^{-1}\right)^{-1}\frac{1}{N}\Sigma \to \frac{1}{N}\operatorname{tr} Q\Sigma$$

as $\Phi^{-1} \to 0$. ■

For more discussion of decision theory see Ferguson (1967). DeGroot (1970), or Berger (1980b).

3.5. IMPROVED ESTIMATION OF THE MEAN

3.5.1. Introduction

The sample mean $\bar{x}$ seems the natural estimator of the population mean μ based on a sample from $N(\mu, \Sigma)$. It is the maximum likelihood estimator, a sufficient statistic when Σ is known, and the minimum variance unbiased estimator. Moreover, it is *equivariant* in the sense that if an arbitrary vector ν is added to each observation vector and to μ, the error of estimation $\overline{(x+\nu)} - (\mu+\nu) = \bar{x} - \mu$ is independent of ν; in other words, the error does not depend on the choice of origin. However, Stein (1956b) showed the startling fact that this conventional estimator is not admissible with respect to the loss function that is the sum of mean squared errors of the components when $\Sigma = I$ and $p \geq 3$. James and Stein (1961) produced an estimator which has a smaller sum of mean squared errors; this estimator will be studied in Section 3.5.2. Subsequent studies have shown that the phenomenon is widespread and the implications imperative.

3.5.2. The James–Stein Estimator

The loss function

$$(1) \qquad L(\mu, m) = (m-\mu)'(m-\mu) = \sum_{i=1}^{p}(m_i - \mu_i)^2 = \|m-\mu\|^2$$

is the sum of mean squared errors of the components of the estimator. We shall show [James and Stein (1961)] that the sample mean is inadmissible by displaying an alternative estimator that has a smaller expected loss for every mean vector $\boldsymbol{\mu}$. We assume that the normal distribution sampled has covariance matrix proportional to $\boldsymbol{I}$ with the constant of proportionality known. It will be convenient to take this constant to be such that $\boldsymbol{Y} = (1/N)\sum_{\alpha=1}^{n} \boldsymbol{X}_\alpha = \bar{\boldsymbol{X}}$ has the distribution $N(\boldsymbol{\mu}, \boldsymbol{I})$. Then the expected loss or risk of the estimator $\boldsymbol{Y}$ is simply $\mathscr{E}\|\boldsymbol{Y} - \boldsymbol{\mu}\|^2 = \operatorname{tr} \boldsymbol{I} = p$. The estimator proposed by James and Stein is (essentially)

$$m(\boldsymbol{y}) = \left(1 - \frac{p-2}{\|\boldsymbol{y} - \boldsymbol{\nu}\|^2}\right)(\boldsymbol{y} - \boldsymbol{\nu}) + \boldsymbol{\nu}, \tag{2}$$

where $\boldsymbol{\nu}$ is an arbitrary fixed vector and $p \geq 3$. This estimator shrinks the observed $\boldsymbol{y}$ toward the specified $\boldsymbol{\nu}$. The amount of shrinkage is negligible if $\boldsymbol{y}$ is very different from $\boldsymbol{\nu}$ and is considerable if $\boldsymbol{y}$ is close to $\boldsymbol{\nu}$. In this sense $\boldsymbol{\nu}$ is a favored point.

Theorem 3.5.1. *With respect to the loss function* (1), *the risk of the estimator* (2) *is less than the risk of the estimator* $\boldsymbol{Y}$ *for* $p \geq 3$.

We shall show that the risk of $\boldsymbol{Y}$ minus the risk of (2) is positive by applying the following lemma due to Stein (1974).

Lemma 3.5.1. *If* $f(x)$ *is a function such that*

$$f(b) - f(a) = \int_a^b f'(x)\,dx \tag{3}$$

for all a *and* b $(a < b)$ *and if*

$$\int_{-\infty}^{\infty} |f'(x)| \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}\,dx < \infty. \tag{4}$$

then

$$\int_{-\infty}^{\infty} f(x)(x-\theta)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}\,dx = \int_{-\infty}^{\infty} f'(x)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}\,dx. \tag{5}$$

Proof of Lemma. We write the left-hand side of (5) as

$$(6)\qquad \int_\theta^\infty [f(x)-f(\theta)](x-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\,dx$$

$$+\int_{-\infty}^\theta [f(x)-f(\theta)](x-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\,dx$$

$$=\int_\theta^\infty\int_\theta^x f'(y)(x-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\,dy\,dx$$

$$-\int_{-\infty}^\theta\int_x^\theta f'(y)(x-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\,dy\,dx$$

$$=\int_\theta^\infty\int_y^\infty f'(y)(x-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\,dx\,dy$$

$$-\int_{-\infty}^\theta\int_{-\infty}^y f'(y)(x-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\,dx\,dy,$$

which yields the right-hand side of (5). Fubini's theorem justifies the interchange of order of integration. (See Problem 3.22.) ■

The lemma can also be derived by integration by parts in special cases.

Proof of Theorem 3.5.1. The difference in risks is

$$(7)\quad \Delta R(\boldsymbol{\mu}) = \mathscr{E}_{\boldsymbol{\mu}}\{\|\boldsymbol{Y}-\boldsymbol{\mu}\|^2 - \|\boldsymbol{m}(\boldsymbol{Y})-\boldsymbol{\mu}\|^2\}$$

$$= \mathscr{E}_{\boldsymbol{\mu}}\left\{\|\boldsymbol{Y}-\boldsymbol{\mu}\|^2 - \left\|\left(1-\frac{p-2}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2}\right)(\boldsymbol{Y}-\boldsymbol{\nu})+\boldsymbol{\nu}-\boldsymbol{\mu}\right\|^2\right\}$$

$$= \mathscr{E}_{\boldsymbol{\mu}}\left\{\sum_{i=1}^p (Y_i-\mu_i)^2 - \sum_{i=1}^p\left[(Y_i-\mu_i)-\frac{p-2}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2}(Y_i-\nu_i)\right]^2\right\}$$

$$= \mathscr{E}_{\boldsymbol{\mu}}\left\{2\frac{p-2}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2}\sum_{i=1}^p (Y_i-\mu_i)(Y_i-\nu_i) - \frac{(p-2)^2}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2}\right\}.$$

Now we use Lemma 3.5.1 with

$$(8)\quad f(y_i) = \frac{y_i - \nu_i}{\sum_{j=1}^{p}(y_j - \nu_j)^2}, \quad f'(y_i) = \frac{1}{\sum_{j=1}^{p}(y_j - \nu_j)^2} - \frac{2(y_i - \nu_i)^2}{\left[\sum_{j=1}^{p}(y_j - \nu_j)^2\right]^2}.$$

[For $p \geq 3$ the condition (4) is satisfied.] Then (7) is

$$(9)\quad \Delta R(\boldsymbol{\mu}) = \mathscr{E}_{\boldsymbol{\mu}}\left\{2(p-2)\sum_{i=1}^{p}\left[\frac{1}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2} - \frac{2(y_i - \nu_i)^2}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^4}\right] - \frac{(p-2)^2}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2}\right\}$$

$$= (p-2)^2\mathscr{E}_{\boldsymbol{\mu}}\frac{1}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2} > 0. \qquad \blacksquare$$

This theorem states that $\boldsymbol{Y}$ is inadmissible for estimating $\boldsymbol{\mu}$ when $p \geq 3$, since the estimator (2) has a smaller risk for every $\boldsymbol{\mu}$ (regardless of the choice of $\boldsymbol{\nu}$).

The risk is the sum of the mean squared errors $\mathscr{E}[m_i(\boldsymbol{Y}) - \mu_i]^2$. Since $Y_1, \dots, Y_p$ are independent and only the distribution of Y_i depends on μ_i, it is puzzling that the improved estimator uses all the Y_j's to estimate μ_i; it seems that irrelevant information is being used. Stein explained the phenomenon by arguing that the sample distance squared of $\boldsymbol{Y}$ from $\boldsymbol{\nu}$, that is, $\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2$, overestimates the squared distance of $\boldsymbol{\mu}$ from $\boldsymbol{\nu}$ and hence that the estimator $\boldsymbol{Y}$ could be improved by bringing it nearer $\boldsymbol{\nu}$ (whatever $\boldsymbol{\nu}$ is). Berger (1980a), following Brown, illustrated by Figure 3.4. The four points x_1, x_2, x_3, x_4 represent a spherical distribution centered at $\boldsymbol{\mu}$. Consider the effects of shrinkage. The average distance of $m(x_1)$ and $m(x_3)$ from $\boldsymbol{\mu}$ is a little greater than that of x_1 and x_3, but $m(x_2)$ and $m(x_4)$ are a little closer to $\boldsymbol{\mu}$ than x_2 and x_4 are if the shrinkage is a certain amount. If $p = 3$, there are two more points (not on the line $\boldsymbol{\nu}, \boldsymbol{\mu}$) that are shrunk closer to $\boldsymbol{\mu}$.

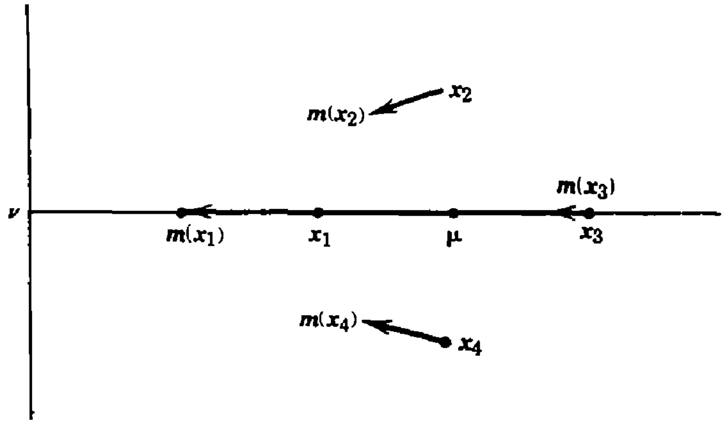

Figure 3.4. Effect of shrinkage.

The risk of the estimator (2) is

$$\mathscr{E}_{\boldsymbol{\mu}}\|\boldsymbol{m}(\boldsymbol{Y})-\boldsymbol{\mu}\|^2 = p-(p-2)^2\,\mathscr{E}_{\boldsymbol{\mu}}\frac{1}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2},\tag{10}$$

where $\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2$ has a noncentral χ^2-distribution with p degrees of freedom and noncentrality parameter $\|\boldsymbol{\mu}-\boldsymbol{\nu}\|^2$. The farther $\boldsymbol{\mu}$ is from $\boldsymbol{\nu}$, the less the improvement due to the James–Stein estimator, but there is always some improvement. The density of $\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2 = V$, say, is (28) of Section 3.3.3, where $\tau^2 = \|\boldsymbol{\mu}-\boldsymbol{\nu}\|^2$. Then

$$\begin{aligned}(11)\quad \mathscr{E}_{\boldsymbol{\mu}}\frac{1}{\|\boldsymbol{Y}-\boldsymbol{\nu}\|^2} &= \mathscr{E}_{\tau^2}V^{-1}\\
&= e^{-\frac12\tau^2}2^{-\frac12 p}\sum_{\beta=0}^{\infty}\left(\frac{\tau^2}{4}\right)^{\beta}\frac{1}{\beta!\,\Gamma(\frac12 p+\beta)}\int_0^{\infty} v^{\frac12 p+\beta-2}e^{-\frac12 v}\,dv\\
&= e^{-\frac12\tau^2}2^{-\frac12 p}\sum_{\beta=0}^{\infty}\left(\frac{\tau^2}{4}\right)^{\beta}\frac{\Gamma(\frac12 p+\beta-1)2^{\frac12 p+\beta-1}}{\beta!\,\Gamma(\frac12 p+\beta)}\\
&= \tfrac12 e^{-\frac12\tau^2}\sum_{\beta=0}^{\infty}\left(\frac{\tau^2}{2}\right)^{\beta}\frac{1}{\beta!\,(\frac12 p+\beta-1)}\end{aligned}$$

for $p \ge 3$. Note that for $\boldsymbol{\mu}=\boldsymbol{\nu}$, that is, $\tau^2 = 0$, (11) is $1/(p-2)$ and the mean squared error (10) is 2. For large p the reduction in risk is considerable.

Table 3.2 gives values of the risk for $p = 10$ and $\sigma^2 = 1$. For example, if $\tau^2 = \|\boldsymbol{\mu}-\boldsymbol{\nu}\|^2$ is 5, the mean squared error of the James–Stein estimator is 8.86, compared to 10 for the natural estimator; this is the case if $\mu_i - \nu_i = 1/\sqrt{2} = 0.707$, $i = 1,\dots,10$, for instance.

Table 3.2†. Average Mean Squared Error of the James–Stein Estimator for $p = 10$ and $\sigma^2 = 1$

$\tau^2 = \|\boldsymbol{\mu}-\boldsymbol{\nu}\|^2$	$\mathscr{E}_{\boldsymbol{\mu}}\|\boldsymbol{m}(\boldsymbol{Y})-\boldsymbol{\mu}\|^2$
0.0	2.00
0.5	4.78
1.0	6.21
2.0	7.51
3.0	8.24
4.0	8.62
5.0	8.86
6.0	9.03

† From Efron and Morris (1977).

An obvious question in using an estimator of this class is how to choose the vector $\boldsymbol{\nu}$ toward which the observed mean vector is shrunk; any $\boldsymbol{\nu}$ yields an estimator better than the natural one. However, as seen from Table 3.2, the improvement is small if $\|\boldsymbol{\mu}-\boldsymbol{\nu}\|$ is very large. Thus, to be effective some knowledge of the position of $\boldsymbol{\mu}$ is necessary. A disadvantage of the procedure is that it is not objective; the choice of $\boldsymbol{\nu}$ is up to the investigator.

A feature of the estimator we have been studying that seems disadvantageous is that for small values of $\|\boldsymbol{Y}-\boldsymbol{\nu}\|$, the multiplier of $\boldsymbol{Y}-\boldsymbol{\nu}$ is negative; that is, the estimator $\boldsymbol{m}(\boldsymbol{Y})$ is in the direction from $\boldsymbol{\nu}$ opposite to that of $\boldsymbol{Y}$. This disadvantage can be overcome and the estimator improved by replacing the factor by 0 when the factor is negative.

Definition 3.5.1. *For any function* $g(u)$, *let*

$$
\begin{aligned}
g^{+}(u) &= g(u), \qquad g(u) \geq 0, \\
&= 0, \qquad\quad\;\; g(u) < 0.
\end{aligned} \tag{12}
$$

Lemma 3.5.2. *When* $\boldsymbol{X}$ *is distributed according to* $N(\boldsymbol{\mu}, \boldsymbol{I})$,

$$
\mathscr{E}_{\boldsymbol{\mu}}\{\|g^{+}(\|\boldsymbol{X}\|)\boldsymbol{X}-\boldsymbol{\mu}\|^2\} \leq \mathscr{E}_{\boldsymbol{\mu}}\{\|g(\|\boldsymbol{X}\|)\boldsymbol{X}-\boldsymbol{\mu}\|^2\}. \tag{13}
$$

Proof. The right-hand side of (13) minus the left-hand side is

$$
\mathscr{E}_{\boldsymbol{\mu}}\{g^2(\|\boldsymbol{X}\|)\|\boldsymbol{X}\|^2 - [g^{+}(\|\boldsymbol{X}\|)]^2\|\boldsymbol{X}\|^2\} \geq 0 \tag{14}
$$

plus 2 times

$$
\begin{aligned}
&\mathscr{E}_{\boldsymbol{\mu}}\boldsymbol{\mu}'\boldsymbol{X}[g^{+}(\|\boldsymbol{X}\|)-g(\|\boldsymbol{X}\|)] \\
&\quad = \|\boldsymbol{\mu}\|\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} y_1[g^{+}(\|\boldsymbol{y}\|)-g(\|\boldsymbol{y}\|)] \\
&\qquad \cdot \frac{1}{(2\pi)^{\frac{1}{2}p}}\exp\left\{-\tfrac{1}{2}\left[\sum_{i=1}^{p} y_i^2 - 2y_1\|\boldsymbol{\mu}\| + \|\boldsymbol{\mu}\|^2\right]\right\} d\boldsymbol{y},
\end{aligned} \tag{15}
$$

where $\boldsymbol{y}'=\boldsymbol{x}'\boldsymbol{P}$, $(\|\boldsymbol{\mu}\|,0,\ldots,0)=\boldsymbol{\mu}'\boldsymbol{P}$, and $\boldsymbol{P}\boldsymbol{P}'=\boldsymbol{I}$. [The first column of $\boldsymbol{P}$ is $(1/\|\boldsymbol{\mu}\|)\boldsymbol{\mu}$.] Then (15) is $\|\boldsymbol{\mu}\|$ times

$$
\begin{aligned}
&e^{-\frac{1}{2}\|\boldsymbol{\mu}\|^2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{0}^{\infty} y_1[g^{+}(\|\boldsymbol{y}\|)-g(\|\boldsymbol{y}\|)][e^{\|\boldsymbol{\mu}\|y_1}-e^{-\|\boldsymbol{\mu}\|y_1}] \\
&\qquad \cdot \frac{1}{(2\pi)^{\frac{1}{2}p}} e^{-\frac{1}{2}\sum_{i=1}^{p} y_i^2}\, dy_1\, dy_2 \cdots dy_p \geq 0
\end{aligned} \tag{16}
$$

(by replacing y_1 by $-y_1$ for $y_1<0$). ■

Theorem 3.5.2. *The estimator*

$$m^+(y) = \left(1 - \frac{p-2}{\|y-\nu\|^2}\right)^+ (y-\nu) + \nu \tag{17}$$

has smaller risk than $m(y)$ defined by (2) and is minimax.

Proof. In Lemma 3.5.2, let $g(u) = 1 - (p-2)/u^2$ and $X = Y - \nu$, and replace μ by $\mu - \nu$. The second assertion in the theorem follows from Theorem 3.4.6. ∎

The theorem shows that $m(Y)$ is not admissible. However, it is known that $m^+(Y)$ is also not admissible, but it is believed that not much further improvement is possible.

This approach is easily extended to the case where one observes $x_1, \dots, x_N$ from $N(\mu, \Sigma)$ with loss function $L(\mu, m) = (m - \mu)'\Sigma^{-1}(m - \mu)$. Let $\Sigma = CC'$ for some nonsingular C, $x_\alpha = Cx_\alpha^*$, $\alpha = 1, \dots, N$, $\mu = C\mu^*$, and $L^*(m^*, \mu^*) = \|m^* - \mu^*\|^2$. Then $x_1^*, \dots, x_N^*$ are observations from $N(\mu^*, I)$, and the problem is reduced to the earlier one. Then

$$\left(1 - \frac{p-2}{N(\bar{x}-\nu)'\Sigma^{-1}(\bar{x}-\nu)}\right)^+ (\bar{x}-\nu) + \nu \tag{18}$$

is a minimax estimator of μ.

3.5.3. Estimation for a General Known Covariance Matrix and an Arbitrary Quadratic Loss Function

Let the parent distribution be $N(\mu, \Sigma)$, where Σ is known, and let the loss function be

$$L(\mu, m) = (m - \mu)'Q(m - \mu), \tag{19}$$

where Q is an arbitrary positive definite matrix which reflects the relative importance of errors in different directions. (If the loss function were singular, the dimensionality of x could be reduced so as to make the loss matrix nonsingular.) Then the sample mean $\bar{x}$ has the distribution $N(\mu, (1/N)\Sigma)$ and risk (expected loss)

$$\mathscr{E}(\bar{x}-\mu)'Q(\bar{x}-\mu) = \mathscr{E}\operatorname{tr} Q(\bar{x}-\mu)(\bar{x}-\mu)' = \frac{1}{N}\operatorname{tr} Q\Sigma, \tag{20}$$

which is constant, not depending on μ.

Several estimators that improve on $\bar{x}$ have been proposed. First we take up an estimator proposed independently by Berger (1975) and Hudson (1974).

Theorem 3.5.3. *Let $r(z)$, $0 \leq z < \infty$, be a nondecreasing differentiable function such that $0 \leq r(z) \leq 2(p-2)$. Then for $p \geq 3$*

$$(21)\qquad m = \left(I - \frac{r\left(N^2(\bar{x}-\nu)'\Sigma^{-1}Q^{-1}\Sigma^{-1}(\bar{x}-\nu)\right)}{N(\bar{x}-\nu)'\Sigma^{-1}Q^{-1}\Sigma^{-1}(\bar{x}-\nu)} Q^{-1}\Sigma^{-1} \right)(\bar{x}-\nu) + \nu$$

has smaller risk than $\bar{x}$ and is minimax.

Proof. There exists a matrix C such that $C'QC = I$ and $(1/N)\Sigma = C\Delta C'$ where Δ is diagonal with diagonal elements $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_p > 0$ (Theorem A.2.2 of the Appendix). Let $\bar{x} = Cy + \nu$ and $\mu = C\mu^* + \nu$. Then y has the distribution $N(\mu^*, \Delta)$, and the transformed loss function is

$$(22)\qquad L^*(m^*, \mu^*) = (m^* - \mu^*)'(m^* - \mu^*) = \|m^* - \mu^*\|^2.$$

The estimator (21) of μ is transformed to the estimator of $\mu^* = C^{-1}(\mu - \nu)$,

$$(23)\qquad m^*(y) = \left(I - \frac{r(y'\Delta^{-2}y)}{y'\Delta^{-2}y}\Delta^{-1} \right) y.$$

We now proceed as in the proof of Theorem 3.5.1. The difference in risks between y and m^* is

(24)

$$\Delta R(\mu^*) = \mathscr{E}_{\mu^*}\left\{\|Y - \mu^*\|^2 - \|m^*(Y) - \mu^*\|^2\right\}$$

$$= \mathscr{E}_{\mu^*}\left\{ 2\frac{r(Y'\Delta^{-2}Y)}{Y'\Delta^{-2}Y} \sum_{i=1}^{p} \frac{1}{\delta_i} Y_i(Y_i - \mu_i^*) - \frac{r^2(Y'\Delta^{-2}Y)}{Y'\Delta^{-2}Y} \right\}.$$

Since $r(z)$ is differentiable, we use Lemma 3.5.1 with $(x - \theta) = (y_i - \mu_i^*)\delta_i$ and

$$(25)\qquad f(y_i) = \frac{r(y'\Delta^{-2}y)}{y'\Delta^{-2}y} y_i,$$

$$(26)\qquad f'(y_i) = \frac{r(y'\Delta^{-2}y)}{y'\Delta^{-2}y} + \frac{2r'(y'\Delta^{-2}y)}{y'\Delta^{-2}y}\frac{y_i^2}{\delta_i^2} - \frac{2r(y'\Delta^{-2}y)}{(y'\Delta^{-2}y)^2}\frac{y_i^2}{\delta_i^2}.$$

Then

(27)

$$\Delta R(\mu^*) = \mathscr{E}_{\mu^*}\left\{2(p-2)\frac{r(Y'\Delta^{-2}Y)}{Y'\Delta^{-2}Y} + 4r'(Y'\Delta^{-2}Y) - \frac{r^2(Y'\Delta^{-2}Y)}{Y'\Delta^{-2}Y}\right\} \geq 0$$

since $r(y'\Delta^{-2}y) \leq 2(p-2)$ and $r'(y'\Delta^{-2}y) \geq 0$. ■

Corollary 3.5.1. *For* $p \geq 3$

(28)

$$\left\{I - \frac{\min\left[p-2, N^2(\bar{x}-\nu)'\Sigma^{-1}Q^{-1}\Sigma^{-1}(\bar{x}-\nu)\right]}{N(\bar{x}-\nu)'\Sigma^{-1}Q^{-1}\Sigma^{-1}(\bar{x}-\nu)}Q^{-1}\Sigma^{-1}\right\}(\bar{x}-\nu)+\nu$$

has smaller risk than $\bar{x}$ *and is minimax.*

Proof. the function $r(z) = \min(p-2, z)$ is differentiable except at $z = p-2$. The function $r(z)$ can be approximated arbitrarily closely by a differentiable function. (For example, the corner at $z = p-2$ can be smoothed by a circular arc of arbitrary small radius.) We shall not give the details of the proof. ■

In canonical form y is shrunk by a scalar times a diagonal matrix. The larger the variance of a component is, the less the effect of the shrinkage.

Berger (1975) has proved these results for a more general density, that is, for a mixture of normals. Berger (1976) has also proved in the case of normality that if

$$r(z) = \frac{z\int_0^{\alpha} u^{\frac{1}{2}p-c+1}e^{-\frac{1}{2}uz}\,du}{\int_0^{\alpha} u^{\frac{1}{2}p-c}e^{-\frac{1}{2}uz}\,du} \tag{29}$$

for $3-\frac{1}{2}p \leq c < 1+\frac{1}{2}p$, where α is the smallest characteristic root of ΣQ, then the estimator m given by (21) is minimax, is admissible if $c < 2$, and is proper Bayes if $c < 1$.

Another approach to minimax estimators has been introduced by Bhattacharya (1966). Let C be such that $C^{-1}(1/N)\Sigma(C^{-1})' = I$ and $C'QC = Q^*$, which is diagonal with diagonal elements $q_1^* \geq q_2^* \geq \cdots \geq q_p^* > 0$. Then $y =$

$C^{-1}\bar{x}$ has the distribution $N(\boldsymbol{\mu}^*, I)$, and the loss function is

$$L^*(\boldsymbol{m}^*, \boldsymbol{\mu}^*) = \sum_{i=1}^{p} q_i^*(m_i^* - \mu_i^*)^2 \tag{30}$$

$$= \sum_{i=1}^{p} \sum_{j=i}^{p} \alpha_j (m_i^* - \mu_i^*)^2$$

$$= \sum_{j=1}^{p} \alpha_j \sum_{i=1}^{j} (m_i^* - \mu_i^*)^2$$

$$= \sum_{j=1}^{p} \alpha_j \|\boldsymbol{m}^{*(j)} - \boldsymbol{\mu}^{*(j)}\|^2,$$

where $\alpha_j = q_j^* - q_{j+1}^*$, $j = 1, \ldots, p-1$, $\alpha_p = q_p^*$, $\boldsymbol{m}^{*(j)} = (m_1^*, \ldots, m_j^*)'$, and $\boldsymbol{\mu}^{*(j)} = (\mu_1^*, \ldots, \mu_j^*)'$, $j = 1, \ldots, p$. This decomposition of the loss function suggests combining minimax estimators of the vectors $\boldsymbol{\mu}^{*(j)}$, $j = 1, \ldots, p$. Let $\boldsymbol{y}^{(j)} = (y_1, \ldots, y_j)'$.

Theorem 3.5.4. *If $\boldsymbol{h}^{(j)}(\boldsymbol{y}^{(j)}) = [h_1^{(j)}(\boldsymbol{y}^{(j)}), \ldots, h_j^{(j)}(\boldsymbol{y}^{(j)})]'$ is a minimax estimator of $\boldsymbol{\mu}^{*(j)}$ under the loss function $\|\boldsymbol{m}^{*(j)} - \boldsymbol{\mu}^{*(j)}\|^2$, $j = 1, \ldots, p$, then*

$$\frac{1}{q_i^*} \sum_{j=i}^{p} \alpha_j h_i^{(j)}(\boldsymbol{y}^{(j)}), \qquad i = 1, \ldots, p, \tag{31}$$

is a minimax estimator of $\mu_1^*, \ldots, \mu_p^*$.

Proof. First consider the randomized estimator defined by

$$\Pr\{G_i(\boldsymbol{y}) = h_i^{(j)}(\boldsymbol{y}^{(j)})\} = \frac{\alpha_j}{q_i^*}, \qquad j = i, \ldots, p, \tag{32}$$

for the ith component. Then the risk of this estimator is

$$\sum_{i=1}^{p} q_i^* \mathscr{E}_{\boldsymbol{\mu}^*}[G_i(\boldsymbol{Y}) - \mu_i^*]^2 = \sum_{i=1}^{p} q_i^* \sum_{j=i}^{p} \frac{\alpha_j}{q_i^*} \mathscr{E}_{\boldsymbol{\mu}^*}[h_i^{(j)}(\boldsymbol{Y}^{(j)}) - \mu_i^*]^2 \tag{33}$$

$$= \sum_{j=1}^{p} \alpha_j \sum_{i=1}^{j} \mathscr{E}_{\boldsymbol{\mu}^*}[h_i^{(j)}(\boldsymbol{Y}^{(j)}) - \mu_i^*]^2$$

$$= \sum_{j=1}^{p} \alpha_j \mathscr{E}_{\boldsymbol{\mu}^*} \|\boldsymbol{h}^{(j)}(\boldsymbol{Y}^{(j)}) - \boldsymbol{\mu}^{*(j)}\|^2$$

$$\le \sum_{j=1}^{p} \alpha_j j = \sum_{j=1}^{p} q_j^*$$

$$= \mathscr{E}_{\boldsymbol{\mu}^*} L^*(\boldsymbol{Y}, \boldsymbol{\mu}^*)^*,$$

and hence the estimator defined by (32) is minimax.

Since the expected value of $G_t(Y)$ with respect to (32) is (31) and the loss function is convex, the risk of the estimator (31) is less than that of the randomized estimator (by Jensen's inequality). ■

3.6. ELLIPTICALLY CONTOURED DISTRIBUTIONS

3.6.1. Observations Elliptically Contoured

Let $x_1, \ldots, x_N$ be N $(= n + 1)$ independent observations on a random vector X with density $|\Lambda|^{-\frac{1}{2}} g[(x_\alpha - \nu)'\Lambda^{-1}(x_\alpha - \nu)]$. The density of the sample is

$$(1) \qquad |\Lambda|^{-\frac{1}{2}N} \prod_{\alpha=1}^{N} g\left[(x-\nu)'\Lambda^{-1}(x-\nu)\right].$$

The sample mean $\bar{x}$ and covariance matrix $S = (1/n)[\sum_{\alpha=1}^{N}(x_\alpha - \mu)(x_\alpha - \mu)' - N(\bar{x} - \mu)(\bar{x} - \mu)']$ are unbiased estimators of the mean $\mu = \nu$ and the covariance matrix $\Sigma = [\mathscr{E} R^2/p]\Lambda$, where $R^2 = (x - \nu)'\Lambda^{-1}(x - \nu)$.

Theorem 3.6.1. *The covariances of the mean and covariance of a sample of N from $|\Lambda|^{-\frac{1}{2}} g[(x - \nu)'\Lambda^{-1}(x - \nu)]$ with $\mathscr{E} R^4 < \infty$ are*

$$(2) \qquad \mathscr{E}(\bar{x} - \mu)(\bar{x} - \mu)' = \frac{1}{N}\Sigma,$$

$$(3) \qquad \mathscr{E}(s_{ij} - \sigma_{ij})(\bar{x} - \mu) = 0, \qquad i, j = 1, \ldots, p,$$

$$(4) \qquad \mathscr{E}(s_{ij} - \sigma_{ij})(s_{kl} - \sigma_{kl}) = \frac{\kappa}{N}(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) + \frac{1}{n}(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}), \qquad i, j, k, l = 1, \ldots, p.$$

Lemma 3.6.1. *The second-order moments of the elements of S are*

$$(5) \qquad \mathscr{E} s_{ij} s_{kl} = \sigma_{ij}\sigma_{kl} + \frac{1}{n}(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) + \frac{\kappa}{N}(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}), \qquad i, j, k, l = 1, \ldots, p$$

Proof of Lemma 3.6.1. We have

$$\begin{aligned}
(6) \qquad \mathscr{E} \sum_{\alpha, \beta=1}^{N} & (x_{i\alpha} - \mu_i)(x_{j\alpha} - \mu_j)(x_{k\beta} - \mu_k)(x_{l\beta} - \mu_l) \\
&= N\mathscr{E}(x_{i\alpha} - \mu_i)(x_{j\alpha} - \mu_j)(x_{k\alpha} - \mu_k)(x_{l\alpha} - \mu_l) \\
&\quad + N(N-1)\mathscr{E}(x_{i\alpha} - \mu_i)(x_{j\alpha} - \mu_j)\mathscr{E}(x_{k\beta} - \mu_k)(x_{l\beta} - \mu_l) \\
&= N(1 + \kappa)(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) + N(N-1)\sigma_{ij}\sigma_{kl},
\end{aligned}$$

$$(7) \quad \mathscr{E} N^2(\bar{x}_i - \mu_i)(\bar{x}_j - \mu_j)(\bar{x}_k - \mu_k)(\bar{x}_l - \mu_l)$$

$$= \frac{1}{N^2} \mathscr{E} \sum_{\alpha, \beta, \gamma, \delta = 1}^{N} (x_{i\alpha} - \mu_i)(x_{j\beta} - \mu_j)(x_{k\gamma} - \mu_\gamma)(x_{l\delta} - \mu_l)$$

$$= \frac{1}{N}(1 + \kappa)(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$$

$$+ \frac{N-1}{N}(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}),$$

$$(8) \quad \mathscr{E} \sum_{\alpha=1}^{N} (x_{i\alpha} - \mu_i)(x_{j\alpha} - \mu_j) \frac{1}{N} \sum_{\beta, \gamma = 1}^{N} (x_{k\beta} - \mu_k)(x_{l\gamma} - \mu_\gamma)$$

$$= (1 + \kappa)(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) + (N-1)\sigma_{ij}\sigma_{kl}. \qquad \blacksquare$$

It will be convenient to use more matrix algebra. Define vec $\boldsymbol{B}$, $\boldsymbol{B} \otimes \boldsymbol{C}$ (the Kronecker product), and $\boldsymbol{K}_{mn}$ (the commutator matrix) by

$$(9) \qquad \operatorname{vec} \boldsymbol{B} = \operatorname{vec}(\boldsymbol{b}_1, \dots, \boldsymbol{b}_n) = \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_n \end{bmatrix},$$

$$(10) \qquad \boldsymbol{B} \otimes \boldsymbol{C} = \begin{bmatrix} b_{11}\boldsymbol{C} & \cdots & b_{1n}\boldsymbol{C} \\ \vdots & & \vdots \\ b_{m1}\boldsymbol{C} & \cdots & b_{mn}\boldsymbol{C} \end{bmatrix},$$

$$(11) \qquad \boldsymbol{K}_{mn} \operatorname{vec} \boldsymbol{B} = \operatorname{vec} \boldsymbol{B}'.$$

See, e.g., Magnus and Neudecker (1979) or Section A.5 of the Appendix. We can rewrite (4) as

$$(12) \quad \mathscr{C}(\operatorname{vec} \boldsymbol{S}) = \mathscr{E}(\operatorname{vec} \boldsymbol{S} - \operatorname{vec} \boldsymbol{\Sigma})(\operatorname{vec} \boldsymbol{S} - \operatorname{vec} \boldsymbol{\Sigma})'$$

$$= \frac{n\kappa + N}{nN}(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{pp})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \frac{\kappa}{N} \operatorname{vec} \boldsymbol{\Sigma} (\operatorname{vec} \boldsymbol{\Sigma})'.$$

Theorem 3.6.2

$$(13) \quad \sqrt{n} \begin{bmatrix} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \\ \operatorname{vec} \boldsymbol{S} - \operatorname{vec} \boldsymbol{\Sigma} \end{bmatrix}$$

$$\xrightarrow{d} N\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & (\kappa + 1)(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{pp})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \kappa \operatorname{vec} \boldsymbol{\Sigma} (\operatorname{vec} \boldsymbol{\Sigma})' \end{pmatrix}\right].$$

This theorem follows from the central limit theorem for independent identically distributed random vectors (with finite fourth moments). The theorem forms the basis for large-sample inference.

3.6.2. Estimation of the Kurtosis Parameter

To apply the large-sample distribution theory derived for normal distributions to problems of inference for elliptically contoured distributions it is necessary to know or estimate the kurtosis parameter κ. Note that

$$(14)\qquad \mathscr{E}\left[(X-\mu)'\Sigma^{-1}(X-\mu)\right]^2 = \left(\frac{p}{\mathscr{E}R^2}\right)^2 \mathscr{E}(Y'Y)^2$$
$$= \frac{p^2\mathscr{E}R^4}{(\mathscr{E}R^2)^2} = p(p+2)(1+\kappa).$$

Since $\bar{x} \xrightarrow{p} \mu$ and $S \xrightarrow{p} \Sigma$,

$$(15)\qquad \frac{1}{N}\sum_{\alpha=1}^{N}\left[(x_\alpha-\bar{x})'S^{-1}(x_\alpha-\bar{x})\right]^2 \xrightarrow{p} p(p+2)(1+\kappa).$$

A consistent estimator of κ is

$$(16)\qquad \hat{\kappa} = \frac{1}{p(p+2)}\frac{1}{N}\sum_{\alpha=1}^{N}\left[(x_\alpha-\bar{x})'S^{-1}(x_\alpha-\bar{x})\right]^2 - 1.$$

Mardia (1970) proposed using M to form a consistent estimator of κ.

3.6.3. Maximum Likelihood Estimation

We have considered using S as an estimator of $\Sigma = (\mathscr{E}R^2/p)\Lambda$. When the parent distribution is normal, S is the sufficient statistic invariant with respect to translations and hence is the efficient unbiased estimator. Now we study other estimators.

We consider first the maximum likelihood estimators of μ and Λ when the form of the density $g(\cdot)$ is known. The logarithm of the likelihood function is

$$(17)\qquad \log L = -\frac{N}{2}\log|\Lambda| + \sum_{\alpha=1}^{N}\log g\left[(x_\alpha-\mu)'\Lambda^{-1}(x_\alpha-\mu)\right].$$

The derivatives of $\log L$ with respect to the components of $\boldsymbol{\mu}$ are

$$\frac{\partial \log L}{\partial \boldsymbol{\mu}} = -2 \sum_{\alpha=1}^{N} \frac{g'[(x_\alpha - \boldsymbol{\mu})'\Lambda^{-1}(x_\alpha - \boldsymbol{\mu})]}{g[(x_\alpha - \boldsymbol{\mu})'\Lambda^{-1}(x_\alpha - \boldsymbol{\mu})]} \Lambda^{-1}(x_\alpha - \boldsymbol{\mu}). \tag{18}$$

Setting the vector of derivatives equal to $\mathbf{0}$ leads to the equation

(19)
$$\sum_{\alpha=1}^{N} \frac{g'[(x_\alpha - \hat{\boldsymbol{\mu}})'\hat{\Lambda}^{-1}(x_\alpha - \hat{\boldsymbol{\mu}})]}{g[(x_\alpha - \hat{\boldsymbol{\mu}})'\hat{\Lambda}^{-1}(x_\alpha - \hat{\boldsymbol{\mu}})]} x_\alpha = \hat{\boldsymbol{\mu}} \sum_{\alpha=1}^{N} \frac{g'[(x_\alpha - \hat{\boldsymbol{\mu}})'\hat{\Lambda}^{-1}(x_\alpha - \hat{\boldsymbol{\mu}})]}{g[(x_\alpha - \hat{\boldsymbol{\mu}})'\hat{\Lambda}^{-1}(x_\alpha - \hat{\boldsymbol{\mu}})]}.$$

Setting equal to 0 the derivatives of $\log L$ with respect to the elements of Λ^{-1} gives

$$\hat{\Lambda} = -\frac{2}{N} \sum_{\alpha=1}^{N} \frac{g'[(x_\alpha - \hat{\boldsymbol{\mu}})'\hat{\Lambda}^{-1}(x_\alpha - \hat{\boldsymbol{\mu}})]}{g[(x_\alpha - \hat{\boldsymbol{\mu}})'\hat{\Lambda}^{-1}(x_\alpha - \hat{\boldsymbol{\mu}})]} (x_\alpha - \hat{\boldsymbol{\mu}})(x_\alpha - \hat{\boldsymbol{\mu}})'. \tag{20}$$

The estimator $\hat{\Lambda}$ is a kind of weighted average of the rank 1 matrices $(x_\alpha - \hat{\boldsymbol{\mu}})(x_\alpha - \hat{\boldsymbol{\mu}})'$. In the normal case the weights are $1/N$. In most cases (19) and (20) cannot be solved explicitly, but the solution may be approximated by iterative methods.

The covariance matrix of the limiting normal distribution of $\sqrt{N}(\operatorname{vec} \hat{\Lambda} - \operatorname{vec} \Lambda)$ is

$$\mathscr{C}(\operatorname{vec} \hat{\Lambda}) = \sigma_{1g}(I_{p^2} + K_{pp})(\Lambda \otimes \Lambda) + \sigma_{2g} \operatorname{vec} \Lambda (\operatorname{vec} \Lambda)', \tag{21}$$

where

$$\sigma_{1g} = \frac{p(p+2)}{4\mathscr{E}\left[\frac{g'(R^2)}{g(R^2)} R^2\right]^2}, \tag{22}$$

$$\sigma_{2g} = \frac{2\sigma_{1g}(1 - \sigma_{1g})}{2 + p(1 - \sigma_{1g})}. \tag{23}$$

See Tyler (1982).

3.6.4. Elliptically Contoured Matrix Distributions

Let

$$Y = \begin{bmatrix} y_1' \\ \vdots \\ y_N' \end{bmatrix} \tag{24}$$

be an $N \times p$ random matrix with density $g(Y'Y) = g(\sum_{\alpha=1}^{N} y_\alpha y_\alpha')$. Note that the density $g(Y'Y)$ is invariant with respect to orthogonal transformations $Y^* = O_N Y$. Such densities are known as *left spherical matrix* densities. An example is the density of N observations from $N(\mathbf{0}, I_p)$,

$$g(Y'Y) = \frac{1}{(2\pi)^{\frac{1}{2}pN}} e^{-\frac{1}{2}\operatorname{tr} Y'Y}. \tag{25}$$

In this example Y is also *right spherical*: $YO_p \overset{d}{=} Y$. When Y is both left spherical and right spherical, it is known as *spherical*. Further, if Y has the density (25), vec Y is spherical; in general if Y has a density, the density is of the form

$$\begin{aligned} g(\mathrm{t} \cdot Y'Y) &= g\left(\sum_{\alpha=1}^{N} \sum_{i=1}^{p} y_{i\alpha}^2 \right) = g(\operatorname{tr} YY') \\ &= g\left[(\operatorname{vec} Y)' \operatorname{vec} Y \right] = g\left[(\operatorname{vec} Y')' \operatorname{vec} Y' \right]. \end{aligned} \tag{26}$$

We call this model *vector-spherical*. Define

$$X = YC' + \varepsilon_N \boldsymbol{\mu}', \tag{27}$$

where $C'\Lambda^{-1}C = I_p$ and $\varepsilon_N' = (1, \dots, 1)$. Since (27) is equivalent to $Y = (X - \varepsilon_N \boldsymbol{\mu}')(C')^{-1}$ and $(C')^{-1}C^{-1} = \Lambda^{-1}$, the matrix X has the density

$$\begin{aligned} &|\Lambda|^{-N/2} g\left[\operatorname{tr}(X - \varepsilon_N \boldsymbol{\mu}') \Lambda^{-1} (X - \varepsilon_N \boldsymbol{\mu}')' \right] \\ &= |\Lambda|^{-N/2} g\left[\sum_{\alpha=1}^{N} (x_\alpha - \boldsymbol{\mu})' \Lambda^{-1} (x_\alpha - \boldsymbol{\mu}) \right]. \end{aligned} \tag{28}$$

From (26) we deduce that vec Y has the representation

$$\operatorname{vec} Y \overset{d}{=} R \operatorname{vec} U, \tag{29}$$

where $w = R^2$ has the density

$$\frac{\pi^{\frac{1}{2}Np}}{\Gamma(Np/2)} w^{\frac{1}{2}Np-1} g(w), \tag{30}$$

vec U has the uniform distribution on $\sum_{\alpha=1}^{N} \sum_{i=1}^{p} u_{i\alpha}^2 = 1$, and R and vec U are independent. The covariance matrix of vec Y is

$$\mathscr{E} \operatorname{vec} Y (\operatorname{vec} Y)' = \frac{\mathscr{E}R^2}{Np} I_{Np} = \frac{\mathscr{E}R^2}{Np} (I_p \otimes I_N). \tag{31}$$

Since $\operatorname{vec} \boldsymbol{FGH} = (\boldsymbol{H}' \otimes \mathbf{F}) \operatorname{vec} \boldsymbol{G}$ for any conformable matrices $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{H}$, we can write (27) as

$$\text{(32)} \qquad \operatorname{vec} \boldsymbol{X} = (\boldsymbol{C} \otimes \boldsymbol{I}_N) \operatorname{vec} \boldsymbol{Y} + \boldsymbol{\mu} \otimes \boldsymbol{\varepsilon}_N .$$

Thus

$$\text{(33)} \qquad \mathscr{E} \operatorname{vec} \boldsymbol{X} = \boldsymbol{\mu} \otimes \boldsymbol{\varepsilon}_N ,$$

$$\text{(34)} \qquad \mathscr{C}(\operatorname{vec} \boldsymbol{X}) = (\boldsymbol{C} \otimes \boldsymbol{I}_N) \mathscr{C}(\operatorname{vec} \boldsymbol{Y})(\boldsymbol{C}' \otimes \boldsymbol{I}_N) = \frac{\mathscr{E}R^2}{Np} \boldsymbol{\Lambda} \otimes \boldsymbol{I}_N ,$$

$$\text{(35)} \qquad \mathscr{E}(\text{row of } \boldsymbol{X}) = \boldsymbol{\mu}' ,$$

$$\text{(36)} \qquad \mathscr{C}(\text{row of } \boldsymbol{X}') = \frac{\mathscr{E}R^2}{Np} \boldsymbol{\Lambda} .$$

The rows of $\boldsymbol{X}$ are uncorrelated (though not necessarily independent). From (32) we obtain

$$\text{(37)} \qquad \operatorname{vec} \boldsymbol{X} \stackrel{d}{=} R(\boldsymbol{C} \otimes \boldsymbol{I}_N) \operatorname{vec} \boldsymbol{U} + \boldsymbol{\mu} \otimes \boldsymbol{\varepsilon}_N ,$$

$$\text{(38)} \qquad \boldsymbol{X} \stackrel{d}{=} R\boldsymbol{U}\boldsymbol{C}' + \boldsymbol{\varepsilon}_N \boldsymbol{\mu}' .$$

Since $\boldsymbol{X} - \boldsymbol{\varepsilon}_N \boldsymbol{\mu}' = (\boldsymbol{X} - \boldsymbol{\varepsilon}_N \bar{\boldsymbol{x}}') + \boldsymbol{\varepsilon}_N (\bar{\boldsymbol{X}} - \boldsymbol{\mu})'$ and $\boldsymbol{\varepsilon}_N'(\boldsymbol{X} - \boldsymbol{\varepsilon}_N \bar{\boldsymbol{x}}') = \mathbf{0}$, we can write the density of $\boldsymbol{X}$ as

$$\text{(39)} \qquad |\boldsymbol{\Lambda}|^{-N/2} g\left[\operatorname{tr} \boldsymbol{\Lambda}^{-1}(\boldsymbol{x} - \boldsymbol{\varepsilon}_N \bar{\boldsymbol{x}}')'(\boldsymbol{X} - \boldsymbol{\varepsilon}_N \bar{\boldsymbol{x}}') + N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \boldsymbol{\Lambda}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right],$$

where $\bar{\boldsymbol{x}} = (1/N)\boldsymbol{X}'\boldsymbol{\varepsilon}_N$. This shows that a sufficient set of statistics for $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ is $\bar{\boldsymbol{x}}$ and $n\boldsymbol{S} = (\boldsymbol{X} - \boldsymbol{\varepsilon}_N \bar{\boldsymbol{x}}')'(\boldsymbol{X} - \boldsymbol{\varepsilon}_N \bar{\boldsymbol{x}}')$, as for the normal distribution. The maximum likelihood estimators can be derived from the following theorem, which will be used later for other models.

Theorem 3.6.3. *Suppose the m-component vector $\boldsymbol{Z}$ has the density $|\boldsymbol{\Phi}|^{-\frac{1}{2}} h[(\boldsymbol{z} - \boldsymbol{\nu})' \boldsymbol{\Phi}^{-1}(\boldsymbol{z} - \boldsymbol{\nu})]$, where $w^{\frac{1}{2}m} h(w)$ has a finite positive maximum at w_h and $\boldsymbol{\Phi}$ is a positive definite matrix. Let Ω be a set in the space of $(\boldsymbol{\nu}, \boldsymbol{\Phi})$ such that if $(\boldsymbol{\nu}, \boldsymbol{\Phi}) \in \Omega$ then $(\boldsymbol{\nu}, c\boldsymbol{\Phi}) \in \Omega$ for all $c > 0$. Suppose that on the basis of an observation $\boldsymbol{z}$ when $h(w) = \text{const}\, e^{-\frac{1}{2}w}$ (i.e., $\boldsymbol{Z}$ has a normal distribution) the maximum likelihood estimator $(\bar{\boldsymbol{\nu}}, \overline{\boldsymbol{\Phi}}) \in \Omega$ exists and is unique with $\overline{\boldsymbol{\Phi}}$ positive definite with probability 1. Then the maximum likelihood estimator of $(\boldsymbol{\nu}, \boldsymbol{\Phi})$ for arbitrary $h(\cdot)$ is*

$$\text{(40)} \qquad \hat{\boldsymbol{\nu}} = \bar{\boldsymbol{\nu}}, \qquad \hat{\boldsymbol{\Phi}} = \frac{m}{w_h} \overline{\boldsymbol{\Phi}},$$

and the maximum of the likelihood is $|\hat{\Phi}|^{-\frac{1}{2}}h(w_h)$ [Anderson, Fang, and Hsu (1986)].

Proof. Let $\Psi = |\Phi|^{-1/m}\Phi$ and

$$d = (z-\nu)'\Phi^{-1}(z-\nu) = \frac{(z-\nu)'\Psi^{-1}(z-\nu)}{|\Phi|^{1/m}}. \tag{41}$$

Then $(\nu, \Phi) \in \Omega$ and $|\Psi| = 1$. The likelihood is

$$\left[(z-\nu)'\Psi^{-1}(z-\nu)\right]^{-\frac{1}{2}m} d^{\frac{1}{2}m} h(d). \tag{42}$$

Under normality $h(d) = (2\pi)^{-\frac{1}{2}m}e^{-\frac{1}{2}d}$, and the maximum of (42) is attained at $\nu = \bar{\nu}$, $\Psi = \bar{\Psi} = |\bar{\Phi}|^{-1/m}\bar{\Phi}$, and $d = m$. For arbitrary $h(\cdot)$ the maximum of (42) is attained at $\hat{\nu} = \bar{\nu}$, $\hat{B} = \bar{B}$, and $\hat{d} = w_h$. Then the maximum likelihood estimator of Φ is

$$\hat{\Phi} = |\hat{\Phi}|^{1/m}\hat{\Psi} = \frac{|\hat{\Phi}|^{1/m}}{|\bar{\Phi}|^{1/m}}\bar{\Phi}. \tag{43}$$

Then (40) follows from (43) by use of (41). ■

Theorem 3.6.4. *Let X $(N \times p)$ have the density* (28), *where $w^{\frac{1}{2}Np}g(w)$ has a finite positive maximum at w_g. Then the maximum likelihood estimators of μ and Λ are*

$$\hat{\mu} = \bar{x}, \qquad \hat{\Lambda} = \frac{Np}{w_g}A, \tag{44}$$

where $A = \sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$.

Corollary 3.6.1. *Let X $(N \times p)$ have the density* (28). *Then the maximum likelihood estimators of ν, $(\lambda_{11}, \dots, \lambda_{pp})$, and ρ_{ij}, $i, j = 1, \dots, p$, are $\bar{x}$, $(p/w_g)(a_{11}, \dots, a_{pp})$, and $a_{ij}/\sqrt{a_{ii}a_{jj}}$, $i, j = 1, \dots, p$.*

Proof. Corollary 3.6.1 follows from Theorem 3.6.3 and Corollary 3.2.1. ■

Theorem 3.6.5. *Let $f(X)$ be a vector-valued function of X $(N \times p)$ such that*

$$f(X + \varepsilon_N \nu') = f(X) \tag{45}$$

for all $\boldsymbol{\nu}$ *and*

$$f(cX) = f(X) \tag{46}$$

for all c. *Then the distribution of* $f(X)$ *where* X *has an arbitrary density* (28) *is the same as its distribution where* X *has the normal density* (28).

Proof. Substitution of the representation (27) into $f(X)$ gives

$$f(X) = f(YC' + \varepsilon_N \boldsymbol{\mu}') = f(YC') \tag{47}$$

by (45). Let $f(X) = h(\operatorname{vec} X)$. Then by (46), $h(cX) = h(X)$ and

$$\begin{aligned} f(YC') &= h\big[(C \otimes I_N)\operatorname{vec} Y\big] = h\big[R(C \otimes I_N)\operatorname{vec} U\big] \\ &= h\big[(C \otimes I_N)\operatorname{vec} U\big]. \qquad \blacksquare \end{aligned} \tag{48}$$

Any statistic satisfying (45) and (46) has the same distribution for all $g(\cdot)$. Hence, if its distribution is known for the normal case, the distribution is valid for all elliptically contoured distributions.

Any function of the sufficient set of statistics that is translation-invariant, that is, that satisfies (45), is a function of S. Thus inference concerning Σ can be based on S.

Corollary 3.6.2. *Let* $f(X)$ *be a vector-valued function of* X $(N \times p)$ *such that* (46) *holds for all* c. *Then the distribution of* $f(X)$ *where* X *has arbitrary density* (28) *with* $\boldsymbol{\mu} = \mathbf{0}$ *is the same as its distribution where* X *has normal density* (28) *with* $\boldsymbol{\mu} = \mathbf{0}$.

Fang and Zhang (1990) give this corollary as Theorem 2.5.8.

PROBLEMS

3.1. (Sec. 3.2) Find $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, and $(\hat{\rho}_{ij})$ for the data given in Table 3.3, taken from Frets (1921).

3.2. (Sec. 3.2) Verify the numerical results of (21).

3.3. (Sec. 3.2) Compute $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, S, and $\hat{\rho}$ for the following pairs of observations: $(34, 55), (12, 29), (33, 75), (44, 89), (89, 62), (59, 69), (50, 41), (88, 67)$. Plot the observations.

3.4. (Sec. 3.2) Use the facts that $|C^*| = \Pi \lambda_i$, $\operatorname{tr} C^* = \Sigma \lambda_i$, and $C^* = I$ if $\lambda_1 = \cdots = \lambda_p = 1$, where $\lambda_1, \dots, \lambda_p$ are the characteristic roots of C^*, to prove Lemma 3.2.2. [*Hint:* Use f as given in (12).]

Table 3.3[†]. Head Lengths and Breadths of Brothers

Head Length, First Son, x_1	Head Breadth, First Son, x_2	Head Length, Second Son, x_3	Head Breadth, Second Son, x_4
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151
179	158	186	148
183	147	174	147
174	150	185	152
190	159	195	157
188	151	187	158
163	137	161	130
195	155	183	158
186	153	173	148
181	145	182	146
175	140	165	137
192	154	185	152
174	143	178	147
176	139	176	143
197	167	200	158
190	163	187	150

[†]These data, used in examples in the first edition of this book, came from Rao (1952), p. 245. Izenman (1980) has indicated some entries were apparently incorrectly copied from Frets (1921) and corrected them (p. 579).

3.5. (Sec. 3.2) Let x_1 be the body weight (in kilograms) of a cat and x_2 the heart weight (in grams). [Data from Fisher (1947b).]

(a) In a sample of 47 female cats the relevant data are

$$\Sigma x_\alpha = \begin{pmatrix} 110.9 \\ 432.5 \end{pmatrix}, \qquad \Sigma x_\alpha x_\alpha' = \begin{pmatrix} 265.13 & 1029.62 \\ 1029.62 & 4064.71 \end{pmatrix}.$$

Find $\hat{\mu}$, $\hat{\Sigma}$, S, and $\hat{\rho}$.

Table 3.4. Four Measurements on Three Species of Iris (in centimeters)

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1

Table 3.4. (*Continued*)

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

(b) In a sample of 97 male cats the relevant data are

$$\Sigma x_\alpha = \begin{pmatrix} 281.3 \\ 1098.3 \end{pmatrix}, \qquad \Sigma x_\alpha x_\alpha' = \begin{pmatrix} 836.75 & 3275.55 \\ 3275.55 & 13056.17 \end{pmatrix}.$$

Find $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, S, and $\hat{\rho}$.

3.6. Find $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, and $(\hat{\rho}_{ij})$ for *Iris setosa* from Table 3.4, taken from Edgar Anderson's famous iris data [Fisher (1936)].

3.7. (Sec. 3.2) *Invariance of the sample correlation coefficient.* Prove that r_{12} is an invariant characteristic of the sufficient statistics $\bar{x}$ and S of a bivariate sample under location and scale transformations ($x_{i\alpha}^* = b_i x_{i\alpha} + c_i$, $b_i > 0$, $i = 1, 2$, $\alpha = 1, \dots, N$) and that every function of $\bar{x}$ and S that is invariant is a function of r_{12}. [*Hint:* See Theorem 2.3.2.]

3.8. (Sec. 3.2) Prove Lemma 3.2.2 by induction. [*Hint:* Let $H_1 = h_{11}$,

$$H_i = \begin{pmatrix} H_{i-1} & h_{(i)} \\ h_{(i)}' & h_{ii} \end{pmatrix}, \qquad i = 2, \dots, p,$$

and use Problem 2.36.]

3.9. (Sec. 3.2) Show that

$$\frac{1}{N(N-1)} \sum_{\alpha<\beta} (x_\alpha - x_\beta)(x_\alpha - x_\beta)' = \frac{1}{N-1} \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'.$$

(Note: When $p = 1$, the left-hand side is the average squared differences of the observations.)

3.10. (Sec. 3.2) *Estimation of Σ when μ is known.* Show that if $x_1, \dots, x_N$ constitute a sample from $N(\mu, \Sigma)$ and μ is known, then $(1/N)\sum_{\alpha=1}^N (x_\alpha - \mu)(x_\alpha - \mu)'$ is the maximum likelihood estimator of Σ.

3.11. (Sec. 3.2) *Estimation of parameters of a complex normal distribution.* Let $z_1, \dots, z_N$ be N observations from the complex normal distributions with mean θ and covariance matrix P. (See Problem 2.64.)

(a) Show that the maximum likelihood estimators of θ and P are

$$\hat{\theta} = \bar{z} = \frac{1}{N}\sum_{\alpha=1}^N z_\alpha, \qquad \hat{P} = \frac{1}{N}\sum_{\alpha=1}^N (z_\alpha - \bar{z})(z_\alpha - \bar{z})^*.$$

(b) Show that $\bar{z}$ has the complex normal distribution with mean θ and covariance matrix $(1/N)P$.

(c) Show that $\bar{z}$ and $\hat{P}$ are independently distributed and that $N\hat{P}$ has the distribution of $\sum_{\alpha=1}^n W_\alpha W_\alpha^*$, where $W_1, \dots, W_n$ are independently distributed, each according to the complex normal distribution with mean 0 and covariance matrix P, and $n = N - 1$.

3.12. (Sec. 3.2) Prove Lemma 3.2.2 by using Lemma 3.2.3 and showing $N \log|C| - \operatorname{tr} CD$ has a maximum at $C = ND^{-1}$ by setting the derivatives of this function with respect to the elements of $C = \Sigma^{-1}$ equal to 0. Show that the function of C tends to $-\infty$ as C tends to a singular matrix or as one or more elements of C tend to ∞ and/or $-\infty$ (nondiagonal elements); for this latter, the equivalent of (13) can be used.

3.13. (Sec. 3.3) Let X_α be distributed according to $N(\gamma c_\alpha, \Sigma)$, $\alpha = 1, \dots, N$, where $\sum c_\alpha^2 > 0$. Show that the distribution of $g = (1/\sum c_\alpha^2)\sum c_\alpha X_\alpha$ is $N[\gamma, (1/\sum c_\alpha^2)\Sigma]$. Show that $E = \sum_\alpha (X_\alpha - g c_\alpha)(X_\alpha - g c_\alpha)'$ is independently distributed as $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$, where $Z_1, \dots, Z_N$ are independent, each with distribution $N(0, \Sigma)$. [*Hint:* Let $Z_\alpha = \sum b_{\alpha\beta} X_\beta$, where $b_{N\beta} = c_\beta / \sqrt{\sum c_\alpha^2}$ and B is orthogonal.]

3.14. (Sec. 3.3) Prove that the power of the test in (19) is a function only of p and $[N_1 N_2/(N_1 + N_2)](\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$, given α.

3.15. (Sec. 3.3) *Efficiency of the mean.* Prove that $\bar{x}$ is efficient for estimating μ.

3.16. (Sec. 3.3) Prove that $\bar{x}$ and S have efficiency $[(N-1)/N]^{p(p+1)/2}$ for estimating μ and Σ.

3.17. (Sec. 3.2) Prove that $\Pr\{|A| = 0\} = 0$ for A defined by (4) when $N > p$. [*Hint:* Argue that if $Z_p^* = (Z_1, \dots, Z_p)$, then $|Z_p^*| \neq 0$ implies $A = Z_p^* Z_p^{*\prime} + \sum_{\alpha=p+1}^{N-1} Z_\alpha Z_\alpha'$ is positive definite. Prove $\Pr\{|Z_j^*| = Z_{jj}|Z_{j-1}^*| + \sum_{i=1}^{j-1} Z_{ij} \operatorname{cof}(Z_{ij}) = 0\} = 0$ by induction, $j = 2, \dots, p$.]

3.18. (Sec. 3.4) Prove

$$I-\Phi(\Phi+\Sigma)^{-1}=\Sigma(\Phi+\Sigma)^{-1}.$$

$$\Phi-\Phi(\Phi+\Sigma)^{-1}\Phi=(\Phi^{-1}+\Sigma^{-1})^{-1}.$$

3.19. (Sec. 3.4) Prove $(1/N)\Sigma_{\alpha=1}^{N}(x_\alpha-\mu)(x_\alpha-\mu)'$ is an unbiased estimator of Σ when μ is known.

3.20. (Sec. 3.4) Show that

$$\Phi\left(\Phi+\frac{1}{N}\Sigma\right)^{-1}x+\frac{1}{N}\Sigma\left(\Phi+\frac{1}{N}\Sigma\right)^{-1}\nu=(\Phi^{-1}+N\Sigma^{-1})^{-1}(N\Sigma^{-1}x+\Phi^{-1}\nu)$$

3.21. (Sec. 3.5) Demonstrate Lemma 3.5.1 using integration by parts.

3.22. (Sec. 3.5) Show that

$$\int_\theta^\infty\int_y^\infty\left|f'(y)(x-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\right|dx\,dy=\int_\theta^\infty|f'(y)|\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-\theta)^2}\,dy,$$

$$\int_{-\infty}^\theta\int_{-\infty}^y\left|f'(y)(\theta-x)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\right|dx\,dy=\int_{-\infty}^\theta|f'(y)|\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-\theta)^2}\,dy.$$

3.23. Let $Z(k)=(Z_{ij}(k))$, where $i=1,\dots,p$, $j=1,\dots,q$ and $k=1,2,\dots$, be a sequence of random matrices. Let one norm of a matrix A be $N_1(A)=\max_{i,j}\operatorname{mod}(a_{ij})$, and another be $N_2(A)=\Sigma_{i,j}a_{ij}^2=\operatorname{tr}AA'$. Some alternative ways of defining stochastic convergence of $Z(k)$ to B $(p\times q)$ are

(a) $N_1(Z(k)-B)$ converges stochastically to 0,
(b) $N_2(Z(k)-B)$ converges stochastically to 0, and
(c) $Z_{ij}(k)-b_{ij}$ converges stochastically to 0, $i=1,\dots,p$, $j=1,\dots,q$.

Prove that these three definitions are equivalent. Note that the definition of $X(k)$ converging stochastically to a is that for every arbitrary positive δ and ε, we can find K large enough so that for $k>K$

$$\Pr\{|X(k)-a|<\delta\}>1-\varepsilon.$$

3.24. (Sec. 3.2) *Covariance matrices with linear structure* [Anderson (1969)]. Let

$$\text{(i)}\qquad \Sigma=\sum_{g=0}^{q}\sigma_g G_g,$$

where $G_0, \dots, G_q$ are given symmetric matrices such that there exists at least one $(q+1)$-tuplet $\sigma_0, \sigma_1, \dots, \sigma_q$ such that (i) is positive definite. Show that the likelihood equations based on N observations are

$$\text{(ii)} \qquad -\frac{N}{2}\operatorname{tr}\Sigma^{-1}G_g + \frac{1}{2}\operatorname{tr}A\Sigma^{-1}G_g\Sigma^{-1} = 0, \qquad g = 0, 1, \dots, q.$$

Show that an iterative (scoring) method can be based on

$$\text{(iii)} \qquad \sum_{h=0}^{q} \operatorname{tr}\hat{\Sigma}_{i-1}^{-1}G_g\hat{\Sigma}_{i-1}^{-1}G_h\hat{\sigma}_h^{(i)} = \frac{1}{N}\operatorname{tr}\hat{\Sigma}_{i-1}^{-1}G_g\hat{\Sigma}_{i-1}^{-1}A, \qquad g = 0, 1, \dots, q,$$

where $\hat{\Sigma}_{i-1}\Sigma_{g=0}^{q}\hat{\sigma}_g^{(i-1)}G_g$.

CHAPTER 4

The Distributions and Uses of Sample Correlation Coefficients

4.1. INTRODUCTION

In Chapter 2, in which the multivariate normal distribution was introduced, it was shown that a measure of dependence between two normal variates is the correlation coefficient $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$. In a conditional distribution of $X_1, \dots, X_q$ given $X_{q+1} = x_{q+1}, \dots, X_p = x_p$, the partial correlation $\rho_{ij \cdot q+1, \dots, p}$ measures the dependence between X_i and X_j. The third kind of correlation discussed was the multiple correlation which measures the relationship between one variate and a set of others. In this chapter we treat the sample equivalents of these quantities; they are point estimates of the population quantities. The distributions of the sample correlations are found. Tests of hypotheses and confidence intervals are developed.

In the cases of joint normal distributions these correlation coefficients are the natural measures of dependence. In the population they are the only parameters except for location (means) and scale (standard deviations) parameters. In the sample the correlation coefficients are derived as the reasonable estimates of the population correlations. Since the sample means and standard deviations are location and scale estimates, the sample correlations (that is, the standardized sample second moments) give all possible information about the population correlations. The sample correlations are the functions of the sufficient statistics that are invariant with respect to location and scale transformations; the population correlations are the functions of the parameters that are invariant with respect to these transformations.

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

In *regression theory* or least squares, one variable is considered random or *dependent*, and the others fixed or *independent*. In correlation theory we consider several variables as random and treat them symmetrically. If we start with a joint normal distribution and hold all variables fixed except one, we obtain the least squares model because the expected value of the random variable in the conditional distribution is a linear function of the variables held fixed. The sample regression coefficients obtained in least squares are functions of the sample variances and correlations.

In testing independence we shall see that we arrive at the same tests in either case (i.e., in the joint normal distribution or in the conditional distribution of least squares). The probability theory under the null hypothesis is the same. The distribution of the test criterion when the null hypothesis is not true differs in the two cases. If all variables may be considered random, one uses correlation theory as given here; if only one variable is random, one uses least squares theory (which is considered in some generality in Chapter 8).

In Section 4.2 we derive the distribution of the sample correlation coefficient, first when the corresponding population correlation coefficient is 0 (the two normal variables being independent) and then for any value of the population coefficient. The Fisher z-transform yields a useful approximate normal distribution. Exact and approximate confidence intervals are developed. In Section 4.3 we carry out the same program for partial correlations, that is, correlations in conditional normal distributions. In Section 4.4 the distributions and other properties of the sample multiple correlation coefficient are studied. In Section 4.5 the asymptotic distributions of these correlations are derived for elliptically contoured distributions. A stochastic representation for a class of such distributions is found.

4.2. CORRELATION COEFFICIENT OF A BIVARIATE SAMPLE

4.2.1. The Distribution When the Population Correlation Coefficient Is Zero; Tests of the Hypothesis of Lack of Correlation

In Section 3.2 it was shown that if one has a sample (of p-component vectors) $x_1, \dots, x_N$ from a normal distribution, the maximum likelihood estimator of the correlation between X_i and X_j (two components of the random vector X) is

$$r_{ij} = \frac{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)^2}\sqrt{\sum_{\alpha=1}^{N}(x_{j\alpha} - \bar{x}_j)^2}}, \tag{1}$$

where $x_{i\alpha}$ is the ith component of $\boldsymbol{x}_\alpha$ and

$$\bar{x}_i = \frac{1}{N}\sum_{\alpha=1}^{N} x_{i\alpha}. \tag{2}$$

In this section we shall find the distribution of r_{ij} when the population correlation between X_i and X_j is zero, and we shall see how to use the sample correlation coefficient to test the hypothesis that the population coefficient is zero.

For convenience we shall treat r_{12}; the same theory holds for each r_{ij}. Since r_{12} depends only on the first two coordinates of each $\boldsymbol{x}_\alpha$, to find the distribution of r_{12} we need only consider the joint distribution of (x_{11}, x_{21}), $(x_{12}, x_{22}), \dots, (x_{1N}, x_{2N})$. We can reformulate the problems to be considered here, therefore, in terms of a bivariate normal distribution. Let $\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_N^*$ be observation vectors from

$$N\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho\\ \sigma_2\sigma_1\rho & \sigma_2^2\end{pmatrix}\right]. \tag{3}$$

We shall consider

$$r = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}}, \tag{4}$$

where

$$a_{ij} = \sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j), \qquad i, j = 1, 2. \tag{5}$$

and $\bar{x}_i$ is defined by (2), $x_{i\alpha}$ being the ith component of $\boldsymbol{x}_\alpha^*$.

From Section 3.3 we see that a_{11}, a_{12}, and a_{22} are distributed like

$$a_{ij} = \sum_{\alpha=1}^{n} z_{i\alpha} z_{j\alpha}, \qquad i, j = 1, 2, \tag{6}$$

where $n = N - 1$, $(z_{1\alpha}, z_{2\alpha})$ is distributed according to

$$N\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho\\ \sigma_2\sigma_1\rho & \sigma_2^2\end{pmatrix}\right], \tag{7}$$

and the pairs $(z_{11}, z_{21}), \dots, (z_{1N}, z_{2N})$ are independently distributed.

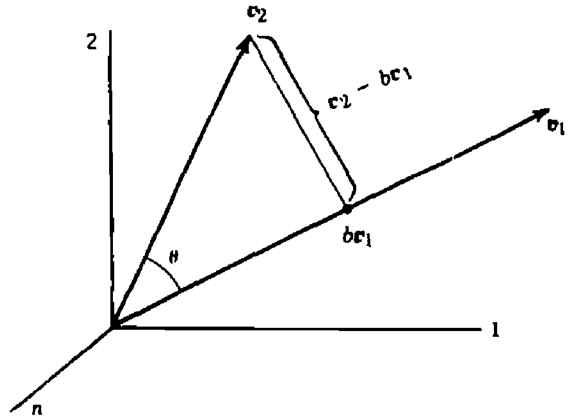

Figure 4.1

Define the n-component vector $v_i = (z_{i1}, \dots, z_{in})'$, $i = 1, 2$. These two vectors can be represented in an n-dimensional space; see Figure 4.1. The correlation coefficient is the cosine of the angle, say θ, between v_1 and v_2. (See Section 3.2.) To find the distribution of $\cos\theta$ we shall first find the distribution of $\cot\theta$. As shown in Section 3.2, if we let $b = v_2'v_1/v_1'v_1$, then $v_2 - bv_1$ is orthogonal to v_1 and

$$\cot\theta = \frac{b\|v_1\|}{\|v_2 - bv_1\|}. \tag{8}$$

If v_1 is fixed, we can rotate coordinate axes so that the first coordinate axis lies along v_1. Then bv_1 has only the first coordinate different from zero, and $v_2 - bv_1$ has this first coordinate equal to zero. We shall show that $\cot\theta$ is proportional to a t-variable when $\rho = 0$.

We use the following lemma.

Lemma 4.2.1. *If $Y_1, \dots, Y_n$ are independently distributed, if $Y_\alpha = (Y_\alpha^{(1)\prime}, Y_\alpha^{(2)\prime})'$ has the density $f(y_\alpha)$, and if the conditional density of $Y_\alpha^{(2)}$ given $Y_\alpha^{(1)} = y_\alpha^{(1)}$ is $f(y_\alpha^{(2)} | y_\alpha^{(1)})$, $\alpha = 1, \dots, n$. then in the conditional distribution of $Y_1^{(2)}, \dots, Y_n^{(2)}$ given $Y_1^{(1)} = y_1^{(1)}, \dots, Y_n^{(1)} = y_n^{(1)}$, the random vectors $Y_1^{(2)}, \dots, Y_n^{(2)}$ are independent and the density of $Y_\alpha^{(2)}$ is $f(y_\alpha^{(2)} | y_\alpha^{(1)})$, $\alpha = 1, \dots, n$.*

Proof. The marginal density of $Y_1^{(1)}, \dots, Y_n^{(1)}$ is $\prod_{\alpha=1}^n f_1(y_\alpha^{(1)})$, where $f_1(y_\alpha^{(1)})$ is the marginal density of $Y_\alpha^{(1)}$, and the conditional density of $Y_1^{(2)}, \dots, Y_n^{(2)}$ given $Y_1^{(1)} = y_1^{(1)}, \dots, Y_n^{(1)} = y_n^{(1)}$ is

$$\frac{\prod_{\alpha=1}^n f(y_\alpha)}{\prod_{\alpha=1}^n f_1(y_\alpha^{(1)})} = \prod_{\alpha=1}^n \frac{f(y_\alpha)}{f_1(y_\alpha^{(1)})} = \prod_{\alpha=1}^n f(y_\alpha^{(2)} | y_\alpha^{(1)}). \tag{9}$$

■

Write $V_i = (Z_{i1}, \dots, Z_{in})'$, $i = 1, 2$, to denote random vectors. The conditional distribution of $Z_{2\alpha}$ given $Z_{1\alpha} = z_{1\alpha}$ is $N(\beta z_{1\alpha}, \sigma^2)$, where $\beta = \rho\sigma_2/\sigma_1$ and $\sigma^2 = \sigma_2^2(1-\rho^2)$. (See Section 2.5.) The density of V_2 given $V_1 = v_1$ is $N(\beta v_1, \sigma^2 I)$ since the $Z_{2\alpha}$ are independent. Let $b = V_2' v_1 / v_1' v_1$ $(= a_{21}/a_{11})$, so that $b v_1'(V_2 - b v_1) = 0$, and let $U = (V_2 - b v_1)'(V_2 - b v_1) = V_2'V_2 - b^2 v_1' v_1$ $(= a_{22} - a_{12}^2/a_{11})$. Then $\cot\theta = b\sqrt{a_{11}/U}$. The rotation of coordinate axes involves choosing an $n \times n$ orthogonal matrix C with first row $(1/c)v_1'$, where $c^2 = v_1' v_1$.

We now apply Theorem 3.3.1 with $X_\alpha = Z_{2\alpha}$. Let $Y_\alpha = \sum_\beta c_{\alpha\beta} Z_{2\beta}$, $\alpha = 1, \dots, n$. Then $Y_1, \dots, Y_n$ are independently normally distributed with variance σ^2 and means

$$\mathscr{E} Y_1 = \sum_{\gamma=1}^{n} c_{1\gamma}\beta z_{1\gamma} = \frac{\beta}{c}\sum_{\gamma=1}^{n} z_{1\gamma}^2 = \beta c, \tag{10}$$

$$\mathscr{E} Y_\alpha = \sum_{\gamma=1}^{n} c_{\alpha\gamma}\beta z_{1\gamma} = \beta c \sum_{\gamma=1}^{n} c_{\alpha\gamma} c_{1\gamma} = 0, \qquad \alpha \neq 1. \tag{11}$$

We have $b = \sum_{\alpha=1}^n Z_{2\alpha} z_{1\alpha} / \sum_{\alpha=1}^n z_{1\alpha}^2 = c\sum_{\alpha=1}^n Z_{2\alpha} c_{1\alpha}/c^2 = Y_1/c$ and, from Lemma 3.3.1,

$$\begin{aligned} U &= \sum_{\alpha=1}^{n} Z_{2\alpha}^2 - b^2 \sum_{\alpha=1}^{n} z_{1\alpha}^2 = \sum_{\alpha=1}^{n} Y_\alpha^2 - Y_1^2 \\ &= \sum_{\alpha=2}^{n} Y_\alpha^2, \end{aligned} \tag{12}$$

which is independent of b. Then U/σ^2 has a χ^2-distribution with $n-1$ degrees of freedom.

Lemma 4.2.2. *If $(Z_{1\alpha}, Z_{2\alpha})$, $\alpha = 1, \dots, n$, are independent, each pair with density (7), then the conditional distributions of $b = \sum_{\alpha=1}^n Z_{2\alpha} Z_{1\alpha}/\sum_{\alpha=1}^n Z_{1\alpha}^2$ and $U/\sigma^2 = \sum_{\alpha=1}^n (Z_{2\alpha} - bZ_{1\alpha})^2/\sigma^2$ given $Z_{1\alpha} = z_{1\alpha}$, $\alpha = 1, \dots, n$, are $N(\beta, \sigma^2/c^2)$ $(c^2 = \sum_{\alpha=1}^n z_{1\alpha}^2)$ and χ^2 with $n-1$ degrees of freedom, respectively; and b and U are independent.*

If $\rho = 0$, then $\beta = 0$, and b is distributed conditionally according to $N(0, \sigma^2/c^2)$, and

$$\frac{cb/\sigma}{\sqrt{\dfrac{U/\sigma^2}{n-1}}} = \frac{cb}{\sqrt{\dfrac{U}{n-1}}} \tag{13}$$

has a conditional t-distribution with $n-1$ degrees of freedom. (See Problem 4.27.) However, this random variable is

$$\sqrt{n-1}\,\frac{\sqrt{a_{11}}\,a_{12}/a_{11}}{\sqrt{a_{22}-a_{12}^2/a_{11}}} = \sqrt{n-1}\,\frac{a_{12}/\sqrt{a_{11}a_{22}}}{\sqrt{1-\left[a_{12}^2/(a_{11}a_{22})\right]}} = \sqrt{n-1}\,\frac{r}{\sqrt{1-r^2}}. \tag{14}$$

Thus $\sqrt{n-1}\,r/\sqrt{1-r^2}$ has a conditional t-distribution with $n-1$ degrees of freedom. The density of t is

$$\frac{\Gamma(\frac{1}{2}n)}{\sqrt{n-1}\,\Gamma[\frac{1}{2}(n-1)]\sqrt{\pi}}\left(1+\frac{t^2}{n-1}\right)^{-\frac{1}{2}n}, \tag{15}$$

and the density of $W=r/\sqrt{1-r^2}$ is

$$\frac{\Gamma(\frac{1}{2}n)}{\Gamma[\frac{1}{2}(n-1)]\sqrt{\pi}}(1+w^2)^{-\frac{1}{2}n}. \tag{16}$$

Since $w=r(1-r^2)^{-\frac{1}{2}}$, we have $dw/dr=(1-r^2)^{-\frac{3}{2}}$. Therefore the density of r is (replacing n by $N-1$)

$$\frac{\Gamma[\frac{1}{2}(N-1)]}{\Gamma[\frac{1}{2}(N-2)]\sqrt{\pi}}(1-r^2)^{\frac{1}{2}(N-4)}. \tag{17}$$

It should be noted that (17) is the conditional density of r for v_1 fixed. However, since (17) does not depend on v_1, it is also the marginal density of r.

Theorem 4.2.1. *Let $X_1,\dots,X_N$ be independent, each with distribution $N(\mu,\Sigma)$. If $\rho_{ij}=0$, the density of r_{ij} defined by* (1) *is* (17).

From (17) we see that the density is symmetric about the origin. For $N>4$, it has a mode at $r=0$ and its order of contact with the r-axis at ± 1 is $\frac{1}{2}(N-5)$ for N odd and $\frac{1}{2}N-3$ for N even. Since the density is even, the odd moments are zero; in particular, the mean is zero. The even moments are found by integration (letting $x=r^2$ and using the definition of the beta function). That $\mathscr{E}r^{2m}=\Gamma[\frac{1}{2}(N-1)]\Gamma(m+\frac{1}{2})/\{\sqrt{\pi}\,\Gamma[\frac{1}{2}(N-1)+m]\}$ and in particular that the variance is $1/(N-1)$ may be verified by the reader.

The most important use of Theorem 4.2.1 is to find significance points for testing the hypothesis that a pair of variables are not correlated. Consider the hypothesis

$$H: \rho_{ij} = 0 \tag{18}$$

for some particular pair (i, j). It would seem reasonable to reject this hypothesis if the corresponding sample correlation coefficient were very different from zero. Now how do we decide what we mean by "very different"?

Let us suppose we are interested in testing H against the alternative hypotheses $\rho_{ij} > 0$. Then we reject H if the sample correlation coefficient r_{ij} is greater than some number r_0. The probability of rejecting H when H is true is

$$\int_{r_0}^{1} k_N(r)\, dr, \tag{19}$$

where $k_N(r)$ is (17), the density of a correlation coefficient based on N observations. We choose r_0 so (19) is the desired significance level. If we test H against alternatives $\rho_{ij} < 0$, we reject H when $r_{ij} < -r_0$.

Now suppose we are interested in alternatives $\rho_{ij} \neq 0$; that is, ρ_{ij} may be either positive or negative. Then we reject the hypothesis H if $r_{ij} > r_1$ or $r_{ij} < -r_1$. The probability of rejection when H is true is

$$\int_{-1}^{-r_1} k_N(r)\, dr + \int_{r_1}^{1} k_N(r)\, dr. \tag{20}$$

The number r_1 is chosen so that (20) is the desired significance level.

The significance points r_1 are given in many books, including Table VI of Fisher and Yates (1942); the index n in Table VI is equal to our $N-2$. Since $\sqrt{N-2}\,r/\sqrt{1-r^2}$ has the t-distribution with $N-2$ degrees of freedom, t-tables can also be used. Against alternatives $\rho_{ij} \neq 0$, reject H if

$$\sqrt{N-2}\,\frac{|r_{ij}|}{\sqrt{1-r_{ij}^2}} > t_{N-2}(\alpha), \tag{21}$$

where $t_{N-2}(\alpha)$ is the two-tailed significance point of the t-statistic with $N-2$ degrees of freedom for significance level α. Against alternatives $\rho_{ij} > 0$, reject H if

$$\sqrt{N-2}\,\frac{r_{ij}}{\sqrt{1-r_{ij}^2}} > t_{N-2}(2\alpha). \tag{22}$$

From (13) and (14) we see that $\sqrt{N-2}\,r/\sqrt{1-r^2}$ is the proper statistic for testing the hypothesis that the regression of V_2 on v_1 is zero. In terms of the original observation $\{x_{i\alpha}\}$, we have

$$(23)\qquad \sqrt{N-2}\,\frac{r}{\sqrt{1-r^2}} = \frac{b\sqrt{\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar{x}_1)^2}}{\sqrt{\sum_{\alpha=1}^{N}\left[x_{2\alpha}-\bar{x}_2-b(x_{1\alpha}-\bar{x}_1)\right]^2/(N-2)}},$$

where $b=\sum_{\alpha=1}^{N}(x_{2\alpha}-\bar{x}_2)(x_{1\alpha}-\bar{x}_1)/\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar{x}_1)^2$ is the least squares regression coefficient of $x_{2\alpha}$ on $x_{1\alpha}$. It is seen that the test of $\rho_{12}=0$ is equivalent to the test that the regression of X_2 on x_1 is zero (i.e., that $\rho_{12}\sigma_2/\sigma_1=0$).

To illustrate this procedure we consider the example given in Section 3.2. Let us test the null hypothesis that the effects of the two drugs are uncorrelated against the alternative that they are positively correlated. We shall use the 5% level of significance. For $N=10$, the 5% significance point (r_0) is 0.5494. Our observed correlation coefficient of 0.7952 is significant; we reject the hypothesis that the effects of the two drugs are independent.

4.2.2. The Distribution When the Population Correlation Coefficient Is Nonzero; Tests of Hypotheses and Confidence Intervals

To find the distribution of the sample correlation coefficient when the population coefficient is different from zero, we shall first derive the joint density of a_{11}, a_{12}, and a_{22}. In Section 4.2.1 we saw that, conditional on $\boldsymbol{v}_1$ held fixed, the random variables $b=a_{12}/a_{11}$ and $U/\sigma^2=(a_{22}-a_{12}^2/a_{11})/\sigma^2$ are distributed independently according to $N(\beta,\sigma^2/c^2)$ and the χ^2-distribution with $n-1$ degrees of freedom, respectively. Denoting the density of the χ^2-distribution by $g_{n-1}(u)$, we write the conditional density of b and U as $n(b|\beta,\sigma^2/a_{11})g_{n-1}(u/\sigma^2)/\sigma^2$. The joint density of V_1, b, and U is $n(v_1|\mathbf{0},\sigma_1^2 I)n(b|\beta,\sigma^2/a_{11})g_{n-1}(u/\sigma^2)/\sigma^2$. The marginal density of $V_1'V_1/\sigma_1^2=a_{11}/\sigma_1^2$ is $g_n(u)$; that is, the density of a_{11} is

$$(24)\qquad \frac{1}{\sigma_1^2}g_n\left(\frac{a_{11}}{\sigma_1^2}\right)=\int\cdots\int_{v_1'v_1=a_{11}} n(v_1|\mathbf{0},\sigma_1^2 I)\,dW,$$

where dW is the proper volume element.

The integration is over the sphere $v_1'v_1=a_{11}$; thus, dW is an element of area on this sphere. (See Problem 7.1 for the use of angular coordinates in defining dW.) Thus the joint density of b, U, and a_{11} is

$$(25)\quad \int_{v_1'v_1=a_{11}}\cdots\int n(b|\beta,\sigma^2/a_{11})g_{n-1}(u/\sigma^2)\frac{1}{\sigma^2}n(v_1|0,\sigma_1^2 I)\,dW$$

$$=\frac{g_n(a_{11}/\sigma_1^2)n(b|\beta,\sigma^2/a_{11})g_{n-1}(u/\sigma^2)}{\sigma_1^2\sigma^2}$$

$$=\frac{(a_{11})^{\frac{1}{2}n-1}}{(2\sigma_1^2)^{\frac{1}{2}n}\Gamma(\frac{1}{2}n)}\exp\left(-\frac{a_{11}}{2\sigma_1^2}\right)\frac{\sqrt{a_{11}}}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{a_{11}}{2\sigma^2}(b-\beta)^2\right]$$

$$\cdot\frac{1}{(2\sigma^2)^{\frac{1}{2}(n-1)}\Gamma[\frac{1}{2}(n-1)]}u^{\frac{1}{2}(n-3)}\exp\left(-\frac{u}{2\sigma^2}\right).$$

Now let $b=a_{12}/a_{11}$, $U=a_{22}-a_{12}^2/a_{11}$. The Jacobian is

$$(26)\quad \left|\frac{\partial(b,u)}{\partial(a_{12},a_{22})}\right|=\begin{vmatrix}\frac{1}{a_{11}} & 0\\ -2\frac{a_{12}}{a_{11}} & 1\end{vmatrix}=\frac{1}{a_{11}}.$$

Thus the density of a_{11}, a_{12}, and a_{22} for $a_{11}\ge 0$, $a_{22}\ge 0$, and $a_{11}a_{22}-a_{12}^2\ge 0$ is

$$(27)\quad \frac{a_{11}^{\frac{1}{2}(n-3)}\left(\dfrac{a_{11}a_{22}-a_{12}^2}{a_{11}}\right)^{\frac{1}{2}(n-3)}e^{-\frac{1}{2}Q}}{2^n\sigma_1^n\sigma_2^n(1-\rho^2)^{\frac{1}{2}n}\sqrt{\pi}\,\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]},$$

where

$$(28)\quad Q=\frac{a_{11}}{\sigma_1^2}+\frac{a_{11}}{\sigma^2}\left(\frac{a_{12}^2}{a_{11}^2}-2\rho\frac{\sigma_1\sigma_2}{\sigma_1^2}\frac{a_{12}}{a_{11}}+\frac{\rho^2\sigma_1^2\sigma_2^2}{\sigma_1^4}\right)+\frac{1}{\sigma^2}\left(a_{22}-\frac{a_{12}^2}{a_{11}}\right)$$

$$=a_{11}\left[\frac{1}{\sigma_1^2}+\frac{\rho^2\sigma_1^2\sigma_2^2}{\sigma_1^4\sigma_2^2(1-\rho^2)}\right]-2a_{12}\frac{\rho\sigma_2}{\sigma_1\sigma_2^2(1-\rho^2)}+\frac{a_{22}}{\sigma_2^2(1-\rho^2)}$$

$$=\frac{1}{1-\rho^2}\left(\frac{a_{11}}{\sigma_1^2}-2\rho\frac{a_{12}}{\sigma_1\sigma_2}+\frac{a_{22}}{\sigma_2^2}\right).$$

The density can be written

$$\frac{|A|^{\frac{1}{2}(n-3)}e^{-\frac{1}{2}Q}}{2^n|\Sigma|^{\frac{1}{2}n}\sqrt{\pi}\,\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]} \tag{29}$$

for A positive definite, and 0 otherwise. This is a special case of the Wishart density derived in Chapter 7.

We want to find the density of

$$r=\frac{a_{12}}{\sqrt{a_{11}a_{22}}}=\frac{a_{12}/(\sigma_1\sigma_2)}{\sqrt{(a_{11}/\sigma_1^2)(a_{22}/\sigma_2^2)}}=\frac{a_{12}^*}{\sqrt{a_{11}^*a_{22}^*}}, \tag{30}$$

where $a_{11}^*=a_{11}/\sigma_1^2$, $a_{22}^*=a_{22}/\sigma_2^2$, and $a_{12}^*=a_{12}/(\sigma_1\sigma_2)$. The transformation is equivalent to setting $\sigma_1=\sigma_2=1$. Then the density of a_{11}, a_{22}, and $r=a_{12}/\sqrt{a_{11}a_{22}}$ $(da_{12}=dr\sqrt{a_{11}a_{22}})$ is

$$\frac{a_{11}^{\frac{1}{2}n-1}a_{22}^{\frac{1}{2}n-1}(1-r^2)^{\frac{1}{2}(n-3)}e^{-\frac{1}{2}Q}}{2^n(1-\rho^2)^{\frac{1}{2}n}\sqrt{\pi}\,\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]}, \tag{31}$$

where

$$Q=\frac{a_{11}-2\rho r\sqrt{a_{11}}\sqrt{a_{22}}+a_{22}}{1-\rho^2}. \tag{32}$$

To find the density of r, we must integrate (31) with respect to a_{11} and a_{22} over the range 0 to ∞. There are various ways of carrying out the integration, which result in different expressions for the density. The method we shall indicate here is straightforward. We expand part of the exponential:

$$\exp\left[\frac{\rho r\sqrt{a_{11}}\sqrt{a_{22}}}{(1-\rho^2)}\right]=\sum_{\alpha=0}^{\infty}\frac{\left(\rho r\sqrt{a_{11}}\sqrt{a_{22}}\right)^\alpha}{\alpha!(1-\rho^2)^\alpha}. \tag{33}$$

Then the density (31) is

$$\frac{(1-r^2)^{\frac{1}{2}(n-3)}}{(1-\rho^2)^{\frac{1}{2}n}2^n\sqrt{\pi}\,\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]}\sum_{\alpha=0}^{\infty}\frac{(\rho r)^\alpha}{\alpha!(1-\rho^2)^\alpha} \tag{34}$$
$$\cdot\left\{\exp\left[-\frac{a_{11}}{2(1-\rho^2)}\right]a_{11}^{(n+\alpha)/2-1}\right\}\left\{\exp\left[-\frac{a_{22}}{2(1-\rho^2)}\right]a_{22}^{(n+\alpha)/2-1}\right\}.$$

Since

$$(35)\quad \int_0^\infty a^{\frac{1}{2}(n+\alpha)-1}\exp\left[-\frac{a}{2(1-\rho^2)}\right]da = \Gamma\left[\tfrac{1}{2}(n+\alpha)\right]\left[2(1-\rho^2)\right]^{\frac{1}{2}(n+\alpha)},$$

the integral of (34) (term-by-term integration is permissible) is

$$(36)\quad \frac{(1-r^2)^{\frac{1}{2}(n-3)}}{(1-\rho^2)^{\frac{1}{2}n}2^n\sqrt{\pi}\,\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]}$$

$$\cdot\sum_{\alpha=0}^{\infty}\frac{(\rho r)^\alpha}{\alpha!(1-\rho^2)^\alpha}\Gamma^2\left[\tfrac{1}{2}(n+\alpha)\right]2^{n+\alpha}(1-\rho^2)^{n+\alpha}$$

$$=\frac{(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{\sqrt{\pi}\,\Gamma(\frac{1}{2}n)\Gamma[\frac{1}{2}(n-1)]}\sum_{\alpha=0}^{\infty}\frac{(2\rho r)^\alpha}{\alpha!}\Gamma^2\left[\tfrac{1}{2}(n+\alpha)\right].$$

The *duplication formula for the gamma function* is

$$(37)\quad \Gamma(2z)=\frac{2^{2z-1}\Gamma(z)(z+\frac{1}{2})}{\sqrt{\pi}}.$$

It can be used to modify the constant in (36).

Theorem 4.2.2. *The correlation coefficient in a sample of N from a bivariate normal distribution with correlation ρ is distributed with density*

$$(38)\quad \frac{2^{n-2}(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{(n-2)!\,\pi}\sum_{\alpha=0}^{\infty}\frac{(2\rho r)^\alpha}{\alpha!}\Gamma^2\left[\tfrac{1}{2}(n+\alpha)\right], \qquad -1\le r\le 1,$$

where $n = N-1$.

The distribution of r was first found by Fisher (1915). He also gave as another form of the density,

$$(39)\quad \frac{(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{\pi(n-2)!}\left[\frac{d^{n-1}}{dx^{n-1}}\left\{\frac{\cos^{-1}(-x)}{\sqrt{1-x^2}}\right\}\bigg|_{x=r\rho}\right].$$

See Problem 4.24.

Hotelling (1953) has made an exhaustive study of the distribution of r. He has recommended the following form:

$$\frac{n-1}{\sqrt{2\pi}}\frac{\Gamma(n)}{\Gamma(n+\frac{1}{2})}(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)} \tag{40}$$

$$\cdot(1-\rho r)^{-n+\frac{1}{2}}F\left(\tfrac{1}{2},\tfrac{1}{2};\, n+\tfrac{1}{2};\, \frac{1+\rho r}{2}\right),$$

where

$$F(a,b;c;x)=\sum_{j=0}^{\infty}\frac{\Gamma(a+j)}{\Gamma(a)}\frac{\Gamma(b+j)}{\Gamma(b)}\frac{\Gamma(c)}{\Gamma(c+j)}\frac{x^j}{j!} \tag{41}$$

is a *hypergeometric function*. (See Problem 4.25.) The series in (40) converges more rapidly than the one in (38). Hotelling discusses methods of integrating the density and also calculates moments of r.

The cumulative distribution of r,

$$\Pr\{r \le r^*\} = F(r^*|N,\rho), \tag{42}$$

has been tabulated by David (1938) for† $\rho = 0(.1).9$, $N = 3(1)25, 50, 100, 200, 400$, and $r^* = -1(.05)1$. (David's n is our N.) It is clear from the density (38) that $F(r^*|N,\rho) = 1 - F(-r^*|N,-\rho)$ because the density for r, ρ is equal to the density for $-r, -\rho$. These tables can be used for a number of statistical procedures.

First, we consider the problem of using a sample to test the hypothesis

$$H\colon \rho = \rho_0. \tag{43}$$

If the alternatives are $\rho > \rho_0$, we reject the hypothesis if the sample correlation coefficient is greater than r_0, where r_0 is chosen so $1 - F(r_0|N,\rho_0) = \alpha$, the significance level. If the alternatives are $\rho < \rho_0$, we reject the hypothesis if the sample correlation coefficient is less than r_0', where r_0' is chosen so $F(r_0'|N,\rho_0) = \alpha$. If the alternatives are $\rho \ne \rho_0$, the region of rejection is $r > r_1$ and $r < r_1'$, where r_1 and r_1' are chosen so $[1 - F(r_1|N,\rho_0)] + F(r_1'|N,\rho_0) = \alpha$. David suggests that r_1 and r_1' be chosen so $[1 - F(r_1|N,\rho_0)] = F(r_1'|N,\rho_0) = \frac{1}{2}\alpha$. She has shown (1937) that for $N \ge 10$, $|\rho| \le 0.8$ this critical region is nearly the region of an unbiased test of H, that is, a test whose power function has its minimum at ρ_0.

It should be pointed out that any test based on r is invariant under transformations of location and scale, that is, $x_{i\alpha}^* = b_i x_{i\alpha} + c_i$, $b_i > 0$, $i = 1,2,$

†$\rho = 0(.1).9$ means $\rho = 0, 0.1, 0.2, \ldots, 0.9$.

Table 4.1. A Power Function

ρ	Probability
-1.0	0.0000
-0.8	0.0000
-0.6	0.0004
-0.4	0.0032
-0.2	0.0147
0.0	0.0500
0.2	0.1376
0.4	0.3215
0.6	0.6235
0.8	0.9279
1.0	1.0000

$\alpha = 1, \dots, N$; and r is essentially the only invariant of the sufficient statistics (Problem 3.7). The above procedure for testing $H: \rho = \rho_0$ against alternatives $\rho > \rho_0$ is uniformly most powerful among all invariant tests. (See Problems 4.16, 4.17, and 4.18.)

As an example suppose one wishes to test the hypothesis that $\rho = 0.5$ against alternatives $\rho \neq 0.5$ at the 5% level of significance using the correlation observed in a sample of 15. In David's tables we find (by interpolation) that $F(0.027|15, 0.5) = 0.025$ and $F(0.805|15, 0.5) = 0.975$. Hence we reject the hypothesis if our sample r is less than 0.027 or greater than 0.805.

Secondly, we can use David's tables to compute the power function of a test of correlation. If the region of rejection of H is $r > r_1$ and $r < r_1'$, the power of the test is a function of the true correlation ρ, namely $[1 - F(r_1|N, \rho) + [F(r_1'|N, \rho)]$; this is the probability of rejecting the null hypothesis when the population correlation is ρ.

As an example consider finding the power function of the test for $\rho = 0$ considered in the preceding section. The rejection region (one-sided) is $r \geq 0.5494$ at the 5% significance level. The probabilities of rejection are given in Table 4.1. The graph of the power function is illustrated in Figure 4.2.

Thirdly, David's computations lead to confidence regions for ρ. For given N, r_1' (defining a significance point) is a function of ρ, say $f_1(\rho)$, and r_1 is another function of ρ, say $f_2(\rho)$, such that

$$\Pr\{f_1(\rho) < r < f_2(\rho) | \rho\} = 1 - \alpha. \tag{44}$$

Clearly, $f_1(\rho)$ and $f_2(\rho)$ are monotonically increasing functions of ρ if r_1 and r_1' are chosen so $1 - F(r_1|N. \rho) = \frac{1}{2}\alpha = F(r_1'|N, \rho)$. If $\rho = f_i^{-1}(r)$ is the

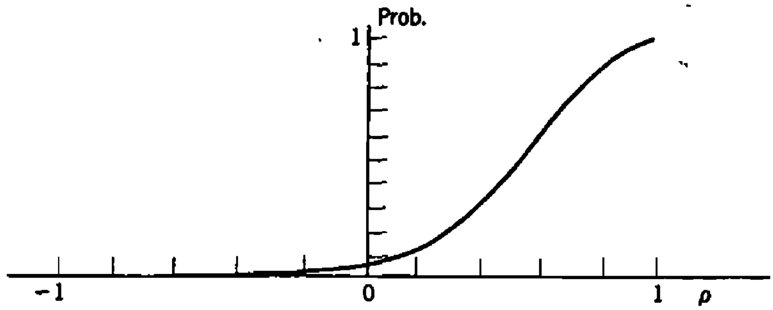

Figure 4.2. A power function.

inverse of $r=f_i(\rho)$, $i=1,2$, then the inequality $f_1(\rho)<r$ is equivalent to[†] $\rho<f_1^{-1}(r)$, and $r<f_2(\rho)$ is equivalent to $f_2^{-1}(r)<\rho$. Thus (44) can be written

$$\Pr\{f_2^{-1}(r)<\rho<f_1^{-1}(r)|\rho\}=1-\alpha. \tag{45}$$

This equation says that the probability is $1-\alpha$ that we draw a sample such that the interval $(f_2^{-1}(r), f_1^{-1}(r))$ covers the parameter ρ. Thus this interval is a confidence interval for ρ with confidence coefficient $1-\alpha$. For a given N and α the curves $r=f_1(\rho)$ and $r=f_2(\rho)$ appear as in Figure 4.3. In testing the hypothesis $\rho=\rho_0$, the intersection of the line $\rho=\rho_0$ and the two curves gives the significance points r_1 and r_1'. In setting up a confidence region for ρ on the basis of a sample correlation r^*, we find the limits $f_2^{-1}(r^*)$ and

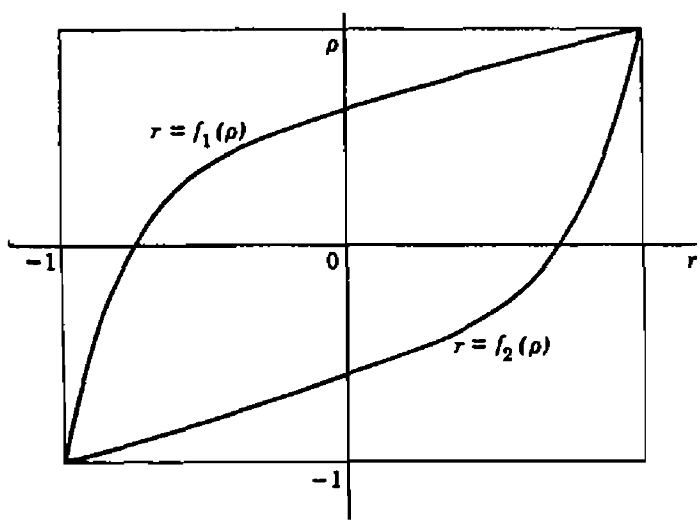

Figure 4.3

[†]The point $(f_1(\rho),\rho)$ on the first curve is to the left of (r,ρ), and the point $(r, f_1^{-1}(r))$ is above (r,ρ).

$f_1^{-1}(r^*)$ by the intersection of the line $r = r^*$ with the two curves. David gives these curves for $\alpha = 0.1$, 0.05, 0.02, and 0.01 for various values of N. One-sided confidence regions can be obtained by using only one inequality above.

The tables of $F(r|N, \rho)$ can also be used instead of the curves for finding the confidence interval. Given the sample value r^*, $f_1^{-1}(r^*)$ is the value of ρ such that $\frac{1}{2}\alpha = \Pr\{r \le r^* | \rho\} = F(r^*|N, \rho)$, and similarly $f_2^{-1}(r^*)$ is the value of ρ such that $\frac{1}{2}\alpha = \Pr\{r \ge r^* | \rho\} = 1 - F(r^*|N, \rho)$. The interval between these two values of ρ, $(f_2^{-1}(r^*), f_1^{-1}(r^*))$, is the confidence interval.

As an example, consider the confidence interval with confidence coefficient 0.95 based on the correlation of 0.7952 observed in a sample of 10. Using Graph II of David, we find the two limits are 0.34 and 0.94. Hence we state that $0.34 < \rho < 0.94$ with confidence 95%.

Definition 4.2.1. *Let $L(x, \boldsymbol{\theta})$ be the likelihood function of the observation vector x and the parameter vector $\boldsymbol{\theta} \in \Omega$. Let a null hypothesis be defined by a proper subset ω of Ω. The* likelihood ratio criterion *is*

$$\lambda(x) = \frac{\sup_{\boldsymbol{\theta} \in \omega} L(x, \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Omega} L(x, \boldsymbol{\theta})}. \tag{46}$$

The likelihood ratio test *is the procedure of rejecting the null hypothesis when $\lambda(x)$ is less than a predetermined constant.*

Intuitively, one rejects the null hypothesis if the density of the observations under the most favorable choice of parameters in the null hypothesis is much less than the density under the most favorable unrestricted choice of the parameters. Likelihood ratio tets have some desirable features; see Lehmann (1959), for example. Wald (1943) has proved some favorable asymptotic properties. For most hypotheses concerning the multivariate normal distribution, likelihood ratio tests are appropriate and often are optimal.

Let us consider the likelihood ratio test of the hypothesis that $\rho = \rho_0$ based on a sample $x_1, \dots, x_N$ from the bivariate normal distribution. The set Ω consists of μ_1, μ_2, σ_1, σ_2, and ρ such that $\sigma_1 > 0$, $\sigma_2 > 0$, $-1 < \rho < 1$. The set ω is the subset for which $\rho = \rho_0$. The likelihood maximized in Ω is (by Lemmas 3.2.2 and 3.2.3)

$$\max_{\Omega} L = \frac{N^N e^{-N}}{(2\pi)^N (1 - r^2)^{\frac{1}{2}N} a_{11}^{N/2} a_{22}^{N/2}}. \tag{47}$$

Under the null hypothesis the likelihood function is

$$\frac{1}{(2\pi)^N(1-\rho_0^2)^{\frac{1}{2}N}(\sigma^2)^N}\exp\left[-\frac{a_{11}/\tau+\tau a_{22}-2\rho_0 a_{12}}{2\sigma^2(1-\rho_0^2)}\right], \tag{48}$$

where $\sigma^2=\sigma_1\sigma_2$ and $\tau=\sigma_1/\sigma_2$. The maximum of (48) with respect to τ occurs at $\hat{\tau}=\sqrt{a_{11}}\,/\sqrt{a_{22}}$. The concentrated likelihood is

$$\frac{1}{(2\pi)^N(1-\rho_0^2)^{\frac{1}{2}N}(\sigma^2)^N}\exp\left[-\frac{\sqrt{a_{11}}\sqrt{a_{22}}\,(1-\rho_0 r)}{\sigma^2(1-\rho_0^2)}\right]; \tag{49}$$

the maximum of (49) occurs at

$$\hat{\sigma}^2=\frac{a_{11}^{\frac{1}{2}}a_{22}^{\frac{1}{2}}(1-\rho_0 r)}{N(1-\rho_0^2)}. \tag{50}$$

The likelihood ratio criterion is, therefore,

$$\frac{\max_\omega L}{\max_\Omega L}=\frac{(1-\rho_0^2)^{\frac{1}{2}N}(1-r^2)^{\frac{1}{2}N}}{(1-\rho_0 r)^N}=\left[\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2}\right]^{\frac{1}{2}N}. \tag{51}$$

The likelihood ratio test is $(1-\rho_0^2)(1-r^2)(1-\rho_0 r)^{-2}<c$, where c is chosen so the probability of the inequality when samples are drawn from normal populations with correlation ρ_0 is the prescribed significance level. The critical region can be written equivalently as

$$(\rho_0^2 c-\rho_0^2+1)r^2-2\rho_0 cr+c-1+\rho_0^2>0, \tag{52}$$

or

$$\begin{aligned} r &> \frac{\rho_0 c+(1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c+1-\rho_0^2}, \\ r &< \frac{\rho_0 c-(1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c+1-\rho_0^2}. \end{aligned} \tag{53}$$

Thus the likelihood ratio test of $H: \rho=\rho_0$ against alternatives $\rho\neq\rho_0$ has a rejection region of the form $r>r_1$ and $r<r_1'$; but r_1 and r_1' are not chosen so that the probability of each inequality is $\alpha/2$ when H is true, but are taken to be of the form given in (53), where c is chosen so that the probability of the two inequalities is α.

4.2.3. The Asymptotic Distribution of a Sample Correlation Coefficient and Fisher's z

In this section we shall show that as the sample size increases, a sample correlation coefficient tends to be normally distributed. The distribution of a particular function of a sample correlation, Fisher's z [Fisher (1921)], which has a variance approximately independent of the population correlation, tends to normality faster.

We are particularly interested in the sample correlation coefficient

$$r(n) = \frac{A_{ij}(n)}{\sqrt{A_{ii}(n)A_{jj}(n)}} \tag{54}$$

for some i and j, $i \neq j$. This can also be written

$$r(n) = \frac{C_{ij}(n)}{\sqrt{C_{ii}(n)C_{jj}(n)}}, \tag{55}$$

where $C_{gh}(n) = A_{gh}(n)/\sqrt{\sigma_{gg}\sigma_{hh}}$. The set $C_{ii}(n)$, $C_{jj}(n)$, and $C_{ij}(n)$ is distributed like the distinct elements of the matrix

$$\sum_{\alpha=1}^{n} \begin{pmatrix} Z_{i\alpha}^* \\ Z_{j\alpha}^* \end{pmatrix} (Z_{i\alpha}^*, Z_{j\alpha}^*) = \sum_{\alpha=1}^{n} \begin{pmatrix} Z_{i\alpha}/\sqrt{\sigma_{ii}} \\ Z_{j\alpha}/\sqrt{\sigma_{jj}} \end{pmatrix} \left(Z_{i\alpha}/\sqrt{\sigma_{ii}}, Z_{j\alpha}/\sqrt{\sigma_{jj}}\right), \tag{56}$$

where the $(Z_{i\alpha}^*, Z_{j\alpha}^*)$ are independent, each with distribution

$$N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right],$$

where

$$\rho = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Let

$$U(n) = \frac{1}{n} \begin{pmatrix} C_{ii}(n) \\ C_{jj}(n) \\ C_{ij}(n) \end{pmatrix}, \tag{57}$$

$$b = \begin{pmatrix} 1 \\ 1 \\ \rho \end{pmatrix}. \tag{58}$$

Then by Theorem 3.4.4 the vector $\sqrt{n}[U(n)-b]$ has a limiting normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{pmatrix}. \tag{59}$$

Now we need the general theorem:

Theorem 4.2.3. *Let $\{U(n)\}$ be a sequence of m-component random vectors and b a fixed vector such that $\sqrt{n}[U(n)-b]$ has the limiting distribution $N(\mathbf{0},\mathbf{T})$ as $n \to \infty$. Let $f(u)$ be a vector-valued function of u such that each component $f_j(u)$ has a nonzero differential at $u=b$, and let $\partial f_j(u)/\partial u_i|_{u=b}$ be the i, jth component of $\mathbf{\Phi}_b$. Then $\sqrt{n}\{f[u(n)]-f(b)\}$ has the limiting distribution $N(\mathbf{0},\mathbf{\Phi}_b'\mathbf{T}\mathbf{\Phi}_b)$.*

Proof. See Serfling (1980), Section 3.3, or Rao (1973), Section 6a.2. A function $g(u)$ is said to have a differential at b or to be totally differentiable at b if the partial derivatives $\partial g(u)/\partial u_i$ exist at $u=b$ and for every $\varepsilon>0$ there exists a neighborhood $N_\varepsilon(b)$ such that

(60)
$$\left| g(u)-g(b)-\sum_{i=1}^{m}\frac{\partial g(u)}{\partial u_i}(u_i-b_i)\right| \le \varepsilon\|u-b\| \qquad \text{for all } u \in N_\varepsilon(b). \quad \blacksquare$$

It is clear that $U(n)$ defined by (57) with b and $\mathbf{T}$ defined by (58) and (59), respectively, satisfies the conditions of the theorem. The function

$$r=\frac{u_3}{\sqrt{u_1u_2}}=u_3u_1^{-\frac12}u_2^{-\frac12} \tag{61}$$

satisfies the conditions; the elements of $\mathbf{\Phi}_b$ are

$$\begin{aligned}
\left.\frac{\partial r}{\partial u_1}\right|_{u=b} &= -\tfrac12 u_3u_1^{-\frac32}u_2^{-\frac12}\big|_{u=b} = -\tfrac12\rho, \\
\left.\frac{\partial r}{\partial u_2}\right|_{u=b} &= -\tfrac12 u_3u_1^{-\frac12}u_2^{-\frac32}\big|_{u=b} = -\tfrac12\rho, \\
\left.\frac{\partial r}{\partial u_3}\right|_{u=b} &= u_1^{-\frac12}u_2^{-\frac12}\big|_{u=b} = 1,
\end{aligned} \tag{62}$$

and $f(b) = \rho$. The variance of the limiting distribution of $\sqrt{n}\,[r(n) - \rho]$ is

$$(63) \qquad (-\tfrac{1}{2}\rho, -\tfrac{1}{2}\rho, 1)\begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{pmatrix}\begin{pmatrix} -\tfrac{1}{2}\rho \\ -\tfrac{1}{2}\rho \\ 1 \end{pmatrix}$$

$$= (\rho - \rho^3, \rho - \rho^3, 1 - \rho^2)\begin{pmatrix} -\tfrac{1}{2}\rho \\ -\tfrac{1}{2}\rho \\ 1 \end{pmatrix}$$

$$= 1 - 2\rho^2 + \rho^4$$

$$= (1 - \rho^2)^2.$$

Thus we obtain the following:

Theorem 4.2.4. *If $r(n)$ is the sample correlation coefficient of a sample of N $(= n+1)$ from a normal distribution with correlation ρ, then $\sqrt{n}\,[r(n) - \rho]/(1-\rho^2)$ [or $\sqrt{N}\,[r(n) - \rho]/(1-\rho^2)$] has the limiting distribution $N(0,1)$.*

It is clear from Theorem 4.2.3 that if $f(x)$ is differentiable at $x = \rho$, then $\sqrt{n}\,[f(r) - f(\rho)]$ is asymptotically normally distributed with mean zero and variance

$$\left(\frac{\partial f}{\partial x}\bigg|_{x=\rho}\right)^2 (1-\rho^2)^2.$$

A useful function to consider is one whose asymptotic variance is constant (here unity) independent of the parameter ρ. This function satisfies the equation

$$(64) \qquad f'(\rho) = \frac{1}{1-\rho^2} = \frac{1}{2}\left(\frac{1}{1+\rho} + \frac{1}{1-\rho}\right).$$

Thus $f(\rho)$ is taken as $\frac{1}{2}[\log(1+\rho) - \log(1-\rho)] = \frac{1}{2}\log(1+\rho)/(1-\rho)]$. The so-called Fisher's z is

$$(65) \qquad z = \tfrac{1}{2}\log\frac{1+r}{1-r} = \tanh^{-1} r,$$

where $r = \tanh z = (e^z - e^{-z})/(e^z + e^{-z})$. Let

$$(66) \qquad \zeta = \tfrac{1}{2}\log\frac{1+\rho}{1-\rho}.$$

Theorem 4.2.5. *Let z be defined by (65), where r is the correlation coefficient of a sample of N $(= n + 1)$ from a bivariate normal distribution with correlation ρ; let ζ be defined by (66). Then $\sqrt{n}(z - \zeta)$ has a limiting normal distribution with mean 0 and variance 1.*

It can be shown that to a closer approximation

$$\mathscr{E}z \sim \zeta + \frac{\rho}{2n}. \tag{67}$$

$$\mathscr{E}(z - \zeta)^2 \sim \frac{1}{n-2} \sim \mathscr{E}\left(z - \zeta - \frac{\rho}{2n}\right)^2. \tag{68}$$

The latter follows from

$$\mathscr{E}(z - \zeta)^2 = \frac{1}{n} + \frac{8 - \rho^2}{4n^2} + \cdots \tag{69}$$

and holds good for ρ^2/n^2 small. Hotelling (1953) gives moments of z to order n^{-3}. An important property of Fisher's z is that the approach to normality is much more rapid than for r. David (1938) makes some comparisons between the tabulated probabilities and the probabilities computed by assuming z is normally distributed. She recommends that for $N > 25$ one take z as normally distributed with mean and variance given by (67) and (68). Konishi (1978a, 1978b, 1979) has also studied z. [Ruben (1966) has suggested an alternative approach, which is more complicated, but possibly more accurate.]

We shall now indicate how Theorem 4.2.5 can be used.

a. Suppose we wish to test the hypothesis $\rho = \rho_0$ on the basis of a sample of N against the alternatives $\rho \neq \rho_0$. We compute r and then z by (65). Let

$$\zeta_0 = \tfrac{1}{2}\log\frac{1 + \rho_0}{1 - \rho_0}. \tag{70}$$

Then a region of rejection at the 5% significance level is

$$\sqrt{N-3}\,|z - \zeta_0| > 1.96. \tag{71}$$

A better region is

$$\sqrt{N-3}\left|z - \zeta_0 - \frac{\frac{1}{2}\rho_0}{N-1}\right| > 1.96. \tag{72}$$

b. Suppose we have a sample of N_1 from one population and a sample of N_2 from a second population. How do we test the hypothesis that the two correlation coefficients are equal, $\rho_1 = \rho_2$? From Theorem 4.2.5 we know that if the null hypothesis is true then $z_1 - z_2$ [where z_1 and z_2 are defined by (65) for the two sample correlation coefficients] is asymptotically normally distributed with mean 0 and variance $1/(N_1 - 3) + 1/(N_2 - 3)$. As a critical region of size 5%, we use

$$\frac{|z_1 - z_2|}{\sqrt{1/(N_1 - 3) + 1/(N_2 - 3)}} > 1.96. \tag{73}$$

c. Under the conditions of paragraph b, assume that $\rho_1 = \rho_2 = \rho$. How do we use the results of both samples to give a joint estimate of ρ? Since z_1 and z_2 have variances $1/(N_1 - 3)$ and $1/(N_2 - 3)$, respectively, we can estimate ζ by

$$\frac{(N_1 - 3) z_1 + (N_2 - 3) z_2}{N_1 + N_2 - 6} \tag{74}$$

and convert this to an estimate of ρ by the inverse of (65).

d. Let r be the sample correlation from N observations. How do we obtain a confidence interval for ρ? We know that approximately

$$\Pr\{-1.96 \le \sqrt{N-3}\,(z - \zeta) \le 1.96\} = 0.95. \tag{75}$$

From this we deduce that $[-1.96/\sqrt{N-3} + z, 1.96/\sqrt{N-3} + z]$ is a confidence interval for ζ. From this we obtain an interval for ρ using the fact $\rho = \tanh \zeta = (e^{\zeta} - e^{-\zeta})/(e^{\zeta} + e^{-\zeta})$, which is a monotonic transformation. Thus a 95% confidence interval is

$$\tanh\left(z - 1.96/\sqrt{N-3}\right) \le \rho \le \tanh\left(z + 1.96/\sqrt{N-3}\right). \tag{76}$$

The *bootstrap* method has been developed to assess the variability of a sample quantity. See Efron (1982). We shall illustrate the method on the sample correlation coefficient, but it can be applied to other quantities studied in this book.

Suppose $x_1, \dots, x_N$ is a sample from some bivariate population not necessarily normal. The approach of the bootstrap is to consider these N vectors as a finite population of size N; a random vector X has the (discrete) probability

$$\Pr\{X = x_\alpha\} = \frac{1}{N}, \qquad \alpha = 1, \dots, N \tag{77}$$

A random sample of size N drawn from this finite population has a probability distribution, and the correlation coefficient calculated from such a sample has a (discrete) probability distribution, say $p_N(r)$. The bootstrap proposes to use this distribution in place of the unobtainable distribution of the correlation coefficient of random samples from the parent population. However, it is prohibitively expensive to compute; instead $p_N(r)$ is estimated by the empirical distribution of r calculated from a large number of random samples from (77). Diaconis and Efron (1983) have given an example of $N = 15$; they find the empirical distribution closely resembles the actual distribution of r (essentially obtainable in this special case). An advantage of this approach is that it is not necessary to assume knowledge of the parent population; a disadvantage is the massive computation.

4.3. PARTIAL CORRELATION COEFFICIENTS; CONDITIONAL DISTRIBUTIONS

4.3.1. Estimation of Partial Correlation Coefficients

Partial correlation coefficients in normal distributions are correlation coefficients in conditional distributions. It was shown in Section 2.5 that if X is distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$(1) \qquad X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

then the conditional distribution of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ is $N[\boldsymbol{\mu}^{(1)} + \boldsymbol{\beta}(x^{(2)} - \boldsymbol{\mu}^{(2)}), \boldsymbol{\Sigma}_{11\cdot 2}]$, where

$$(2) \qquad \boldsymbol{\beta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1},$$

$$(3) \qquad \boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

The partial correlations of $X^{(1)}$ given $x^{(2)}$ are the correlations calculated in the usual way from $\boldsymbol{\Sigma}_{11\cdot 2}$. In this section we are interested in statistical problems concerning these correlation coefficients.

First we consider the problem of estimation on the basis of a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. What are the maximum likelihood estimators of the partial correlations of $X^{(1)}$ (of q components), $\rho_{ij\cdot q+1,\dots,p}$? We know that the maximum likelihood estimator of Σ is $(1/N)A$, where

$$(4)\qquad A=\sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})'$$
$$=\sum_{\alpha=1}^{N}\begin{pmatrix}x_\alpha^{(1)}-\bar{x}^{(1)}\\ x_\alpha^{(2)}-\bar{x}^{(2)}\end{pmatrix}\left(x_\alpha^{(1)\prime}-\bar{x}^{(1)\prime},x_\alpha^{(2)\prime}-\bar{x}^{(2)\prime}\right)$$
$$=\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}$$

and $\bar{x}=(1/N)\Sigma_{\alpha=1}^N x_\alpha=(\bar{x}^{(1)\prime}\ \bar{x}^{(2)\prime})'$. The correspondence between Σ and $\Sigma_{11\cdot2}$, $\boldsymbol{\beta}$, and Σ_{22} is one-to-one by virtue of (2) and (3) and

$$(5)\qquad \Sigma_{12}=\boldsymbol{\beta}\Sigma_{22},$$

$$(6)\qquad \Sigma_{11}=\Sigma_{11\cdot2}+\boldsymbol{\beta}\Sigma_{22}\boldsymbol{\beta}'.$$

We can now apply Corollary 3.2.1 to the effect that maximum likelihood estimators of functions of parameters are those functions of the maximum likelihood estimators of those parameters.

Theorem 4.3.1. *Let $x_1,\dots,x_N$ be a sample from $N(\boldsymbol{\mu},\Sigma)$, where $\boldsymbol{\mu}$ and Σ are partitioned as in* (1). *Define A by* (4) *and $(\bar{x}^{(1)\prime}\ \bar{x}^{(2)\prime})=(1/N)\Sigma_{\alpha=1}^N(x_\alpha^{(1)\prime}\ x_\alpha^{(2)\prime})$. Then the maximum likelihood estimators of $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $\boldsymbol{\beta}$, $\Sigma_{11\cdot2}$, and Σ_{22} are $\hat{\boldsymbol{\mu}}^{(1)}=\bar{x}^{(1)}$, $\hat{\boldsymbol{\mu}}^{(2)}=\bar{x}^{(2)}$,*

$$(7)\qquad \hat{\boldsymbol{\beta}}=A_{12}A_{22}^{-1},\qquad \hat{\Sigma}_{11\cdot2}=\frac{1}{N}\left(A_{11}-A_{12}A_{22}^{-1}A_{21}\right),$$

and $\hat{\Sigma}_{22}=(1/N)A_{22}$, respectively.

In turn, Corollary 3.2.1 can be used to obtain the maximum likelihood estimators of $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $\boldsymbol{\beta}$, Σ_{22}, $\sigma_{ii\cdot q+1,\dots,p}$, $i=1,\dots,q$, and $\rho_{ij\cdot q+1,\dots,p}$, $i,j=1,\dots,q$. It follows that the maximum likelihood estimators of the partial correlation coefficients are

$$(8)\qquad \hat{\rho}_{ij\cdot q+1,\dots,p}=\frac{\hat{\sigma}_{ij\cdot q+1,\dots,p}}{\sqrt{\hat{\sigma}_{ii\cdot q+1,\dots,p}\hat{\sigma}_{jj\cdot q+1,\dots,p}}},\qquad i,j=1,\dots,q,$$

where $\hat{\sigma}_{ij\cdot q+1,\dots,p}$ is the i,jth element of $\hat{\Sigma}_{11\cdot2}$.

Theorem 4.3.2. *Let $x_1, \ldots, x_N$ be a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The maximum likelihood estimators of $\rho_{ij\cdot q+1,\ldots,p}$, the partial correlations of the first q components conditional on the last $p-q$ components, are given by*

$$(9)\qquad \hat{\rho}_{ij\cdot q+1,\ldots,p} = \frac{a_{ij\cdot q+1,\ldots,p}}{\sqrt{a_{ii\cdot q+1,\ldots,p}\,a_{jj\cdot q+1,\ldots,p}}}, \qquad i,j = 1,\ldots,q,$$

where

$$(10)\qquad (a_{ij\cdot q+1,\ldots,p}) = A_{11} - A_{12}A_{22}^{-1}A_{21} = A_{11\cdot 2}.$$

The estimator $\hat{\rho}_{ij\cdot q+1,\ldots,p}$, denoted by $r_{ij\cdot q+1,\ldots,p}$, is called the *sample partial correlation coefficient between X_i and X_j holding $X_{q+1}, \ldots, X_p$ fixed*. It is also called the sample partial correlation coefficient between X_i and X_j having taken account of $X_{q+1}, \ldots, X_p$. Note that the calculations can be done in terms of (r_{ij}).

The matrix $A_{11\cdot 2}$ can also be represented as

$$(11)\qquad A_{11\cdot 2} = \sum_{\alpha=1}^{N}\left[x_\alpha^{(1)} - \bar{x}^{(1)} - \hat{\boldsymbol{\beta}}\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)\right]\left[x_\alpha^{(1)} - \bar{x}^{(1)} - \hat{\boldsymbol{\beta}}\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)\right]'$$
$$= A_{11} - \hat{\boldsymbol{\beta}}A_{22}\hat{\boldsymbol{\beta}}'.$$

The vector $x_\alpha^{(1)} - \bar{x}^{(1)} - \hat{\boldsymbol{\beta}}(x_\alpha^{(2)} - \bar{x}^{(2)})$ is the residual of $x_\alpha^{(1)}$ from its regression on $x_\alpha^{(2)}$ and 1. The partial correlations are simple correlations between these residuals. The definition can be used also when the distributions involved are not normal.

Two geometric interpretations of the above theory can be given. In p-dimensional space $x_1, \ldots, x_N$ represent N points. The sample regression function

$$(12)\qquad x^{(1)} = \bar{x}^{(1)} + \hat{\boldsymbol{\beta}}(x^{(2)} - \bar{x}^{(2)})$$

is a $(p-q)$-dimensional hyperplane which is the intersection of q $(p-1)$-dimensional hyperplanes,

$$(13)\qquad x_i = \bar{x}_i + \sum_{j=q+1}^{p} \hat{\beta}_{ij}(x_j - \bar{x}_j), \qquad i = 1,\ldots,q,$$

where x_i, x_j are running variables. Here $\hat{\beta}_{ij}$ is an element of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1} = A_{12}A_{22}^{-1}$. The ith row of $\hat{\boldsymbol{\beta}}$ is $(\hat{\beta}_{i,q+1}, \ldots, \hat{\beta}_{ip})$. Each right-hand side of (13) is the least squares *regression* function of x_i on $x_{q+1}, \ldots, x_p$; that is, if we project the points $x_1, \ldots, x_N$ on the coordinate hyperplane of $x_i, x_{q+1}, \ldots, x_p$,

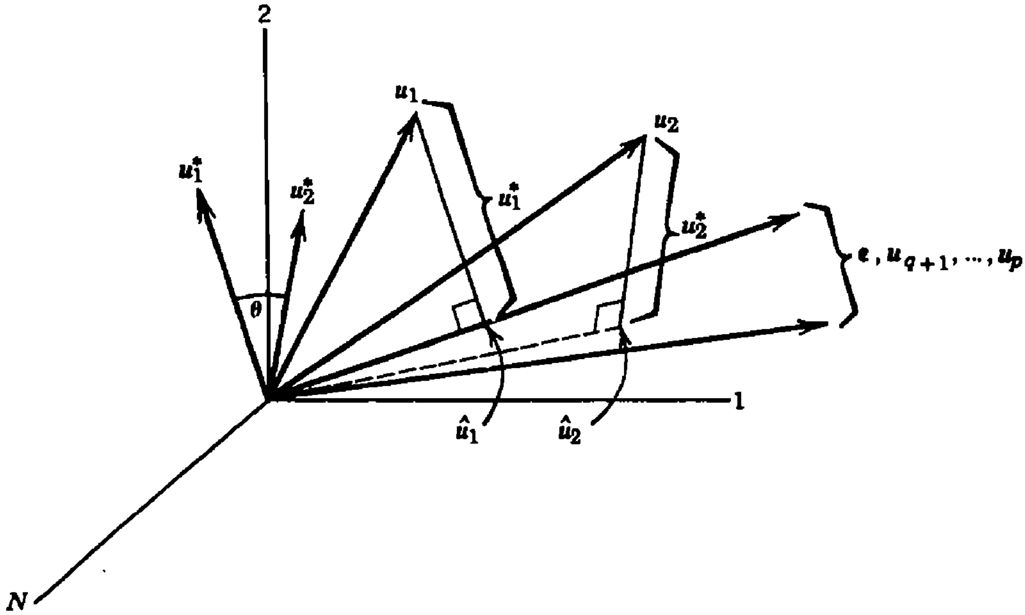

Figure 4.4

then (13) is the regression plane. The point with coordinates

$$x_i = \bar{x}_i + \sum_{j=q+1}^{p} \hat{\beta}_{ij}(x_{j\alpha} - \bar{x}_j), \qquad i = 1, \dots, q, \tag{14}$$

$$x_j = x_{j\alpha}, \qquad j = q+1, \dots, p,$$

is on the hyperplane (13). The difference in the ith coordinate of x_α and the point (14) is $y_{i\alpha} = x_{i\alpha} - [\bar{x}_i + \sum_{j=q+1}^{p} \hat{\beta}_{ij}(x_{j\alpha} - \bar{x}_j)]$ for $i = 1, \dots, q$ and 0 for the other coordinates. Let $y'_\alpha = (y_{1\alpha}, \dots, y_{q\alpha})$. These points can be represented as N points in a q-dimensional space. Then $A_{11\cdot 2} = \sum_{\alpha=1}^{N} y_\alpha y'_\alpha$.

We can also interpret the sample as p points in N-space (Figure 4.4). Let $u_j = (x_{j1}, \dots, x_{jN})'$ be the jth point, and let $\varepsilon = (1, \dots, 1)'$ be another point. The point with coordinates $\bar{x}_i, \dots, \bar{x}_i$ is $\bar{x}_i \varepsilon$. The projection of u_i on the hyperplane spanned by $u_{q+1}, \dots, u_p, \varepsilon$ is

$$\hat{u}_i = \bar{x}_i \varepsilon + \sum_{j=q+1}^{p} \hat{\beta}_{ij}(u_j - \bar{x}_j \varepsilon); \tag{15}$$

this is the point on the hyperplane that is at a minimum distance from u_i. Let u_i^* be the vector from $\hat{u}_i$ to u_i, that is, $u_i - \hat{u}_i$, or, equivalently, this vector translated so that one endpoint is at the origin. The set of vectors $u_1^*, \dots, u_q^*$ are the projections of $u_1, \dots, u_q$ on the hyperplane orthogonal to

$u_{q+1}, \ldots, u_p, \varepsilon$. Then $u_i^{*\prime} u_i^* = a_{ii \cdot q+1,\ldots,p}$, the length squared of u_i^* (i.e., the square of the distance of u from $\hat{u}_i$). Then $u_i^{*\prime} u_j^* / \sqrt{u_i^{*\prime} u_i^* u_j^{*\prime} u_j^*} = r_{ij \cdot q+1,\ldots,p}$ is the cosine of the angle between u_i^* and u_j^*.

As an example of the use of partial correlations we consider some data [Hooker (1907)] on yield of hay (X_1) in hundredweights per acrc, spring rainfall (X_2) in inches, and accumulated temperature above 42°F in the spring (X_3) for an English area over 20 years. The estimates of μ_i, σ_i $(= \sqrt{\sigma_{ii}})$, and ρ_{ij} are

$$\hat{\boldsymbol{\mu}} = \bar{x} = \begin{pmatrix} 28.02 \\ 4.91 \\ 594 \end{pmatrix},$$

$$\begin{pmatrix} \hat{\sigma}_1 \\ \hat{\sigma}_2 \\ \hat{\sigma}_3 \end{pmatrix} = \begin{pmatrix} 4.42 \\ 1.10 \\ 85 \end{pmatrix}, \tag{16}$$

$$\begin{pmatrix} 1 & \hat{\rho}_{12} & \hat{\rho}_{13} \\ \hat{\rho}_{21} & 1 & \hat{\rho}_{23} \\ \hat{\rho}_{31} & \hat{\rho}_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1.00 & 0.80 & -0.40 \\ 0.80 & 1.00 & -0.56 \\ -0.40 & -0.56 & 1.00 \end{pmatrix}.$$

From the correlations we observe that yield and rainfall are positively related, yield and temperature are negatively related, and rainfall and temperature are negatively related. What interpretation is to be given to the apparent negative relation between yield and temperature? Does high temperature tend to cause low yield, or is high temperature associated with low rainfall and hence with low yield? To answer this question we consider the correlation between yield and temperature when rainfall is held fixed; that is, we use the data given above to estimate the partial correlation between X_1 and X_3 with X_2 held fixed. It is†

$$r_{13 \cdot 2} = \frac{\hat{\sigma}_{13 \cdot 2}}{\sqrt{\hat{\sigma}_{11 \cdot 2} \hat{\sigma}_{33 \cdot 2}}} = 0.097. \tag{17}$$

Thus, if the effect of rainfall is removed, yield and temperature are positively correlated. The conclusion is that both high raninfall and high temperature increase hay yield, but in most years high rainfall occurs with low temperature and vice versa.

†We compute with $\hat{\boldsymbol{\Sigma}}$ as if it were $\boldsymbol{\Sigma}$.

4.3.2. The Distribution of the Sample Partial Correlation Coefficient

In order to test a hypothesis about a population partial correlation coefficient we want the distribution of the sample partial correlation coefficient. The partial correlations are computed from $A_{11\cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ (as indicated in Theorem 4.3.1) in the same way that correlations are computed from A. To obtain the distribution of a simple correlation we showed that A was distributed as $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$, where $Z_1, \dots, Z_{N-1}$ are distributed independently according to $N(\mathbf{0}, \Sigma)$ and independent of $\bar{X}$ (Theorem 3.3.2). Here we want to show that $A_{11\cdot 2}$ is distributed as $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$, where $U_1, \dots, U_{N-1-(p-q)}$ are distributed independently according to $N(\mathbf{0}, \Sigma_{11\cdot 2})$ and independently of $\hat{\boldsymbol{\beta}}$. The distribution of a partial correlation coefficient will follow from the characterization of the distribution of $A_{11\cdot 2}$. We state the theorem in a general form; it will be used in Chapter 8, where we treat regression in detail. The following corollary applies it to $A_{11\cdot 2}$, expressed in terms of residuals.

Theorem 4.3.3. *Suppose $Y_1, \dots, Y_m$ are independent with Y_α distributed according to $N(\Gamma w_\alpha, \Phi)$, where w_α is an r-component vector. Let $H = \sum_{\alpha=1}^{m} w_\alpha w_\alpha'$, assumed nonsingular, $G = \sum_{\alpha=1}^{m} Y_\alpha w_\alpha' H^{-1}$, and*

$$C = \sum_{\alpha=1}^{m} (Y_\alpha - Gw_\alpha)(Y_\alpha - Gw_\alpha)' = \sum_{\alpha=1}^{m} Y_\alpha Y_\alpha' - GHG'. \tag{18}$$

Then C is distributed as $\sum_{\alpha=1}^{m-r} U_\alpha U_\alpha'$, where $U_1, \dots, U_{m-r}$ are independently distributed according to $N(\mathbf{0}, \Phi)$ and independently of G.

Proof. The rows of $Y = (Y_1, \dots, Y_m)$ are random vectors in an m-dimensional space, and the rows of $W = (w_1, \dots, w_m)$ are fixed vectors in that space. The idea of the proof is to rotate coordinate axes so that the last r axes are in the space spanned by the rows of W. Let $E_2 = FW$, where F is a square matrix such that $FHF' = I$. Then

$$\begin{aligned} E_2 E_2' &= FWW'F' = F \sum_{\alpha=1}^{m} w_\alpha w_\alpha' F' \\ &= FHF' = I. \end{aligned} \tag{19}$$

Thus the m-component rows of E_2 are orthogonal and of unit length. It is possible to find an $(m-r) \times m$ matrix E_1 such that

$$E = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} \tag{20}$$

is orthogonal. (See Appendix, Lemma A.4.2.) Now let $U = YE'$ (i.e., $U_\alpha = \sum_{\beta=1}^m e_{\alpha\beta} Y_\beta$). By Theorem 3.3.1 the columns of $U = (U_1, \dots, U_m)$ are independently and normally distributed, each with covariance matrix $\mathbf{\Phi}$. The means are given by

$$\mathscr{E}U = \mathscr{E}YE' = \Gamma W E' = \Gamma F^{-1} E_2 (E_1' \quad E_2') = (\mathbf{0} \quad \Gamma F^{-1}) \tag{21}$$

by orthogonality of E. To complete the proof we need to show that C transforms to $\sum_{\alpha=1}^{m-r} U_\alpha U_\alpha'$. We have

$$\sum_{\alpha=1}^m Y_\alpha Y_\alpha' = YY' = UEE'U' = UU' = \sum_{\alpha=1}^m U_\alpha U_\alpha'. \tag{22}$$

Note that

$$\begin{aligned} G = YW'H^{-1} &= UEE_2'(F^{-1})'F'F \\ &= U\begin{pmatrix} E_1 \\ E_2 \end{pmatrix} E_2' F \\ &= U\begin{pmatrix} \mathbf{0} \\ I \end{pmatrix} F = U^{(2)} F, \end{aligned} \tag{23}$$

where $U^{(2)} = (U_{m-r+1}, \dots, U_m)$. Then

$$GHG' = U^{(2)} FHF' U^{(2)\prime} = U^{(2)} U^{(2)\prime} = \sum_{\alpha=m-r+1}^m U_\alpha U_\alpha'. \tag{24}$$

Thus C is

$$\sum_{\alpha=1}^m Y_\alpha Y_\alpha' - GHG' = \sum_{\alpha=1}^m U_\alpha U_\alpha' - \sum_{\alpha=m-r+1}^m U_\alpha U_\alpha' = \sum_{\alpha=1}^{m-r} U_\alpha U_\alpha'. \tag{25}$$

This proves the theorem. ∎

It follows from the above considerations that when $\Gamma = \mathbf{0}$, the $\mathscr{E}U = \mathbf{0}$, and we obtain the following:

Corollary 4.3.1. *If $\Gamma = \mathbf{0}$, the matrix GHG' defined in Theorem 4.3.3 is distributed as $\sum_{\alpha=m-r+1}^m U_\alpha U_\alpha'$, where $U_{m-r+1}, \dots, U_m$ are independently distributed, each according to $N(\mathbf{0}, \mathbf{\Phi})$.*

We now find the distribution of $A_{11\cdot 2}$ in the same form. It was shown in Theorem 3.3.1 that A is distributed as $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$, where $Z_1, \dots, Z_{N-1}$ are independent, each with distribution $N(\mathbf{0}, \Sigma)$. Let Z_α be partitioned into two subvectors of q and $p-q$ components, respectively:

$$(26) \qquad Z_\alpha = \begin{pmatrix} Z_\alpha^{(1)} \\ Z_\alpha^{(2)} \end{pmatrix}.$$

Then $A_{ij} = \sum_{\alpha=1}^{N} Z_\alpha^{(i)} Z_\alpha^{(j)\prime}$. By Lemma 4.2.1, conditionally on $Z_1^{(2)} = z_1^{(2)}, \dots, Z_{N-1}^{(2)} = z_{N-1}^{(2)}$, the random vectors $Z_1^{(1)}, \dots, Z_{N-1}^{(1)}$ are independently distributed, with $Z_\alpha^{(1)}$ distributed according to $N(\boldsymbol{\beta} z_\alpha^{(2)}, \Sigma_{11\cdot 2})$, where $\boldsymbol{\beta} = \Sigma_{12}\Sigma_{22}^{-1}$ and $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Now we apply Theorem 4.3.3 with $Z_\alpha^{(1)} = Y_\alpha$, $z_\alpha^{(2)} = w_\alpha$, $N-1 = m$, $p - q = r$, $\boldsymbol{\beta} = \Gamma$, $\Sigma_{11\cdot 2} = \Phi$, $A_{11} = \sum_{\alpha=1}^{N-1} Y_\alpha Y_\alpha'$, $A_{12}A_{22}^{-1} = G$, $A_{22} = H$. We find that the conditional distribution of $A_{11} - (A_{12}A_{22}^{-1})A_{22}(A_{22}^{-1}A_{12}') = A_{11\cdot 2}$ given $Z_\alpha^{(2)} = z_\alpha^{(2)}$, $\alpha = 1, \dots, N-1$, is that of $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$, where $U_1, \dots, U_{N-1-(p-q)}$ are independent, each with distribution $N(\mathbf{0}, \Sigma_{11\cdot 2})$. Since this distribution does not depend on $\{z_\alpha^{(2)}\}$, we obtain the following theorem:

Theorem 4.3.4. *The matrix $A_{11\cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ is distributed as $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$, where $U_1, \dots, U_{N-1-(p-q)}$ are independently distributed, each according to $N(\mathbf{0}, \Sigma_{11\cdot 2})$, and independently of A_{12} and A_{22}.*

Corollary 4.3.2. *If $\Sigma_{12} = 0$ ($\boldsymbol{\beta} = 0$), then $A_{11\cdot 2}$ is distributed as $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$ and $A_{12}A_{22}^{-1}A_{21}$ is distributed as $\sum_{\alpha=N-(p-q)}^{N-1} U_\alpha U_\alpha'$, where $U_1, \dots, U_{N-1}$ are independently distributed, each according to $N(\mathbf{0}, \Sigma_{11\cdot 2})$.*

Now it follows that the distribution of $r_{ij\cdot q+1,\dots,p}$ based on N observations is the same as that of a simple correlation coefficient based on $N-(p-q)$ observations with a corresponding population correlation value of $\rho_{ij\cdot q+1,\dots,p}$.

Theorem 4.3.5. *If the cdf of r_{ij} based on a sample of N from a normal distribution with correlation ρ_{ij} is denoted by $F(r|N, \rho_{ij})$, then the cdf of the sample partial correlation $r_{ij\cdot q+1,\dots,p}$ based on a sample of N from a normal distribution with partial correlation coefficient $\rho_{ij\cdot q+1,\dots,p}$ is $F[r|N-(p-q), \rho_{ij\cdot q+1,\dots,p}]$.*

This distribution was derived by Fisher (1924).

4.3.3. Tests of Hypotheses and Confidence Regions for Partial Correlation Coefficients

Since the distribution of a sample partial correlation $r_{ij\cdot q+1,\dots,p}$ based on a sample of N from a distribution with population correlation $\rho_{ij\cdot q+1,\dots,p}$

equal to a certain value, ρ, say, is the same as the distribution of a simple correlation r based on a sample of size $N-(p-q)$ from a distribution with the corresponding population correlation of ρ, all statistical inference procedures for the simple population correlation can be used for the partial correlation. The procedure for the partial correlation is exactly the same except that N is replaced by $N-(p-q)$. To illustrate this rule we give two examples.

Example 1. Suppose that on the basis of a sample of size N we wish to obtain a confidence interval for $\rho_{ij\cdot q+1,\dots,p}$. The sample partial correlation is $r_{ij\cdot q+1,\dots,p}$. The procedure is to use David's charts for $N-(p-q)$. In the example at the end of Section 4.3.1, we might want to find a confidence interval for $\rho_{12\cdot 3}$ with confidence coefficient 0.95. The sample partial correlation is $r_{12\cdot 3}=0.759$. We use the chart (or table) for $N-(p-q)=20-1=19$. The interval is $0.50<\rho_{12\cdot 3}<0.88$.

Example 2. Suppose that on the basis of a sample of size N we use Fisher's z for an approximate significance test of $\rho_{ij\cdot q+1,\dots,p}=\rho_0$ against two-sided alternatives. We let

$$
\begin{aligned}
z &= \tfrac{1}{2}\log\frac{1+r_{ij\cdot q+1,\dots,p}}{1-r_{ij\cdot q+1,\dots,p}}, \\
\zeta_0 &= \tfrac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}.
\end{aligned}
\tag{27}
$$

Then $\sqrt{N-(p-q)-3}\,(z-\zeta_0)$ is compared with the significance points of the standardized normal distribution. In the example at the end of Section 4.3.1, we might wish to test the hypothesis $\rho_{13\cdot 2}=0$ at the 0.05 level. Then $\zeta_0=0$ and $\sqrt{20-1-3}\,(0.0973)=0.3892$. This value is clearly nonsignificant $(|0.3892|<1.96)$, and hence the data do not indicate rejection of the null hypothesis.

To answer the question whether two variables x_1 and x_2 are related when both may be related to a vector $x^{(2)}=(x_3,\dots,x_p)$, two approaches may be used. One is to consider the regression of x_1 on x_2 and $x^{(2)}$ and test whether the regression of x_1 on x_2 is 0. Another is to test whether $\rho_{12\cdot 3,\dots,p}=0$. Problems 4.43–4.47 show that these approaches lead to exactly the same test.

4.4. THE MULTIPLE CORRELATION COEFFICIENT

4.4.1. Estimation of the Multiple Correlation Coefficient

The population multiple correlation between one variate and a set of variates was defined in Section 2.5. For the sake of convenience in this section we shall treat the case of the multiple correlation between X_1 and the vector

$X^{(2)} = (X_2, \dots, X_p)'$; we shall not need subscripts on R. The variables can always be numbered so that the desired multiple correlation is this one (any irrelevant variables being omitted). Then the multiple correlation in the population is

$$(1) \qquad \bar{R} = \frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}}{\sqrt{\sigma_{11}\boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}}} = \sqrt{\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}}{\sigma_{11}}} = \sqrt{\frac{\boldsymbol{\sigma}'_{(1)}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{(1)}}{\sigma_{11}}},$$

where $\boldsymbol{\beta}$, $\boldsymbol{\sigma}_{(1)}$, and $\boldsymbol{\Sigma}_{22}$ are defined by

$$(2) \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}'_{(1)} \\ \boldsymbol{\sigma}_{(1)} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

$$(3) \qquad \boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{(1)}.$$

Given a sample $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ $(N > p)$, we estimate $\boldsymbol{\Sigma}$ by $S = [N/(N-1)]\hat{\boldsymbol{\Sigma}}$ or

$$(4) \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}A = \frac{1}{N}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})' = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\boldsymbol{\sigma}}'_{(1)} \\ \hat{\boldsymbol{\sigma}}_{(1)} & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix},$$

and we estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\sigma}}_{(1)} = A_{22}^{-1}\boldsymbol{a}_{(1)}$. We define the *sample multiple correlation coefficient* by

$$(5) \qquad R = \sqrt{\frac{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\Sigma}}_{22}\hat{\boldsymbol{\beta}}}{\hat{\sigma}_{11}}} = \sqrt{\frac{\hat{\boldsymbol{\sigma}}'_{(1)}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\sigma}}_{(1)}}{\hat{\sigma}_{11}}} = \sqrt{\frac{\boldsymbol{a}'_{(1)}A_{22}^{-1}\boldsymbol{a}_{(1)}}{a_{11}}}.$$

That this is the maximum likelihood estimator of $\bar{R}$ is justified by Corollary 3.2.1, since we can define $\bar{R}, \boldsymbol{\sigma}_{(1)}, \boldsymbol{\Sigma}_{22}$ as a one-to-one transformation of $\boldsymbol{\Sigma}$. Another expression for R [see (16) of Section 2.5] follows from

$$(6) \qquad 1 - R^2 = \frac{|\hat{\boldsymbol{\Sigma}}|}{\hat{\sigma}_{11}|\hat{\boldsymbol{\Sigma}}_{22}|} = \frac{|A|}{a_{11}|A_{22}|}.$$

The quantities R and $\hat{\boldsymbol{\beta}}$ have properties in the sample that are similar to those $\bar{R}$ and $\boldsymbol{\beta}$ have in the population. We have analogs of Theorems 2.5.2, 2.5.3, and 2.5.4. Let $\hat{x}_{1\alpha} = \bar{x}_1 + \hat{\boldsymbol{\beta}}'(\boldsymbol{x}_\alpha^{(2)} - \bar{\boldsymbol{x}}^{(2)})$, and $x_{1\alpha}^* = x_{1\alpha} - \hat{x}_{1\alpha}$ be the residual.

Theorem 4.4.1. *The residuals $x_{1\alpha}^*$ are uncorrelated in the sample with the components of $\boldsymbol{x}_\alpha^{(2)}$, $\alpha = 1, \dots, N$. For every vector $\boldsymbol{a}$*

$$(7) \qquad \sum_{\alpha=1}^{N}\left[x_{1\alpha} - \bar{x}_1 - \hat{\boldsymbol{\beta}}'\left(\boldsymbol{x}_\alpha^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)\right]^2 \le \sum_{\alpha=1}^{N}\left[x_{1\alpha} - \bar{x}_1 - \boldsymbol{a}'\left(\boldsymbol{x}_\alpha^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)\right]^2.$$

The sample correlation between $x_{1\alpha}$ and $a'x_\alpha^{(2)}$, $\alpha = 1, \dots, N$, is maximized for $a = \hat{\boldsymbol{\beta}}$, and that maximum correlation is R.

Proof. Since the sample mean of the residuals is 0, the vector of sample covariances between $x_{1\alpha}^*$ and $x_\alpha^{(2)}$ is proportional to

$$(8)\qquad \sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar{x}_1)-\hat{\boldsymbol{\beta}}'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)' = a'_{(1)}-\hat{\boldsymbol{\beta}}'A_{22} = \mathbf{0}.$$

The right-hand side of (7) can be written as the left-hand side plus

$$(9)\qquad \sum_{\alpha=1}^{N}\left[(\hat{\boldsymbol{\beta}}-a)'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]^2 = (\hat{\boldsymbol{\beta}}-a)'\sum_{\alpha=1}^{N}\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)'(\hat{\boldsymbol{\beta}}-a),$$

which is 0 if and only if $a = \hat{\boldsymbol{\beta}}$. To prove the third assertion we consider the vector a for which $\sum_{\alpha=1}^{N}[a'(x_\alpha^{(2)}-\bar{x}^{(2)})]^2 = \sum_{\alpha=1}^{N}[\hat{\boldsymbol{\beta}}'(x_\alpha^{(2)}-\bar{x}^{(2)})]^2$, since the correlation is unchanged when the linear function is mulitplied by a positive constant. From (7) we obtain

$$(10)\qquad \begin{aligned} a_{11} &- 2\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar{x}_1)\hat{\boldsymbol{\beta}}'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)+\sum_{\alpha=1}^{N}\left[\hat{\boldsymbol{\beta}}'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]^2 \\ &\le a_{11}-2\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar{x}_1)a'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)+\sum_{\alpha=1}^{N}\left[a'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]^2, \end{aligned}$$

from which we deduce

$$(11)\qquad \begin{aligned} \frac{\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar{x}_1)\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)'a}{\sqrt{a_{11}}\sqrt{\sum_{\alpha=1}^{N}\left[a'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]^2}} &\le \frac{\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar{x}_1)\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)'\hat{\boldsymbol{\beta}}}{\sqrt{a_{11}}\sqrt{\sum_{\alpha=1}^{N}\left[\hat{\boldsymbol{\beta}}'\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]^2}} \\ &= \frac{a'_{(1)}\hat{\boldsymbol{\beta}}}{\sqrt{a_{11}}\sqrt{\hat{\boldsymbol{\beta}}'A_{22}\hat{\boldsymbol{\beta}}}}, \end{aligned}$$

which is (5). ∎

Thus $\bar{x}_1 + \hat{\boldsymbol{\beta}}'(x_\alpha^{(2)}-\bar{x}^{(2)})$ is the best linear predictor of $x_{1\alpha}$ in the sample, and $\hat{\boldsymbol{\beta}}'x_\alpha^{(2)}$ is the linear function of $x_\alpha^{(2)}$ that has maximum sample correlation with $x_{1\alpha}$. The minimum sum of squares of deviations [the left-hand side of (7)] is

$$\sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar{x}_1)-\hat{\boldsymbol{\beta}}'\left(\boldsymbol{x}_\alpha^{(2)}-\bar{\boldsymbol{x}}^{(2)}\right)\right]^2 = a_{11}-\hat{\boldsymbol{\beta}}'A_{22}\hat{\boldsymbol{\beta}} \tag{12}$$

$$= a_{11}-\boldsymbol{a}'_{(1)}A_{22}^{-1}\boldsymbol{a}_{(1)}$$

$$= a_{11\cdot2}$$

as defined in Section 4.3 with $q=1$. The maximum likelihood estimator of $\sigma_{11\cdot2}$ is $\hat{\sigma}_{11\cdot2}=a_{11\cdot2}/N$. It follows that

$$\hat{\sigma}_{11\cdot2}=(1-R^2)\hat{\sigma}_{11}. \tag{13}$$

Thus $1-R^2$ measures the proportional reduction in the variance by using residuals. We can say that R^2 is the fraction of the variance explained by $\boldsymbol{x}^{(2)}$. The larger R^2 is, the more the variance is decreased by use of the explanatory variables in $\boldsymbol{x}^{(2)}$.

In p-dimensional space $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ represent N points. The sample regression function $x_1=\bar{x}_1+\hat{\boldsymbol{\beta}}'(\boldsymbol{x}^{(2)}-\bar{\boldsymbol{x}}^{(2)})$ is the $(p-1)$-dimensional hyperplane that minimizes the squared deviations of the points from the hyperplane, the deviations being calculated in the x_1-direction. The hyperplane goes through the point $\bar{\boldsymbol{x}}$.

In N-dimensional space the rows of $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_N)$ represent p points. The N-component vector with αth component $x_{i\alpha}-\bar{x}_i$ is the projection of the vector with αth component $x_{i\alpha}$ on the plane orthogonal to the equiangular line. We have p such vectors; $\boldsymbol{a}'(\boldsymbol{x}_\alpha^{(2)}-\bar{\boldsymbol{x}}^{(2)})$ is the αth component of a vector in the hyperplane spanned by the last $p-1$ vectors. Since the right-hand side of (7) is the squared distance between the first vector and the linear combination of the last $p-1$ vectors, $\hat{\boldsymbol{\beta}}'(\boldsymbol{x}_\alpha^{(2)}-\bar{\boldsymbol{x}}^{(2)})$ is a component of the vector which minimizes this squared distance. The interpretation of (8) is that the vector with αth component $(x_{1\alpha}-\bar{x}_1)-\hat{\boldsymbol{\beta}}'(\boldsymbol{x}_\alpha^{(2)}-\bar{\boldsymbol{x}}^{(2)})$ is orthogonal to each of the last $p-1$ vectors. Thus the vector with αth component $\hat{\boldsymbol{\beta}}'(\boldsymbol{x}_\alpha^{(2)}-\bar{\boldsymbol{x}}^{(2)})$ is the projection of the first vector on the hyperplane. See Figure 4.5. The length squared of the projection vector is

$$\sum_{\alpha=1}^{N}\left[\hat{\boldsymbol{\beta}}'\left(\boldsymbol{x}_\alpha^{(2)}-\bar{\boldsymbol{x}}^{(2)}\right)\right]^2=\hat{\boldsymbol{\beta}}'A_{22}\hat{\boldsymbol{\beta}}=\boldsymbol{a}'_{(1)}A_{22}^{-1}\boldsymbol{a}_{(1)}, \tag{14}$$

and the length squared of the first vector is $\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar{x}_1)^2=a_{11}$. Thus R is the cosine of the angle between the first vector and its projection.

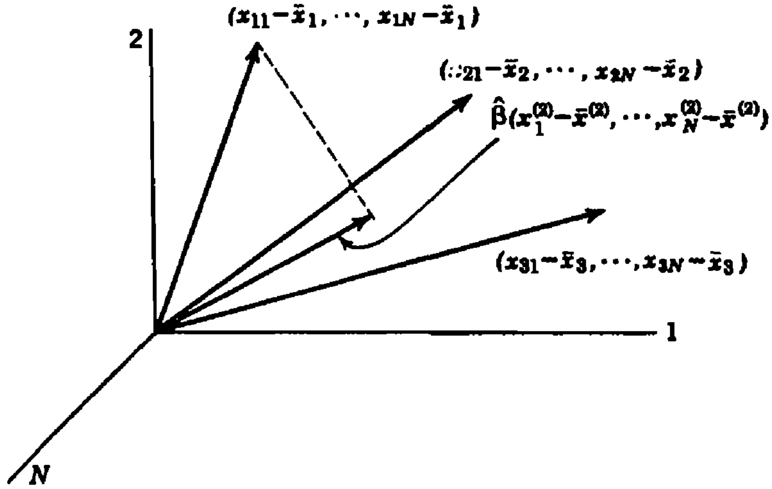

Figure 4.5

In Section 3.2 we saw that the simple correlation coefficient is the cosine of the angle between the two vectors involved (in the plane orthogonal to the equiangular line). The property of R that it is the maximum correlation between $x_{1\alpha}$ and linear combinations of the components of $x_\alpha^{(2)}$ corresponds to the geometric property that R is the cosine of the smallest angle between the vector with components $x_{1\alpha}-\bar{x}_1$ and a vector in the hyperplane spanned by the other $p-1$ vectors.

The geometric interpretations are in terms of the vectors in the $(N-1)$-dimensional hyperplane orthogonal to the equiangular line. It was shown in Section 3.3 that the vector $(x_{i1}-\bar{x}_i, \dots, x_{iN}-\bar{x}_i)$ in this hyperplane can be designated as $(z_{i1}, \dots, z_{i,N-1})$, where the $z_{i\alpha}$ are the coordinates referred to an $(N-1)$-dimensional coordinate system in the hyperplane. It was shown that the new coordinates are obtained from the old by the transformation $z_{i\alpha} = \sum_{\beta=1}^N b_{\alpha\beta} x_{i\beta}$, $\alpha = 1, \dots, N$, where $B = (b_{\alpha\beta})$ is an orthogonal matrix with last row $(1/\sqrt{N}, \dots, 1/\sqrt{N})$. Then

$$a_{ij} = \sum_{\alpha=1}^{N} (x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j) = \sum_{\alpha=1}^{N-1} z_{i\alpha} z_{j\alpha}. \tag{15}$$

It will be convenient to refer to the multiple correlation defined in terms of $z_{i\alpha}$ as the *multiple correlation without subtracting the means*.

The population multiple correlation $\bar{R}$ is essentially the only function of the parameters μ and Σ that is invariant under changes of location, changes of scale of X_1, and nonsingular linear transformations of $X^{(2)}$, that is, transformations $X_1^* = cX_1 + d$, $X^{(2)*} = CX^{(2)} + d$. Similarly, the sample multiple correlation coefficient R is essentially the only function of $\bar{x}$ and $\hat{\Sigma}$, the sufficient set of statistics for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, that is invariant under these transformations. Just as the simple correlation r is a measure of association between two scalar variables in a sample, the multiple correlation R is a measure of association between a scalar variable and a vector variable in a sample.

4.4.2. Distribution of the Sample Multiple Correlation Coefficient When the Population Multiple Correlation Coefficient Is Zero

From (5) we have

$$R^2 = \frac{\boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)}}{a_{11}}; \tag{16}$$

then

$$1 - R^2 = 1 - \frac{\boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)}}{a_{11}} = \frac{a_{11} - \boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)}}{a_{11}} = \frac{a_{11\cdot 2}}{a_{11}}, \tag{17}$$

and

$$\frac{R^2}{1 - R^2} = \frac{\boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)}}{a_{11\cdot 2}}. \tag{18}$$

For $q = 1$, Corollary 4.3.2 states that when $\boldsymbol{\beta} = \mathbf{0}$, that is, when $\bar{R} = 0$, $a_{11\cdot 2}$ is distributed as $\sum_{\alpha=1}^{N-p} V_\alpha^2$ and $\boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)}$ is distributed as $\sum_{\alpha=N-p+1}^{N-1} V_\alpha^2$, where $V_1, \dots, V_{N-1}$ are independent, each with distribution $N(0, \sigma_{11\cdot 2})$. Then $a_{11\cdot 2}/\sigma_{11\cdot 2}$ and $\boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)}/\sigma_{11\cdot 2}$ are distributed independently as χ^2-variables with $N - p$ and $p - 1$ degrees of freedom, respectively. Thus

$$\begin{aligned} \frac{R^2}{1 - R^2} \cdot \frac{N - p}{p - 1} &= \frac{\boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)}/\sigma_{11\cdot 2}}{a_{11\cdot 2}/\sigma_{11\cdot 2}} \cdot \frac{N - p}{p - 1} \\ &= \frac{\chi_{p-1}^2}{\chi_{N-p}^2} \cdot \frac{N - p}{p - 1} \\ &= F_{p-1, N-p} \end{aligned} \tag{19}$$

has the F-distribution with $p - 1$ and $N - p$ degrees of freedom. The density of F is

(20)

$$\frac{\Gamma[\frac{1}{2}(N-1)]}{\Gamma[\frac{1}{2}(p-1)]\Gamma[\frac{1}{2}(N-p)]} \left(\frac{p-1}{N-p}\right)^{\frac{1}{2}(p-1)} f^{\frac{1}{2}(p-1)-1} \left(1 + \frac{p-1}{N-p} f\right)^{-\frac{1}{2}(N-1)}.$$

Thus the density of

$$R = \sqrt{\frac{\dfrac{p-1}{N-p}F_{p-1,N-p}}{1+\dfrac{p-1}{N-p}F_{p-1,N-p}}} \tag{21}$$

is

$$2\frac{\Gamma[\frac{1}{2}(N-1)]}{\Gamma[\frac{1}{2}(p-1)]\Gamma[\frac{1}{2}(N-p)]}R^{p-2}(1-R^2)^{\frac{1}{2}(N-p)-1}, \qquad 0 \le R \le 1. \tag{22}$$

Theorem 4.4.2. *Let R be the sample multiple correlation coefficient [defined by (5)] between X_1 and $X^{(2)\prime} = (X_2, \dots, X_p)$ based on a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\bar{R} = 0$ [that is, if $(\sigma_{12}, \dots, \sigma_{1p})' = \mathbf{0} = \boldsymbol{\beta}$], then $[R^2/(1-R^2)]\cdot[(N-p)/(p-1)]$ is distributed as F with $p-1$ and $N-p$ degrees of freedom.*

It should be noticed that $p-1$ is the number of components of $X^{(2)}$ and that $N-p = N-(p-1)-1$. If the multiple correlation is between a component X_i and q other components, the numbers are q and $N-q-1$.

It might be observed that $R^2/(1-R^2)$ is the quantity that arises in regression (or least squares) theory for testing the hypothesis that the regression of X_1 on $X_2, \dots, X_p$ is zero.

If $\bar{R} \neq 0$, the distribution of R is much more difficult to derive. This distribution will be obtained in Section 4.4.3.

Now let us consider the statistical problem of testing the hypothesis $H: \bar{R} = 0$ on the basis of a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. [$\bar{R}$ is the population multiple correlation between X_1 and $(X_2, \dots, X_p)$.] Since $\bar{R} \ge 0$, the alternatives considered are $\bar{R} > 0$.

Let us derive the likelihood ratio test of this hypothesis. The likelihood function is

$$L(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \frac{1}{(2\pi)^{\frac{1}{2}pN}|\boldsymbol{\Sigma}^*|^{\frac{1}{2}N}}\exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha - \boldsymbol{\mu}^*)'\boldsymbol{\Sigma}^{*-1}(\boldsymbol{x}_\alpha - \boldsymbol{\mu}^*)\right] \tag{23}$$

The observations are given; L is a function of the indeterminates $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$. Let ω be the region in the parameter space Ω specified by the null hypothesis. The likelihood ratio criterion is

$$\lambda = \frac{\max_{\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* \in \omega} L(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)}{\max_{\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* \in \Omega} L(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)}. \tag{24}$$

Here Ω is the space of $\boldsymbol{\mu}^*$, $\boldsymbol{\Sigma}^*$ positive definite, and ω is the region in this space where $\bar{R} = \sqrt{\boldsymbol{\sigma}'_{(1)}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{(1)}} \big/ \sqrt{\sigma_{11}} = 0$, that is, where $\boldsymbol{\sigma}'_{(1)}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{(1)} = 0$. Because $\boldsymbol{\Sigma}_{22}^{-1}$ is positive definite, this condition is equivalent to $\boldsymbol{\sigma}_{(1)} = \mathbf{0}$. The maximum of $L(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ over Ω occurs at $\boldsymbol{\mu}^* = \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}^* = \hat{\boldsymbol{\Sigma}} = (1/N)A = (1/N)\sum_{\alpha=1}^N (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ and is

$$\max_{\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* \in \Omega} L(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \frac{N^{-\frac{1}{2}pN} e^{\frac{1}{2}pN}}{(2\pi)^{\frac{1}{2}pN} |A|^{\frac{1}{2}N}}. \tag{25}$$

In ω the likelihood function is

$$L\left(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* | \boldsymbol{\sigma}_{(1)}^* = \mathbf{0}\right) = \frac{1}{(2\pi)^{\frac{1}{2}N} \sigma_{11}^{*\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^N (x_{1\alpha} - \mu_1^*)^2 / \sigma_{11}^*\right] \tag{26}$$
$$\cdot \frac{1}{(2\pi)^{\frac{1}{2}(p-1)N} |\boldsymbol{\Sigma}_{22}^*|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^N \left(\boldsymbol{x}_\alpha^{(2)} - \boldsymbol{\mu}^{(2)*}\right)' \boldsymbol{\Sigma}_{22}^{*-1} \left(\boldsymbol{x}_\alpha^{(2)} - \boldsymbol{\mu}^{(2)*}\right)\right].$$

The first factor is maximized at $\mu_1^* = \hat{\mu}_1 = \bar{x}_1$ and $\sigma_{11}^* = \hat{\sigma}_{11}^* = (1/N)a_{11}$, and the second factor is maximized at $\boldsymbol{\mu}^{(2)*} = \hat{\boldsymbol{\mu}}^{(2)} = \bar{\boldsymbol{x}}^{(2)}$ and $\boldsymbol{\Sigma}_{22}^* = \hat{\boldsymbol{\Sigma}}_{22} = (1/N)A_{22}$. The value of the maximized function is

$$\max_{\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* \in \omega} L(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \frac{N^{\frac{1}{2}N} e^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}N} a_{11}^{\frac{1}{2}N}} \cdot \frac{N^{\frac{1}{2}(p-1)N} e^{-\frac{1}{2}(p-1)N}}{(2\pi)^{\frac{1}{2}(p-1)N} |A_{22}|^{\frac{1}{2}N}}. \tag{27}$$

Thus the likelihood ratio criterion is [see (6)]

$$\lambda = \frac{|A|^{\frac{1}{2}N}}{a_{11}^{\frac{1}{2}N} |A_{22}|^{\frac{1}{2}N}} = (1 - R^2)^{\frac{1}{2}N}. \tag{28}$$

The likelihood ratio test consists of the critical region $\lambda < \lambda_0$, where λ_0 is chosen so the probability of this inequality when $\bar{R} = 0$ is the significance level α. An equivalent test is

$$1 - \lambda^{2/N} = R^2 > 1 - \lambda_0^{2/N}. \tag{29}$$

Since $[R^2/(1 - R^2)][(N-p)/(p-1)]$ is a monotonic function of R, an equivalent test involves this ratio being larger than a constant. When $\bar{R} = 0$, this ratio has an $F_{p-1, N-p}$-distribution. Hence, the critical region is

$$\frac{R^2}{1 - R^2} \cdot \frac{N-p}{p-1} > F_{p-1, N-p}(\alpha), \tag{30}$$

where $F_{p-1, N-p}(\alpha)$ is the (upper) significance point corresponding to the α significance level.

Theorem 4.4.3. *Given a sample* $x_1, \dots, x_N$ *from* $N(\mu, \Sigma)$, *the likelihood ratio test at significance level* α *for the hypothesis* $\bar{R} = 0$, *where* $\bar{R}$ *is the population multiple correlation coefficient between* X_1 *and* $(X_2, \dots, X_p)$, *is given by* (30), *where* R *is the sample multiple correlation coefficient defined by* (5).

As an example consider the data given at the end of Section 4.3.1. The sample multiple correlation coefficient is found from

$$(31)\quad 1 - R^2 = \frac{\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix}} = \frac{\begin{vmatrix} 1.00 & 0.80 & -0.40 \\ 0.80 & 1.00 & -0.56 \\ -0.40 & -0.56 & 1.00 \end{vmatrix}}{\begin{vmatrix} 1.00 & -0.56 \\ -0.56 & 1.00 \end{vmatrix}} = 0.357.$$

Thus R is 0.802. If we wish to test the hypothesis at the 0.01 level that hay yield is independent of spring rainfall and temperature, we compare the observed $[R^2/(1-R^2)][(20-3)/(3-1)] = 15.3$ with $F_{2,17}(0.01) = 6.11$ and find the result significant; that is, we reject the null hypothesis.

The test of independence between X_1 and $(X_2, \dots, X_p) = X^{(2)\prime}$ is equivalent to the test that if the regression of X_1 on $x^{(2)}$ (that is, the conditional expected value of X_1 given $X_2 = x_2, \dots, X_p = x_p$) is $\mu_1 + \beta'(x^{(2)} - \mu^{(2)})$, the vector of regression coefficients is $\mathbf{0}$. Here $\hat{\beta} = A_{22}^{-1} a_{(1)}$ is the usual least squares estimate of β with expected value β and covariance matrix $\sigma_{11\cdot 2} A_{22}^{-1}$ (when the $X_\alpha^{(2)}$ are fixed), and $a_{11\cdot 2}/(N-p)$ is the usual estimate of $\sigma_{11\cdot 2}$. Thus [see (18)]

$$(32)\qquad \frac{R^2}{1-R^2} \cdot \frac{N-p}{p-1} = \frac{\hat{\beta}' A_{22} \hat{\beta}}{a_{11\cdot 2}} \cdot \frac{N-p}{p-1}$$

is the usual F-statistic for testing the hypothesis that the regression of X_1 on $x_2, \dots, x_p$ is 0. In this book we are primarily interested in the multiple correlation coefficient as a measure of association between one variable and a vector of variables when both are random. We shall not treat problems of univariate regression. In Chapter 8 we study regression when the dependent variable is a vector.

Adjusted Multiple Correlation Coefficient

The expression (17) is the ratio of $a_{11\cdot 2}$, the sum of squared deviations from the fitted regression, to a_{11}, the sum of squared deviations around the mean. To obtain unbiased estimators of σ_{11} when $\beta = \mathbf{0}$ we would divide these quantities by their numbers of degrees of freedom, $N-p$ and $N-1$, respectively. Accordingly we can define an *adjusted multiple correlation coefficient* R^* by

$$(33) \qquad 1 - R^{*2} = \frac{a_{11\cdot 2}/(N-p)}{a_{11}/(N-1)} = \frac{N-1}{N-p}(1-R^2),$$

which is equivalent to

$$(34) \qquad R^{*2} = R^2 - \frac{p-1}{N-p}(1-R^2).$$

This quantity is smaller than R^2 (unless $p = 1$ or $R^2 = 1$). A possible merit to it is that it takes account of p; the idea is that the larger p is relative to N, the greater the tendency of R^2 to be large by chance.

4.4.3. Distribution of the Sample Multiple Correlation Coefficient When the Population Multiple Correlation Coefficient Is Not Zero

In this subsection we shall find the distribution of R when the null hypothesis $\bar{R} = 0$ is not true. We shall find that the distribution depends only on the population multiple correlation coefficient $\bar{R}$.

First let us consider the conditional distribution of $R^2/(1-R^2) = a'_{(1)}A_{22}^{-1}a_{(1)}/a_{11\cdot 2}$ given $Z_\alpha^{(2)} = z_\alpha^{(2)}$, $\alpha = 1, \ldots, n$. Under these conditions $Z_{11}, \ldots, Z_{1n}$ are independently distributed, $Z_{1\alpha}$ according to $N(\boldsymbol{\beta}' z_\alpha^{(2)}, \sigma_{11\cdot 2})$, where $\boldsymbol{\beta} = \Sigma_{22}^{-1}\boldsymbol{\sigma}_{(1)}$ and $\sigma_{11\cdot 2} = \sigma_{11} - \boldsymbol{\sigma}'_{(1)}\Sigma_{22}^{-1}\boldsymbol{\sigma}_{(1)}$. The conditions are those of Theorem 4.3.3 with $Y_\alpha = Z_{1\alpha}$, $\Gamma = \boldsymbol{\beta}'$, $w_\alpha = z_\alpha^{(2)}$, $r = p-1$, $\Phi = \sigma_{11\cdot 2}$, $m = n$. Then $a_{11\cdot 2} = a_{11} - a'_{(1)}A_{22}^{-1}a_{(1)}$ corresponds to $\sum_{\alpha=1}^m Y_\alpha Y'_\alpha - GHG'$, and $a_{11\cdot 2}/\sigma_{11\cdot 2}$ has a χ^2-distribution with $n-(p-1)$ degrees of freedom. $a'_{(1)}A_{22}^{-1}a_{(1)} = (A_{22}^{-1}a_{(1)})'A_{22}(A_{22}^{-1}a_{(1)})$ corresponds to GHG' and is distributed as $\sum_\alpha U_\alpha^2$, $\alpha = n-(p-1)+1, \ldots, n$, where $\operatorname{Var}(U_\alpha) = \sigma_{11\cdot 2}$ and

$$(35) \qquad \mathscr{E}(U_{n-p+2}, \ldots, U_n) = \Gamma F^{-1},$$

where $FHF' = I$ $[H = F^{-1}(F')^{-1}]$. Then $a'_{(1)}A_{22}^{-1}a_{(1)}/\sigma_{11\cdot 2}$ is distributed as $\sum_\alpha (U_\alpha/\sqrt{\sigma_{11\cdot 2}})^2$, where $\operatorname{Var}(U_\alpha/\sqrt{\sigma_{11\cdot 2}}) = 1$ and

$$(36) \qquad \sum_{\alpha=n-p+2}^{n} \left(\frac{\mathscr{E}U_\alpha}{\sqrt{\sigma_{11\cdot 2}}}\right)^2 = \frac{1}{\sigma_{11\cdot 2}}\Gamma F^{-1}(\Gamma F^{-1})' = \frac{\Gamma H \Gamma'}{\sigma_{11\cdot 2}}$$

$$= \frac{\boldsymbol{\beta}' A_{22}\boldsymbol{\beta}}{\sigma_{11\cdot 2}}.$$

Thus (conditionally) $a'_{(1)}A_{22}^{-1}a_{(1)}/\sigma_{11\cdot 2}$ has a noncentral χ^2-distribution with

$p-1$ degrees of freedom and noncentrality parameter $\boldsymbol{\beta}' A_{22} \boldsymbol{\beta} / \sigma_{11 \cdot 2}$. (See Theorem 5.4.1.) We are led to the following theorem:

Theorem 4.4.4. *Let R be the sample multiple correlation coefficient between $X_{(1)}$ and $\boldsymbol{X}^{(2)\prime} = (X_2, \ldots, X_p)$ based on N observations $(x_{11}, \boldsymbol{x}_1^{(2)}), \ldots, (x_{1N}, \boldsymbol{x}_N^{(2)})$. The conditional distribution of $[R^2/(1-R^2)][N-p)/(p-1)]$ given $\boldsymbol{x}_\alpha^{(2)}$ fixed is noncentral F with $p-1$ and $N-p$ degrees of freedom and noncentrality parameter $\boldsymbol{\beta}' A_{22} \boldsymbol{\beta} / \sigma_{11 \cdot 2}$.*

The conditional density (from Theorem 5.4.1) of $F = [R^2/(1-R^2)][(N-p)/(p-1)]$ is

$$
(37) \quad \frac{(p-1)\exp\left[-\frac{1}{2}\boldsymbol{\beta}' A_{22} \boldsymbol{\beta} / \sigma_{11 \cdot 2}\right]}{(N-p)\Gamma\left[\frac{1}{2}(N-p)\right]}
\cdot \sum_{\alpha=0}^{\infty} \frac{\left(\dfrac{\boldsymbol{\beta}' A_{22} \boldsymbol{\beta}}{2\sigma_{11 \cdot 2}}\right)^{\alpha} \left[\dfrac{(p-1)f}{N-p}\right]^{\frac{1}{2}(p-1)+\alpha-1} \Gamma\left[\frac{1}{2}(N-1)+\alpha\right]}{\alpha!\,\Gamma\left[\frac{1}{2}(p-1)+\alpha\right]\left[1+\dfrac{(p-1)f}{N-p}\right]^{\frac{1}{2}(N-1)+\alpha}},
$$

and the conditional density of $W = R^2$ is $(df = [(N-p)/(p-1)](1-w)^{-2}\,dw)$

$$
(38) \quad \frac{\exp\left[-\frac{1}{2}\boldsymbol{\beta}' A_{22} \boldsymbol{\beta} / \sigma_{11 \cdot 2}\right]}{\Gamma\left[\frac{1}{2}(N-p)\right]}(1-w)^{\frac{1}{2}(N-p)-1}
\cdot \sum_{\alpha=0}^{\infty} \frac{\left(\dfrac{\boldsymbol{\beta}' A_{22} \boldsymbol{\beta}}{2\sigma_{11 \cdot 2}}\right)^{\alpha} w^{\frac{1}{2}(p-1)+\alpha-1}\,\Gamma\left[\frac{1}{2}(N-1)+\alpha\right]}{\alpha!\,\Gamma\left[\frac{1}{2}(p-1)+\alpha\right]}.
$$

To obtain the unconditional density we need to multiply (38) by the density of $\boldsymbol{Z}_1^{(2)}, \ldots, \boldsymbol{Z}_n^{(2)}$ to obtain the joint density of W and $\boldsymbol{Z}_1^{(2)}, \ldots, \boldsymbol{Z}_n^{(2)}$ and then integrate with respect to the latter set to obtain the marginal density of W. We have

$$
(39) \quad \frac{\boldsymbol{\beta}' A_{22} \boldsymbol{\beta}}{\sigma_{11 \cdot 2}} = \frac{\boldsymbol{\beta}' \sum_{\alpha=1}^{n} \boldsymbol{z}_\alpha^{(2)} \boldsymbol{z}_\alpha^{(2)\prime} \boldsymbol{\beta}}{\sigma_{11 \cdot 2}}
= \sum_{\alpha=1}^{n} \left(\frac{\boldsymbol{\beta}' \boldsymbol{z}_\alpha^{(2)}}{\sqrt{\sigma_{11 \cdot 2}}}\right)^2.
$$

Since the distribution of $\mathbf{Z}_\alpha^{(2)}$ is $N(\mathbf{0}, \mathbf{\Sigma}_{22})$, the distribution of $\boldsymbol{\beta}'\mathbf{Z}_\alpha^{(2)}/\sqrt{\sigma_{11\cdot 2}}$ is normal with mean zero and variance

$$(40) \qquad \mathscr{E}\left(\frac{\boldsymbol{\beta}'\mathbf{Z}_\alpha^{(2)}}{\sqrt{\Sigma_{11\cdot 2}}}\right)^2 = \frac{\mathscr{E}\boldsymbol{\beta}'\mathbf{Z}_\alpha^{(2)}\mathbf{Z}_\alpha^{(2)\prime}\boldsymbol{\beta}}{\sigma_{11\cdot 2}}$$

$$= \frac{\boldsymbol{\beta}'\mathbf{\Sigma}_{22}\boldsymbol{\beta}}{\sigma_{11} - \boldsymbol{\beta}'\mathbf{\Sigma}_{22}\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}'\mathbf{\Sigma}_{22}\boldsymbol{\beta}/\sigma_{11}}{1 - \boldsymbol{\beta}'\mathbf{\Sigma}_{22}\boldsymbol{\beta}/\sigma_{11}}$$

$$= \frac{\bar{R}^2}{1 - \bar{R}^2}.$$

Thus $(\boldsymbol{\beta}'A_{22}\boldsymbol{\beta}/\sigma_{11\cdot 2})/[\bar{R}^2/(1-\bar{R}^2)]$ has a χ^2-distribution with n degrees of freedom. Let $\bar{R}^2/(1-\bar{R}^2) = \phi$. Then $\boldsymbol{\beta}'A_{22}\boldsymbol{\beta}/\sigma_{11\cdot 2} = \phi\chi_n^2$. We compute

$$(41) \qquad \mathscr{E}e^{-\frac{1}{2}\phi\chi_n^2}\left(\frac{\phi\chi_n^2}{2}\right)^\alpha$$

$$= \frac{\phi^\alpha}{2^\alpha}\int_0^\infty u^\alpha e^{-\frac{1}{2}\phi u}\frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)}u^{\frac{1}{2}n-1}e^{-\frac{1}{2}u}\,du$$

$$= \frac{\phi^\alpha}{2^\alpha}\int_0^\infty \frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)}u^{\frac{1}{2}n+\alpha-1}e^{-\frac{1}{2}(1+\phi)u}\,du$$

$$= \frac{\phi^\alpha}{(1+\phi)^{\frac{1}{2}n+\alpha}}\frac{\Gamma\left(\frac{1}{2}n+\alpha\right)}{\Gamma\left(\frac{1}{2}n\right)}\int_0^\infty \frac{1}{2^{\frac{1}{2}n+\alpha}\Gamma\left(\frac{1}{2}n+\alpha\right)}v^{\frac{1}{2}n+\alpha-1}e^{-\frac{1}{2}v}\,dv$$

$$= \frac{\phi^\alpha}{(1+\phi)^{\frac{1}{2}n+\alpha}}\frac{\Gamma\left(\frac{1}{2}n+\alpha\right)}{\Gamma\left(\frac{1}{2}n\right)}.$$

Applying this result to (38), we obtain as the density of R^2

$$(42) \qquad \frac{(1-R^2)^{\frac{1}{2}(n-p-1)}(1-\bar{R}^2)^{\frac{1}{2}n}}{\Gamma\left[\frac{1}{2}(n-p+1)\right]\Gamma\left(\frac{1}{2}n\right)}\sum_{\mu=0}^\infty \frac{(\bar{R}^2)^\mu (R^2)^{\frac{1}{2}(p-1)+\mu-1}\Gamma^2\left(\frac{1}{2}n+\mu\right)}{\mu!\,\Gamma\left[\frac{1}{2}(p-1)+\mu\right]}.$$

Fisher (1928) found this distribution. It can also be written

$$(43) \qquad \frac{\Gamma\left(\frac{1}{2}n\right)(1-\bar{R}^2)^{\frac{1}{2}n}}{\Gamma\left[\frac{1}{2}(n-p+1)\right]\Gamma\left[\frac{1}{2}(p-1)\right]}(R^2)^{\frac{1}{2}(p-3)}(1-R^2)^{\frac{1}{2}(n-p-1)}$$

$$\cdot F\left[\tfrac{1}{2}n, \tfrac{1}{2}n; \tfrac{1}{2}(p-1); R^2\bar{R}^2\right],$$

where F is the hypergeometric function defined in (41) of Section 4.2.

Another form of the density can be obtained when $n-p+1$ is even. We have

$$(44)\quad \sum_{\mu=0}^{\infty} \frac{(R^2\bar{R}^2)^{\mu}}{\mu!} \frac{\Gamma^2(\frac{1}{2}n+\mu)}{\Gamma[\frac{1}{2}(p-1)+\mu]}$$

$$= \sum_{\mu=0}^{\infty} \frac{(R^2\bar{R}^2)^{\mu}}{\mu!} \Gamma(\tfrac{1}{2}n+\mu)\left(\frac{\partial}{\partial t}\right)^{\frac{1}{2}(n-p+1)} t^{\frac{1}{2}n+\mu-1}\Bigg|_{t=1}$$

$$= \left(\frac{\partial}{\partial t}\right)^{\frac{1}{2}(n-p+1)} t^{\frac{1}{2}n-1} \sum_{\mu=0}^{\infty} \frac{(t\bar{R}^2R^2)^{\mu}}{\mu!} \frac{\Gamma(\frac{1}{2}n+\mu)}{\Gamma(\frac{1}{2}n)}\Bigg|_{t=1} \Gamma(\tfrac{1}{2}n)$$

$$= \Gamma(\tfrac{1}{2}n)\left(\frac{\partial}{\partial t}\right)^{\frac{1}{2}(n-p+1)} t^{\frac{1}{2}n-1}(1-tR^2\bar{R}^2)^{-\frac{1}{2}n}\Bigg|_{t=1}.$$

The density is therefore

$$(45)\quad \frac{(1-\bar{R}^2)^{\frac{1}{2}n}(R^2)^{\frac{1}{2}(p-3)}(1-R^2)^{\frac{1}{2}(n-p-1)}}{\Gamma[\frac{1}{2}(n-p+1)]} \cdot \left(\frac{\partial}{\partial t}\right)^{\frac{1}{2}(n-p+1)} t^{\frac{1}{2}n-1}(1-tR^2\bar{R}^2)^{-\frac{1}{2}n}\Bigg|_{t=1}.$$

Theorem 4.4.5. *The density of the square of the multiple correlation coefficient, R^2, between X_1 and $X_2, \dots, X_p$ based on a sample of $N=n+1$ is given by (42) or (43) [or (45) in the case of $n-p+1$ even], where $\bar{R}^2$ is the corresponding population multiple correlation coefficient.*

The moments of R are

$$(46)\quad \mathscr{E}R^h = \frac{(1-\bar{R}^2)^{\frac{1}{2}n}}{\Gamma[\frac{1}{2}(n-p+1)]\Gamma(\frac{1}{2}n)} \sum_{\mu=0}^{\infty} \frac{(\bar{R}^2)^{\mu}\Gamma^2(\frac{1}{2}n+\mu)}{\Gamma[\frac{1}{2}(p-1)+\mu]\,\mu!} \cdot \int_0^1 (1-R^2)^{\frac{1}{2}(n-p+1)-1}(R^2)^{\frac{1}{2}(p+h-1)+\mu-1}\,d(R^2)$$

$$= \frac{(1-\bar{R}^2)^{\frac{1}{2}n}}{\Gamma(\frac{1}{2}n)} \sum_{\mu=0}^{\infty} \frac{(\bar{R}^2)^{\mu}\Gamma^2(\frac{1}{2}n+\mu)\Gamma[\frac{1}{2}(p+h-1)+\mu]}{\mu!\Gamma[\frac{1}{2}(p-1)+\mu]\Gamma[\frac{1}{2}(n+h)+\mu]}.$$

The sample multiple correlation tends to overestimate the population multiple correlation. The sample multiple correlation is the maximum sample correlation between x_1 and linear combinations of $\boldsymbol{x}^{(2)}$ and hence is greater than the sample correlation between x_1 and $\boldsymbol{\beta}'\boldsymbol{x}^{(2)}$; however, the latter is the simple sample correlation corresponding to the simple population correlation between x_1 and $\boldsymbol{\beta}'\boldsymbol{x}^{(2)}$, which is $\bar{R}$, the population multiple correlation.

Suppose R_1 is the multiple correlation in the first of two samples and $\hat{\boldsymbol{\beta}}_1$ is the estimate of $\boldsymbol{\beta}$; then the simple correlation between x_1 and $\hat{\boldsymbol{\beta}}_1' x^{(2)}$ in the second sample will tend to be less than R_1 and in particular will be less than R_2, the multiple correlation in the second sample. This has been called "the shrinkage of the multiple correlation."

Kramer (1963) and Lee (1972) have given tables of the upper significance points of R. Gajjar (1967), Gurland (1968), Gurland and Milton (1970), Khatri (1966), and Lee (1917b) have suggested approximations to the distributions of $R^2/(1-R^2)$ and obtained large-sample results.

4.4.4. Some Optimal Properties of the Multiple Correlation Test

Theorem 4.4.6. *Given the observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, of all tests of $\bar{R} = 0$ at a given significance level based on $\bar{\boldsymbol{x}}$ and $A = \sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$ that are invariant with respect to transformations*

$$\begin{aligned} &\bar{x}_1^* = c\bar{x}_1 + d, \qquad \bar{\boldsymbol{x}}^{(2)*} = C\bar{\boldsymbol{x}}^{(2)} + \boldsymbol{d}, \\ &a_{11}^* = c^2 a_{11}, \qquad \boldsymbol{a}_{(1)}^* = cC\boldsymbol{a}_{(1)}, \qquad A_{22}^* = CA_{22}C', \end{aligned} \tag{47}$$

any critical rejection region given by R greater than a constant is uniformly most powerful.

Proof. The multiple correlation coefficient R is invariant under the transformation, and any function of the sufficient statistics that is invariant is a function of R. (See Problem 4.34.) Therefore, any invariant test must be based on R. The Neyman–Pearson fundamental lemma applied to testing the null hypothesis $\bar{R} = 0$ against a specific alternative $\bar{R} = \bar{R}_0 > 0$ tells us the most powerful test at a given level of significance is based on the ratio of the density of R for $\bar{R} = \bar{R}_0$, which is (42) times $2R$ [because (42) is the density of R^2], to the density for $R = 0$, which is (22). The ratio is a positive constant times

$$\sum_{\mu=0}^{\infty} \frac{\left(\bar{R}_0^2\right)^{\mu} \Gamma^2\left(\frac{1}{2}n + \mu\right)}{\mu!\,\Gamma\left[\frac{1}{2}(p-1) + \mu\right]} R^{p-2+2\mu}. \tag{48}$$

Since (48) is an increasing function of R for $R \geq 0$, the set of R for which (48) is greater than a constant is an interval of R greater than a constant. ∎

Theorem 4.4.7. *On the basis of observations $x_1, \ldots, x_N$ from $N(\mu, \Sigma)$, of all tests of $\bar{R} = 0$ at a given significance level with power depending only on $\bar{R}$, the test with critical region given by R greater than a constant is uniformly most powerful.*

Theorem 4.4.7 follows from Theorem 4.4.6 in the same way that Theorem 5.6.4 follows from Theorem 5.6.1.

4.5. ELLIPTICALLY CONTOURED DISTRIBUTIONS

4.5.1. Observations Elliptically Contoured

Suppose $x_1, \ldots, x_N$ are N independent observations on a random p-vector X with density

$$(1) \qquad |\Lambda|^{-\frac{1}{2}} g\left[(x-\nu)'\Lambda^{-1}(x-\nu)\right].$$

The sample covariance matrix S is an unbiased estimator of the covariance matrix $\Sigma = [\mathscr{E}R^2/p]\Lambda$, where $R^2 = (X-\nu)'\Lambda^{-1}(X-\nu)$ and $\mathscr{E}R^2 < \infty$. An estimator of $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} = \lambda_{ij}/\sqrt{\lambda_{ii}\lambda_{jj}}$ is $r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$, $i, j = 1, \ldots, p$. The small-sample distribution of r_{ij} is in general difficult to obtain, but the asymptotic distribution can be obtained from the limiting normal distribution of $\sqrt{N}(S - \Sigma)$ given in (13) of Section 3.6.

First we prove a general theorem on asymptotic distributions of functions of the sample covariance matrix S using Theorems 4.2.3 and 3.6.5. Define

$$(2) \qquad s = \operatorname{vec} S, \qquad \boldsymbol{\sigma} = \operatorname{vec} \Sigma.$$

Theorem 4.5.1. *Let $f(s)$ be a vector-valued function such that each component of $f(s)$ has a nonzero differential at $s = \boldsymbol{\sigma}$. Suppose S is the covariance of a sample from (1) such that $\mathscr{E}R^4 < \infty$. Then*

$$(3) \qquad \sqrt{N}[f(s) - f(\boldsymbol{\sigma})] = \frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}'}\sqrt{N}(s - \boldsymbol{\sigma}) + o_p(1)$$
$$\xrightarrow{d} N\left\{\mathbf{0}, \frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}'}[2(1+\kappa)(\Sigma \otimes \Sigma) + \kappa\boldsymbol{\sigma}\boldsymbol{\sigma}']\left(\frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}'}\right)'\right\}.$$

Corollary 4.5.1. *If*

$$(4) \qquad f(cs) = f(s)$$

for all $c > 0$ and all positive definite S and the conditions of Theorem 4.5.1 hold, then

$$(5) \qquad \sqrt{N}[f(s) - f(\boldsymbol{\sigma})] \xrightarrow{d} N\left[\mathbf{0}, 2(1+\kappa)\frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}'}(\Sigma \otimes \Sigma)\left(\frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}'}\right)'\right].$$

Proof. From (4) we deduce

$$(6) \qquad 0 = \frac{\partial f(cs)}{\partial c} = \frac{\partial f(cs)}{\partial s'} \frac{\partial (cs)}{\partial c} = \frac{\partial f(cs)}{\partial s'} s.$$

That is,

$$(7) \qquad \frac{\partial f(\sigma)}{\partial \sigma'} \sigma = 0. \qquad \blacksquare$$

The conclusion of Corollary 4.5.1 can be framed as

$$(8) \qquad \frac{\sqrt{N}}{\sqrt{1+\kappa}} [f(s) - f(\sigma)] \xrightarrow{d} N\left[0, 2\frac{\partial f(\sigma)}{\partial \sigma'} (\Sigma \otimes \Sigma) \left(\frac{\partial f(\sigma)}{\partial \sigma'}\right)'\right].$$

The limiting normal distribution in (8) holds in particular when the sample is drawn from the normal distribution. The corollary holds true if κ is replaced by a consistent estimator $\hat{\kappa}$. For example, a consistent estimator of $1 + \hat{\kappa}$ given by (16) of Section 3.6 is

$$(9) \qquad 1 + \hat{\kappa} = \sum_{\alpha=1}^{N} \left[(x_\alpha - \bar{x})' S^{-1} (x_\alpha - \bar{x})\right]^2 \Big/ [Np(p+2)].$$

A sample correlation such as $f(s) = r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$ or a set of such correlations is a function of S that is invariant under scale transformations; that is, it satisfies (4).

Corollary 4.5.2. *Under the conditions of Theorem* 4.5.1,

$$(10) \qquad \sqrt{\frac{N}{1+\hat{\kappa}}} \frac{(r_{ij} - \rho_{ij})}{\sqrt{1 - r_{ij}^2}} \xrightarrow{d} N(0,1).$$

As in the case of the observations normally distributed,

$$(11) \qquad \sqrt{\frac{N}{1+\hat{\kappa}}} \left(\frac{1}{2}\log\frac{1+r_{ij}}{1-r_{ij}} - \frac{1}{2}\log\frac{1+\rho_{ij}}{1-\rho_{ij}}\right) \xrightarrow{d} N(0,1).$$

Of course, any improvement of (11) over (10) depends on the distribution samples.

Partial correlations such as $r_{ij.q+1,\dots,p}$, $i,j = 1,\dots,q$, are also invariant functions of S.

Corollary 4.5.3. *Under the conditions of Theorem* 4.5.1,

$$(12) \qquad \sqrt{\frac{N}{1+\hat{\kappa}}} \left(r_{ij.q+1,\dots,p} - \rho_{ij.q+1,\dots,p}\right) \xrightarrow{d} N(0,1).$$

Now let us consider the asymptotic distribution of R^2, the square of the multiple correlation, when $\bar{R}^2$, the square of the population multiple correlation, is 0. We use the notation of Section 4.4. $\bar{R}^2 = 0$ is equivalent to $\boldsymbol{\sigma}_{(1)} = \mathbf{0}$. Since the sample and population multiple correlation coefficients between X_1 and $X^{(2)} = (X_2, \ldots, X_p)'$ are invariant with respect to linear transformations (47) of Section 4.4, for purposes of studying the distribution of R^2 we can assume $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. In that case $s_{11} \xrightarrow{p} 1$, $\boldsymbol{s}_{(1)} \xrightarrow{p} \mathbf{0}$, and $\boldsymbol{S}_{22} \xrightarrow{p} \boldsymbol{I}_{p-1}$. Furthermore, for $k, i \neq 1$ and $j = l = 1$, Lemma 3.6.1 gives

$$\mathscr{E}\boldsymbol{s}_{(1)}\boldsymbol{s}'_{(1)} = \left(\frac{1}{n} + \frac{\kappa}{N}\right)\boldsymbol{I}_{p-1}. \tag{13}$$

Theorem 4.5.2. *Under the conditions of Theorem* 4.5.1

$$\sqrt{\frac{N}{1+\kappa}}\,\boldsymbol{s}_{(1)} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}_{p-1}). \tag{14}$$

Corollary 4.5.4. *Under the conditions of Theorem* 4.5.1

$$\frac{NR^2}{1+\hat{\kappa}} = \frac{N\boldsymbol{s}'_{(1)}\boldsymbol{S}_{22}^{-1}\boldsymbol{s}_{(1)}}{(1+\hat{\kappa})s_{11}} \xrightarrow{d} \chi^2_{p-1}. \tag{15}$$

4.5.2. Elliptically Contoured Matrix Distributions

Now let us turn to the model

$$|\boldsymbol{\Lambda}|^{-N/2} g\left[\operatorname{tr}(\boldsymbol{X} - \boldsymbol{\varepsilon}_N\boldsymbol{\nu}')\boldsymbol{\Lambda}^{-1}(\boldsymbol{X} - \boldsymbol{\varepsilon}_N\boldsymbol{\nu}')'\right] \tag{16}$$

based on the vector spherical model $g(\operatorname{tr} \boldsymbol{Y}'\boldsymbol{Y})$. The unbiased estimators of $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma} = (\mathscr{E}R^2/p)\boldsymbol{\Lambda}$ are $\bar{\boldsymbol{x}} = (1/N)\boldsymbol{X}'\boldsymbol{\varepsilon}_N$ and $\boldsymbol{S} = (1/n)\boldsymbol{A}$, where $\boldsymbol{A} = (\boldsymbol{X} - \boldsymbol{\varepsilon}_n\bar{\boldsymbol{x}}')'(\boldsymbol{X} - \boldsymbol{\varepsilon}_N\bar{\boldsymbol{x}}')$.

Since

$$(\boldsymbol{X} - \boldsymbol{\varepsilon}_N\boldsymbol{\nu}')'(\boldsymbol{X} - \boldsymbol{\varepsilon}_N\boldsymbol{\nu}') = \boldsymbol{A} + N(\bar{\boldsymbol{x}} - \boldsymbol{\nu})(\bar{\boldsymbol{x}} - \boldsymbol{\nu})', \tag{17}$$

$\boldsymbol{A}$ and $\bar{\boldsymbol{x}}$ are a complete set of sufficient statistics.

The maximum likelihood estimators of $\boldsymbol{\nu}$ and $\boldsymbol{\Lambda}$ are $\hat{\boldsymbol{\nu}} = \bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Lambda}} = (p/w_g)\boldsymbol{A}$. The maximum likelihood estimator of $\rho_{ij} = \lambda_{ij}/\sqrt{\lambda_{ii}\lambda_{jj}} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ is $\hat{\rho}_{ij} = a_{ij}/\sqrt{a_{ii}a_{jj}} = s_{ij}/\sqrt{s_{ii}s_{jj}}$ (Theorem 3.6.4).

The sample correlation r_{ij} is a function $f(\boldsymbol{X})$ that satisfies the conditions (45) and (46) of Theorem 3.6.5 and hence has the same distribution for an arbitrary density $g[\operatorname{tr}(\cdot)]$ as for the normal density $g[\operatorname{tr}(\cdot)] = \text{const } e^{-\frac{1}{2}\operatorname{tr}(\cdot)}$. Similarly, a partial correlation $r_{ij\cdot q+1,\ldots,p}$ and a multiple correlation R^2 satisfy the conditions, and the conclusion holds.

Theorem 4.5.3. *When X has the vector elliptical density* (16), *the distributions of r_{ij}, $r_{ij.q+1}$, and R^2 are the distributions derived for normally distributed observations.*

It follows from Theorem 4.5.3 that the asymptotic distributions of r_{ij}, $r_{ij.q+1,\dots,p}$, and R^2 are the same as for sampling from normal distributions.

The class of left spherical matrices Y with densities is the class of $g(Y'Y)$. Let $X = YC' + \varepsilon_N \nu'$, where $C'\Lambda^{-1}C = I$, that is, $\Lambda = CC'$. Then X has the density

$$(18) \qquad |C|^{-N} g\left[C^{-1}(X-\varepsilon_N\nu')'(X-\varepsilon_N\nu')(C')^{-1}\right].$$

We now find a stochastic representation of the matrix Y.

Lemma 4.5.1. *Let $V = (v_1,\dots,v_p)$, where v_i is an N-component vector, $i = 1,\dots,p$. Define recursively $w_1 = v_1$,*

$$(19) \qquad w_i = v_i - \sum_{j=1}^{i-1} \frac{v_i' w_j}{w_j' w_j} w_j, \qquad i = 2,\dots,p.$$

Let $u_i = w_i/\|w_i\|$. Then $\|u_i\| = 1$, $i = 1,\dots,p$, and $u_i'u_j = 0$, $i \neq j$. Further,

$$(20) \qquad V = UT',$$

where $U = (u_1,\dots,u_p)$; $t_{ii} = \|w_i\|$, $i = 1,\dots,p$; $t_{ij} = v_i'w_j/\|w_j\| = v_i'u_j$, $j = 1,\dots,i-1$, $i = 1,\dots,p$; and $t_{ij} = 0$, $i < j$.

The proof of the lemma is given in the first part of Section 7.2 and as the Gram–Schmidt orthogonalization in the Appendix (Section A.5.1). This lemma generalizes the construction in Section 3.2; see Figure 3.1. See also Figure 7.1.

Note that T is lower triangular, $U'U = I_p$, and $V'V = TT'$. The last equation, $t_{ii} \geq 0$, $i = 1,\dots,p$, and $t_{ij} = 0$, $i < j$, can be solved uniquely for T. Thus T is a function of $V'V$ (and the restrictions).

Let Y $(N \times p)$ have the density $g(Y'Y)$, and let O_N be an orthogonal $N \times N$ matrix. Then $Y^* = O_N Y$ has the density $g(Y^{*\prime}Y^*)$. Hence $Y^* = O_N Y \overset{d}{=} Y$. Let $Y^* = U^* T^{*\prime}$, where $t_{ii}^* > 0$, $i = 1,\dots,p$, and $t_{ij}^* = 0$, $i < j$. From $Y^{*\prime}Y^* = Y'Y$ it follows that $T^*T^{*\prime} = TT'$ and hence $T^* = T$, $Y^* = U^*T$, and $U^* = O_N U \overset{d}{=} U$. Let the space of U $(N \times p)$ such that $U'U = I_p$ be denoted $O(N \times p)$.

Definition 4.5.1. *If U $(N \times p)$ satisfies $U'U = I_p$ and $O_N U \overset{d}{=} U$ for all orthogonal O_N, then U is* uniformly distributed on $O(N \times p)$.

The space of U satisfying $U'U = I_p$ is known as a *Steifel manifold.* The probability measure of Definition 4.5.1 is known as the *Haar invariant distribution.* The property $O_N U \overset{d}{=} U$ for all orthogonal O_N defines the (normalized) measure uniquely [Halmos (1956)].

Theorem 4.5.4. *If* Y $(N \times p)$ *has the density* $g(Y'Y)$, *then* U *defined by* $Y = UT'$, $U'U = I_p$, $t_{ii} > 0$, $i = 1, \dots, p$, *and* $t_{ij} = 0$, $i < j$, *is* uniformly distributed on $O(N \times p)$.

The proof of Corollary 7.2.1 shows that for arbitrary $g(\cdot)$ the density of T is

$$\prod_{i=1}^{p} \{C[\tfrac{1}{2}(N+1-i)]t_{ii}^{N-i}\} g(\operatorname{tr} TT'), \tag{21}$$

where $C(\cdot)$ is defined in (8) of Section 2.7.

The *stochastic representation* of Y $(N \times p)$ with density $g(Y'Y)$ is

$$Y = UT', \tag{22}$$

where U $(N \times p)$ is uniformly distributed on $O(N \times p)$ and T is lower triangular with positive diagonal elements and has density (21).

Theorem 4.5.5. *Let* $f(X)$ *be a vector-valued function of* X $(N \times p)$ *such that*

$$f(X + \varepsilon_N \nu') = f(X) \tag{23}$$

for all ν *and*

$$f(XG') = f(X) \tag{24}$$

for all G $(p \times p)$. *Then the distribution of* $f(X)$ *where* X *has an arbitrary density* (18) *is the same as the distribution of* $f(X)$ *where* X *has the normal density* (18).

Proof. From (23) we find that $f(X) = f(YC')$, and from (24) we find $f(YC') = f(UT'C') = f(U)$, which is the same for arbitrary and normal densities (18). ∎

Corollary 4.5.5. *Let* $f(X)$ *be a vector-valued function of* X $(N \times p)$ *with the density* (18), *where* $\nu = \mathbf{0}$. *Suppose* (24) *holds for all* G $(p \times p)$. *Then the distribution of* $f(X)$ *for an arbitrary density* (18) *is the same as the distribution of* $f(X)$ *when* X *has the normal density* (18).

The condition (24) of Corollary 4.5.5 is that $f(X)$ is invariant with respect to linear transformations $X \to XG$.

The density (18) can be written as

$$|C|^{-1} g\{C^{-1}[A + N(\bar{x} - \nu)(\bar{x} - \nu)'](C')^{-1}\}, \tag{25}$$

which shows that A and $\bar{x}$ are a complete set of sufficient statistics for $\Lambda = CC'$ and ν.

PROBLEMS

4.1. (Sec. 4.2.1) Sketch

$$k_N(r) = \frac{\Gamma[\frac{1}{2}(N-1)]}{\Gamma(\frac{1}{2}N - 1)\sqrt{\pi}} (1 - r^2)^{\frac{1}{2}(N-4)}$$

for (a) $N = 3$, (b) $N = 4$, (c) $N = 5$, and (d) $N = 10$.

4.2. (Sec. 4.2.1) Using the data of Problem 3.1, test the hypothesis that X_1 and X_2 are independent against all alternatives of dependence at significance level 0.01.

4.3. (Sec. 4.2.1) Suppose a sample correlation of 0.65 is observed in a sample of 10. Test the hypothesis of independence against the alternatives of positive correlation at significance level 0.05.

4.4. (Sec. 4.2.2) Suppose a sample correlation of 0.65 is observed in a sample of 20. Test the hypothesis that the population correlation is 0.4 against the alternatives that the population correlation is greater than 0.4 at significance level 0.05.

4.5. (Sec. 4.2.1) Find the significance points for testing $\rho = 0$ at the 0.01 level with $N = 15$ observations against alternatives (a) $\rho \neq 0$, (b) $\rho > 0$, and (c) $\rho < 0$.

4.6. (Sec. 4.2.2) Find significance points for testing $\rho = 0.6$ at the 0.01 level with $N = 20$ observations against alternatives (a) $\rho \neq 0.6$, (b) $\rho > 0.6$, and (c) $\rho < 0.6$.

4.7. (Sec. 4.2.2) Tablulate the power function at $\rho = -1(0.2)1$ for the tests in Problem 4.5. Sketch the graph of each power function.

4.8. (Sec. 4.2.2) Tablulate the power function at $\rho = -1(0.2)1$ for the tests in Problem 4.6. Sketch the graph of each power function.

4.9. (Sec. 4.2.2) Using the data of Problem 3.1, find a (two-sided) confidence interval for ρ_{12} with confidence coefficient 0.99.

4.10. (Sec. 4.2.2) Suppose $N = 10$, $r = 0.795$. Find a one-sided confidence interval for ρ [of the form $(r_0, 1)$] with confidence coefficient 0.95.

4.11. (Sec. 4.2.3) Use Fisher's z to test the hypothesis $\rho = 0.7$ against alternatives $\rho \neq 0.7$ at the 0.05 level with $r = 0.5$ and $N = 50$.

4.12. (Sec. 4.2.3) Use Fisher's z to test the hypothesis $\rho_1 = \rho_2$ against the alternatives $\rho_1 \neq \rho_2$ at the 0.01 level with $r_1 = 0.5, N_1 = 40, r_2 = 0.6, N_2 = 40$.

4.13. (Sec. 4.2.3) Use Fisher's z to estimate ρ based on sample correlations of -0.7 $(N = 30)$ and of -0.6 $(N = 40)$.

4.14. (Sec. 4.2.3) Use Fisher's z to obtain a confidence interval for ρ with confidence 0.95 based on a sample correlation of 0.65 and a sample size of 25.

4.15. (Sec. 4.2.2). Prove that when $N = 2$ and $\rho = 0$, $\Pr\{r = 1\} = \Pr\{r = -1\} = \frac{1}{2}$.

4.16. (Sec. 4.2) Let $k_N(r, \rho)$ be the density of the sample correlation coefficient r for a given value of ρ and N. Prove that r has a monotone likelihood ratio; that is, show that if $\rho_1 > \rho_2$, then $k_N(r, \rho_1)/k_N(r, \rho_2)$ is monotonically increasing in r. [*Hint:* Using (40), prove that if

$$F\left[\tfrac{1}{2}, \tfrac{1}{2}; n + \tfrac{1}{2}; \tfrac{1}{2}(1 + \rho r)\right] = \sum_{\alpha=0}^{\infty} c_\alpha (1 + \rho r)^\alpha = g(r, \rho)$$

has a monotone ratio, then $k_N(r, \rho)$ does. Show

$$\frac{\partial^2}{\partial\rho\,\partial r} \log g(r, \rho) = \frac{\sum_{\alpha,\beta=0}^{\infty} c_\alpha c_\beta \left[(\alpha - \beta)^2 r\rho + (\alpha + \beta)\right](1 + r\rho)^{\alpha+\beta-2}}{2\left[\sum_{\alpha=0}^{\infty} c_\alpha (1 + r\rho)^\alpha\right]^2};$$

if $(\partial^2/\partial\rho\,\partial r)\log g(r, \rho) > 0$, then $g(r, \rho)$ has a monotone ratio. Show the numerator of the above expression is positive by showing that for each α the sum on β is positive; use the fact that $c_{\alpha+1} < \frac{1}{2}c_\alpha$.]

4.17. (Sec. 4.2) Show that of all tests of ρ_0 against a specific ρ_1 $(> \rho_0)$ based on r, the procedures for which $r > c$ implies rejection are the best. [*Hint:* This follows from Problem 4.16.]

4.18. (Sec. 4.2) Show that of all tests of $\rho = \rho_0$ against $\rho > \rho_0$ based on r, a procedure for which $r > c$ implies rejection is uniformly most powerful.

4.19. (Sec. 4.2) Prove r has a monotone likelihood ratio for $r > 0$, $\rho > 0$ by proving $h(r) = k_N(r, \rho_1)/k_N(r, \rho_2)$ is monotonically increasing for $\rho_1 > \rho_2$. Here $h(r)$ is a constant times $(\sum_{\alpha=0}^{\infty} c_\alpha \rho_1^\alpha r^\alpha)/(\sum_{\alpha=0}^{\infty} c_\alpha \rho_2^\alpha r^\alpha)$. In the numerator of $h'(r)$, show that the coefficient of r^β is positive.

4.20. (Sec. 4.2) Prove that if Σ is diagonal, then the sets r_{ij} and a_{ii} are independently distributed. [*Hint:* Use the facts that r_{ij} is invariant under scale transformations and that the density of the observations depends only on the a_{ii}.]

4.21. (Sec. 4.2.1) Prove that if $\rho = 0$

$$\mathscr{E} r^{2m} = \frac{\Gamma[\frac{1}{2}(N-1)]\Gamma(m+\frac{1}{2})}{\sqrt{\pi}\,\Gamma[\frac{1}{2}(N-1)+m]}.$$

4.22. (Sec. 4.2.2) Prove $f_1(\rho)$ and $f_2(\rho)$ are monotonically increasing functions of ρ.

4.23. (Sec. 4.2.2) Prove that the density of the sample correlation r [given by (38)] is

$$\frac{n-1}{\pi}(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}\int_0^1 \frac{x^{n-1}\,dx}{(1-\rho r x)^n\sqrt{1-x^2}}.$$

[*Hint:* Expand $(1-\rho r x)^{-n}$ in a power series, integrate, and use the duplication formula for the gamma function.]

4.24. (Sec. 4.2) Prove that (39) is the density of r. [*Hint:* From Problem 2.12 show

$$\int_0^\infty\int_0^\infty e^{-\frac{1}{2}(y^2-2xyz+z^2)}\,dy\,dz = \frac{\cos^{-1}(-x)}{\sqrt{1-x^2}}.$$

Then argue

$$\int_0^\infty\int_0^\infty (yz)^{n-1}\,e^{-\frac{1}{2}(y^2-2xyz+z^2)}\,dy\,dz = \frac{d^{n-1}}{dx^{n-1}}\,\frac{\cos^{-1}(-x)}{\sqrt{1-x^2}}.$$

Finally show that the integral of (31) with respect to a_{11} ($=y^2$) and a_{22} ($=z^2$) is (39).]

4.25. (Sec. 4.2) Prove that (40) is the density of r. [*Hint:* In (31) let $a_{11} = ue^{-v}$ and $a_{22} = ue^{v}$; show that the density of v ($0 \le v < \infty$) and r ($-1 \le r \le 1$) is

$$\frac{n-1}{\pi\sqrt{2}}(1-\rho^2)^{\frac{1}{2}n}(1-\rho r)^{-n+\frac{1}{2}}(1-r^2)^{\frac{1}{2}(n-3)}v^{-\frac{1}{2}}(1-v)^{n-1}\left[1-\tfrac{1}{2}(1+\rho r)v\right]^{-\frac{1}{2}}$$

Use the expansion

$$(1-y)^{-\frac{1}{2}} = \sum_{j=0}^{\infty}\frac{\Gamma(j+\frac{1}{2})}{\Gamma(\frac{1}{2})j!}y^j.$$

Show that the integral is (40).]

4.26. (Sec. 4.2) Prove for integer h

$$\mathscr{E} r^{2h+1} = \frac{(1-\rho^2)^{\frac{1}{2}n}}{\sqrt{\pi}\,\Gamma(\frac{1}{2}n)} \sum_{\beta=0}^{\infty} \frac{(2\rho)^{2\beta+1}}{(2\beta+1)!} \frac{\Gamma^2[\frac{1}{2}(n+1)+\beta]\Gamma(h+\beta+\frac{3}{2})}{\Gamma(\frac{1}{2}n+h+\beta+1)},$$

$$\mathscr{E} r^{2h} = \frac{(1-\rho^2)^{\frac{1}{2}n}}{\sqrt{\pi}\,\Gamma(\frac{1}{2}n)} \sum_{\beta=0}^{\infty} \frac{(2\rho)^{2\beta}}{(2\beta)!} \frac{\Gamma^2(\frac{1}{2}n+\beta)\Gamma(h+\beta+\frac{1}{2})}{\Gamma(\frac{1}{2}n+h+\beta)}.$$

4.27. (Sec. 4.2) *The t-distribution.* Prove that if X and Y are independently distributed, X having the distribution $N(0,1)$ and Y having the χ^2-distribution with m degrees of freedom, then $W = X/\sqrt{Y/m}$ has the density

$$\frac{\Gamma[\frac{1}{2}(m+1)]}{\sqrt{m}\,\sqrt{\pi}\,\Gamma(\frac{1}{2}m)}\left(1+\frac{t^2}{m}\right)^{-\frac{1}{2}(m+1)}.$$

[*Hint:* In the joint density of X and Y, let $x = tw^{\frac{1}{2}}m^{-\frac{1}{2}}$ and integrate out w.]

4.28. (Sec. 4.2) Prove

$$\mathscr{E} r = \frac{(1-\rho^2)^{\frac{1}{2}n}}{\Gamma(\frac{1}{2}n)} \sum_{\beta=0}^{\infty} \frac{\rho^{2\beta+1}\Gamma^2[\frac{1}{2}(n+1)+\beta]}{\beta!\,\Gamma[\frac{1}{2}n+\beta+1]}.$$

[*Hint:* Use Problem 4.26 and the duplication formula for the gamma function.]

4.29. (Sec. 4.2) Show that $\sqrt{n}(r_{ij}-\rho_{ij})$, $(i,j)=(1,2),(1,3),(2,3)$, have a joint limiting distribution with variances $(1-\rho_{ij}^2)^2$ and covariances of r_{ij} and r_{ik}, $j \neq k$ being $\frac{1}{2}(2\rho_{jk}-\rho_{ij}\rho_{jk})(1-\rho_{ij}^2-\rho_{ik}^2-\rho_{jk}^2)+\rho_{jk}^2$.

4.30. (Sec. 4.3.2) Find a confidence interval for $\rho_{13\cdot 2}$ with confidence 0.95 based on $r_{13\cdot 2} = 0.097$ and $N = 20$.

4.31. (Sec. 4.3.2) Use Fisher's z to test the hypothesis $\rho_{12\cdot 34} = 0$ against alternatives $\rho_{12\cdot 34} \neq 0$ at significance level 0.01 with $r_{12\cdot 34} = 0.14$ and $N = 40$.

4.32. (Sec. 4.3) Show that the inequality $r_{12\cdot 3}^2 \leq 1$ is the same as the inequality $|r_{ij}| \geq 0$, where $|r_{ij}|$ denotes the determinant of the 3×3 correlation matrix.

4.33. (Sec. 4.3) *Invariance of the sample partial correlation coefficient.* Prove that $r_{12\cdot 3,\dots,p}$ is invariant under the transformations $x_{i\alpha}^* = a_i x_{i\alpha} + b_i' x_\alpha^{(3)} + c_i$, $a_i > 0$, $i = 1,2$, $x_\alpha^{(3)*} = Cx_\alpha^{(3)} + b$, $\alpha = 1,\dots,N$, where $x_\alpha^{(3)} = (x_{3\alpha},\dots,x_{p\alpha})'$, and that any function of $\bar{x}$ and $\hat{\Sigma}$ that is invariant under these transformations is a function of $r_{12\cdot 3,\dots,p}$.

4.34. (Sec. 4.4) *Invariance of the sample multiple correlation coefficient.* Prove that R is a function of the sufficient statistics $\bar{x}$ and S that is invariant under changes of location and scale of $x_{1\alpha}$ and nonsingular linear transformations of $x_\alpha^{(2)}$ (that is, $x_{1\alpha}^* = cx_{1\alpha} + d$, $x_\alpha^{(2)*} = Cx_\alpha^{(2)} + d$, $\alpha = 1,\dots,N$) and that every function of $\bar{x}$ and S that is invariant is a function of R.

4.35. (Sec. 4.4) Prove that conditional on $Z_{1\alpha} = z_{1\alpha}$, $\alpha = 1, \dots, n$, $R^2/(1-R^2)$ is distributed like $T^2/(N^*-1)$, where $T^2 = N^* \bar{x}' S^{-1} \bar{x}$ based on $N^* = n$ observations on a vector X with $p^* = p-1$ components, with mean vector $(c/\sigma_{11})\sigma'_{(1)}$ ($nc^2 = \sum z_{1\alpha}^2$) and covariance matrix $\Sigma_{22\cdot 1} = \Sigma_{22} - (1/\sigma_{11})\sigma_{(1)}\sigma'_{(1)}$. [*Hint:* The conditional distribution of $Z_\alpha^{(2)}$ given $Z_{1\alpha} = z_{1\alpha}$ is $N[(1/\sigma_{11})\sigma_{(1)} z_{1\alpha}, \Sigma_{22\cdot 1}]$. There is an $n \times n$ orthogonal matrix B which carries $(z_{11}, \dots, z_{1n})$ into $(c, \dots, c)$ and $(Z_{i1}, \dots, Z_{in})$ into $(Y_{i1}, \dots, Y_{in})$, $i = 2, \dots, p$. Let the new X'_α be $(Y_{2\alpha}, \dots, Y_{p\alpha})$.]

4.36. (Sec. 4.4) Prove that the noncentrality parameter in the distribution in Problem 4.35 is $(a_{11}/\sigma_{11})\bar{R}^2/(1-\bar{R}^2)$.

4.37. (Sec. 4.4) Find the distribution of $R^2/(1-R^2)$ by multiplying the density of Problem 4.35 by the density of a_{11} and integrating with respect to a_{11}.

4.38. (Sec. 4.4) Show that the density of r^2 derived from (38) of Section 4.2 is identical with (42) in Section 4.4 for $p = 2$. [*Hint:* Use the duplication formula for the gamma function.]

4.39. (Sec. 4.4) Prove that (30) is the uniformly most powerful test of $\bar{R} = 0$ based on r. [*Hint:* Use the Neyman–Pearson fundamental lemma.]

4.40. (Sec. 4.4) Prove that (47) is the unique unbiased estimator of $\bar{R}^2$ based on R^2.

4.41. The estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in Problem 3.1 are

$$\bar{x} = (\,185.72 \quad 151.12 \qquad 183.84 \quad 149.24\,)',$$

$$S = \left(\begin{array}{cc:cc} 95.2933 & 52.8683 & 69.6617 & 46.1117 \\ 52.8683 & 54.3600 & 51.3117 & 35.0533 \\ \hdashline 69.6617 & 51.3117 & 100.8067 & 56.5400 \\ 46.1117 & 35.0533 & 56.5400 & 45.0233 \end{array}\right).$$

(a) Find the estimates of the parameters of the conditional distribution of (x_3, x_4) given (x_1, x_2); that is, find $S_{21}S_{11}^{-1}$ and $S_{22\cdot 1} = S_{22} - S_{21}S_{11}^{-1}S_{12}$.
(b) Find the partial correlation $r_{34\cdot 12}$.
(c) Use Fisher's z to find a confidence interval for $\rho_{34\cdot 12}$ with confidence 0.95.
(d) Find the sample multiple correlation coefficients between x_3 and (x_1, x_2) and between x_4 and (x_1, x_2).
(e) Test the hypotheses that x_3 is independent of (x_1, x_2) and x_4 is independent of (x_1, x_2) at significance levels 0.05.

4.42. Let the components of X correspond to scores on tests in arithmetic speed (X_1), arithmetic power (X_2), memory for words (X_3), memory for meaningful symbols (X_4), and memory for meaningless symbols (X_5). The observed correlations in a sample of 140 are [Kelley (1928)]

$$\begin{pmatrix} 1.0000 & 0.4248 & 0.0420 & 0.0215 & 0.0573 \\ 0.4248 & 1.0000 & 0.1487 & 0.2489 & 0.2843 \\ 0.0420 & 0.1487 & 1.0000 & 0.6693 & 0.4662 \\ 0.0215 & 0.2489 & 0.6693 & 1.0000 & 0.6915 \\ 0.0573 & 0.2843 & 0.4662 & 0.6915 & 1.0000 \end{pmatrix}.$$

(a) Find the partial correlation between X_4 and X_5, holding X_3 fixed.

(b) Find the partial correlation between X_1 and X_2, holding X_3, X_4, and X_5 fixed.

(c) Find the multiple correlation between X_1 and the set X_3, X_4, and X_5.

(d) Test the hypothesis at the 1% significance level that arithmetic speed is independent of the three memory scores.

4.43. (Sec. 4.3) Prove that if $\rho_{ij\cdot q+1,\dots,p} = 0$, then $\sqrt{N-2-(p-q)}\, r_{ij\cdot q+1,\dots,p}/\sqrt{1-r^2_{ij\cdot q+1,\dots,p}}$ is distributed according to the t-distribution with $N-2-(p-q)$ degrees of freedom.

4.44. (Sec. 4.3) Let $X' = (X_1, X_2, X^{(2)\prime})$ have the distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The conditional distribution of X_1 given $X_2 = x_2$ and $X^{(2)} = x^{(2)}$ is .

$$N\left[\mu_1 + \gamma_2(x_2 - \mu_2) + \boldsymbol{\gamma}'(\boldsymbol{x}^{(2)} - \boldsymbol{\mu}^{(2)}),\, \sigma_{11\cdot 2,\dots,p}\right],$$

where

$$\begin{pmatrix} \sigma_{22} & \boldsymbol{\sigma}'_{(2)} \\ \boldsymbol{\sigma}_{(2)} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \gamma_2 \\ \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \sigma_{12} \\ \boldsymbol{\sigma}_{(1)} \end{pmatrix}.$$

The estimators of γ_2 and $\boldsymbol{\gamma}$ are defined by

$$\begin{pmatrix} a_{22} & \boldsymbol{a}'_{(2)} \\ \boldsymbol{a}_{(2)} & A_{22} \end{pmatrix} \begin{pmatrix} c_2 \\ \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} a_{12} \\ \boldsymbol{a}_{(1)} \end{pmatrix}.$$

Show $c_2 = a_{12\cdot 3,\dots,p}/a_{22\cdot 3,\dots,p}$. [*Hint:* Solve for $\boldsymbol{c}$ in terms of c_2 and the a's, and substitute.]

4.45. (Sec. 4.3) In the notation of Problem 4.44, prove

$$\begin{aligned} a_{11\cdot 2,\dots,p} &= a_{11} - \boldsymbol{a}'_{(1)} A_{22}^{-1} \boldsymbol{a}_{(1)} - c_2^2\left(a_{22} - \boldsymbol{a}'_{(2)} A_{22}^{-1} \boldsymbol{a}_{(2)}\right) \\ &= a_{11\cdot 3,\dots,p} - c_2^2 a_{22\cdot 3,\dots,p}. \end{aligned}$$

Hint: Use

$$a_{11\cdot 2,\ldots,p} = a_{11} - (c_2 \quad c')\begin{pmatrix} a_{22} & a'_{(2)} \\ a_{(2)} & A_{22} \end{pmatrix}\begin{pmatrix} c_2 \\ c \end{pmatrix}.$$

4.46. (Sec. 4.3) Prove that $1/a_{22\cdot 3,\ldots,p}$ is the element in the upper left-hand corner of

$$\begin{pmatrix} a_{22} & a'_{(2)} \\ a_{(2)} & A_{22} \end{pmatrix}^{-1}.$$

4.47. (Sec. 4.3) Using the results in Problems 4.43–4.46, prove that the test for $\rho_{12\cdot 3,\ldots,p} = 0$ is equivalent to the usual t-test for $\gamma_2 = 0$.

4.48. *Missing observations.* Let $X = (Y' \quad Z')'$, where Y has p components and Z has q components, be distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yz} \\ \boldsymbol{\Sigma}_{zy} & \boldsymbol{\Sigma}_{zz} \end{pmatrix}.$$

Let M observations be made on X, and $N - M$ additional observations be made on Y. Find the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. [Anderson (1957).] [*Hint:* Express the likelihood function in terms of the marginal density of Y and the conditional density of Z given Y.]

4.49. Suppose X is distributed according to $N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}.$$

Show that on the basis of one observation, $x' = (x_1, x_2, x_3)$, we can obtain a confidence interval for ρ (with confidence coefficient $1 - \alpha$) by using as endpoints of the interval the solutions in t of

$$\left[x_2^2 + \chi_3^2(\alpha)\right]t^2 - 2(x_1x_2 + x_2x_3)t + x_1^2 + x_2^2 + x_3^2 - \chi_3^2(\alpha) = 0,$$

where $\chi_3^2(\alpha)$ is the significance point of the χ^2-distribution with three degrees of freedom at significance level α.

CHAPTER 5

The Generalized T^2-Statistic

5.1. INTRODUCTION

One of the most important groups of problems in univariate statistics relates to the mean of a given distribution when the variance of the distribution is unknown. On the basis of a sample one may wish to decide whether the mean is equal to a number specified in advance, or one may wish to give an interval within which the mean lies. The statistic usually used in univariate statistics is the difference between the mean of the sample $\bar{x}$ and the hypothetical population mean μ divided by the sample standard deviation s. If the distribution sampled is $N(\mu, \sigma^2)$, then

$$t = \sqrt{N}\,\frac{\bar{x} - \mu}{s} \tag{1}$$

has the well-known t-distribution with $N - 1$ degrees of freedom, where N is the number of observations in the sample. On the basis of this fact, one can set up a test of the hypothesis $\mu = \mu_0$, where μ_0 is specified, or one can set up a confidence interval for the unknown parameter μ.

The multivariate analog of the square of t given in (1) is

$$T^2 = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' S^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}), \tag{2}$$

where $\bar{\boldsymbol{x}}$ is the mean vector of a sample of N, and S is the sample covariance matrix. It will be shown how this statistic can be used for testing hypotheses about the mean vector $\boldsymbol{\mu}$ of the population and for obtaining confidence regions for the unknown $\boldsymbol{\mu}$. The distribution of T^2 will be obtained when $\boldsymbol{\mu}$ in (2) is the mean of the distribution sampled and when $\boldsymbol{\mu}$ is different from

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

the population mean. Hotelling (1931) proposed the T^2-statistic for two samples and derived the distribution when $\boldsymbol{\mu}$ is the population mean.

In Section 5.3 various uses of the T^2-statistic are presented, including simultaneous confidence intervals for all linear combinations of the mean vector. A James–Stein estimator is given when $\boldsymbol{\Sigma}$ is unknown. The power function of the T^2-test is treated in Section 5.4, and the multivariate Behrens–Fisher problem in Section 5.5. In Section 5.6, optimum properties of the T^2-test are considered, with regard to both invariance and admissibility. Stein's criterion for admissibility in the general exponential family is proved and applied. The last section is devoted to inference about the mean in elliptically contoured distributions.

5.2. DERIVATION OF THE GENERALIZED T^2-STATISTIC AND ITS DISTRIBUTION

5.2.1. Derivation of the T^2-Statistic As a Function of the Likelihood Ratio Criterion

Although the T^2-statistic has many uses, we shall begin our discussion by showing that the likelihood ratio test of the hypothesis $H: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ on the basis of a sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is based on the T^2-statistic given in (2) of Section 5.1. Suppose we have N observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ $(N > p)$. The likelihood function is

$$(1) \quad L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{1}{2}pN} |\boldsymbol{\Sigma}|^{-\frac{1}{2}N} \exp\left[-\frac{1}{2} \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}) \right].$$

The observations are given; L is a function of the indeterminates $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. (We shall not distinguish in notation between the indeterminates and the parameters.) The likelihood ratio criterion is

$$(2) \qquad \lambda = \frac{\max_{\boldsymbol{\Sigma}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})},$$

that is, the numerator is the maximum of the likelihood function for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ in the parameter space restricted by the null hypothesis ($\boldsymbol{\mu} = \boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}$ positive definite), and the denominator is the maximum over the entire parameter space ($\boldsymbol{\Sigma}$ positive definite). When the parameters are unrestricted, the maximum occurs when $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ are defined by the maximum likelihood estimators

(Section 3.2) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$,

$$\hat{\boldsymbol{\mu}}_\Omega = \bar{\boldsymbol{x}}, \tag{3}$$

$$\hat{\boldsymbol{\Sigma}}_\Omega = \frac{1}{N} \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'. \tag{4}$$

When $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, the likelihood function is maximized at

$$\hat{\boldsymbol{\Sigma}}_\omega = \frac{1}{N} \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \boldsymbol{\mu}_0)(\boldsymbol{x}_\alpha - \boldsymbol{\mu}_0)' \tag{5}$$

by Lemma 3.2.2. Furthermore, by Lemma 3.2.2

$$\max_{\boldsymbol{\Sigma},\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}pN} |\hat{\boldsymbol{\Sigma}}_\Omega|^{\frac{1}{2}N}} e^{-\frac{1}{2}pN}, \tag{6}$$

$$\max_{\boldsymbol{\Sigma}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}pN} |\hat{\boldsymbol{\Sigma}}_\omega|^{\frac{1}{2}N}} e^{-\frac{1}{2}pN}. \tag{7}$$

Thus the likelihood ratio criterion is

$$\begin{aligned} \lambda &= \frac{|\hat{\boldsymbol{\Sigma}}_\Omega|^{\frac{1}{2}N}}{|\hat{\boldsymbol{\Sigma}}_\omega|^{\frac{1}{2}N}} = \frac{|\sum(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'|^{\frac{1}{2}N}}{|\sum(\boldsymbol{x}_\alpha - \boldsymbol{\mu}_0)(\boldsymbol{x}_\alpha - \boldsymbol{\mu}_0)'|^{\frac{1}{2}N}} \\ &= \frac{|A|^{\frac{1}{2}N}}{|A + N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'|^{\frac{1}{2}N}}, \end{aligned} \tag{8}$$

where

$$A = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})' = (N-1)S. \tag{9}$$

Application of Corollary A.3.1 of the Appendix shows

$$\begin{aligned} \lambda^{2/N} &= \frac{|A|}{\left|A + \left[\sqrt{N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right]\left[\sqrt{N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right]'\right|} \\ &= \frac{1}{1 + N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' A^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)} \\ &= \frac{1}{1 + T^2/(N-1)}, \end{aligned} \tag{10}$$

where

$$T^2 = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' S^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) = (N-1) N (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' A^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0). \tag{11}$$

The likelihood ratio test is defined by the critical region (region of rejection)

$$\lambda \le \lambda_0, \tag{12}$$

where λ_0 is chosen so that the probability of (12) when the null hypothesis is true is equal to the significance level. If we take the $\frac{1}{2}N$th root of both sides of (12) and invert, subtract 1, and multiply by $N-1$, we obtain

$$T^2 \ge T_0^2, \tag{13}$$

where

$$T_0^2 = (N-1)\left(\lambda_0^{-2/N} - 1\right). \tag{14}$$

Theorem 5.2.1. *The likelihood ratio test of the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ for the distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by* (13), *where T^2 is defined by* (11), *$\bar{\boldsymbol{x}}$ is the mean of a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{S}$ is the covariance matrix of the sample, and T_0^2 is chosen so that the probability of* (13) *under the null hypothesis is equal to the chosen significance level.*

The Student t-test has the property that when testing $\mu = 0$ it is invariant with respect to scale transformations. If the scalar random variable X is distributed according to $N(\mu, \sigma^2)$, then $X^* = cX$ is distributed according to $N(c\mu, c^2\sigma^2)$, which is in the same class of distributions, and the hypothesis $\mathscr{E}X = 0$ is equivalent to $\mathscr{E}X^* = \mathscr{E}cX = 0$. If the observations x_α are transformed similarly ($x_\alpha^* = cx_\alpha$), then, for $c > 0$, t^* computed from x_α^* is the same as t computed from x_α. Thus, whatever the unit of measurement the statistical result is the same.

The generalized T^2-test has a similar property. If the vector random variable $\boldsymbol{X}$ is distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}^* = \boldsymbol{C}\boldsymbol{X}$ (for $|\boldsymbol{C}| \ne 0$) is distributed according to $N(\boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')$, which is in the same class of distributions. The hypothesis $\mathscr{E}\boldsymbol{X} = \boldsymbol{0}$ is equivalent to the hypothesis $\mathscr{E}\boldsymbol{X}^* = \mathscr{E}\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}$. If the observations $\boldsymbol{x}_\alpha$ are transformed in the same way, $\boldsymbol{x}_\alpha^* = \boldsymbol{C}\boldsymbol{x}_\alpha$, then T^{*2} computed on the basis of $\boldsymbol{x}_\alpha^*$ is the same as T^2 computed on the basis of $\boldsymbol{x}_\alpha$. This follows from the facts that $\bar{\boldsymbol{x}}^* = \boldsymbol{C}\bar{\boldsymbol{x}}$ and $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}'$ and the following lemma:

Lemma 5.2.1. *For any $p \times p$ nonsingular matrices $\boldsymbol{C}$ and $\boldsymbol{H}$ and any vector $\boldsymbol{k}$,*

$$\boldsymbol{k}'\boldsymbol{H}^{-1}\boldsymbol{k} = (\boldsymbol{C}\boldsymbol{k})'(\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}')^{-1}(\boldsymbol{C}\boldsymbol{k}). \tag{15}$$

Proof. The right-hand side of (15) is

$$(16)\qquad (Ck)'(CHC')^{-1}(Ck) = k'C'(C')^{-1}H^{-1}C^{-1}Ck = k'H^{-1}k. \qquad \blacksquare$$

We shall show in Section 5.6 that of all tests invariant with respect to such transformations, (13) is the uniformly most powerful.

We can give a geometric interpretation of the $\frac{1}{2}N$th root of the likelihood ratio criterion,

$$(17)\qquad \lambda^{2/N} = \frac{|\sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'|}{|\sum_{\alpha=1}^{N}(x_\alpha - \mu_0)(x_\alpha - \mu_0)'|},$$

in terms of parallelotopes. (See Section 7.5.) In the p-dimensional representation the numerator of $\lambda^{2/N}$ is the sum of squares of volumes of all parallelotopes with principal edges p vectors, each with one endpoint at $\bar{x}$ and the other at an x_α. The denominator is the sum of squares of volumes of all parallelotopes with principal edges p vectors, each with one endpoint at μ_0 and the other at x_α. If the sum of squared volumes involving vectors emanating from $\bar{x}$, the "center" of the x_α, is much less than that involving vectors emanating from μ_0, then we reject the hypothesis that μ_0 is the mean of the distribution.

There is also an interpretation in the N-dimensional representation. Let $y_i = (x_{i1}, \dots, x_{iN})'$ be the ith vector. Then

$$(18)\qquad \sqrt{N}\bar{x}_i = \sum_{\alpha=1}^{N} \frac{1}{\sqrt{N}} x_{i\alpha}$$

is the distance from the origin of the projection of y_i on the equiangular line (with direction cosines $1/\sqrt{N}, \dots, 1/\sqrt{N}$). The coordinates of the projection are $(\bar{x}_i, \dots, \bar{x}_i)$. Then $(x_{i1} - \bar{x}_i, \dots, x_{iN} - \bar{x}_i)$ is the projection of y_i on the plane through the origin perpendicular to the equiangular line. The numerator of $\lambda^{2/N}$ is the square of the p-dimensional volume of the parallelotope with principal edges, the vectors $(x_{i1} - \bar{x}_i, \dots, x_{iN} - \bar{x}_i)$. A point $(x_{i1} - \mu_{0i}, \dots, x_{iN} - \mu_{0i})$ is obtained from y_i by translation parallel to the equiangular line (by a distance $\sqrt{N}\mu_{0i}$). The denominator of $\lambda^{2/N}$ is the square of the volume of the parallelotope with principal edges these vectors. Then $\lambda^{2/N}$ is the ratio of these squared volumes.

5.2.2. The Distribution of T^2

In this subsection we will find the distribution of T^2 under general conditions, including the case when the null hypothesis is not true. Let $T^2 = Y'S^{-1}Y$ where Y is distributed according to $N(\nu, \Sigma)$ and nS is distributed independently as $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ with $Z_1, \dots, Z_n$ independent, each with distribution

$N(\mathbf{0}, \mathbf{\Sigma})$. The T^2 defined in Section 5.2.1 is a special case of this with $\boldsymbol{Y} = \sqrt{N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)$ and $\boldsymbol{\nu} = \sqrt{N}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ and $n = N - 1$. Let $\boldsymbol{D}$ be a nonsingular matrix such that $\boldsymbol{D\Sigma D}' = \boldsymbol{I}$, and define

$$(19) \qquad \boldsymbol{Y}^* = \boldsymbol{DY}, \qquad \boldsymbol{S}^* = \boldsymbol{DSD}', \qquad \boldsymbol{\nu}^* = \boldsymbol{D\nu}.$$

Then $T^2 = \boldsymbol{Y}^{*\prime}\boldsymbol{S}^{*-1}\boldsymbol{Y}^*$ (by Lemma 5.2.1), where $\boldsymbol{Y}^*$ is distributed according to $N(\boldsymbol{\nu}^*, \boldsymbol{I})$ and $n\boldsymbol{S}^*$ is distributed independently as $\sum_{\alpha=1}^n \boldsymbol{Z}_\alpha^* \boldsymbol{Z}_\alpha^{*\prime} = \sum_{\alpha=1}^n \boldsymbol{DZ}_\alpha(\boldsymbol{DZ}_\alpha)'$ with the $\boldsymbol{Z}_\alpha^* = \boldsymbol{DZ}_\alpha$ independent, each with distribution $N(\mathbf{0}, \boldsymbol{I})$. We note $\boldsymbol{\nu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu} = \boldsymbol{\nu}^{*\prime}(\boldsymbol{I})^{-1}\boldsymbol{\nu}^* = \boldsymbol{\nu}^{*\prime}\boldsymbol{\nu}^*$ by Lemma 5.2.1.

Let the first row of a $p \times p$ orthogonal matrix $\boldsymbol{Q}$ be defined by

$$(20) \qquad q_{1i} = \frac{Y_i^*}{\sqrt{\boldsymbol{Y}^{*\prime}\boldsymbol{Y}^*}}, \qquad i = 1, \ldots, p;$$

this is permissible because $\sum_{i=1}^p q_{1i}^2 = 1$. The other $p - 1$ rows can be defined by some arbitrary rule (Lemma A.4.2 of the Appendix). Since $\boldsymbol{Q}$ depends on $\boldsymbol{Y}^*$, it is a random matrix. Now let

$$(21) \qquad \begin{aligned} \boldsymbol{U} &= \boldsymbol{QY}^*, \\ \boldsymbol{B} &= \boldsymbol{Q}n\boldsymbol{S}^*\boldsymbol{Q}'. \end{aligned}$$

From the way $\boldsymbol{Q}$ was defined,

$$(22) \qquad \begin{aligned} U_1 &= \sum q_{1i} Y_i^* = \sqrt{\boldsymbol{Y}^{*\prime}\boldsymbol{Y}^*}, \\ U_j &= \sum q_{ji} Y_i^* = \sqrt{\boldsymbol{Y}^{*\prime}\boldsymbol{Y}^*} \sum q_{ji} q_{1i} = 0, \qquad j \neq 1. \end{aligned}$$

Then

$$(23) \qquad \frac{T^2}{n} = \boldsymbol{U}'\boldsymbol{B}^{-1}\boldsymbol{U} = (U_1, 0, \ldots, 0) \begin{pmatrix} b^{11} & b^{12} & \cdots & b^{1p} \\ b^{21} & b^{22} & \cdots & b^{2p} \\ \vdots & \vdots & & \vdots \\ b^{p1} & b^{p2} & \cdots & b^{pp} \end{pmatrix} \begin{pmatrix} U_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= U_1^2 b^{11},$$

where $(b^{ij}) = \boldsymbol{B}^{-1}$. By Theorem A.3.3 of the Appendix, $1/b^{11} = b_{11} - \boldsymbol{b}_{(1)}'\boldsymbol{B}_{22}^{-1}\boldsymbol{b}_{(1)} = b_{11\cdot 2, \ldots, p}$, where

$$(24) \qquad \boldsymbol{B} = \begin{pmatrix} b_{11} & \boldsymbol{b}_{(1)}' \\ \boldsymbol{b}_{(1)} & \boldsymbol{B}_{22} \end{pmatrix},$$

and $T^2/n = U_1^2/b_{11\cdot 2, \ldots, p} = \boldsymbol{Y}^{*\prime}\boldsymbol{Y}^*/b_{11\cdot 2, \ldots, p}$. The conditional distribution of $\boldsymbol{B}$ given $\boldsymbol{Q}$ is that of $\sum_{\alpha=1}^n \boldsymbol{V}_\alpha \boldsymbol{V}_\alpha'$, where conditionally the $\boldsymbol{V}_\alpha = \boldsymbol{QZ}_\alpha^*$ are independent, each with distribution $N(\mathbf{0}, I)$. By Theorem 4.3.3 $b_{11\cdot 2,\dots,p}$ is conditionally distributed as $\sum_{\alpha=1}^{n-(p-1)} W_\alpha^2$, where conditionally the W_α are independent, each with the distribution $N(0,1)$; that is, $b_{11\cdot 2,\dots,p}$ is conditionally distributed as χ^2 with $n-(p-1)$ degrees of freedom. Since the conditional distribution of $b_{11\cdot 2,\dots,p}$ does not depend on $\boldsymbol{Q}$, it is unconditionally distributed as χ^2. The quantity $\boldsymbol{Y}^{*\prime}\boldsymbol{Y}^*$ has a noncentral χ^2-distribution with p degrees of freedom and noncentrality parameter $\boldsymbol{\nu}^{*\prime}\boldsymbol{\nu}^* = \boldsymbol{\nu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}$. Then T^2/n is distributed as the ratio of a noncentral χ^2 and an independent χ^2.

Theorem 5.2.2. *Let $T^2 = \boldsymbol{Y}'\boldsymbol{S}^{-1}\boldsymbol{Y}$, where $\boldsymbol{Y}$ is distributed according to $N(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ and $n\boldsymbol{S}$ is independently distributed as $\sum_{\alpha=1}^{n} \boldsymbol{Z}_\alpha \boldsymbol{Z}_\alpha'$ with $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ independent, each with distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. Then $(T^2/n)[n-p+1)/p]$ is distributed as a noncentral F with p and $n-p+1$ degrees of freedom and noncentrality parameter $\boldsymbol{\nu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}$. If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is central F.*

We shall call this the T^2-distribution with n degrees of freedom.

Corollary 5.2.1. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ be a sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $T^2 = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)$. The distribution of $[T^2/(N-1)][(N-p)/p]$ is noncentral F with p and $N-p$ degrees of freedom and noncentrality parameter $N(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$. If $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, then the F-distribution is central.*

The above derivation of the T^2-distribution is due to Bowker (1960). The noncentral F-density and tables of the distribution are discussed in Section 5.4.

For large samples the distribution of T^2 given by Corollary 5.2.1 is approximately valid even if the parent distribution is not normal; in this sense the T^2-test is a robust procedure.

Theorem 5.2.3. *Let $\{\boldsymbol{X}_\alpha\}$, $\alpha = 1, 2, \dots$, be a sequence of independently identically distributed random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; let $\bar{\boldsymbol{X}}_N = (1/N)\sum_{\alpha=1}^{N} \boldsymbol{X}_\alpha$, $\boldsymbol{S}_N = [1/(N-1)]\sum_{\alpha=1}^{N}(\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}}_N)(\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}}_N)'$, and $T_N^2 = N(\bar{\boldsymbol{X}}_N - \boldsymbol{\mu}_0)'\boldsymbol{S}_N^{-1}(\bar{\boldsymbol{X}}_N - \boldsymbol{\mu}_0)$. Then the limiting distribution of T_N^2 as $N \to \infty$ is the χ^2-distribution with p degrees of freedom if $\boldsymbol{\mu} = \boldsymbol{\mu}_0$.*

Proof. By the central limit theorem (Theorem 4.2.3) the limiting distribution of $\sqrt{N}(\bar{\boldsymbol{X}}_N - \boldsymbol{\mu})$ is $N(\mathbf{0}, \boldsymbol{\Sigma})$. The sample covariance matrix converges stochastically to $\boldsymbol{\Sigma}$. Then the limiting distribution of T_N^2 is the distribution of $\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$, where $\boldsymbol{Y}$ has the distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. The theorem follows from Theorem 3.3.3. ■

When the null hypothesis is true, T^2/n is distributed as χ_p^2/χ_{n-p+1}^2, and $\lambda^{2/N}$ given by (10) has the distribution of $\chi_{n-p+1}^2/(\chi_{n-p+1}^2 + \chi_p^2)$. The density of $V = \chi_a^2/(\chi_a^2 + \chi_b^2)$, when χ_a^2 and χ_b^2 are independent, is

$$(25) \qquad \frac{\Gamma[\frac{1}{2}(a+b)]}{\Gamma(\frac{1}{2}a)\Gamma(\frac{1}{2}b)} v^{\frac{1}{2}a-1}(1-v)^{\frac{1}{2}b-1} = \beta(v; \tfrac{1}{2}a, \tfrac{1}{2}b);$$

this is the density of the *beta distribution* with parameters $\frac{1}{2}a$ and $\frac{1}{2}b$ (Problem 5.27). Thus the distribution of $\lambda^{2/N} = (1 + T^2/n)^{-1}$ is the beta distribution with parameters $\frac{1}{2}p$ and $\frac{1}{2}(n-p+1)$.

5.3. USES OF THE T^2-STATISTIC

5.3.1. Testing the Hypothesis That the Mean Vector Is a Given Vector

The likelihood ratio test of the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ on the basis of a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is equivalent to

$$(1) \qquad T^2 \geq T_0^2$$

as given in Section 5.2.1. If the significance level is α, then the $100\alpha\%$ point of the F-distribution is taken, that is,

$$(2) \qquad T_0^2 = \frac{(N-1)p}{N-p} F_{p,N-p}(\alpha) = T_{p,N-1}^2(\alpha),$$

say. The choice of significance level may depend on the power of the test. We shall discuss this in Section 5.4.

The statistic T^2 is computed from $\bar{x}$ and A. The vector $A^{-1}(\bar{x} - \boldsymbol{\mu}_0) = \boldsymbol{b}$ is the solution of $A\boldsymbol{b} = \bar{\boldsymbol{x}} - \boldsymbol{\mu}_0$. Then $T^2/(N-1) = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{b}$.

Note that $T^2/(N-1)$ is the nonzero root of

$$(3) \qquad |N(\bar{x} - \boldsymbol{\mu}_0)(\bar{x} - \boldsymbol{\mu}_0)' - \lambda A| = 0.$$

Lemma 5.3.1. *If $\boldsymbol{v}$ is a vector of p components and if $\boldsymbol{B}$ is a nonsingular $p \times p$ matrix, then $\boldsymbol{v}'\boldsymbol{B}^{-1}\boldsymbol{v}$ is the nonzero root of*

$$(4) \qquad |\boldsymbol{v}\boldsymbol{v}' - \lambda \boldsymbol{B}| = 0.$$

Proof. The nonzero root, say λ_1, of (4) is associated with a characteristic vector $\boldsymbol{\beta}$ satisfying

$$(5) \qquad \boldsymbol{v}\boldsymbol{v}'\boldsymbol{\beta} = \lambda_1 \boldsymbol{B}\boldsymbol{\beta}.$$

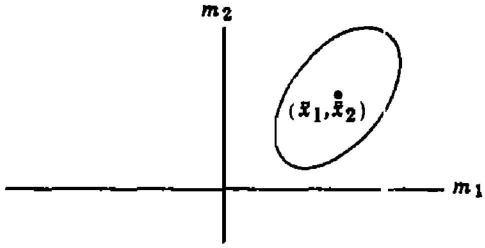

Figure 5.1. A confidence ellipse.

Since $\lambda_1 \neq 0$, $\boldsymbol{\nu}'\boldsymbol{\beta} \neq \mathbf{0}$. Multiplying on the left by $\boldsymbol{\nu}'B^{-1}$, we obtain

$$(\boldsymbol{\nu}'B^{-1}\boldsymbol{\nu})(\boldsymbol{\nu}'\boldsymbol{\beta}) = \lambda_1(\boldsymbol{\nu}'\boldsymbol{\beta}). \qquad \blacksquare \tag{6}$$

In the case above $\boldsymbol{\nu} = \sqrt{N}(\bar{x} - \boldsymbol{\mu}_0)$ and $B = A$.

5.3.2. A Confidence Region for the Mean Vector

If $\boldsymbol{\mu}$ is the mean of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the probability is $1-\alpha$ of drawing a sample of N with mean $\bar{x}$ and covariance matrix S such that

$$N(\bar{x} - \boldsymbol{\mu})'S^{-1}(\bar{x} - \boldsymbol{\mu}) \leq T^2_{p,N-1}(\alpha). \tag{7}$$

Thus, if we compute (7) for a particular sample, we have confidence $1-\alpha$ that (7) is a true statement concerning $\boldsymbol{\mu}$. The inequality

$$N(\bar{x} - \boldsymbol{m})'S^{-1}(\bar{x} - \boldsymbol{m}) \leq T^2_{p,N-1}(\alpha) \tag{8}$$

is the interior and boundary of an ellipsoid in the p-dimensional space of $\boldsymbol{m}$ with center at $\bar{x}$ and with size and shape depending on S^{-1} and α. See Figure 5.1. We state that $\boldsymbol{\mu}$ lies within this ellipsoid with confidence $1-\alpha$. Over random samples (8) is a random ellipsoid.

5.3.3. Simultaneous Confidence Intervals for All Linear Combinations of the Mean Vector

From the confidence region (8) for $\boldsymbol{\mu}$ we can obtain confidence intervals for linear functions $\boldsymbol{\gamma}'\boldsymbol{\mu}$ that hold simultaneously with a given confidence coefficient.

Lemma 5.3.2 (Generalized Cauchy–Schwarz Inequality). *For a positive definite matrix S,*

$$(\boldsymbol{\gamma}'\boldsymbol{y})^2 \leq \boldsymbol{\gamma}'S\boldsymbol{\gamma}\,\boldsymbol{y}'S^{-1}\boldsymbol{y}. \tag{9}$$

Proof. Let $b = \gamma' y / \gamma' S \gamma$. Then

$$
\begin{aligned}
(10) \qquad 0 &\le (y - bS\gamma)' S^{-1} (y - bS\gamma) \\
&= y' S^{-1} y - b\gamma' S S^{-1} y - y' S^{-1} S \gamma b + b^2 \gamma' S S^{-1} S \gamma \\
&= y' S^{-1} y - \frac{(\gamma' y)^2}{\gamma' S \gamma},
\end{aligned}
$$

which yields (9). ■

When $y = \bar{x} - \mu$, then (9) implies that

$$
\begin{aligned}
(11) \qquad |\gamma'(\bar{x} - \mu)| &\le \sqrt{\gamma' S \gamma (\bar{x} - \mu)' S^{-1} (\bar{x} - \mu)} \\
&\le \sqrt{\gamma' S \gamma}\, \sqrt{T^2_{p, N-1}(\alpha)/N}
\end{aligned}
$$

holds for all γ with probability $1 - \alpha$. Thus we can assert with confidence $1 - \alpha$ that the unknown parameter vector satisfies simultaneously for all γ the inequalities

$$
(12) \qquad |\gamma' \bar{x} - \gamma' m| \le \sqrt{\gamma' S \gamma}\, \sqrt{T^2_{p, N-1}(\alpha)/N}.
$$

The confidence region (8) can be explored by setting γ in (12) equal to simple vectors such as $(1, 0, \dots, 0)'$ to obtain m_1, $(1, -1, 0, \dots, 0)$ to yield $m_1 - m_2$, and so on. It should be noted that if only one linear function $\gamma' \mu$ were of interest, $\sqrt{T^2_{p, N-1}(\alpha)} = \sqrt{npF_{p, n-p+1}(\alpha)/(n - p + 1)}$ would be replaced by $t_n(\alpha)$.

5.3.4. Two-Sample Problems

Another situation in which the T^2-statistic is used is one in which the null hypothesis is that the mean of one normal population is equal to the mean of the other where the covariance matrices are assumed equal but unknown. Suppose $y_1^{(i)}, \dots, y_{N_i}^{(i)}$ is a sample from $N(\mu^{(i)}, \Sigma)$, $i = 1, 2$. We wish to test the null hypothesis $\mu^{(1)} = \mu^{(2)}$. The vector $\bar{y}^{(i)}$ is distributed according to $N[\mu^{(i)}, (1/N_i)\Sigma]$. Consequently $\sqrt{N_1 N_2/(N_1 + N_2)}\,(\bar{y}^{(1)} - \bar{y}^{(2)})$ is distributed according to $N(\mathbf{0}, \Sigma)$ under the null hypothesis. If we let

$$
(13) \qquad S = \frac{1}{N_1 + N_2 - 2} \left\{ \sum_{\alpha=1}^{N_1} \left(y_\alpha^{(1)} - \bar{y}^{(1)}\right)\left(y_\alpha^{(1)} - \bar{y}^{(1)}\right)' + \sum_{\alpha=1}^{N_2} \left(y_\alpha^{(2)} - \bar{y}^{(2)}\right)\left(y_\alpha^{(2)} - \bar{y}^{(2)}\right)' \right\},
$$

then $(N_1+N_2-2)S$ is distributed as $\sum_{\alpha=1}^{N_1+N_2-2} Z_\alpha Z'_\alpha$, where Z_α is distributed according to $N(\mathbf{0}, \mathbf{\Sigma})$. Thus

$$T^2 = \frac{N_1 N_2}{N_1+N_2}\left(\bar{y}^{(1)}-\bar{y}^{(2)}\right)' S^{-1}\left(\bar{y}^{(1)}-\bar{y}^{(2)}\right) \tag{14}$$

is distributed as T^2 with N_1+N_2-2 degrees of freedom. The critical region is

$$T^2 > \frac{(N_1+N_2-2)p}{N_1+N_2-p-1} F_{p, N_1+N_2-p-1}(\alpha) \tag{15}$$

with significance level α.

A confidence region for $\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}$ with confidence level $1-\alpha$ is the set of vectors $\boldsymbol{m}$ satisfying

$$\begin{aligned}(16)\qquad &\left(\bar{y}^{(1)}-\bar{y}^{(2)}-\boldsymbol{m}\right)' S^{-1}\left(\bar{y}^{(1)}-\bar{y}^{(2)}-\boldsymbol{m}\right)\\ &\leq \frac{N_1+N_2}{N_1 N_2} T^2_{p, N_1+N_2-2}(\alpha)\\ &= \frac{N_1+N_2}{N_1 N_2}\,\frac{(N_1+N_2-2)p}{N_1+N_2-p-1} F_{p, N_1+N_2-p-1}(\alpha).\end{aligned}$$

Simultaneous confidence intervals are

$$|\boldsymbol{\gamma}'(\bar{y}^{(1)}-\bar{y}^{(2)}) - \boldsymbol{\gamma}'\boldsymbol{m}| \leq \sqrt{\boldsymbol{\gamma}' S \boldsymbol{\gamma}}\sqrt{\frac{N_1+N_2}{N_1 N_2} T^2_{p, N_1+N_2-2}(\alpha)}\,. \tag{17}$$

An example may be taken from Fisher (1936). Let x_1 = sepal length, x_2 = sepal width, x_3 = petal length, x_4 = petal width. Fifty observations are taken from the population *Iris versicolor* (1) and 50 from the population *Iris setosa* (2). See Table 3.4. The data may be summarized (in centimeters) as

$$\bar{x}^{(1)} = \begin{pmatrix} 5.936 \\ 2.770 \\ 4.260 \\ 1.326 \end{pmatrix}, \tag{18}$$

$$\bar{x}^{(2)} = \begin{pmatrix} 5.006 \\ 3.428 \\ 1.462 \\ 0.246 \end{pmatrix}, \tag{19}$$

$$(20) \qquad 98S = \begin{pmatrix} 19.1434 & 9.0356 & 9.7634 & 3.2394 \\ 9.0356 & 11.8658 & 4.6232 & 2.4746 \\ 9.7634 & 4.6232 & 12.2978 & 3.8794 \\ 3.2394 & 2.4746 & 3.8794 & 2.4604 \end{pmatrix}.$$

The value of $T^2/98$ is 26.334, and $(T^2/98) \times \frac{95}{4} = 625.5$. This value is highly significant compared to the F-value for 4 and 95 degrees of freedom of 3.52 at the 0.01 significance level.

Simultaneous confidence intervals for the differences of component means $\mu_i^{(1)} - \mu_i^{(2)}$, $i = 1,2,3,4$, are 0.930 ± 0.337, -0.658 ± 0.265, -2.798 ± 0.270, and 1.080 ± 0.121. In each case 0 does not lie in the interval. [Since $t_{98}(.01) < T_{4,98}(.01)$, a univariate test on any component would lead to rejection of the null hypothesis.] The last two components show the most significant differences from 0.

5.3.5. A Problem of Several Samples

After considering the above example, Fisher considers a third sample drawn from a population assumed to have the same covariance matrix. He treats the same measurements on 50 *Iris virginica* (Table 3.4). There is a theoretical reason for believing the gene structures of these three species to be such that the mean vectors of the three populations are related as

$$(21) \qquad 3\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(3)} + 2\boldsymbol{\mu}^{(2)},$$

where $\boldsymbol{\mu}^{(3)}$ is the mean vector of the third population.

This is a special case of the following general problem. Let $\{x_\alpha^{(i)}\}$, $\alpha = 1, \dots, N_i$, $i = 1, \dots, q$, be samples from $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i = 1, \dots, q$, respectively. Let us test the hypothesis

$$(22) \qquad H: \sum_{i=1}^{q} \beta_i \boldsymbol{\mu}^{(i)} = \boldsymbol{\mu},$$

where $\beta_1, \dots, \beta_q$ are given scalars and $\boldsymbol{\mu}$ is a given vector. The criterion is

$$(23) \qquad T^2 = c\left(\sum_{i=1}^{q} \beta_i \bar{x}^{(i)} - \boldsymbol{\mu}\right)' S^{-1} \left(\sum_{i=1}^{q} \beta_i \bar{x}^{(i)} - \boldsymbol{\mu}\right),$$

where

$$(24) \qquad \bar{x}^{(i)} = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} x_\alpha^{(i)},$$

$$\left(\sum_{i=1}^{q} N_i - q\right) S = \sum_{i=1}^{q} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)', \tag{25}$$

$$\frac{1}{c} = \sum_{i=1}^{q} \frac{\beta_i^2}{N_i}. \tag{26}$$

This T^2 has the T^2-distribution with $\sum_{i=1}^{q} N_i - q$ degrees of freedom.

Fisher actually assumes in his example that the covariance matrices of the three populations may be different. Hence he uses the technique described in Section 5.5.

5.3.6. A Problem of Symmetry

Consider testing the hypothesis $H: \mu_1 = \mu_2 = \cdots = \mu_p$ on the basis of a sample $x_1, \dots, x_N$ from $N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}' = (\mu_1, \dots, \mu_p)$. Let C be any $(p-1) \times p$ matrix of rank $p-1$ such that

$$C\boldsymbol{\varepsilon} = \mathbf{0}, \tag{27}$$

where $\boldsymbol{\varepsilon}' = (1, \dots, 1)$. Then

$$y_\alpha = Cx_\alpha, \qquad \alpha = 1, \dots, N, \tag{28}$$

has mean $C\boldsymbol{\mu}$ and covariance matrix $C\Sigma C'$. The hypothesis H is $C\boldsymbol{\mu} = \mathbf{0}$. The statistic to be used is

$$T^2 = N\bar{y}'S^{-1}\bar{y}, \tag{29}$$

where

$$\bar{y} = \frac{1}{N} \sum_{\alpha=1}^{N} y_\alpha = C\bar{x}, \tag{30}$$

$$\begin{aligned} S &= \frac{1}{N-1} \sum_{\alpha=1}^{N} (y_\alpha - \bar{y})(y_\alpha - \bar{y})' \\ &= \frac{1}{N-1} C \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'C'. \end{aligned} \tag{31}$$

This statistic has the T^2-distribution with $N-1$ degrees of freedom for a $(p-1)$-dimensional distribution. This T^2-statistic is invariant under any linear transformation in the $p-1$ dimensions orthogonal to $\boldsymbol{\varepsilon}$. Hence the statistic is independent of the choice of C.

An example of this sort has been given by Rao (1948b). Let N be the amount of cork in a boring from the north into a cork tree; let E, S, and W be defined similarly. The set of amounts in four borings on one tree is considered as an observation from a 4-variate normal distribution. The question is whether the cork trees have the same amount of cork on each side. We make a transformation

$$
\begin{aligned}
y_1 &= N - E - W + S,\\
y_2 &= S - W, \qquad\qquad (32)\\
y_3 &= N - S.
\end{aligned}
$$

The number of observations is 28. The vector of means is

$$
\bar{y} = \begin{pmatrix} 8.86 \\ 4.50 \\ 0.86 \end{pmatrix}; \tag{33}
$$

the covariance matrix for y is

$$
S = \begin{pmatrix} 128.72 & 61.41 & -21.02 \\ 61.41 & 56.93 & -28.30 \\ -21.02 & -28.30 & 63.53 \end{pmatrix}. \tag{34}
$$

The value of $T^2/(N-1)$ is 0.768. The statistic $0.768 \times 25/3 = 6.402$ is to be compared with the F-significance point with 3 and 25 degrees of freedom. It is significant at the 1% level.

5.3.7. Improved Estimation of the Mean

In Section 3.5 we considered estimation of the mean when the covariance matrix was known and showed that the Stein-type estimation based on this knowledge yielded lower quadratic risks than did the sample mean. In particular, if the loss is $(\boldsymbol{m} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})$, then

$$
\left(1 - \frac{p-2}{N(\bar{\boldsymbol{x}} - \boldsymbol{\nu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\nu})}\right)^{+}(\bar{\boldsymbol{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu} \tag{35}
$$

is a minimax estimator of $\boldsymbol{\mu}$ for any $\boldsymbol{\nu}$ and has a smaller risk than $\bar{\boldsymbol{x}}$ when $p \geq 3$. When $\boldsymbol{\Sigma}$ is unknown, we consider replacing it by an estimator, namely, a multiple of $A = nS$.

Theorem 5.3.1. *When the loss is* $(\boldsymbol{m} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})$, *the estimator for* $p \geq 3$ *given by*

$$
\left(1 - \frac{a}{N(\bar{\boldsymbol{x}} - \boldsymbol{\nu})'A^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\nu})}\right)(\bar{\boldsymbol{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu} \tag{36}
$$

has smaller risk than $\bar{\boldsymbol{x}}$ *and is minimax for* $0 < a < 2(p-2)/(n-p+3)$, *and the risk is minimized for* $a = (p-2)/(n-p+3)$.

Proof. As in the case when Σ is known (Section 3.5.2), we can make a transformation that carries $(1/N)\Sigma$ to I. Then the problem is to estimate $\boldsymbol{\mu}$ based on Y with the distribution $N(\boldsymbol{\mu}, I)$ and $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, where $Z_1, \dots, Z_n$ are independently distributed, each according to $N(\mathbf{0}, I)$, and the loss function is $(\boldsymbol{m} - \boldsymbol{\mu})'(\boldsymbol{m} - \boldsymbol{\mu})$. (We have dropped a factor of N.) The difference in risks is

(37)

$$\begin{aligned}\Delta R(\boldsymbol{\mu}) &= \mathscr{E}_{\boldsymbol{\mu}}\left\{\|Y-\boldsymbol{\mu}\|^2 - \left\|\left(1 - \frac{a}{(Y-\boldsymbol{\nu})'A^{-1}(Y-\boldsymbol{\nu})}\right)(Y-\boldsymbol{\nu}) + \boldsymbol{\nu} - \boldsymbol{\mu}\right\|^2\right\} \\ &= \mathscr{E}_{\boldsymbol{\mu}}\left\{2\frac{a}{(Y-\boldsymbol{\nu})'A^{-1}(Y-\boldsymbol{\nu})}\sum_{i=1}^{p}(Y_i-\mu_i)(Y_i-\nu_i)\right. \\ &\qquad \left. - \frac{a^2}{\left[(Y-\boldsymbol{\nu})'A^{-1}(Y-\boldsymbol{\nu})\right]^2}\|Y-\boldsymbol{\nu}\|^2\right\}.\end{aligned}$$

The proof of Theorem 5.2.2 shows that $(Y-\boldsymbol{\nu})'A^{-1}(Y-\boldsymbol{\nu})$ is distributed as $\|Y-\boldsymbol{\nu}\|^2/\chi^2_{n-p+1}$, where the χ^2_{n-p+1} is independent of Y. Then the difference in risks is

$$\begin{aligned}(38)\quad \Delta R(\boldsymbol{\mu}) &= \mathscr{E}_{\boldsymbol{\mu}}\left\{\frac{2a\chi^2_{n-p+1}}{\|Y-\boldsymbol{\nu}\|^2}\sum_{i=1}^{p}(Y_i-\mu_i)(Y_i-\nu_i) - \frac{a^2\left(\chi^2_{n-p+1}\right)^2}{\|Y-\boldsymbol{\nu}\|^2}\right\} \\ &= \mathscr{E}_{\boldsymbol{\mu}}\left\{\frac{2a(p-2)\chi^2_{n-p+1}}{\|Y-\boldsymbol{\nu}\|^2} - \frac{a^2\left(\chi^2_{n-p+1}\right)^2}{\|Y-\boldsymbol{\nu}\|^2}\right\} \\ &= \left\{2(p-2)(n-p+1)a\right. \\ &\qquad \left. - \left[2(n-p+1) + (n-p+1)^2\right]a^2\right\}\mathscr{E}_{\boldsymbol{\mu}}\frac{1}{\|Y-\boldsymbol{\nu}\|^2}.\end{aligned}$$

The factor in braces is $n-p+1$ times $2(p-2)a - (n-p+3)a^2$, which is positive for $0 < a < 2(p-2)/(n-p+3)$ and is maximized for $a = (p-2)/(n-p+3)$. ∎

The improvement over the risk of Y is $(n-p+1)(p-2)^2/(n-p+3)\cdot \mathscr{E}_{\boldsymbol{\mu}}\|Y-\boldsymbol{\nu}\|^{-2}$, as compared to the improvement $(p-2)^2\mathscr{E}_{\boldsymbol{\mu}}\|Y-\boldsymbol{\nu}\|^{-2}$ of $m(y)$ of Section 3.5 when Σ is known.

Corollary 5.3.1. *The estimator for $p \geq 3$*

$$(39) \qquad \left(1 - \frac{a}{N(\bar{x}-\nu)'A^{-1}(\bar{x}-\nu)}\right)^{+}(\bar{x}-\nu)+\nu$$

has smaller risk than (36) *and is minimax for* $0 < a < 2(p-2)/(n-p+3)$.

Proof. This corollary follows from Theorem 5.3.1 and Lemma 3.5.2. ∎

The risk of (39) is not necessarily minimized at $a = (p-2)/(n-p+3)$, but that value seems like a good choice. This is the estimator (18) of Section 3.5 with $\mathbf{\Sigma}$ replaced by $[1/(n-p+3)]A$.

When the loss function is $(\boldsymbol{m}-\boldsymbol{\mu})'\boldsymbol{Q}(\boldsymbol{m}-\boldsymbol{\mu})$, where $\boldsymbol{Q}$ is an arbitrary positive definite matrix, it is harder to present a uniformly improved estimator that is attractive. The estimators of Section 3.5 can be used with $\mathbf{\Sigma}$ replaced by an estimate.

5.4. THE DISTRIBUTION OF T^2 UNDER ALTERNATIVE HYPOTHESES; THE POWER FUNCTION

In Section 5.2.2 we showed that $(T^2/n)(N-p)/p$ has a noncentral F-distribution. In this section we shall discuss the noncentral F-distribution, its tabulation, and applications to procedures based on T^2.

The noncentral F-distribution is defined as the distribution of the ratio of a noncentral χ^2 and an independent χ^2 divided by the ratio of corresponding degrees of freedom. Let V have the noncentral χ^2-distribution with p degrees of freedom and noncentrality parameter τ^2 (as given in Theorem 3.3.5), and let W be independently distributed as χ^2 with m degrees of freedom. We shall find the density of $F = (V/p)/(W/m)$, which is the noncentral F with noncentrality parameter τ^2. The joint density of V and W is (28) of Section 3.3 multiplied by the density of W, which is $2^{-\frac{1}{2}m}\Gamma^{-1}(\frac{1}{2}m)w^{\frac{1}{2}m-1}e^{-\frac{1}{2}w}$. The joint density of F and W ($dv = pw\,df/m$) is

$$(1) \qquad \frac{e^{-\frac{1}{2}\tau^2}}{2^{\frac{1}{2}(p+m)}\Gamma(\frac{1}{2}m)}e^{-\frac{1}{2}w(1+pf/m)}$$

$$\cdot\frac{p}{m}\sum_{\beta=0}^{\infty}\left(\frac{\tau^2}{4}\right)^{\beta}\frac{1}{\beta!\Gamma(\frac{1}{2}p+\beta)}\left(\frac{pf}{m}\right)^{\frac{1}{2}p+\beta-1}w^{\frac{1}{2}(p+m)+\beta-1}.$$

The marginal density, obtained by integrating (1) with respect to w from 0 to ∞, is

$$(2) \qquad \frac{pe^{-\frac{1}{2}\tau^2}}{m\Gamma(\frac{1}{2}m)}\sum_{\beta=0}^{\infty}\frac{(\tau^2/2)^{\beta}(pf/m)^{\frac{1}{2}p+\beta-1}\Gamma[\frac{1}{2}(p+m)+\beta]}{\beta!\Gamma(\frac{1}{2}p+\beta)(1+pf/m)^{\frac{1}{2}(p+m)+\beta}}.$$

Theorem 5.4.1. *If V has a noncentral χ^2-distribution with p degrees of freedom and noncentrality parameter τ^2, and W has an independent χ^2-distribution with m degrees of freedom, then $F = (V/p)/(W/m)$ has the density* (2).

The density (2) is the density of the *noncentral F-distribution.*

If $T^2 = N(\bar{x} - \boldsymbol{\mu}_0)'S^{-1}(\bar{x} - \boldsymbol{\mu}_0)$ is based on a sample of N from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(T^2/n)(N-p)/p$ has the noncentral F-distribution with p and $N-p$ degrees of freedom and noncentrality parameter $N(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) = \tau^2$. From (2) we find that the density of T^2 is

$$(3)\qquad \frac{e^{-\frac{1}{2}\tau^2}}{(N-1)\Gamma\left[\frac{1}{2}(N-p)\right]} \sum_{\beta=0}^{\infty} \frac{(\tau^2/2)^{\beta}\left[t^2/(N-1)\right]^{\frac{1}{2}p+\beta-1}\Gamma\left(\frac{1}{2}N+\beta\right)}{\beta!\,\Gamma\left(\frac{1}{2}p+\beta\right)\left[1+t^2/(N-1)\right]^{\frac{1}{2}N+\beta}}$$

$$= \frac{\Gamma\left(\frac{1}{2}N\right)}{(N-1)\Gamma\left[\frac{1}{2}(N-p)\right]\Gamma\left(\frac{1}{2}p\right)}\left(\frac{t^2}{N-1}\right)^{\frac{1}{2}p-1}\left(1+\frac{t^2}{N-1}\right)^{-\frac{1}{2}N}$$

$$\cdot e^{-\frac{1}{2}\tau^2}\,{}_1F_1\left(\tfrac{1}{2}N; \tfrac{1}{2}p; \frac{\tau^2 t^2}{2(N-1)}\right),$$

where

$$(4)\qquad {}_1F_1(a;b;x) = \sum_{\beta=0}^{\infty} \frac{\Gamma(a+\beta)\Gamma(b)x^{\beta}}{\Gamma(a)\Gamma(b+\beta)\beta!}.$$

The density (3) is the density of the *noncentral T^2-distribution.*

Tables have been given by Tang (1938) of the probability of accepting the null hypothesis (that is, the probability of Type II error) for various values of τ^2 and for significance levels 0.05 and 0.01. His number of degrees of freedom f_1 is our p [1(1)8], his f_2 is our $n-p+1$ [2, 4(1)30, 60, ∞], and his noncentrality parameter ϕ is related to our τ^2 by

$$(5)\qquad \phi = \frac{\tau}{\sqrt{p+1}}$$

[$1(\frac{1}{2})3(1)8$]. His accompanying tables of significance points are for $T^2/(T^2 + N - 1)$.

As an example, suppose $p = 4$, $n - p + 1 = 20$, and consider testing the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$ at the 1% level of significance. We would like to know the probability, say, that we accept the null hypothesis when $\phi = 2.5$ ($\tau^2 = 31.25$). It is 0.227. If we think the disadvantage of accepting the null hypothesis when N, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are such that $\tau^2 = 31.25$ is less than the disadvantage of rejecting the null hypothesis when it is true, then we may find it reasonable to conduct the test as assumed. However, if the disadvantage of one type of error is about equal to that of the other, it would seem reasonable to bring down the probability of a Type II error. Thus, if we use a significance level of 5%, the probability of Type II error (for $\phi = 2.5$) is only 0.043.

Lehmer (1944) has computed tables of ϕ for given significance level and given probability of Type II error. Here tables can be used to see what value of τ^2 is needed to make the probability of acceptance of the null hypothesis sufficiently low when $\boldsymbol{\mu} \neq \mathbf{0}$. For instance, if we want to be able to reject the hypothesis $\boldsymbol{\mu} = \mathbf{0}$ on the basis of a sample for a given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we may be able to choose N so that $N\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \tau^2$ is sufficiently large. Of course, the difficulty with these considerations is that we usually do not know exactly the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (and hence of τ^2) for which we want the probability of rejection at a certain value.

The distribution of T^2 when the null hypothesis is not true was derived by different methods by Hsu (1938) and Bose and Roy (1938).

5.5. THE TWO-SAMPLE PROBLEM WITH UNEQUAL COVARIANCE MATRICES

If the covariance matrices are not the same, the T^2-test for equality of mean vectors has a probability of rejection under the null hypothesis that depends on these matrices. If the difference between the matrices is small or if the sample sizes are large, there is no practical effect. However, if the covariance matrices are quite different and/or the sample sizes are relatively small, the nominal significance level may be distorted. Hence we develop a procedure with assigned significance level. Let $\{\boldsymbol{x}_\alpha^{(i)}\}$, $\alpha = 1, \dots, N_i$, be samples from $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We wish to test the hypothesis $H: \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$. The mean $\bar{\boldsymbol{x}}^{(1)}$ of the first sample is normally distributed with expected value

$$\mathscr{E}\bar{\boldsymbol{x}}^{(1)} = \boldsymbol{\mu}^{(1)} \tag{1}$$

and covariance matrix

$$\mathscr{E}(\bar{\boldsymbol{x}}^{(1)} - \boldsymbol{\mu}^{(1)})(\bar{\boldsymbol{x}}^{(1)} - \boldsymbol{\mu}^{(1)})' = \frac{1}{N_1}\boldsymbol{\Sigma}_1. \tag{2}$$

Similarly, the mean $\bar{\boldsymbol{x}}^{(2)}$ of the second sample is normally distributed with expected value

$$\mathscr{E}\bar{\boldsymbol{x}}^{(2)} = \boldsymbol{\mu}^{(2)} \tag{3}$$

and covariance matrix

$$\mathscr{E}(\bar{x}^{(2)}-\mu^{(2)})(\bar{x}^{(2)}-\mu^{(2)})' = \frac{1}{N_2}\Sigma_2. \tag{4}$$

Thus $\bar{x}^{(1)}-\bar{x}^{(2)}$ has mean $\mu^{(1)}-\mu^{(2)}$ and covariance matrix $(1/N_1)\Sigma_1 + (1/N_2)\Sigma_2$. We cannot use the technique of Section 5.2, however, because

$$\sum_{\alpha=1}^{N_1}\left(x_\alpha^{(1)}-\bar{x}^{(1)}\right)\left(x_\alpha^{(1)}-\bar{x}^{(1)}\right)' + \sum_{\alpha=1}^{N_2}\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)' \tag{5}$$

does not have the Wishart distribution with covariance matrix a multiple of $(1/N_1)\Sigma_1 + (1/N_2)\Sigma_2$.

If $N_1 = N_2 = N$, say, we can use the T^2-test in an obvious way. Let $y_\alpha = x_\alpha^{(1)} - x_\alpha^{(2)}$ (assuming the numbering of the observations in the two samples is independent of the observations themselves). Then y_α is normally distributed with mean $\mu^{(1)}-\mu^{(2)}$ and covariance matrix $\Sigma_1+\Sigma_2$, and $y_1,\dots,y_N$ are independent. Let $\bar{y} = (1/N)\sum_{\alpha=1}^N y_\alpha = \bar{x}^{(1)}-\bar{x}^{(2)}$, and define S by

$$\begin{aligned}(N-1)S &= \sum_{\alpha=1}^{N}(y_\alpha-\bar{y})(y_\alpha-\bar{y})' \\ &= \sum_{\alpha=1}^{N}\left(x_\alpha^{(1)}-x_\alpha^{(2)}-\bar{x}^{(1)}+\bar{x}^{(2)}\right)\left(x_\alpha^{(1)}-x_\alpha^{(2)}-\bar{x}^{(1)}+\bar{x}^{(2)}\right)'.\end{aligned} \tag{6}$$

Then

$$T^2 = N\bar{y}'S^{-1}\bar{y} \tag{7}$$

is suitable for testing the hypothesis $\mu^{(1)}-\mu^{(2)} = \mathbf{0}$, and has the T^2-distribution with $N-1$ degrees of freedom. It should be observed that if we had known $\Sigma_1 = \Sigma_2$, we would have used a T^2-statistic with $2N-2$ degrees of freedom; thus we have lost $N-1$ degrees of freedom in constructing a test which is independent of the two covariance matrices. If $N_1 = N_2 = 50$ as in the example in Section 5.3.4, then $T_{4,49}^2(.01) = 15.93$ as compared to $T_{4,98}^2(.01) = 14.52$.

Now let us turn our attention to the case of $N_1 \neq N_2$. For convenience, let $N_1 < N_2$. Then we define

$$y_\alpha = x_\alpha^{(1)} - \sqrt{\frac{N_1}{N_2}}\,x_\alpha^{(2)} + \frac{1}{\sqrt{N_1N_2}}\sum_{\beta=1}^{N_1}x_\beta^{(2)} - \frac{1}{N_2}\sum_{\gamma=1}^{N_2}x_\gamma^{(2)}, \qquad \alpha = 1,\dots,N_1 \tag{8}$$

The expected value of y_α is

$$(9) \qquad \mathscr{E} y_\alpha = \mu^{(1)} - \sqrt{\frac{N_1}{N_2}}\,\mu^{(2)} + \frac{N_1}{\sqrt{N_1 N_2}}\mu^{(2)} - \frac{N_2}{N_2}\mu^{(2)} = \mu^{(1)} - \mu^{(2)}.$$

The covariance matrix of y_α and y_β is

$$(10) \qquad \mathscr{E}(y_\alpha - \mathscr{E}y_\alpha)(y_\beta - \mathscr{E}y_\beta)' = \delta_{\alpha\beta}\left(\Sigma_1 + \frac{N_1}{N_2}\Sigma_2\right).$$

Thus a suitable statistic for testing $\mu^{(1)} - \mu^{(2)} = 0$, which has the T^2-distribution with $N_1 - 1$ degrees of freedom, is

$$(11) \qquad T^2 = N_1 \bar{y}' S^{-1} \bar{y},$$

where

$$(12) \qquad \bar{y} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} y_\alpha = \bar{x}^{(1)} - \bar{x}^{(2)}$$

and

$$(13) \qquad (N_1 - 1)S = \sum_{\alpha=1}^{N_1}(y_\alpha - \bar{y})(y_\alpha - \bar{y})' = \sum_{\alpha=1}^{N_1}(u_\alpha - \bar{u})(u_\alpha - \bar{u})',$$

where $\bar{u} = (1/N_1)\sum_{\alpha=1}^{N_1} u_\alpha$ and $u_\alpha = x_\alpha^{(1)} - \sqrt{N_1/N_2}\,x_\alpha^{(2)}$, $\alpha = 1, \dots, N_1$.

This procedure was suggested by Scheffé (1943) in the univariate case. Scheffé showed that in the univariate case this technique gives the shortest confidence intervals obtained by using the t-distribution. The advantage of the method is that $\bar{x}^{(1)} - \bar{x}^{(2)}$ is used, and this statistic is most relevant to $\mu^{(1)} - \mu^{(2)}$. The sacrifice of observations in estimating a covariance matrix is not so important. Bennett (1951) gave the extension of the procedure to the multivariate case.

This approach can be used for more general cases. Let $\{x_\alpha^{(i)}\}$, $\alpha = 1, \dots, N_i$, $i = 1, \dots, q$, be samples from $N(\mu^{(i)}, \Sigma_i)$, $i = 1, \dots, q$, respectively. Consider testing the hypothesis

$$(14) \qquad H: \sum_{i=1}^{q} \beta_i \mu^{(i)} = \mu,$$

where $\beta_1, \dots, \beta_q$ are given scalars and μ is a given vector. If the N_i are unequal, take N_1 to be the smallest. Let

$$(15) \quad y_\alpha = \beta_1 x_\alpha^{(1)} + \sum_{i=2}^{q} \beta_i \sqrt{\frac{N_1}{N_i}}\left(x_\alpha^{(i)} - \frac{1}{N_1}\sum_{\beta=1}^{N_1} x_\beta^{(i)} + \frac{1}{\sqrt{N_1 N_i}}\sum_{\gamma=1}^{N_i} x_\gamma^{(i)}\right).$$

Then $\mathscr{E}y_\alpha = \sum_{i=1}^q \beta_i \boldsymbol{\mu}^{(i)}$, and

$$\mathscr{E}(y_\alpha - \mathscr{E}y_\alpha)(y_\beta - \mathscr{E}y_\beta)' = \delta_{\alpha\beta}\left(\sum_{i=1}^q \frac{\beta_i^2 N_1}{N_i}\Sigma_i\right). \tag{16}$$

Let $\bar{y}$ and S be defined by

$$\bar{y} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} y_\alpha = \sum_{i=1}^q \beta_i \bar{x}^{(i)}, \qquad \bar{x}^{(i)} = \frac{1}{N_i}\sum_{\beta=1}^{N_i} x_\beta^{(i)}, \tag{17}$$

$$(N_1 - 1)S = \sum_{\alpha=1}^{N_1}(y_\alpha - \bar{y})(y_\alpha - \bar{y})'. \tag{18}$$

Then

$$T^2 = N_1(\bar{y} - \boldsymbol{\mu})'S^{-1}(\bar{y} - \boldsymbol{\mu}) \tag{19}$$

is suitable for testing H, and when the hypothesis is true, this statistic has the T^2-distribution for dimension p with $N_1 - 1$ degrees of freedom. If we let $u_\alpha = \sum_{i=1}^q \beta_i\sqrt{N_1/N_i}\,x_\alpha^{(i)}$, $\alpha = 1, \dots, N_1$, then S can be defined as

$$(N_1 - 1)S = \sum_{\alpha=1}^{N_1}(u_\alpha - \bar{u})(u_\alpha - \bar{u})'. \tag{20}$$

Another problem that is amenable to this kind of treatment is testing the hypothesis that two subvectors have equal means. Let $x = (x^{(1)\prime}, x^{(2)\prime})'$ be distributed normally with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}^{(1)\prime}, \boldsymbol{\mu}^{(2)\prime})'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{21}$$

We assume that $x^{(1)}$ and $x^{(2)}$ are each of q components. Then $y = x^{(1)} - x^{(2)}$ is distributed normally with mean $\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$ and covariance matrix $\Sigma_y = \Sigma_{11} - \Sigma_{21} - \Sigma_{12} + \Sigma_{22}$. To test the hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ we use a T^2-statistic $N\bar{y}'S_y^{-1}\bar{y}$, where the mean vector and covariance matrix of the sample are partitioned similarly to $\boldsymbol{\mu}$ and Σ.

5.6. SOME OPTIMAL PROPERTIES OF THE T^2-TEST

5.6.1. Optimal Invariant Tests

In this section we shall indicate that the T^2-test is the best in certain classes of tests and sketch briefly the proofs of these results.

The hypothesis $\boldsymbol{\mu} = \mathbf{0}$ is to be tested on the basis of the N observations $x_1, \dots, x_N$ from $N(\boldsymbol{\mu}, \Sigma)$. First we consider the class of tests based on the statistics $A = \sum(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$ and $\bar{x}$ which are invariant with respect to the transformations $A^* = CAC'$ and $\bar{x}^* = C\bar{x}$, where C is nonsingular. The transformation $x_\alpha^* = Cx_\alpha$ leaves the problem invariant; that is, in terms of x_α^* we test the hypothesis $\mathscr{E}x_\alpha^* = 0$ given that $x_1^*, \dots, x_N^*$ are N observations from a multivariate normal population. It seems reasonable that we require a solution that is also invariant with respect to these transformations; that is, we look for a critical region that is not changed by a nonsingular linear transformation. (The definition of the region is the same in different coordinate systems.)

Theorem 5.6.1. *Given the observations $x_1, \dots, x_N$ from $N(\mu, \Sigma)$, of all tests of $\mu = 0$ based on $\bar{x}$ and $A = \sum(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$ that are invariant with respect to transformations $\bar{x}^* = C\bar{x}$, $A^* = CAC'$ (C nonsingular), the T^2-test is uniformly most powerful.*

Proof. First, as we have seen in Section 5.2.1, any test based on T^2 is invariant. Second, this function is essentially the only invariant, for if $f(\bar{x}, A)$ is invariant, then $f(\bar{x}, A) = f(\bar{x}^*, I)$, where only the first coordinate of $\bar{x}^*$ is different from zero and it is $\sqrt{\bar{x}'A^{-1}\bar{x}}$. (There is a matrix C such that $C\bar{x} = \bar{x}^*$ and $CAC' = I$.) Thus $f(\bar{x}, A)$ depends only on $\bar{x}'A^{-1}\bar{x}$. Thus an invariant test must be based on $\bar{x}'A^{-1}\bar{x}$. Third, we can apply the Neyman–Pearson fundamental lemma to the distribution of T^2 [(3) of Section 5.4] to find the uniformly most powerful test based on T^2 against a simple alternative $\tau^2 = N\mu'\Sigma^{-1}\mu$. The most powerful test of $\tau^2 = 0$ is based on the ratio of (3) of Section 5.4 to (3) with $\tau^2 = 0$. The critical region is

(1)
$$c < e^{-\frac{1}{2}\tau^2} \sum_{\alpha=0}^{\infty} \frac{(\tau^2/2)^\alpha (t^2/n)^{\frac{1}{2}p+\alpha-1}(1+t^2/n)^{-\frac{1}{2}(n+1)-\alpha}\Gamma[\frac{1}{2}(n+1)+\alpha]}{\alpha!\Gamma(\frac{1}{2}p+\alpha)}$$
$$\Bigg/ \frac{(t^2/n)^{\frac{1}{2}p-1}(1+t^2/n)^{-\frac{1}{2}(n+1)}\Gamma[\frac{1}{2}(n+1)]}{\Gamma(\frac{1}{2}p)}$$
$$= \frac{\Gamma(\frac{1}{2}p)}{\Gamma[\frac{1}{2}(n+1)]} e^{-\frac{1}{2}\tau^2} \sum_{\alpha=0}^{\infty} \frac{(\tau^2/2)^\alpha \Gamma[\frac{1}{2}(n+1)+\alpha]}{\alpha!\Gamma(\frac{1}{2}p+\alpha)} \left(\frac{t^2/n}{1+t^2/n}\right)^\alpha.$$

The right-hand side of (1) is a strictly increasing function of $(t^2/n)/(1+t^2/n)$, hence of t^2. Thus the inequality is equivalent to $t^2 > k$ for k suitably chosen. Since this does not depend on the alternative τ^2, the test is uniformly most powerful invariant. ■

Definition 5.6.1. *A* critical function $\psi(\bar{x}, A)$ *is a function with values between* 0 *and* 1 *(inclusive) such that* $\mathscr{E}\psi(\bar{x}, A) = \varepsilon$, *the significance level, when* $\boldsymbol{\mu} = \mathbf{0}$.

A randomized test consists of rejecting the hypothesis with probability $\psi(\boldsymbol{x}, \boldsymbol{B})$ when $\bar{x} = \boldsymbol{x}$ and $A = \boldsymbol{B}$. A nonrandomized test is defined when $\psi(\bar{x}, A)$ takes on only the values 0 and 1. Using the form of the Neyman–Pearson lemma appropriate for critical functions, we obtain the following corollary:

Corollary 5.6.1. *On the basis of observations* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ *from* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *of all randomized tests based on* $\bar{x}$ *and* A *that are invariant with respect to transformations* $\bar{x}^* = C\bar{x}$, $A^* = CAC'$ *(C nonsingular), the* T^2*-test is uniformly most powerful.*

Theorem 5.6.2. *On the basis of observations* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ *from* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *of all tests of* $\boldsymbol{\mu} = \mathbf{0}$ *that are invariant with respect to transformations* $\boldsymbol{x}_\alpha^* = C\boldsymbol{x}_\alpha$ *(C nonsingular), the* T^2*-test is a uniformly most powerful test; that is, the* T^2*-test is at least as powerful as any other invariant test.*

Proof. Let $\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ be the critical function of an invariant test. Then

$$(2) \qquad \mathscr{E}[\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)] = \mathscr{E}_{\bar{x}, A}\{\mathscr{E}[\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)|\bar{x}, A]\}.$$

Since $\bar{x}, A$ are sufficient statistics for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, the expectation $\mathscr{E}[\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)|\bar{x}, A]$ depends only on $\bar{x}, A$. It is invariant and has the same power as $\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$. Thus each test in this larger class can be replaced by one in the smaller class (depending only on $\bar{x}$ and A) that has identical power. Corollary 5.6.1 completes the proof. ∎

Theorem 5.6.3. *Given observations* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ *from* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *of all tests of* $\boldsymbol{\mu} = \mathbf{0}$ *based on* $\bar{x}$ *and* $A = \sum(\boldsymbol{x}_\alpha - \bar{x})(\boldsymbol{x}_\alpha - \bar{x})'$ *with power depending only on* $N\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, *the* T^2*-test is uniformly most powerful.*

Proof. We wish to reduce this theorem to Theorem 5.6.1 by identifying the class of tests with power depending on $N\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ with the class of invariant tests. We need the following definition:

Definition 5.6.2. *A test* $\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ *is said to be* almost invariant *if*

$$(3) \qquad \psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) = \psi(C\boldsymbol{x}_1, \dots, C\boldsymbol{x}_N)$$

for all $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ *except for a set of* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ *of Lebesgue measure zero; this exception set may depend on* C.

It is clear that Theorems 5.6.1 and 5.6.2 hold if we extend the definition of invariant test to mean that (3) holds except for a fixed set of $x_1, \dots, x_N$ of measure 0 (the set not depending on C). It has been shown by Hunt and Stein [Lehmann (1959)] that in our problem almost invariance implies invariance (in the broad sense).

Now we wish to argue that if $\psi(\bar{x}, A)$ has power depending only on $N\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, it is almost invariant. Since the power of $\psi(\bar{x}, A)$ depends only on $N\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, the power is

$$(4)\qquad \mathscr{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\psi(\bar{x}, A) \equiv \mathscr{E}_{C^{-1}\boldsymbol{\mu},C^{-1}\boldsymbol{\Sigma}(C^{-1})'}\psi(\bar{x}, A)$$
$$\equiv \mathscr{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\psi(C\bar{x}, CAC').$$

The second and third terms of (4) are merely different ways of writing the same integral. Thus

$$(5)\qquad \mathscr{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}[\psi(\bar{x}, A) - \psi(C\bar{x}, CAC')] \equiv 0,$$

identically in $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Since $\bar{x}, A$ are a complete sufficient set of statistics for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ (Theorem 3.4.2), $f(\bar{x}, A) = \psi(\bar{x}, A) - \psi(C\bar{x}, CAC') = 0$ almost everywhere. Theorem 5.6.3 follows. ∎

As Theorem 5.6.2 follows from Theorem 5.6.1, so does the following theorem from Theorem 5.6.3:

Theorem 5.6.4. *On the basis of observations $x_1, \dots, x_N$ from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, of all tests of $\boldsymbol{\mu} = \mathbf{0}$ with power depending only on $N\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, the T^2-test is a uniformly most powerful test.*

Theorem 5.6.4 was first proved by Simaika (1941). The results and proofs given in this section follow Lehmann (1959). Hsu (1945) has proved an optimal property of the T^2-test that involves averaging the power over $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

5.6.2. Admissible Tests

We now turn to the question of whether the T^2-test is a good test compared to all possible tests; the comparison in the previous section was to the restricted class of invariant tests. The main result is that the T^2-test is admissible in the class of all tests; that is, there is no other procedure that is better.

Definition 5.6.3. *A test T^* of the null hypothesis $H_0: \omega \in \Omega_0$ against the alternative $\omega \in \Omega_1$ (disjoint from Ω_0) is* admissible *if there exists no other test T*

such that

$$\Pr\{\text{Reject } H_0 | T, \omega\} \le \Pr\{\text{Reject } H_0 | T^*, \omega\}, \qquad \omega \in \Omega_0, \tag{6}$$

$$\Pr\{\text{Reject } H_0 | T, \omega\} \ge \Pr\{\text{Reject } H_0 | T^*, \omega\}, \qquad \omega \in \Omega_1, \tag{7}$$

with strict inequality for at least one ω.

The admissibility of the T^2-test follows from a theorem of Stein (1956a) that applies to any exponential family of distributions.

An *exponential family* of distributions $(\mathscr{Y}, \mathscr{B}, m, \Omega, P)$ consists of a finite-dimensional Euclidean space $\mathscr{Y}$, a measure m on the σ-algebra $\mathscr{B}$ of all ordinary Borel sets of $\mathscr{Y}$, a subset Ω of the adjoint space $\mathscr{Y}'$ (the linear space of all real-valued linear functions on $\mathscr{Y}$) such that

$$\psi(\omega) = \int_{\mathscr{Y}} e^{\omega' y}\, dm(y) < \infty, \qquad \omega \in \Omega, \tag{8}$$

and P, the function on Ω to the set of probability measures on $\mathscr{B}$ given by

$$P_\omega(A) = \frac{1}{\psi(\omega)} \int_A e^{\omega' y}\, dm(y), \qquad A \in \mathscr{B}.$$

The family of normal distributions $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ constitutes an exponential family, for the density can be written

$$n(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{e^{-\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{(2\pi)^{\frac{1}{2}p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1})\boldsymbol{x} + \operatorname{tr}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1})\boldsymbol{x}\boldsymbol{x}'}. \tag{9}$$

We map from $\mathscr{X}$ to $\mathscr{Y}$; the vector $\boldsymbol{y} = (\boldsymbol{y}^{(1)\prime}, \boldsymbol{y}^{(2)\prime})'$ is composed of $\boldsymbol{y}^{(1)} = \boldsymbol{x}$ and $\boldsymbol{y}^{(2)} = (x_1^2, 2x_1x_2, \dots, 2x_1x_p, x_2^2, \dots, x_p^2)'$. The vector $\boldsymbol{\omega} = (\boldsymbol{\omega}^{(1)\prime}, \boldsymbol{\omega}^{(2)\prime})'$ is composed of $\boldsymbol{\omega}^{(1)} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\omega}^{(2)} = -\frac{1}{2}(\sigma^{11}, \sigma^{12}, \dots, \sigma^{1p}, \sigma^{22}, \dots, \sigma^{pp})'$, where $(\sigma^{ij}) = \boldsymbol{\Sigma}^{-1}$; the transformation of parameters is one to one. The measure $m(A)$ of a set $A \in \mathscr{B}$ is the ordinary Lebesgue measure of the set of $\boldsymbol{x}$ that maps into the set A. (Note that the probability measure in $\mathscr{Y}$ is not defined by a density.)

Theorem 5.6.5 (Stein). *Let* $(\mathscr{Y}, \mathscr{B}, m, \Omega, P)$ *be an exponential family and* Ω_0 *a nonempty proper subset of* Ω. (i) *Let* A *be a subset of* $\mathscr{Y}$ *that is closed and convex.* (ii) *Suppose that for every vector* $\boldsymbol{\omega} \in \mathscr{Y}'$ *and real* c *for which* $\{\boldsymbol{y}|\boldsymbol{\omega}'\boldsymbol{y} > c\}$ *and* A *are disjoint, there exists* $\boldsymbol{\omega}_1 \in \Omega$ *such that for arbitrarily large* λ *the vector* $\boldsymbol{\omega}_1 + \lambda\boldsymbol{\omega} \in \Omega - \Omega_0$. *Then the test with acceptance region* A *is admissible for testing the hypothesis that* $\boldsymbol{\omega} \in \Omega_0$ *against the alternative* $\boldsymbol{\omega} \in \Omega - \Omega_0$.

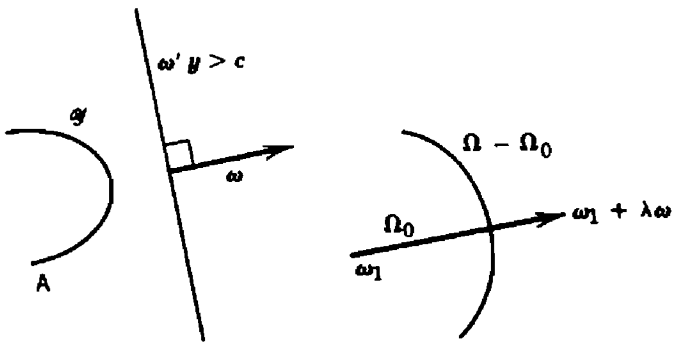

Figure 5.2

The conditions of the theorem are illustrated in Figure 5.2, which is drawn simultaneously in the space $\mathscr{Y}$ and the set Ω.

Proof. The critical function of the test with acceptance region A is $\phi_A(y) = 0$. $y \in A$, and $\phi_A(y) = 1$, $y \notin A$. Suppose $\phi(y)$ is the critical function of a better test, that is,

$$\int \phi(y)\, dP_\omega(y) \le \int \phi_A(y)\, dP_\omega(y), \qquad \omega \in \Omega_0, \tag{10}$$

$$\int \phi(y)\, dP_\omega(y) \ge \int \phi_A(y)\, dP_\omega(y), \qquad \omega \in \Omega - \Omega_0, \tag{11}$$

with strict inequality for some ω; we shall show that this assumption leads to a contradiction. Let $B = \{y \mid \phi(y) < 1\}$. (If the competing test is nonrandomized, B is its acceptance region.) Then

$$\{y \mid \phi_A(y) - \phi(y) > 0\} = \bar{A} \cap B, \tag{12}$$

where $\bar{A}$ is the complement of A. The m-measure of the set (12) is positive; otherwise $\phi_A(y) = \phi(y)$ almost everywhere, and (10) and (11) would hold with equality for all ω. Since A is convex, there exists an ω and a c such that the intersection of $\bar{A} \cap B$ and $\{y \mid \omega' y > c\}$ has positive m-measure. (Since A is closed, $\bar{A}$ is open and it can be covered with a denumerable collection of open spheres, for example, with rational radii and centers with rational coordinates. Because there is a hyperplane separating A and each sphere, there exists a denumerable collection of open half-spaces H_j disjoint from A that covers $\bar{A}$. Then at least one half-space has an intersection with $\bar{A} \cap B$ with positive m-measure.) By hypothesis there exists $\omega_1 \in \Omega$ and an arbitrarily large λ such that

$$\omega_\lambda = \omega_1 + \lambda\omega \in \Omega - \Omega_0. \tag{13}$$

Then

$$
\begin{aligned}
(14)\quad & \int [\phi_A(y) - \phi(y)]\, dP_{\omega_\lambda}(y) \\
&= \frac{1}{\psi(\omega_\lambda)} \int [\phi_A(y) - \phi(y)] e^{\omega_\lambda' y}\, dm(y) \\
&= \frac{\psi(\omega_1)}{\psi(\omega_\lambda)} \int [\phi_A(y) - \phi(y)] e^{\lambda\omega' y}\, dP_{\omega_1}(y) \\
&= \frac{\psi(\omega_1)}{\psi(\omega_\lambda)} e^{\lambda c} \int [\phi_A(y) - \phi(y)] e^{\lambda(\omega' y - c)}\, dP_{\omega_1}(y) \\
&= \frac{\psi(\omega_1)}{\psi(\omega_\lambda)} e^{\lambda c} \left\{ \int_{\omega' y > c} [\phi_A(y) - \phi(y)] e^{\lambda(\omega' y - c)}\, dP_{\omega_1}(y) \right. \\
&\qquad \left. + \int_{\omega' y \le c} [\phi_A(y) - \phi(y)] e^{\lambda(\omega' y - c)}\, dP_{\omega_1}(y) \right\}.
\end{aligned}
$$

For $\omega' y > c$ we have $\phi_A(y) = 1$ and $\phi_A(y) - \phi(y) \ge 0$, and $\{y \mid \phi_A(y) - \phi(y) > 0\}$ has positive measure; therefore, the first integral in the braces approaches ∞ as $\lambda \to \infty$. The second integral is bounded because the integrand is bounded by 1, and hence the last expression is positive for sufficiently large λ. This contradicts (11). ■

This proof was given by Stein (1956a). It is a generalization of a theorem of Birnhaum (1955).

Corollary 5.6.2. *If the conditions of Theorem 5.6.5 hold except that A is not necessarily closed, but the boundary of A has m-measure 0, then the conclusion of Theorem 5.6.5 holds.*

Proof. The closure of A is convex (Problem 5.18), and the test with acceptance region equal to the closure of A differs from A by a set of probability 0 for all $\omega \in \Omega$. Furthermore,

$$
\begin{aligned}
(15)\qquad A \cap \{y \mid \omega' y > c\} = \varnothing \quad &\Rightarrow \quad A \subset \{y \mid \omega' y \le c\} \\
&\Rightarrow \quad \text{closure } A \subset \{y \mid \omega' y \le c\}.
\end{aligned}
$$

Then Theorem 5.6.5 holds with A replaced by the closure of A. ■

Theorem 5.6.6. *Based on observations $x_1, \ldots, x_N$ from $N(\mu, \Sigma)$, Hotelling's T^2-test is admissible for testing the hypothesis $\mu = 0$.*

Proof. To apply Theorem 5.6.5 we put the distribution of the observations into the form of an exponential family. By Theorems 3.3.1 and 3.3.2 we can transform $x_1, \dots, x_N$ to $z_\alpha = \sum_{\beta=1}^N c_{\alpha\beta} x_\beta$, where $(c_{\alpha\beta})$ is orthogonal and $z_N = \sqrt{N}\bar{x}$. Then the density of $z_1, \dots, z_N$ (with respect to Lebesgue measure) is

$$\frac{e^{-\frac{1}{2}N\mu'\Sigma^{-1}\mu}}{(2\pi)^{\frac{1}{2}pN}|\Sigma|^{\frac{1}{2}N}} \exp\left[\sqrt{N}\,\mu'\Sigma^{-1}z_N + \operatorname{tr}\left(-\tfrac{1}{2}\Sigma^{-1}\right)\sum_{\alpha=1}^{N} z_\alpha z_\alpha'\right]. \tag{16}$$

The vector $y = (y^{(1)\prime}, y^{(2)\prime})'$ is composed of $y^{(1)} = z_N$ $(=\sqrt{N}\bar{x})$ and $y^{(2)} = (b_{11}, 2b_{12}, \dots, 2b_{1p}, b_{22}, \dots, b_{pp})'$, where

$$B = \sum_{\alpha=1}^{N} z_\alpha z_\alpha' \left(= \sum_{\alpha=1}^{N} x_\alpha x_\alpha'\right). \tag{17}$$

The vector $\omega = (\omega^{(1)\prime}, \omega^{(2)\prime})'$ is composed of $\omega^{(1)} = \sqrt{N}\Sigma^{-1}\mu$ and $\omega^{(2)} = -\frac{1}{2}(\sigma^{11}, \sigma^{12}, \dots, \sigma^{1p}, \sigma^{22}, \dots, \sigma^{pp})'$. The measure $m(A)$ is the Lebesgue measure of the set of $z_1, \dots, z_N$ that maps into the set A.

Lemma 5.6.1. *Let $B = A + N\bar{x}\bar{x}'$. Then*

$$N\bar{x}'A^{-1}\bar{x} = \frac{N\bar{x}'B^{-1}\bar{x}}{1 - N\bar{x}'B^{-1}\bar{x}}. \tag{18}$$

Proof of Lemma. If we let $B = A + \sqrt{N}\bar{x}\sqrt{N}\bar{x}'$ in (10) of Section 5.2, we obtain by Corollary A.3.1

$$\begin{aligned}\frac{1}{1 + T^2/(N-1)} &= \lambda^{2/N} = \frac{|B - \sqrt{N}\bar{x}\sqrt{N}\bar{x}'|}{|B|} \\ &= 1 - N\bar{x}'B^{-1}\bar{x}. \qquad \blacksquare\end{aligned} \tag{19}$$

Thus the acceptance region of a T^2-test is

$$A = \left\{z_N, B \mid z_N'B^{-1}z_N \le k,\ B \text{ positive definite}\right\} \tag{20}$$

for a suitable k.

The function $z_N'B^{-1}z_N$ is convex in (z, B) for B positive definite (Problem 5.17). Therefore, the set $z_N'B^{-1}z_N \le k$ is convex. This shows that the set A is convex. Furthermore, the closure of A is convex (Problem 5.18), and the probability of the boundary of A is 0.

Now consider the other condition of Theorem 5.6.5. Suppose A is disjoint with the half-space

$$c < \omega'y = \nu'z_N - \tfrac{1}{2}\operatorname{tr}\Lambda B, \tag{21}$$

where Λ is a symmetric matrix and B is positive semidefinite. We shall take $\Lambda_1 = I$. We want to show that $\boldsymbol{\omega}_1 + \lambda\boldsymbol{\omega} \in \Omega - \Omega_0$; that is, that $\boldsymbol{\nu}_1 + \lambda\boldsymbol{\nu} \neq 0$ (which is trivial) and $\Lambda_1 + \lambda\Lambda$ is positive definite for $\lambda > 0$. This is the case when Λ is positive semidefinite. Now we shall show that a half-space (21) disjoint with A and Λ not positive semidefinite implies a contradiction. If Λ is not positive semidefinite, it can be written (by Corollary A.4.1 of the Appendix)

$$\Lambda = D\begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} D', \tag{22}$$

where D is nonsingular. If Λ is not positive semidefinite, $-I$ is not vacuous, because its order is the number of negative characteristic roots of Λ. Let $z_N = (1/\gamma)z_0$ and

$$B = (D')^{-1}\begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \gamma I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix} D^{-1}. \tag{23}$$

Then

$$\boldsymbol{\omega}'\boldsymbol{y} = \frac{1}{\gamma}\boldsymbol{\nu}'z_0 + \frac{1}{2}\operatorname{tr}\begin{bmatrix} -I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \gamma I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{24}$$

which is greater than c for sufficiently large γ. On the other hand

$$z_N' B^{-1} z_N = \frac{1}{\gamma^2} z_0' D\begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \gamma^{-1} I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix} D' z_0, \tag{25}$$

which is less than k for sufficiently large γ. This contradicts the fact that (20) and (21) are disjoint. Thus the conditions of Theorem 5.6.5 are satisfied and the theorem is proved. ■

This proof is due to Stein.

An alternative proof of admissibility is to show that the T^2-test is a proper Bayes procedure. Suppose an arbitrary random vector X has density $f(\boldsymbol{x}|\boldsymbol{\omega})$ for $\boldsymbol{\omega} \in \Omega$. Consider testing the null hypothesis $H_0 : \boldsymbol{\omega} \in \Omega_0$ against the alternative $H_1 : \boldsymbol{\omega} \in \Omega - \Omega_0$. Let Π_0 be a prior finite measure on Ω_0, and Π_1 a prior finite measure on Ω_1. Then the Bayes procedure (with 0–1 loss function) is to reject H_0 if

$$\frac{\int f(x|\omega)\Pi_1(d\omega)}{\int f(x|\omega)\Pi_0(d\omega)} \geq c \tag{26}$$

for some c $(0 \leq c \leq \infty)$. If equality in (26) occurs with probability 0 for all $\omega \in \Omega_0$, then the Bayes procedure is unique and hence admissible. Since the measures are finite, they can be normed to be probability measures. For the T^2-test of $H_0: \boldsymbol{\mu} = \mathbf{0}$ a pair of measures is suggested in Problem 5.15. (This pair is not unique.) The reader can verify that with these measures (26) reduces to the complement of (20).

Among invariant tests it was shown that the T^2-test is uniformly most powerful; that is, it is most powerful against every value of $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ among invariant tests of the specified significance level. We can ask whether the T^2-test is "best" against a specified value of $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ among all tests. Here "best" can be taken to mean admissible *minimax*; and "minimax" means maximizing with respect to procedures the minimum with respect to parameter values of the power. This property was shown in the simplest case of $p = 2$ and $N = 3$ by Giri, Kiefer, and Stein (1963). The property for general p and N was announced by Šalaevskiĭ (1968). He has furnished a proof for the case of $p = 2$ [Šalaevskiĭ (1971)], but has not given a proof for $p > 2$.

Giri and Kiefer (1964) have proved the T^2-test is locally minimax (as $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \to 0$) and asymptotically (logarithmically) minimax as $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \to \infty$.

5.7. ELLIPTICALLY CONTOURED DISTRIBUTIONS

5.7.1. Observations Elliptically Contoured

When $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ constitute a sample of N from

$$|\boldsymbol{\Lambda}|^{-\frac{1}{2}} g\left[(\boldsymbol{x}-\boldsymbol{\nu})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{\nu})\right], \tag{1}$$

the sample mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ are unbiased estimators of the distribution mean $\boldsymbol{\mu} = \boldsymbol{\nu}$ and covariance matrix $\boldsymbol{\Sigma} = (\mathscr{E}R^2/p)\boldsymbol{\Lambda}$, where $R^2 = (\boldsymbol{X}-\boldsymbol{\nu})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{X}-\boldsymbol{\nu})$ has finite expectation. The T^2-statistic, $T^2 = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})$, can be used for tests and confidence regions for $\boldsymbol{\mu}$ when $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Lambda}$) is unknown, but the small-sample distribution of T^2 in general is difficult to obtain. However, the limiting distribution of T^2 when $N \to \infty$ is obtained from the facts that $\sqrt{N}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{S} \xrightarrow{p} \boldsymbol{\Sigma}$ (Theorem 3.6.2).

Theorem 5.7.1. *Let $x_1, \ldots, x_N$ be a sample from* (1). *Assume $\mathscr{E}R^2 < \infty$. Then $T^2 \xrightarrow{d} \chi_p^2$.*

Proof. Theorem 3.6.2 implies that $N(\bar{x} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{x} - \boldsymbol{\mu}) \xrightarrow{d} \chi_p^2$ and $N(\bar{x} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{x} - \boldsymbol{\mu}) - T^2 \xrightarrow{p} 0$. ■

Theorem 5.7.1 implies that the procedures in Section 5.3 can be done on an asymptotic basis for elliptically contoured distributions. For example, to test the null hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, reject the null hypothesis if

$$N(\bar{x} - \boldsymbol{\mu}_0)'S^{-1}(\bar{x} - \boldsymbol{\mu}_0) \geq \chi_p^2(\alpha), \tag{2}$$

where $\chi_p^2(\alpha)$ is the α-significance point of the χ^2-distribution with p degrees of freedom the limiting probability of (2) when the null hypothesis is true and $N \to \infty$ is α. Similarly the confidence region $N(\bar{x} - m)'S^{-1}(\bar{x} - m) \leq \chi_p^2(\alpha)$ has limiting confidence $1 - \alpha$.

5.7.2. Elliptically Contoured Matrix Distributions

Let X $(N \times p)$ have the density

$$|C|^{-N} g\left[C^{-1}(X - \varepsilon_N \boldsymbol{\nu}')'(X - \varepsilon_N \boldsymbol{\nu}')(C')^{-1}\right] \tag{3}$$

based on the left spherical density $g(Y'Y)$. Here Y has the representation $Y \stackrel{d}{=} UR'$, where U $(N \times p)$ has the uniform distribution on $O(N \times p)$, R is lower triangular, and U and R are independent. Then $X \stackrel{d}{=} \varepsilon_N \boldsymbol{\nu}' + UR'C'$. The T^2-criterion to test the hypothesis $\boldsymbol{\nu} = \mathbf{0}$ is $N\bar{x}'S^{-1}\bar{x}$, which is invariant with respect to transformations $X \to XG$. By Corollary 4.5.5 we obtain the following theorem.

Theorem 5.7.2. *Suppose X has the density* (3) *with $\boldsymbol{\nu} = \mathbf{0}$ and $T^2 = N\bar{x}'S^{-1}\bar{x}$. Then $[T^2/(N-1)][(N-p)/p]$ has the distribution of $F_{p,N-p} = (\chi_p^2/p)/[\chi_{N-p}^2/(N-p)]$.*

Thus the tests of hypotheses and construction of confidence regions at stated significance and confidence levels are valid for left spherical distributions.

The T^2-criterion for $H: \boldsymbol{\nu} = \mathbf{0}$ is

$$T^2 = N\bar{x}'S^{-1}\bar{x} \stackrel{d}{=} N\bar{u}'S_u^{-1}\bar{u}, \tag{4}$$

since $X \stackrel{d}{=} UR'C'$,

$$\bar{x}' = \frac{1}{N}\varepsilon_N' X \stackrel{d}{=} \left(\frac{1}{N}\varepsilon_N' U\right)R'C' = \bar{u}'(CR)', \tag{5}$$

and

$$(6)\quad S = \frac{1}{N-1}(X'X - N\bar{x}\bar{x}') = \frac{1}{N-1}[CRU'URC' - CR\bar{u}\bar{u}'(C'R)']$$
$$= CRS_u(CR)'.$$

5.7.3. Linear Combinations

Läuter, Glimm, and Kropf (1996a, 1996b, 1996c) have observed that a statistician can use $X'X = CRR'C'$ when $\nu = 0$ to determine a $p \times q$ matrix D and base a T-test on the transform $Z = XD$. Specifically, define

$$(7)\qquad \bar{z}' = \frac{1}{N}\varepsilon_N' Z = \bar{x}'D,$$

$$(8)\qquad S_Z = \frac{1}{N-1}(Z'Z - N\bar{z}\bar{z}') = D'SD,$$

$$(9)\qquad T_D^2 = N\bar{z}'S_Z^{-1}\bar{z}'.$$

Since $Q_N Z \overset{d}{=} Q_N UR'C' \overset{d}{=} UR'C' = Z$, the matrix Z is based on the left-spherical YD and hence has the representation $Z = VR^{*\prime}$, where V $(N \times q)$ has the uniform distribution on $O(N \times p)$, independent of $R^{*\prime}$ (upper triangular) having the distribution derived from $R^* R^{*\prime} = Z'Z$. The distribution of $T^2/(N-1)$ is $F_{q, N-q}\, q/(N-q)$.

The matrix D can also involve prior information as well as knowledge of $X'X$. If p is large, q can be small; the power of the test based on T_D^2 may be more powerful than a test based on T^2.

Läuter, Glimm, and Kropf give several examples of choosing D. One of them is to chose D $(p \times 1)$ as $[\mathrm{Diag}(X'X)^{-\frac{1}{2}}]\varepsilon_p$, where Diag A is a diagonal matrix with ith diagonal element a_{ii}. The statistic T_D^2 is called the *standardized sum* statistic.

PROBLEMS

5.1. (Sec. 5.2) Let x_α be distributed according to $N(\mu + \beta(z_\alpha - \bar{z}), \Sigma)$, $\alpha = 1, \dots, N$, where $\bar{z} = (1/N)\Sigma z_\alpha$. Let $b = [1/\Sigma(z_\alpha - \bar{z})^2]\Sigma x_\alpha(z_\alpha - \bar{z})$, $(N-2)S = \Sigma[x_\alpha - \bar{x} - b(z_\alpha - \bar{z})][x_\alpha - \bar{x} - b(z_\alpha - \bar{z})]'$, and $T^2 = \Sigma(z_\alpha - \bar{z})^2 b'S^{-1}b$. Show that T^2 has the T^2-distribution with $N-2$ degrees of freedom. [*Hint:* See Problem 3.13.]

5.2. (Sec. 5.2.2) Show that $T^2/(N-1)$ can be written as $R^2/(1-R^2)$ with the correspondences given in Table 5.1.

Table 5.1

Section 5.2	Section 4.4
$x_{0\alpha} = 1/\sqrt{N}$	$z_{1\alpha}$
x_α	$z_\alpha^{(2)}$
$\sqrt{N}\bar{x}$	$a_{(1)} = \Sigma z_{1\alpha} z_\alpha^{(2)}$
$B = \Sigma x_\alpha x_\alpha'$	$A_{22} = \Sigma z_\alpha^{(2)} z_\alpha^{(2)\prime}$
$1 = \Sigma x_{0\alpha}^2$	$a_{11} = \Sigma z_{1\alpha}^2$
$\dfrac{T^2}{N-1}$	$\dfrac{R^2}{1-R^2}$
p	$p-1$
N	n

5.3. (Sec. 5.2.2) Let

$$\frac{R^2}{1-R^2} = \frac{\Sigma u_\alpha x_\alpha'(\Sigma x_\alpha x_\alpha')^{-1}\Sigma u_\alpha x_\alpha}{\Sigma u_\alpha^2 - \Sigma u_\alpha x_\alpha'(\Sigma x_\alpha x_\alpha')^{-1}\Sigma u_\alpha x_\alpha},$$

where $u_1, \dots, u_N$ are N numbers and $x_1, \dots, x_N$ are independent, each with the distribution $N(\mathbf{0}, \Sigma)$. Prove that the distribution of $R^2/(1-R^2)$ is independent of $u_1, \dots, u_N$. [*Hint:* There is an orthogonal $N \times N$ matrix C that carries $(u_1, \dots, u_N)$ into a vector proportional to $(1/\sqrt{N}, \dots, 1/\sqrt{N})$.]

5.4. (Sec. 5.2.2) Use Problems 5.2 and 5.3 to show that $[T^2/(N-1)][(N-p)/p]$ has the $F_{p,N-p}$-distribution (under the null hypothesis). [*Note:* This is the analysis that corresponds to Hotelling's geometric proof (1931).]

5.5. (Sec. 5.2.2) Let $T^2 = N\bar{x}'S^{-1}\bar{x}$, where $\bar{x}$ and S are the mean vector and covariance matrix of a sample of N from $N(\mu, \Sigma)$. Show that T^2 is distributed the same when μ is replaced by $\lambda = (\tau, 0, \dots, 0)'$, where $\tau^2 = \mu'\Sigma^{-1}\mu$, and Σ is replaced by I.

5.6. (Sec. 5.2.2) Let $u = [T^2/(N-1)]/[1 + T^2/(N-1)]$. Show that $u = \gamma V'(VV')^{-1}V\gamma'$, where $\gamma = (1/\sqrt{N}, \dots, 1/\sqrt{N})$ and

$$V = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pN} \end{pmatrix}.$$

5.7. (Sec. 5.2.2) Let

$$\boldsymbol{\nu}_1^* = \boldsymbol{\nu}_1,$$

$$\boldsymbol{\nu}_i^* = \boldsymbol{\nu}_i - \frac{\boldsymbol{\nu}_i \boldsymbol{\nu}_1'}{\boldsymbol{\nu}_1 \boldsymbol{\nu}_1'} \boldsymbol{\nu}_1 = \boldsymbol{\nu}_1 \left(\boldsymbol{I} - \frac{1}{\boldsymbol{\nu}_1 \boldsymbol{\nu}_1'} \boldsymbol{\nu}_1' \boldsymbol{\nu}_1 \right), \qquad i \neq 1,$$

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma} - \frac{\boldsymbol{\gamma} \boldsymbol{\nu}_1'}{\boldsymbol{\nu}_1 \boldsymbol{\nu}_1'} \boldsymbol{\nu}_1,$$

$$V^* = \begin{pmatrix} \boldsymbol{\nu}_1^* \\ \vdots \\ \boldsymbol{\nu}_p^* \end{pmatrix}.$$

Prove that $U = s + (1 - s)w$, where

$$s = \frac{(\boldsymbol{\gamma} \boldsymbol{\nu}_1^{*\prime})^2}{\boldsymbol{\nu}_1^* \boldsymbol{\nu}_1^{*\prime}} = \frac{(\boldsymbol{\gamma} \boldsymbol{\nu}_1')^2}{\boldsymbol{\nu}_1 \boldsymbol{\nu}_1'},$$

$$w = \frac{1}{\boldsymbol{\gamma}^* \boldsymbol{\gamma}^{*\prime}} \boldsymbol{\gamma}^* \begin{pmatrix} \boldsymbol{\nu}_2^* \\ \vdots \\ \boldsymbol{\nu}_p^* \end{pmatrix}' \begin{pmatrix} \boldsymbol{\nu}_2^* \boldsymbol{\nu}_2^{*\prime} & \cdots & \boldsymbol{\nu}_2^* \boldsymbol{\nu}_p^{*\prime} \\ \vdots & & \vdots \\ \boldsymbol{\nu}_p^* \boldsymbol{\nu}_2^{*\prime} & \cdots & \boldsymbol{\nu}_p^* \boldsymbol{\nu}_p^{*\prime} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\nu}_2^* \\ \vdots \\ \boldsymbol{\nu}_p^* \end{pmatrix} \boldsymbol{\gamma}^{*\prime}.$$

Hint: $EV = V^*$, where

$$E = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\dfrac{\boldsymbol{\nu}_2 \boldsymbol{\nu}_1'}{\boldsymbol{\nu}_1 \boldsymbol{\nu}_1'} & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ -\dfrac{\boldsymbol{\nu}_p \boldsymbol{\nu}_1'}{\boldsymbol{\nu}_1 \boldsymbol{\nu}_1'} & 0 & \cdots & 1 \end{pmatrix}.$$

5.8. (Sec. 5.2.2) Prove that w has the distribution of the square of a multiple correlation between one vector and $p - 1$ vectors in $(N - 1)$-space without subtracting means; that is, it has density

$$\frac{\Gamma[\frac{1}{2}(N-1)]}{\Gamma[\frac{1}{2}(N-p)]\Gamma[\frac{1}{2}(p-1)]} w^{\frac{1}{2}(p-1)-1}(1-w)^{\frac{1}{2}(N-p)-1}.$$

[*Hint:* The transformation of Problem 5.7 is a projection of $\boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_p, \boldsymbol{\gamma}$ on the $(N - 1)$-space orthogonal to $\boldsymbol{\nu}_1$.]

5.9. (Sec. 5.2.2) Verify that $r = s/(1 - s)$ multiplied by $(N - 1)/1$ has the noncentral F-distribution with 1 and $N - 1$ degrees of freedom and noncentrality parameter $N\tau^2$.

5.10. (Sec. 5.2.2) From Problems 5.5–5.9, verify Corollary 5.2.1.

5.11. (Sec. 5.3) Use the data in Section 3.2 to test the hypothesis that neither drug has a soporific effect at significance level 0.01.

5.12. (Sec. 5.3) Using the data in Section 3.2, give a confidence region for $\boldsymbol{\mu}$ with confidence coefficient 0.95.

5.13. (Sec. 5.3) Prove the statement in Section 5.3.6 that the T^2-statistic is independent of the choice of C.

5.14. (Sec. 5.5) Use the data of Problem 4.41 to test the hypothesis that the mean head length and breadth of first sons are equal to those of second sons at significance level 0.01.

5.15. (Sec. 5.6.2) *T^2-test as a Bayes procedure* [Kiefer and Schwartz (1965)]. Let $x_1, \dots, x_N$ be independently distributed, each according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let Π_0 be defined by $[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = [\mathbf{0}, (I + \boldsymbol{\eta\eta}')^{-1}]$ with $\boldsymbol{\eta}$ having a density proportional to $|I + \boldsymbol{\eta\eta}'|^{-\frac{1}{2}N}$, and let Π_1 be defined by $[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = [(I + \boldsymbol{\eta\eta}')^{-1}\boldsymbol{\eta}, (I + \boldsymbol{\eta\eta}')^{-1}]$ with $\boldsymbol{\eta}$ having a density proportional to

$$|I + \boldsymbol{\eta\eta}'|^{-\frac{1}{2}N} \exp\left[\tfrac{1}{2}N\boldsymbol{\eta}'(I + \boldsymbol{\eta\eta}')^{-1}\boldsymbol{\eta}\right].$$

(a) Show that the measures are finite for $N > p$ by showing $\boldsymbol{\eta}'(I + \boldsymbol{\eta\eta}')^{-1}\boldsymbol{\eta} \le 1$ and verifying that the integral of $|I + \boldsymbol{\eta\eta}'|^{-\frac{1}{2}N} = (1 + \boldsymbol{\eta\eta}')^{-\frac{1}{2}N}$ is finite.

(b) Show that the inequality (26) is equivalent to $N\bar{x}'(\sum_{\alpha=1}^{N} x_\alpha x_\alpha')^{-1}\bar{x} \ge k$. Hence the T^2-test is Bayes and thus admissible.

5.16. (Sec. 5.6.2) Let $g(t) = f[ty_1 + (1-t)y_2]$, where $f(y)$ is a real-valued function of the vector y. Prove that if $g(t)$ is convex, then $f(y)$ is convex.

5.17. (Sec. 5.6.2) Show that $z'B^{-1}z$ is a convex function of (z, B), where B is a positive definite matrix. [*Hint:* Use Problem 5.16.]

5.18. (Sec. 5.6.2) Prove that if the set A is convex, then the closure of A is convex.

5.19. (Sec. 5.3) Let $\bar{x}$ and S be based on N observations from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let x be an additional observation from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that $x - \bar{x}$ is distributed according to

$$N[\mathbf{0}, (1 + 1/N)\boldsymbol{\Sigma}].$$

Verify that $[N/(N+1)](x-\bar{x})'S^{-1}(x-\bar{x})$ has the T^2-distribution with $N-1$ degrees of freedom. Show how this statistic can be used to give a prediction region for x based on $\bar{x}$ and S (i.e., a region such that one has a given confidence that the next observation will fall into it).

5.20. (Sec. 5.3) Let $x_\alpha^{(i)}$ be observations from $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_i)$, $\alpha = 1, \ldots, N_i$, $i = 1, 2$. Find the likelihood ratio criterion for testing the hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$.

5.21. (Sec. 5.4) Prove that $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is larger for $\boldsymbol{\mu}' = (\mu_1, \mu_2)$ than for $\boldsymbol{\mu} = \mu_1$ by verifying

$$\frac{1}{1-\rho^2}\left(\frac{\mu_1^2}{\sigma_1^2} - 2\rho\frac{\mu_1\mu_2}{\sigma_1\sigma_2} + \frac{\mu_2^2}{\sigma_2^2}\right) = \frac{\mu_1^2}{\sigma_1^2} + \frac{(\mu_2 - \rho\sigma_2\mu_1/\sigma_1)^2}{(1-\rho^2)\sigma_2^2}.$$

Discuss the power of the test $\mu_1 = 0$ compared to the power of the test $\mu_1 = 0$, $\mu_2 = 0$.

5.22. (Sec. 5.3)

(a) Using the data of Section 5.3.4, test the hypothesis $\mu_1^{(1)} = \mu_1^{(2)}$.
(b) Test the hypothesis $\mu_1^{(1)} = \mu_1^{(2)}$, $\mu_2^{(1)} = \mu_2^{(2)}$.

5.23. (Sec. 5.4) Let

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Prove $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \geq \boldsymbol{\mu}^{(1)\prime}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}^{(1)}$. Give a condition for strict inequality to hold. [*Hint:* This is the vector analog of Problem 5.21.]

5.24. Let $X^{(i)\prime} = (Y^{(i)\prime}, Z^{(i)\prime})$, $i = 1, 2$, where $Y^{(i)}$ has p components and $Z^{(i)}$ has q components, be distributed according to $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu}^{(i)} = \begin{pmatrix} \boldsymbol{\mu}_y^{(i)} \\ \boldsymbol{\mu}_z^{(i)} \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yz} \\ \boldsymbol{\Sigma}_{zy} & \boldsymbol{\Sigma}_{zz} \end{pmatrix}, \qquad i = 1, 2.$$

Find the likelihood ratio criterion (or equivalent T^2-criterion) for testing $\boldsymbol{\mu}_z^{(1)} = \boldsymbol{\mu}_z^{(2)}$ given $\boldsymbol{\mu}_y^{(1)} = \boldsymbol{\mu}_y^{(2)}$ on the basis of a sample of N_i on $X^{(i)}$, $i = 1, 2$. [*Hint:* Express the likelihood in terms of the marginal density of $Y^{(i)}$ and the conditional density of $Z^{(i)}$ given $Y^{(i)}$.]

5.25. Find the distribution of the criterion in the preceding problem under the null hypothesis.

5.26. (Sec. 5.5) Suppose $x_\alpha^{(g)}$ is an observation from $N(\boldsymbol{\mu}^{(g)}, \boldsymbol{\Sigma}_g)$, $\alpha = 1, \ldots, N_g$, $g = 1, \ldots, q$.

(a) Show that the hypothesis $\boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(q)}$ is equivalent to $\mathscr{E}\mathbf{y}_\alpha^{(i)} = \mathbf{0}$, $i = 1, \dots, q-1$, where

$$\mathbf{y}_\alpha^{(i)} = a_1^{(i)} \mathbf{x}_\alpha^{(1)} + \sum_{g=2}^{q} a_g^{(i)} \left(\frac{N_1}{N_g} \right)^{\frac{1}{2}} \left[\mathbf{x}_\alpha^{(g)} - \frac{1}{N_1} \sum_{\beta=1}^{N_1} \mathbf{x}_\beta^{(g)} + \frac{1}{(N_1 N_g)^{\frac{1}{2}}} \sum_{\beta=1}^{N_g} \mathbf{x}_\beta^{(g)} \right],$$

$$\alpha = 1, \dots, N_1, \quad i = 1, \dots, q-1;$$

$N_1 \le N_g$, $g = 2, \dots, q$; and $(a_1^{(i)}, \dots, a_q^{(i)})$, $i = 1, \dots, q-1$, are linearly independent.

(b) Show how to construct a T^2-test of the hypothesis using $(\bar{\mathbf{y}}^{(1)\prime}, \dots, \bar{\mathbf{y}}^{(q-1)\prime})'$ yielding an F-statistic with $(q-1)p$ and $N-(q-1)p$ degrees of freedom [Anderson (1963b)].

5.27. (Sec. 5.2) Prove (25) is the density of $V = \chi_a^2/(\chi_a^2 + \chi_b^2)$. [*Hint:* In the joint density of $U = \chi_a^2$ and $W = \chi_b^2$ make the transformation $u = vw(1-v)^{-1}$, $w = w$ and integrate out w.]

CHAPTER 6

Classification of Observations

6.1. THE PROBLEM OF CLASSIFICATION

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several categories on the basis of these measurements. The investigator cannot identify the individual with a category directly but must use these measurements. In many cases it can be assumed that there are a finite number of categories or populations from which the individual may have come and each population is characterized by a probability distribution of the measurements. Thus an individual is considered as a random observation from this population. The question is: Given an individual with certain measurements, from which population did the person arise?

The problem of classification may be considered as a problem of "statistical decision functions." We have a number of hypotheses: Each hypothesis is that the distribution of the observation is a given one. We must accept one of these hypotheses and reject the others. If only two populations are admitted, we have an elementary problem of testing one hypothesis of a specified distribution against another.

In some instances, the categories are specified beforehand in the sense that the probability distributions of the measurements are assumed completely known. In other cases, the form of each distribution may be known, but the parameters of the distribution must be estimated from a sample from that population.

Let us give an example of a problem of classification. Prospective students applying for admission into college are given a battery of tests; the vector of

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

scores is a set of measurements x. The prospective student may be a member of one population consisting of those students who will successfully complete college training or, rather, have potentialities for successfully completing training, or the student may be a member of the other population, those who will not complete the college course successfully. The problem is to classify a student applying for admission on the basis of his scores on the entrance examination.

In this chapter we shall develop the theory of classification in general terms and then apply it to cases involving the normal distribution. In Section 6.2 the problem of classification with two populations is defined in terms of decision theory, and in Section 6.3 Bayes and admissible solutions are obtained. In Section 6.4 the theory is applied to two known normal populations, differing with respect to means, yielding the population linear discriminant function. When the parameters are unknown, they are replaced by estimates (Section 6.5). An alternative procedure is maximum likelihood. In Section 6.6 the probabilities of misclassification by the two methods are evaluated in terms of asymptotic expansions of the distributions. Then these developments are carried out for several populations. Finally, in Section 6.10 linear procedures for the two populations are studied when the covariance matrices are different and the parameters are known.

6.2. STANDARDS OF GOOD CLASSIFICATION

6.2.1. Preliminary Considerations

In constructing a procedure of classification, it is desired to minimize the probability of misclassification, or, more specifically, it is desired to minimize on the average the bad effects of misclassification. Now let us make this notion precise. For convenience we shall now consider the case of only two categories. Later we shall treat the more general case. This section develops the ideas of Section 3.4 in more detail for the problem of two decisions.

Suppose an individual is an observation from either population π_1 or population π_2. The classification of an observation depends on the vector of measurements $x' = (x_1, \dots, x_p)$ on that individual. We set up a rule that if an individual is characterized by certain sets of values of $x_1, \dots, x_p$ that person will be classified as from π_1, if other values, as from π_2.

We can think of an observation as a point in a p-dimensional space. We divide this space into two regions. If the observation falls in R_1, we classify it as coming from population π_1, and if it falls in R_2 we classify it as coming from population π_2.

In following a given classification procedure, the statistician can make two kinds of errors in classification. If the individual is actually from π_1, the

Table 6.1

		Statistician's Decision	
		π_1	π_2
Population	π_1	0	$C(2\|1)$
	π_2	$C(1\|2)$	0

statistician can classify him or her as coming from population π_2; if from π_2, the statistician can classify him or her as from π_1. We need to know the relative undesirability of these two kinds of misclassification. Let the *cost* of the first type of misclassification be $C(2|1)$ (>0), and let the cost of misclassifying an individual from π_2 as from π_1 be $C(1|2)$ (>0). These costs may be measured in any kind of units. As we shall see later, it is only the ratio of the two costs that is important. The statistician may not know these costs in each case, but will often have at least a rough idea of them.

Table 6.1 indicates the costs of correct and incorrect classification. Clearly, a good classification procedure is one that minimizes in some sense or other the cost of misclassification.

6.2.2. Two Cases of Two Populations

We shall consider ways of defining "minimum cost" in two cases. In one case we shall suppose that we have a priori probabilities of the two populations. Let the probability that an observation comes from population π_1 be q_1 and from population π_2 be q_2 ($q_1 + q_2 = 1$). The probability properties of population π_1 are specified by a distribution function. For convenience we shall treat only the case where the distribution has a density, although the case of discrete probabilities lends itself to almost the same treatment. Let the density of population π_1 be $p_1(x)$ and that of π_2 be $p_2(x)$. If we have a region R_1 of classification as from π_1, the probability of correctly classifying an observation that actually is drawn from population π_1 is

$$P(1|1,R) = \int_{R_1} p_1(x)\,dx, \tag{1}$$

where $dx = dx_1 \cdots dx_p$, and the probability of misclassification of an observation from π_1 is

$$P(2|1,R) = \int_{R_2} p_1(x)\,dx. \tag{2}$$

Similarly, the probability of correctly classifying an observation from π_2 is

$$P(2|2,R) = \int_{R_2} p_2(x)\,dx, \tag{3}$$

and the probability of misclassifying such an observation is

$$P(1|2, R) = \int_{R_1} p_2(x)\, dx. \tag{4}$$

Since the probability of drawing an observation from π_1 is q_1, the probability of drawing an observation from π_1 and correctly classifying it is $q_1 P(1|1, R)$; that is, this is the probability of the situation in the upper left-hand corner of Table 6.1. Similarly, the probability of drawing an observation from π_1 and misclassifying it is $q_1 P(2|1, R)$. The probability associated with the lower left-hand corner of Table 6.1 is $q_2 P(1|2, R)$, and with the lower right-hand corner is $q_2 P(2|2, R)$.

What is the average or expected loss from costs of misclassification? It is the sum of the products of costs of misclassifications with their respective probabilities of occurrence:

$$C(2|1)P(2|1, R)q_1 + C(1|2)P(1|2, R)q_2. \tag{5}$$

It is this average loss that we wish to minimize. That is, we want to divide our space into regions R_1 and R_2 such that the expected loss is as small as possible. A procedure that minimizes (5) for given q_1 and q_2 is called a *Bayes procedure*.

In the example of admission of students, the undesirability of misclassification is, in one instance, the expense of teaching a student who will not complete the course successfully and is, in the other instance, the undesirability of excluding from college a potentially good student.

The other case we shall treat is that in which there are no known a priori probabilities. In this case the expected loss if the observation is from π_1 is

$$C(2|1)P(2|1, R) = r(1, R); \tag{6}$$

the expected loss if the observation is from π_2 is

$$C(1|2)P(1|2, R) = r(2, R). \tag{7}$$

We do not know whether the observation is from π_1 or from π_2, and we do not know probabilities of these two instances.

A procedure R is at least as good as a procedure R^* if $r(1, R) \leq r(1, R^*)$ and $r(2, R) \leq r(2, R^*)$; R is better than R^* if at least one of these inequalities is a strict inequality. Usually there is no one procedure that is better than all other procedures or is at least as good as all other procedures. A procedure R is called *admissible* if there is no procedure better than R; we shall be interested in the entire class of admissible procedures. It will be shown that under certain conditions this class is the same as the class of Bayes procedures. A class of procedures is *complete* if for every procedure outside the class there is one in the class which is better; a class is called *essentially complete* if for every procedure outside the class there is one in the class which is at least as good. A *minimal complete class* (if it exists) is a complete class such that no proper subset is a complete class; a similar definition holds for a minimal essentially complete class. Under certain conditions we shall show that the admissible class is minimal complete. To simplify the discussion we shall consider procedures the same if they only differ on sets of probability zero. In fact, throughout the next section we shall make statements which are meant to hold *except for sets of probability zero* without saying so explicitly.

A principle that usually leads to a unique procedure is the minimax principle. A procedure is *minimax* if the maximum expected loss, $r(i, R)$, is a minimum. From a conservative point of view, this may be considered an optimum procedure. For a general discussion of the concepts in this section and the next see Wald (1950), Blackwell and Girshick (1954), Ferguson (1967), DeGroot (1970), and Berger (1980b).

6.3. PROCEDURES OF CLASSIFICATION INTO ONE OF TWO POPULATIONS WITH KNOWN PROBABILITY DISTRIBUTIONS

6.3.1. The Case When A Priori Probabilities Are Known

We now turn to the problem of choosing regions R_1 and R_2 so as to minimize (5) of Section 6.2. Since we have a priori probabilities, we can define joint probabilities of the population and the observed set of variables. The probability that an observation comes from π_1 and that each variate is less than the corresponding component in y is

$$\int_{-\infty}^{y_p} \cdots \int_{-\infty}^{y_1} q_1 p_1(x)\, dx_1 \cdots dx_p. \tag{1}$$

We can also define the conditional probability that an observation came from a certain population given the values of the observed variates. For instance, the conditional probability of coming from population π_1, given an observation x, is

$$\frac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)}. \tag{2}$$

Suppose for a moment that $C(1|2) = C(2|1) = 1$. Then the expected loss is

$$q_1 \int_{R_2} p_1(x)\, dx + q_2 \int_{R_1} p_2(x)\, dx. \tag{3}$$

This is also the probability of a misclassification; hence we wish to minimize the probability of misclassification.

For a given observed point x we minimize the probability of a misclassification by assigning the population that has the higher conditional probability. If

$$\frac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)} \geq \frac{q_2 p_2(x)}{q_1 p_1(x) + q_2 p_2(x)}, \tag{4}$$

we choose population π_1. Otherwise we choose population π_2. Since we minimize the probability of misclassification at each point, we minimize it over the whole space. Thus the rule is

$$\begin{aligned} &R_1\colon q_1 p_1(x) \geq q_2 p_2(x), \\ &R_2\colon q_1 p_1(x) < q_2 p_2(x). \end{aligned} \tag{5}$$

If $q_1 p_1(x) = q_2 p_2(x)$, the point could be classified as either from π_1 or π_2; we have arbitrarily put it into R_1. If $q_1 p_1(x) + q_2 p_2(x) = 0$ for a given x, that point also may go into either region.

Now let us prove formally that (5) is the best procedure. For any procedure $R^* = (R_1^*, R_2^*)$, the probability of misclass'fication is

$$\begin{aligned} &q_1 \int_{R_2^*} p_1(x)\,dx + q_2 \int_{R_1^*} p_2(x)\,dx \\ &\quad = \int_{R_2^*} [q_1 p_1(x) - q_2 p_2(x)]\,dx + q_2 \int p_2(x)\,dx. \end{aligned} \tag{6}$$

On the right-hand side the second term is a given number; the first term is minimized if R_2^* includes the points x such that $q_1 p_1(x) - q_2 p_2(x) < 0$ and excludes the points for which $q_1 p_1(x) - q_2 p_2(x) > 0$. If

$$\Pr\left\{ \frac{p_1(x)}{p_2(x)} = \frac{q_2}{q_1} \,\middle|\, \pi_i \right\} = 0, \qquad i = 1, 2. \tag{7}$$

then the Bayes procedure is unique except for sets of probability zero.

Now we notice that mathematically the problem was: given nonnegative constants q_1 and q_2 and nonnegative functions $p_1(x)$ and $p_2(x)$, choose regions R_1 and R_2 so as to minimize (3). The solution is (5). If we wish to minimize (5) of Section 6.2, which can be written

$$[C(2|1)q_1] \int_{R_2} p_1(x)\,dx + [C(1|2)q_2] \int_{R_1} p_2(x)\,dx, \tag{8}$$

we choose R_1 and R_2 according to

$$
\begin{aligned}
(9)\qquad & R_1\colon [C(2|1)q_1]\,p_1(\boldsymbol{x}) \ge [C(1|2)q_2]\,p_2(\boldsymbol{x}),\\
& R_2\colon [C(2|1)q_1]\,p_1(\boldsymbol{x}) < [C(1|2)q_2]\,p_2(\boldsymbol{x}),
\end{aligned}
$$

since $C(2|1)q_1$ and $C(1|2)q_2$ are nonnegative constants. Another way of writing (9) is

$$
\begin{aligned}
(10)\qquad & R_1\colon \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} \ge \frac{C(1|2)q_2}{C(2|1)q_1},\\
& R_2\colon \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} < \frac{C(1|2)q_2}{C(2|1)q_1}.
\end{aligned}
$$

Theorem 6.3.1. *If q_1 and q_2 are a priori probabilities of drawing an observation from population π_1 with density $p_1(\boldsymbol{x})$ and π_2 with density $p_2(\boldsymbol{x})$, respectively, and if the cost of misclassifying an observation from π_1 as from π_2 is $C(2|1)$ and an observation from π_2 as from π_1 is $C(1|2)$, then the regions of classification R_1 and R_2, defined by* (10), *minimize the expected cost. If*

$$
(11)\qquad \Pr\left\{ \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} = \frac{q_2C(1|2)}{q_1C(2|1)} \,\middle|\, \pi_i \right\} = 0, \qquad i = 1,2,
$$

then the procedure is unique except for sets of probability zero.

6.3.2. The Case When No Set of A Priori Probabilities Is Known

In many instances of classification the statistician cannot assign a priori probabilities to the two populations. In this case we shall look for the class of admissible procedures, that is, the set of procedures that cannot be improved upon.

First, let us prove that a Bayes procedure is admissible. Let $R = (R_1, R_2)$ be a Bayes procedure for a given q_1, q_2; is there a procedure $R^* = (R_1^*, R_2^*)$ such that $P(1|2, R^*) \le P(1|2, R)$ and $P(2|1, R^*) \le P(2|1, R)$ with at least one strict inequality? Since R is a Bayes procedure,

$$
(12)\qquad q_1P(2|1, R) + q_2P(1|2, R) \le q_1P(2|1, R^*) + q_2P(1|2, R^*).
$$

This inequality can be written

$$
(13)\qquad q_1[P(2|1, R) - P(2|1, R^*)] \le q_2[P(1|2, R^*) - P(1|2, R)].
$$

Suppose $0 < q_1 < 1$. Then if $P(1|2, R^*) < P(1|2, R)$, the right-hand side of (13) is less than zero and therefore $P(2|1, R) < P(2|1, R^*)$. Then $P(2|1, R^*) < P(2|1, R)$ similarly implies $P(1|2, R) < P(1|2, R^*)$. Thus R^* is not better than R, and R is admissible. If $q_1 = 0$, then (13) implies $0 \le P(1|2, R^*) - P(1|2, R)$. For a Bayes procedure, R_1 includes only points for which $p_2(\boldsymbol{x}) = 0$. Therefore, $P(1|2, R) = 0$ and if R^* is to be better $P(1|2, R^*) = 0$. If $\Pr\{p_2(\boldsymbol{x}) = 0|\pi_1\} = 0$, then $P(2|1, R) = \Pr\{p_2(\boldsymbol{x}) > 0|\pi_1\} = 1$. If $P(1|2, R^*) = 0$, then R_1^* contains only points for which $p_2(\boldsymbol{x}) = 0$. Then $P(2|1, R^*) = \Pr\{R_2^*|\pi_1\} = \Pr\{p_2(\boldsymbol{x}) > 0|\pi_1\} = 1$, and R^* is not better than R.

Theorem 6.3.2. *If* $\Pr\{p_2(\boldsymbol{x}) = 0|\pi_1\} = 0$ *and* $\Pr\{p_1(\boldsymbol{x}) = 0|\pi_2\} = 0$, *then every Bayes procedure is admissible.*

Now let us prove the converse, namely, that every admissible procedure is a Bayes procedure. We assume[†]

$$\Pr\left\{\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} = k \,\middle|\, \pi_i\right\} = 0, \qquad i = 1, 2, \quad 0 \le k \le \infty. \tag{14}$$

Then for any q_1 the Bayes procedure is unique. Moreover, the cdf of $p_1(\boldsymbol{x})/p_2(\boldsymbol{x})$ for π_1 and π_2 is continuous.

Let R be an admissible procedure. Then there exists a k such that

$$\begin{aligned} P(2|1, R) &= \Pr\left\{\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} \le k \,\middle|\, \pi_1\right\} \\ &= P(2|1, R^*), \end{aligned} \tag{15}$$

where R^* is the Bayes procedure corresponding to $q_2/q_1 = k$ [i.e., $q_1 = 1/(1+k)$]. Since R is admissible, $P(1|2, R) \le P(1|2, R^*)$. However, since by Theorem 6.3.2 R^* is admissible, $P(1|2, R) \ge P(1|2, R^*)$; that is, $P(1|2, R) = P(1|2, R^*)$. Therefore, R is also a Bayes procedure; by the uniqueness of Bayes procedures R is the same as R^*.

Theorem 6.3.3. *If* (14) *holds, then every admissible procedure is a Bayes procedure.*

The proof of Theorem 6.3.3 shows that the class of Bayes procedures is complete. For if R is any procedure outside the class, we construct a Bayes procedure R^* so that $P(2|1, R) = P(2|1, R^*)$. Then, since R^* is admissible, $P(1|2, R) \ge P(1|2, R^*)$. Furthermore, the class of Bayes procedures is minimal complete since it is identical with the class of admissible procedures.

[†] $p_1(\boldsymbol{x})/p_2(\boldsymbol{x}) = \infty$ means $p_2(\boldsymbol{x}) = 0$.

Theorem 6.3.4. *If* (14) *holds, the class of Bayes procedures is minimal complete.*

Finally, let us consider the minimax procedure. Let $P(i|j, q_1) = P(i|j, R)$, where R is the Bayes procedure corresponding to q_1. $P(i|j, q_1)$ is a continuous function of q_1. $P(2|1, q_1)$ varies from 1 to 0 as q_1 goes from 0 to 1; $P(1|2, q_1)$ varies from 0 to 1. Thus there is a value of q_1, say q_1^*, such that $P(2|1, q_1^*) = P(1|2, q_1^*)$. This is the minimax solution, for if there were another procedure R^* such that $\max\{P(2|1, R^*), P(1|2, R^*)\} \le P(2|1, q_1^*) = P(1|2, q_1^*)$, that would contradict the fact that every Bayes solution is admissible.

6.4. CLASSIFICATION INTO ONE OF TWO KNOWN MULTIVARIATE NORMAL POPULATIONS

Now we shall use the general procedure outlined above in the case of two multivariate normal populations with equal covariance matrices, namely, $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}^{(i)\prime} = (\mu_1^{(i)}, \dots, \mu_p^{(i)})$ is the vector of means of the ith population, $i = 1, 2$, and $\boldsymbol{\Sigma}$ is the matrix of variances and covariances of each population. [The approach was first used by Wald (1944).] Then the ith density is

$$(1) \qquad p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{1}{2}p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^{(i)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(i)})\right].$$

The ratio of densities is

$$(2) \qquad \begin{aligned} \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} &= \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(1)})\right]}{\exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(2)})\right]} \\ &= \exp\left\{-\tfrac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(1)})\right.\right. \\ &\qquad \left.\left. -(\boldsymbol{x} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(2)})\right]\right\}. \end{aligned}$$

The region of classification into π_1, R_1, is the set of $\boldsymbol{x}$'s for which (2) is greater than or equal to k (for k suitably chosen). Since the logarithmic function is monotonically increasing, the inequality can be written in terms of the logarithm of (2) as

$$(3) \quad -\tfrac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(1)}) - (\boldsymbol{x} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(2)})\right] \ge \log k.$$

The left-hand side of (3) can be expanded as

$$(4)\quad -\tfrac{1}{2}\big[x'\Sigma^{-1}x - x'\Sigma^{-1}\mu^{(1)} - \mu^{(1)\prime}\Sigma^{-1}x - \mu^{(1)\prime}\Sigma^{-1}\mu^{(1)} \\ -x'\Sigma^{-1}x + x'\Sigma^{-1}\mu^{(2)} + \mu^{(2)\prime}\Sigma^{-1}x - \mu^{(2)\prime}\Sigma^{-1}\mu^{(2)}\big].$$

By rearrangement of the terms we obtain

$$(5)\qquad x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}).$$

The first term is the well-known *discriminant function*. It is a function of the components of the observation vector.

The following theorem is now a direct consequence of Theorem 6.3.1.

Theorem 6.4.1. *If π_i has the density (1), $i = 1, 2$, the best regions of classification are given by*

$$(6)\qquad \begin{aligned} R_1&: x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \ge \log k, \\ R_2&: x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) < \log k. \end{aligned}$$

If a priori probabilities q_1 and q_2 are known, then k is given by

$$(7)\qquad k = \frac{q_2 C(1|2)}{q_1 C(2|1)}.$$

In the particular case of the two populations being equally likely and the costs being equal, $k = 1$ and $\log k = 0$. Then the region of classification into π_1 is

$$(8)\qquad R_1: x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \ge \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}).$$

If we do not have a priori probabilities, we may select $\log k = c$, say, on the basis of making the expected losses due to misclassification equal. Let X be a random observation. Then we wish to find the distribution of

$$(9)\qquad U = X'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$$

on the assumption that X is distributed according to $N(\mu^{(1)}, \Sigma)$ and then on the assumption that X is distributed according to $N(\mu^{(2)}, \Sigma)$. When X is distributed according to $N(\mu^{(1)}, \Sigma)$, U is normally distributed with mean

$$(10)\qquad \begin{aligned} \mathscr{E}_1 U &= \mu^{(1)\prime}\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \\ &= \tfrac{1}{2}(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \end{aligned}$$

and variance

$$\text{(11)}\quad \begin{aligned}\operatorname{Var}_1(U) &= \mathscr{E}_1(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}^{(1)})(X - \boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})\\ &= (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}).\end{aligned}$$

The Mahalanobis squared distance between $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ is

$$\text{(12)}\qquad (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) = \Delta^2,$$

say. Then U is distributed according to $N(\frac{1}{2}\Delta^2, \Delta^2)$ if X is distributed according to $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$. If X is distributed according to $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$, then

$$\text{(13)}\quad \begin{aligned}\mathscr{E}_2 U &= \boldsymbol{\mu}^{(2)\prime}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) - \tfrac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})\\ &= \tfrac{1}{2}(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})\\ &= -\tfrac{1}{2}\Delta^2.\end{aligned}$$

The variance is the same as when X is distributed according to $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ because it depends only on the second-order moments of X. Thus U is distributed according to $N(-\frac{1}{2}\Delta^2, \Delta^2)$.

The probability of misclassification if the observation is from π_1 is

$$\text{(14)}\quad P(2|1) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}\,\Delta} e^{-\frac{1}{2}(z - \frac{1}{2}\Delta^2)^2/\Delta^2}\,dz = \int_{-\infty}^{(c-\frac{1}{2}\Delta^2)/\Delta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,dy,$$

and the probability of misclassification if the observation is from π_2 is

$$\text{(15)}\quad P(1|2) = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi}\,\Delta} e^{-\frac{1}{2}(z + \frac{1}{2}\Delta^2)^2/\Delta^2}\,dz = \int_{(c+\frac{1}{2}\Delta^2)/\Delta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,dy.$$

Figure 6.1 indicates the two probabilities as the shaded portions in the tails

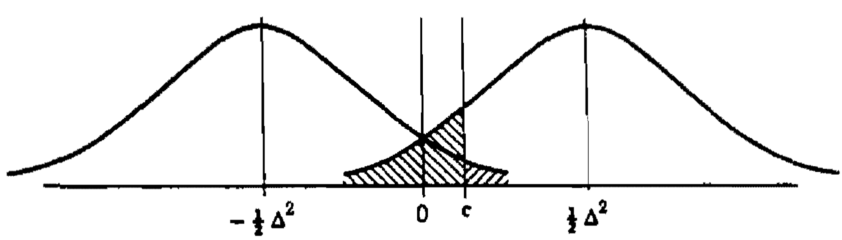

Figure 6.1

For the minimax solution we choose c so that

$$(16)\quad C(1|2)\int_{(c+\frac{1}{2}\Delta^2)/\Delta}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\,dy = C(2|1)\int_{-\infty}^{(c-\frac{1}{2}\Delta^2)/\Delta}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\,dy,$$

Theorem 6.4.2. *If the π_i have densities* (1), $i = 1,2$, *the minimax regions of classification are given by* (6) *where* $c = \log k$ *is chosen by the condition* (16) *with* $C(i|j)$ *the two costs of misclassification.*

It should be noted that if the costs of misclassification are equal, $c = 0$ and the probability of misclassification is

$$(17)\qquad \int_{\Delta/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\,dy.$$

In case the costs of misclassification are unequal, c could be determined to sufficient accuracy by a trial-and-error method with the normal tables.

Both terms in (5) involve the vector

$$(18)\qquad \boldsymbol{\delta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}).$$

This is obtained as the solution of

$$(19)\qquad \boldsymbol{\Sigma}\boldsymbol{\delta} = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$$

by an efficient computing method. The *discriminant function* $\boldsymbol{x}'\boldsymbol{\delta}$ is the linear function that maximizes

$$(20)\qquad \frac{[\mathscr{E}_1(\boldsymbol{X}'\boldsymbol{d}) - \mathscr{E}_2(\boldsymbol{X}'\boldsymbol{d})]^2}{\operatorname{Var}(\boldsymbol{X}'\boldsymbol{d})}$$

for all choices of $\boldsymbol{d}$. The numerator of (20) is

$$(21)\qquad [\boldsymbol{\mu}^{(1)\prime}\boldsymbol{d} - \boldsymbol{\mu}^{(2)\prime}\boldsymbol{d}]^2 = \boldsymbol{d}'[(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})']\boldsymbol{d};$$

the denominator is

$$(22)\qquad \boldsymbol{d}'\mathscr{E}(\boldsymbol{X} - \mathscr{E}\boldsymbol{X})(\boldsymbol{X} - \mathscr{E}\boldsymbol{X})'\boldsymbol{d} = \boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{d}.$$

We wish to maximize (21) with respect to $\boldsymbol{d}$, holding (22) constant. If λ is a Lagrange multiplier, we ask for the maximum of

$$(23)\qquad \boldsymbol{d}'[(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})']\boldsymbol{d} - \lambda(\boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{d} - 1).$$

The derivatives of (23) with respect to the components of d are set equal to zero to obtain

$$(24) \qquad 2\left[(\mu^{(1)}-\mu^{(2)})(\mu^{(1)}-\mu^{(2)})'\right]d = 2\lambda\Sigma d.$$

Since $(\mu^{(1)}-\mu^{(2)})'d$ is a scalar, say ν, we can write (24) as

$$(25) \qquad \mu^{(1)}-\mu^{(2)} = \frac{\lambda}{\nu}\Sigma d.$$

Thus the solution is proportional to δ.

We may finally note that if we have a sample of N from either π_1 or π_2, we use the mean of the sample and classify it as from $N[\mu^{(1)},(1/N)\Sigma]$ or $N[\mu^{(2)},(1/N)\Sigma]$.

6.5. CLASSIFICATION INTO ONE OF TWO MULTIVARIATE NORMAL POPULATIONS WHEN THE PARAMETERS ARE ESTIMATED

6.5.1. The Criterion of Classification

Thus far we have assumed that the two populations are known exactly. In most applications of this theory the populations are not known, but must be inferred from samples, one from each population. We shall now treat the case in which we have a sample from each of two normal populations and we wish to use that information in classifying another observation as coming from one of the two populations.

Suppose that we have a sample $x_1^{(1)},\dots,x_{N_1}^{(1)}$ from $N(\mu^{(1)},\Sigma)$ and a sample $x_1^{(2)},\dots,x_{N_2}^{(2)}$ from $N(\mu^{(2)},\Sigma)$. In one terminology these are "training samples." On the basis of this information we wish to classify the observation x as coming from π_1 to π_2. Clearly, our best estimate of $\mu^{(1)}$ is $\bar{x}^{(1)}=\sum_1^{N_1}x_\alpha^{(1)}/N_1$, of $\mu^{(2)}$ is $\bar{x}^{(2)}=\sum_1^{N_2}x_\alpha^{(2)}/N_2$, and of Σ is S defined by

$$(1) \qquad (N_1+N_2-2)S = \sum_{\alpha=1}^{N_1}\left(x_\alpha^{(1)}-\bar{x}^{(1)}\right)\left(x_\alpha^{(1)}-\bar{x}^{(1)}\right)' + \sum_{\alpha=1}^{N_2}\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)'.$$

We substitute these estimates for the parameters in (5) of Section 6.4 to obtain

$$(2) \qquad W(x) = x'S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) - \tfrac{1}{2}(\bar{x}^{(1)}+\bar{x}^{(2)})'S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}).$$

The first term of (2) is the discriminant function based on two samples [suggested by Fisher (1936)]. It is the linear function that has greatest *variance between samples* relative to the *variance within samples* (Problem 6.12). We propose that (2) be used as the criterion of classification in the same way that (5) of Section 6.4 is used.

When the populations are known, we can argue that the classification criterion is the best in the sense that its use minimizes the expected loss in the case of known a priori probabilities and generates the class of admissible procedures when a priori probabilities are not known. We cannot justify the use of (2) in the same way. However, it seems intuitively reasonable that (2) should give good results. Another criterion is indicated in Section 6.5.5.

Suppose we have a sample $x_1, \dots, x_N$ from either π_1 or π_2, and we wish to classify the sample as a whole. Then we define S by

$$(3) \quad (N_1 + N_2 + N - 3)S = \sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)' + \sum_{\alpha=1}^{N_2} \left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)' - \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})',$$

where

$$(4) \qquad \bar{x} = \frac{1}{N} \sum_{\alpha=1}^{N} x_\alpha .$$

Then the criterion is

$$(5) \qquad \left[\bar{x} - \tfrac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})\right]' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}).$$

The larger N is, the smaller are the probabilities of misclassification.

6.5.2. On the Distribution of the Criterion

Let

$$(6) \qquad \begin{aligned} W &= X'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - \tfrac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) \\ &= \left[X - \tfrac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})\right]'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) \end{aligned}$$

for random X, $\bar{X}^{(1)}$, $\bar{X}^{(2)}$, and S.

The distribution of W is extremely complicated. It depends on the sample sizes and the unknown Δ^2. Let

$$Y_1 = c_1\left[X - (N_1 + N_2)^{-1}\left(N_1 \bar{X}^{(1)} + N_2 \bar{X}^{(2)}\right)\right], \tag{7}$$

$$Y_2 = c_2(\bar{X}^{(1)} - \bar{X}^{(2)}), \tag{8}$$

where $c_1 = \sqrt{(N_1 + N_2)/(N_1 + N_2 + 1)}$ and $c_2 = \sqrt{N_1 N_2/(N_1 + N_2)}$. Then Y_1 and Y_2 are independently normally distributed with covariance matrix $\boldsymbol{\Sigma}$. The expected value of Y_2 is $c_2(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$, and the expected value of Y_1 is $c_1[N_2/(N_1 + N_2)](\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$ if X is from π_1 and $-c_1[N_1/(N_1 + N_2)](\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$ if X is from π_2. Let $Y = (Y_1 \ \ Y_2)$ and

$$M = Y'S^{-1}Y = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}. \tag{9}$$

Then

$$W = \sqrt{\frac{N_1 + N_2 + 1}{N_1 N_2}}\, m_{12} + \frac{N_1 - N_2}{2N_1 N_2} m_{22}. \tag{10}$$

The density of M has been given by Sitgreaves (1952). Anderson (1951a) and Wald (1944) have also studied the distribution of W.

If $N_1 = N_2$, the distribution of W for X from π_1 is the same as that of $-W$ for X from π_2. Thus, if $W \geq 0$ is the region of classification as π_1, then the probability of misclassifying X when it is from π_1 is equal to the probability of misclassifying it when it is from π_2.

6.5.3. The Asymptotic Distribution of the Criterion

In the case of large samples from $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$, we can apply limiting distribution theory. Since $\bar{X}^{(1)}$ is the mean of a sample of N_1 independent observations from $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$, we know that

$$\operatorname*{plim}_{N_1 \to \infty} \bar{X}^{(1)} = \boldsymbol{\mu}^{(1)}. \tag{11}$$

The explicit definition of (11) is as follows: Given arbitrary positive δ and ε, we can find N large enough so that for $N_1 \geq N$

$$\Pr\left\{|\bar{X}_i^{(1)} - \mu_i^{(1)}| < \delta,\ i = 1, \ldots, p\right\} > 1 - \varepsilon. \tag{12}$$

(See Problem 3.23.) This can be proved by using the Tchebycheff inequality. Similarly,

$$\operatorname*{plim}_{N_2\to\infty} \bar{X}^{(2)} = \boldsymbol{\mu}^{(2)}, \tag{13}$$

and

$$\operatorname{plim} S = \boldsymbol{\Sigma} \tag{14}$$

as $N_1 \to \infty$, $N_2 \to \infty$ or as both $N_1, N_2 \to \infty$. From (14) we obtain

$$\operatorname{plim} S^{-1} = \boldsymbol{\Sigma}^{-1}, \tag{15}$$

since the probability limits of sums, differences, products, and quotients of random variables are the sums, differences, products, and quotients of their probability limits as long as the probability limit of each denominator is different from zero [Cramér (1946), p. 254]. Furthermore,

$$\operatorname*{plim}_{N_1, N_2\to\infty} S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}), \tag{16}$$

$$\operatorname*{plim}_{N_1, N_2\to\infty} (\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) = (\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}). \tag{17}$$

It follows then that the limiting distribution of W is the distribution of U. For sufficiently large samples from π_1 and π_2 we can use the criterion as if we knew the population exactly and make only a small error. [The result was first given by Wald (1944).]

Theorem 6.5.1. *Let W be given by (6) with $\bar{X}^{(1)}$ the mean of a sample of N_1 from $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$, $\bar{X}^{(2)}$ the mean of a sample of N_2 from $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$, and S the estimate of $\boldsymbol{\Sigma}$ based on the pooled sample. The limiting distribution of W as $N_1 \to \infty$ and $N_2 \to \infty$ is $N(\frac{1}{2}\Delta^2, \Delta^2)$ if X is distributed according to $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$ and is $N(-\frac{1}{2}\Delta^2, \Delta^2)$ if X is distributed according to $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$.*

6.5.4. Another Derivation of the Criterion

A convenient mnemonic derivation of the criterion is the use of regression of a dummy variate [given by Fisher (1936)]. Let

$$y_\alpha^{(1)} = \frac{N_2}{N_1 + N_2}, \quad \alpha = 1, \dots, N_1, \qquad y_\alpha^{(2)} = \frac{-N_1}{N_1 + N_2}, \quad \alpha = 1, \dots, N_2. \tag{18}$$

Then formally find the regression on the variates $x_\alpha^{(i)}$ by choosing b to minimize

$$\sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}\left[y_\alpha^{(i)} - b'\left(x_\alpha^{(i)} - \bar{x}\right)\right]^2, \tag{19}$$

where

$$\bar{x} = \frac{N_1\bar{x}^{(1)} + N_2\bar{x}^{(2)}}{N_1 + N_2}. \tag{20}$$

The *normal equations* are

$$\begin{aligned}\sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)} - \bar{x}\right)\left(x_\alpha^{(i)} - \bar{x}\right)'b &= \sum_{i=1}^{2}\sum_{\alpha=1}^{N_i} y_\alpha^{(i)}\left(x_\alpha^{(i)} - \bar{x}\right)\\ &= \frac{N_1N_2}{N_1+N_2}\left[\left(\bar{x}^{(1)} - \bar{x}\right) - \left(\bar{x}^{(2)} - \bar{x}\right)\right]\\ &= \frac{N_1N_2}{N_1+N_2}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right).\end{aligned} \tag{21}$$

The matrix multiplying b can be written as

$$\begin{aligned}&\sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)} - \bar{x}\right)\left(\bar{x}_\alpha^{(i)} - \bar{x}\right)'\\ &\quad= \sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)'\\ &\qquad + N_1\left(\bar{x}^{(1)} - \bar{x}\right)\left(\bar{x}^{(1)} - \bar{x}\right)' + N_2\left(\bar{x}^{(2)} - \bar{x}\right)\left(\bar{x}^{(2)} - \bar{x}\right)'\\ &\quad= \sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)'\\ &\qquad + \frac{N_1N_2}{N_1+N_2}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)'.\end{aligned} \tag{22}$$

Thus (21) can be written as

$$Ab = \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)\left[\frac{N_1N_2}{N_1+N_2} - \frac{N_1N_2}{N_1+N_2}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)'b\right], \tag{23}$$

where

$$A = \sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)'. \tag{24}$$

Since $(\bar{x}^{(1)} - \bar{x}^{(2)})'b$ is a scalar, we see that the solution b of (23) is proportional to $S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$.

6.5.5. The Likelihood Ratio Criterion

Another criterion which can be used in classification is the likelihood ratio criterion. Consider testing the composite null hypothesis that $x, x_1^{(1)}, \dots, x_{N_1}^{(1)}$ are drawn from $N(\mu^{(1)}, \Sigma)$ and $x_1^{(2)}, \dots, x_{N_2}^{(2)}$ are drawn from $N(\mu^{(2)}, \Sigma)$ against the composite alternative hypothesis that $x_1^{(1)}, \dots, x_{N_1}^{(1)}$ are drawn from $N(\mu^{(1)}, \Sigma)$ and $x, x_1^{(2)}, \dots, x_{N_2}^{(2)}$ are drawn from $N(\mu^{(2)}, \Sigma)$, with $\mu^{(1)}$, $\mu^{(2)}$, and Σ unspecified. Under the first hypothesis the maximum likelihood estimators of $\mu^{(1)}$, $\mu^{(2)}$, and Σ are

$$\hat{\mu}_1^{(1)} = \frac{N_1 \bar{x}^{(1)} + x}{N_1 + 1}, \tag{25}$$

$$\hat{\mu}_1^{(2)} = \bar{x}^{(2)},$$

$$\hat{\Sigma}_1 = \frac{1}{N_1 + N_2 + 1} \left[\sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \hat{\mu}_1^{(1)}\right)\left(x_\alpha^{(1)} - \hat{\mu}_1^{(1)}\right)' + \left(x - \hat{\mu}_1^{(1)}\right)\left(x - \hat{\mu}_1^{(1)}\right)' \right.$$
$$\left. + \sum_{\alpha=1}^{N_2} \left(x_\alpha^{(2)} - \hat{\mu}_1^{(2)}\right)\left(x_\alpha^{(2)} - \hat{\mu}_1^{(2)}\right)' \right].$$

Since

$$\sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \hat{\mu}_1^{(1)}\right)\left(x_\alpha^{(1)} - \hat{\mu}_1^{(1)}\right)' + \left(x - \hat{\mu}_1^{(1)}\right)\left(x - \hat{\mu}_1^{(1)}\right)' \tag{26}$$
$$= \sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)' + N_1\left(\bar{x}^{(1)} - \hat{\mu}_1^{(1)}\right)\left(\bar{x}^{(1)} - \hat{\mu}_1^{(1)}\right)'$$
$$+ \left(x - \hat{\mu}_1^{(1)}\right)\left(x - \hat{\mu}_1^{(1)}\right)'$$
$$= \sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)' + \frac{N_1}{N_1 + 1}\left(x - \bar{x}^{(1)}\right)\left(x - \bar{x}^{(1)}\right)',$$

we can write $\hat{\mathbf{\Sigma}}_1$ as

$$\hat{\mathbf{\Sigma}}_1 = \frac{1}{N_1 + N_2 + 1}\left[A + \frac{N}{N_1 + 1}(x - \bar{x}^{(1)})(x - \bar{x}^{(1)})'\right], \tag{27}$$

where A is given by (24). Under the assumptions of the alternative hypothesis we find (by considerations of symmetry) that the maximum likelihood estimators of the parameters are

$$\begin{aligned} \hat{\boldsymbol{\mu}}_2^{(1)} &= \bar{x}^{(1)}, \\ \hat{\boldsymbol{\mu}}_2^{(2)} &= \frac{N_2 \bar{x}^{(2)} + x}{N_2 + 1}, \\ \hat{\mathbf{\Sigma}}_2 &= \frac{1}{N_1 + N_2 + 1}\left[A + \frac{N_2}{N_2 + 1}(x - \bar{x}^{(2)})(x - \bar{x}^{(2)})'\right]. \end{aligned} \tag{28}$$

The likelihood ratio criterion is, therefore, the $(N_1 + N_2 + 1)/2$th power of

$$\frac{|\hat{\mathbf{\Sigma}}_2|}{|\hat{\mathbf{\Sigma}}_1|} = \frac{\left|A + \frac{N_2}{N_2 + 1}(x - \bar{x}^{(2)})(x - \bar{x}^{(2)})'\right|}{\left|A + \frac{N_1}{N_1 + 1}(x - \bar{x}^{(1)})(x - \bar{x}^{(1)})'\right|}. \tag{29}$$

This ratio can also be written (Corollary A.3.1)

$$\begin{aligned} & \frac{1 + \frac{N_2}{N_2 + 1}(x - \bar{x}^{(2)})'A^{-1}(x - \bar{x}^{(2)})}{1 + \frac{N_1}{N_1 + 1}(x - \bar{x}^{(1)})'A^{-1}(x - \bar{x}^{(1)})} \\ & \qquad = \frac{n + \frac{N_2}{N_2 + 1}(x - \bar{x}^{(2)})'S^{-1}(x - \bar{x}^{(2)})}{n + \frac{N_1}{N_1 + 1}(x - \bar{x}^{(1)})'S^{-1}(x - \bar{x}^{(1)})}, \end{aligned} \tag{30}$$

where $n = N_1 + N_2 - 2$. The region of classification into π_1 consists of those points for which the ratio (30) is greater than or equal to a given number K_n. It can be written

$$\begin{aligned} R_1 : n &+ \frac{N_2}{N_2 + 1}(x - \bar{x}^{(2)})'S^{-1}(x - \bar{x}^{(2)}) \\ &\geq K_n\left[n + \frac{N_1}{N_1 + 1}(x - \bar{x}^{(1)})'S^{-1}(x - \bar{x}^{(1)})\right]. \end{aligned} \tag{31}$$

If $K_n = 1 + 2c/n$ and N_1 and N_2 are large, the region (31) is approximately $W(x) \geq c$.

If we take $K_n = 1$, the rule is to classify as π_1 if (30) is greater than 1 and as π_2 if (30) is less than 1. This is the *maximum likelihood* rule. Let

$$Z = \frac{1}{2}\left[\frac{N_2}{N_2+1}(x-\bar{x}^{(2)})'S^{-1}(x-\bar{x}^{(2)}) - \frac{N_1}{N_1+1}(x-\bar{x}^{(1)})'S^{-1}(x-\bar{x}^{(1)})\right]. \tag{32}$$

Then the maximum likelihood rule is to classify as π_1 if $Z > 0$ and π_2 if $Z < 0$. Roughly speaking, assign x to π_1 or π_2 according to whether the distance to $\bar{x}^{(1)}$ is less or greater than the distance to $\bar{x}^{(2)}$. The difference between W and Z is

$$W - Z = \frac{1}{2}\left[\frac{1}{N_2+1}(x-\bar{x}^{(2)})'S^{-1}(x-\bar{x}^{(2)}) - \frac{1}{N_1+1}(x-\bar{x}^{(1)})'S^{-1}(x-\bar{x}^{(1)})\right], \tag{33}$$

which has the probability limit 0 as $N_1, N_2 \to \infty$. The probabilities of misclassification with W are equivalent asymptotically to those with Z for large samples.

Note that for $N_1 = N_2$, $Z = [N_1/(N_1+1)]W$. Then the symmetric test based on the cutoff $c = 0$ is the same for Z and W.

6.5.6. Invariance

The classification problem is invariant with respect to transformations

$$\begin{aligned} x_\alpha^{(1)*} &= Bx_\alpha^{(1)} + c, & \alpha &= 1, \dots, N_1, \\ x_\alpha^{(2)*} &= Bx_\alpha^{(2)} + c, & \alpha &= 1, \dots, N_2, \\ x^* &= Bx + c, \end{aligned} \tag{34}$$

where B is nonsingular and c is a vector. This transformation induces the following transformation on the sufficient statistics:

$$\begin{aligned} \bar{x}^{(1)*} &= B\bar{x}^{(1)} + c, & \bar{x}^{(2)*} &= B\bar{x}^{(2)} + c, \\ x^* &= Bx + c, & S^* &= BSB', \end{aligned} \tag{35}$$

with the same transformations on the parameters, $\mu^{(1)}$, $\mu^{(2)}$, and Σ. (Note that $\mathscr{E}x = \mu^{(1)}$ or $\mu^{(2)}$.) Any invariant of the parameters is a function of

$\Delta^2 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$. There exists a matrix $\boldsymbol{B}$ and a vector $\boldsymbol{c}$ such that

$$\begin{aligned}(36)\qquad \boldsymbol{\mu}^{(1)*} &= \boldsymbol{B}\boldsymbol{\mu}^{(1)} + \boldsymbol{c} = \boldsymbol{0}, \qquad \boldsymbol{\mu}^{(2)*} = \boldsymbol{B}\boldsymbol{\mu}^{(2)} + \boldsymbol{c} = (\Delta, 0, \ldots, 0)', \\ \boldsymbol{\Sigma}^* &= \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}' = \boldsymbol{I}.\end{aligned}$$

Therefore, Δ^2 is the minimal invariant of the parameters. The elements of M defined by (9) are invariant and are the minimal invariants of the sufficient statistics. Thus invariant procedures depend on M, and the distribution of M depends only on Δ^2. The statistics W and Z are invariant.

6.6. PROBABILITIES OF MISCLASSIFICATION

6.6.1. Asymptotic Expansions of the Probabilities of Misclassification Using W

We may want to know the probabilities of misclassification before we draw the two samples for determining the classification rule, and we may want to know the (conditional) probabili.ies of misclassification after drawing the samples. As observed earlier, the exact distributions of W and Z are very difficult to calculate. Therefore, we treat asymptotic expansions of their probabilities as N_1 and N_2 increase. The background is that the limiting distribution of W and Z is $N(\frac{1}{2}\Delta^2, \Delta^2)$ if x is from π_1 and is $N(-\frac{1}{2}\Delta^2, \Delta^2)$ if x is from π_2.

Okamoto (1963) obtained the asymptotic expansion of the distribution of W to terms of order n^{-2}, and Siotani and Wang (1975, 1977) to terms of order n^{-3}. [Bowker and Sitgreaves (1961) treated the case of $N_1 = N_2$.] Let $\Phi(\cdot)$ and $\phi(\cdot)$ be the cdf and density of $N(0,1)$, respectively.

Theorem 6.6.1. *As $N_1 \to \infty$, $N_2 \to \infty$, and $N_1/N_2 \to$ a positive limit ($n = N_1 + N_2 - 2$),*

$$\begin{aligned}(1)\quad &\Pr\left\{\frac{W - \frac{1}{2}\Delta^2}{\Delta} \le u \,\middle|\, \pi_1\right\} \\ &\quad = \Phi(u) - \phi(u)\left\{\frac{1}{2N_1\Delta^2}\left[u^3 + (p-3)u - p\Delta\right]\right. \\ &\qquad + \frac{1}{2N_2\Delta^2}\left[u^3 + 2\Delta u^2 + (p - 3 + \Delta^2)u + (p-2)\Delta\right] \\ &\qquad \left. + \frac{1}{4n}\left[4u^3 + 4\Delta u^2 + (6p - 6 + \Delta^2)u + 2(p-1)\Delta\right]\right\} + O(n^{-2}),\end{aligned}$$

and $\Pr\{-(W + \frac{1}{2}\Delta^2)/\Delta \le u \mid \pi_2\}$ *is* (1) *with N_1 and N_2 interchanged.*

The rule using W is to assign the observation x to π_1 if $W(x) > c$ and to π_2 if $W(x) \le c$. The probabilities of misclassification are given by Theorem 6.6.1 with $u = (c - \frac{1}{2}\Delta^2)/\Delta$ and $u = -(c + \frac{1}{2}\Delta^2)/\Delta$, respectively. For $c = 0$, $u = -\frac{1}{2}\Delta$. If $N_1 = N_2$, this defines an exact minimax procedure [Das Gupta (1965)].

Corollary 6.6.1

$$\Pr\left\{W \le 0 \,|\, \pi_1, \lim_{n\to\infty} \frac{N_1}{N_2} = 1\right\} \tag{2}$$

$$= \Phi(-\tfrac{1}{2}\Delta) + \frac{1}{n}\phi(\tfrac{1}{2}\Delta)\left[\frac{p-1}{\Delta} + \frac{p}{4}\Delta\right] + o(n^{-1})$$

$$= \Pr\left\{W \ge 0 \,|\, \pi_2, \lim_{n\to\infty} \frac{N_1}{N_2} = 1\right\}.$$

Note that the correction term is positive, as far as this correction goes; that is, the probability of misclassification is greater than the value of the normal approximation. The correction term (to order n^{-1}) increases with p for given Δ and decreases with Δ for given p.

Since Δ is usually unknown, it is relevant to Studentize W. The sample Mahalanobis squared distance

$$D^2 = (\bar{x}^{(1)} - \bar{x}^{(2)})' S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) \tag{3}$$

is an estimator of the population Mahalanobis squared distance Δ^2. The expectation of D^2 is

$$\mathscr{E} D^2 = \frac{n}{n-p-1}\left[\Delta^2 + p\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\right]. \tag{4}$$

See Problem 6.14. If N_1 and N_2 are large, this is approximately Δ^2.

Anderson (1973b) showed the following:

Theorem 6.6.2. *If $N_1/N_2 \to$ a positive limit as $n \to \infty$,*

$$\Pr\left\{\frac{W - \frac{1}{2}D^2}{D} \le u \,\middle|\, \pi_1\right\} \tag{5}$$

$$= \Phi(u) - \phi(u)\left\{\frac{1}{N_1}\left(\frac{u}{2} - \frac{p-1}{\Delta}\right) + \frac{1}{n}\left[\frac{u^3}{4} + \left(p - \frac{3}{4}\right)u\right]\right\} + O(n^{-2}),$$

$$(6)\quad \Pr\left\{-\frac{W+\frac{1}{2}D^2}{D} \le u \,\middle|\, \pi_2\right\}$$
$$= \Phi(u) - \phi(u)\left\{\frac{1}{N_2}\left(\frac{u}{2} - \frac{p-1}{\Delta}\right) + \frac{1}{n}\left[\frac{u^3}{4} + \left(p - \frac{3}{4}\right)u\right]\right\} + O(n^{-2}).$$

Usually, one is interested in $u \le 0$ (small probabilities of error). Then the correction term is positive; that is, the normal approximation underestimates the probability of misclassification.

One may want to choose the cutoff point c so that one probability of misclassification is controlled. Let α be the desired $\Pr\{W < c \mid \pi_1\}$. Anderson (1973b, 1973c) derived the following theorem:

Theorem 6.6.3. *Let u_0 be such that $\Phi(u_0) = \alpha$, and let*

$$(7)\qquad u = u_0 - \frac{1}{N_1}\left[\frac{p-1}{D} - \frac{1}{2}u_0\right] + \frac{1}{n}\left[\left(p - \frac{3}{4}\right)u_0 + \frac{1}{4}u_0^3\right].$$

Then as $N_1 \to \infty$, $N_2 \to \infty$, and $N_1/N_2 \to$ a positive limit,

$$(8)\qquad \Pr\left\{\frac{W - \frac{1}{2}D^2}{D} \le u \,\middle|\, \pi_1\right\} = \alpha + O(n^{-2}).$$

Then $c = Du + \frac{1}{2}D^2$ will attain the desired probability α to within $O(n^{-2})$.

We now turn to evaluating the probabilities of misclassification after the two samples have been drawn. Conditional on $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, and S, the random variable W is normally distributed with conditional mean

$$(9)\qquad \mathscr{E}\left(W \mid \pi_i, \bar{x}^{(1)}, \bar{x}^{(2)}, S\right) = \left[\boldsymbol{\mu}^{(i)} - \tfrac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})\right]' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$$
$$= \mu^{(i)}(\bar{x}^{(1)}, \bar{x}^{(2)}, S)$$

when x is from π_i, $i = 1, 2$, and conditional variance

$$(10)\qquad \mathscr{V}\left(W \mid \bar{x}^{(1)}, \bar{x}^{(2)}, S\right) = (\bar{x}^{(1)} - \bar{x}^{(2)})' S^{-1} \Sigma S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)})$$
$$= \sigma^2(\bar{x}^{(1)}, \bar{x}^{(2)}, S).$$

Note that these means and variance are functions of the samples with probability limits

$$(11)\qquad \begin{aligned} \operatorname*{plim}_{N_1, N_2 \to \infty} \mu^{(i)}(\bar{x}^{(1)}, \bar{x}^{(2)}, S) &= (-1)^{i-1}\tfrac{1}{2}\Delta^2, \\ \operatorname*{plim}_{N_1, N_2 \to \infty} \sigma^2(\bar{x}^{(1)}, \bar{x}^{(2)}, S) &= \Delta^2. \end{aligned}$$

For large N_1 and N_2 the conditional probabilities of misclassification are close to the limiting normal probabilities (with high probability relative to $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, and S).

When c is the cutoff point, probabilities of misclassification conditional on $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, and S are

$$P(2|1,c,\bar{x}^{(1)},\bar{x}^{(2)},S)=\Phi\left[\frac{c-\mu^{(1)}(\bar{x}^{(1)},\bar{x}^{(2)},S)}{\sigma(\bar{x}^{(1)},\bar{x}^{(2)},S)}\right], \tag{12}$$

$$P(1|2,c,\bar{x}^{(1)},\bar{x}^{(2)},S)=1-\Phi\left[\frac{c-\mu^{(2)}(\bar{x}^{(1)},\bar{x}^{(2)},S)}{\sigma(\bar{x}^{(1)},\bar{x}^{(2)},S)}\right]. \tag{13}$$

In (12) write c as $Du_1+\frac{1}{2}D^2$. Then the argument of $\Phi(\cdot)$ in (12) is $u_1D/\sigma+(\bar{x}^{(1)}-\bar{x}^{(2)})'S^{-1}(\bar{x}^{(1)}-\boldsymbol{\mu}^{(1)})/\sigma$; the first term converges in probability to u_1, the second term tends to 0 as $N_1\to\infty$, $N_2\to\infty$, and (12) to $\Phi(u_1)$. In (13) write c as $Du_2-\frac{1}{2}D^2$. Then the argument of $\Phi(\cdot)$ in (13) is $u_2D/\sigma+(\bar{x}^{(1)}-\bar{x}^{(2)})'S^{-1}(\bar{x}^{(2)}-\boldsymbol{\mu}^{(2)})/\sigma$. The first term converges in probability to u_2 and the second term to 0; (13) converges to $1-\Phi(u_2)$.

For given $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, and S the (conditional) probabilities of misclassification (12) and (13) are functions of the parameters $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $\boldsymbol{\Sigma}$ and can be estimated. Consider them when $c=0$. Then (12) and (13) converge in probability to $\Phi(-\frac{1}{2}\Delta)$; that suggests $\Phi(-\frac{1}{2}D)$ as an estimator of (12) and (13). A better estimator is $\Phi(-\frac{1}{2}\bar{D})$, where $\bar{D}^2=(n-p-1)D^2/n$, which is closer to being an unbiased estimator of Δ^2. [See (4).] McLachlan (1973, 1974a, 1974b, 1974c) gave an estimator of (12) whose bias is of order n^{-2}; it is

$$\Phi(-\tfrac{1}{2}D)+\phi(\tfrac{1}{2}D)\left\{\frac{p-D}{N_1D}+\frac{1}{32n}\left[-D^3+4(4p-1)D\right]\right\}. \tag{14}$$

[McLachlan gave (14) to terms of order n^{-1}.] McLachlan explored the properties of these and other estimators, as did Lachenbruch and Mickey (1968).

Now consider (12) with $c=Du_1+\frac{1}{2}D^2$; u_1 might be chosen to control $P(2|1)$ conditional on $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, S. This conditional probability as a function of $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, S is a random variable whose distribution may be approximated. McLachlan showed the following:

Theorem 6.6.4. *As $N_1\to\infty$, $N_2\to\infty$, and $N_1/N_2\to$ a positive limit,*

$$\Pr\left\{\sqrt{n}\,\frac{P(2|1,Du_1+\frac{1}{2}D^2,\bar{x}^{(1)},\bar{x}^{(2)},S)-\Phi(u_1)}{\phi(u_2)\left[\frac{1}{2}u_1^2+n/N_1\right]^{\frac{1}{2}}}\le x\right\} \tag{15}$$

$$=\Phi\left[x-\frac{(p-1)n/N_1-(p-\frac{3}{4}+n/N_1)u_1-u_1^3/4}{\sqrt{n}\left[\frac{1}{2}u_1^2+n/N_1\right]^{\frac{1}{2}}}\right]+O(n^{-2}).$$

McLachlan (1977) gave a method of selecting u_1 so that the probability of one misclassification is less than a preassigned δ with a preassigned confidence level $1 - \varepsilon$.

6.6.2. Asymptotic Expansions of the Probabilities of Misclassification Using Z

We now turn our attention to Z defined by (32) of Section 6.5. The results are parallel to those for W. Memon and Okamoto (1971) expanded the distribution of Z to terms of order n^{-2}, and Siotani and Wang (1975), (1977) to terms of order n^{-3}.

Theorem 6.6.5. *As $N_1 \to \infty$, $N_2 \to \infty$, and N_1/N_2 approaches a positive limit,*

$$\Pr\left\{\frac{Z - \frac{1}{2}\Delta^2}{\Delta} \le u \,\middle|\, \pi_1\right\} \tag{16}$$

$$= \Phi(u) - \phi(u)\left\{\frac{1}{2N_1\Delta^2}\left[u^3 + \Delta u^2 + (p-3)u - \Delta\right]\right.$$

$$+ \frac{1}{2N_2\Delta^2}\left[u^3 + \Delta u^2 + (p - 3 - \Delta^2)u - \Delta^3 - \Delta\right]$$

$$\left. + \frac{1}{4n}\left[4u^3 + 4\Delta u^2 + (6p - 6 + \Delta^2)u + 2(p-1)\Delta\right]\right\} + O(n^{-2}),$$

and $\Pr\{-(Z + \frac{1}{2}\Delta^2)/\Delta \le u | \pi_2\}$ *is* (16) *with N_1 and N_2 interchanged.*

When $c = 0$, then $u = -\frac{1}{2}\Delta$. If $N_1 = N_2$, the rule with Z is identical to the rule with W, and the probability of misclassification is given by (2).

Fujikoshi and Kanazawa (1976) proved

Theorem 6.6.6

$$\Pr\left\{\frac{Z - \frac{1}{2}D^2}{D} \le u \,\middle|\, \pi_1\right\} \tag{17}$$

$$= \Phi(u) - \phi(u)\left\{\frac{1}{2N_1\Delta}\left[u^2 + \Delta u - (p-1)\right]\right.$$

$$- \frac{1}{2N_2\Delta}\left[u^2 + 2\Delta u + p - 1 + \Delta^2\right]$$

$$\left. + \frac{1}{4n}\left[u^3 + (4p-3)u\right]\right\} + O(n^{-2}),$$

$$(18)\quad \Pr\left\{-\frac{Z+\frac{1}{2}D^2}{D} \le u \,\middle|\, \pi_2\right\}$$
$$= \Phi(u) - \phi(u)\left\{-\frac{1}{2N_1\Delta}\left[u^2 + 2\Delta u + p - 1 + \Delta^2\right]\right.$$
$$\left. + \frac{1}{2N_2\Delta}\left[u^2 + \Delta u - (p-1)\right] + \frac{1}{4n}\left[u^3 + (4p-3)u\right]\right\} + O(n^{-2})$$

Kanazawa (1979) showed the following:

Theorem 6.6.7. *Let u_0 be such that $\Phi(u_0) = \alpha$, and let*

$$(19)\quad u = u_0 + \frac{1}{2N_1 D}\left[u_0^2 + Du_0 - (p-1)\right]$$
$$- \frac{1}{2N_2 D}\left[u_0^2 + Du_0 + (p-1) - D^2\right]$$
$$+ \frac{1}{4n}\left[u_0^3 + (4p-5)u_0\right].$$

Then as $N_1 \to \infty$, $N_2 \to \infty$, and $N_1/N_2 \to$ a positive limit,

$$(20)\quad \Pr\left\{\frac{Z - \frac{1}{2}D^2}{D} \le u\right\} = \alpha + O(n^{-2}).$$

Now consider the probabilities of misclassification after the samples have been drawn. The conditional distribution of Z is not normal; Z is quadratic in x unless $N_1 = N_2$. We do not have expressions equivalent to (12) and (13). Siotani (1980) showed the following:

Theorem 6.6.8. *As $N_1 \to \infty$, $N_2 \to \infty$, and $N_1/N_2 \to$ a positive limit,*

$$(21)\quad \Pr\left\{2\sqrt{\frac{N_1N_2}{N_1+N_2}}\,\frac{P(2|1,0,\bar{x}^{(1)},\bar{x}^{(2)},S) - \Phi(-\frac{1}{2}\Delta)}{\phi(\frac{1}{2}\Delta)} \le x\right\}$$
$$= \Phi\left[x - 2\sqrt{\frac{N_1N_2}{N_1+N_2}}\left\{\frac{1}{16N_1\Delta}\left[4(p-1) - \Delta^2\right]\right.\right.$$
$$\left.\left. + \frac{1}{16N_2}\left[4(p-1) + 3\Delta^2\right] - \frac{(p-1)\Delta}{4n}\right\}\right] + O(n^{-2}).$$

It is also possible to obtain a similar expression for $P(2|1, Du_1 + \frac{1}{2}D^2, \bar{x}^{(1)}, \bar{x}^{(2)}, S)$ for Z and a confidence interval. See Siotani (1980).

6.7. CLASSIFICATION INTO ONE OF SEVERAL POPULATIONS

Let us now consider the problem of classifying an observation into one of several populations. We shall extend the consideration of the previous sections to the cases of more than two populations. Let $\pi_1, \dots, \pi_m$ be m populations with density functions $p_1(x), \dots, p_m(x)$, respectively. We wish to divide the space of observations into m mutually exclusive and exhaustive regions $R_1, \dots, R_m$. If an observation falls into R_i, we shall say that it *comes from* π_i. Let the cost of misclassifying an observation from π_i as coming from π_j be $C(j|i)$. The probability of this misclassification is

$$(1) \qquad P(j|i, R) = \int_{R_j} p_i(x)\, dx.$$

Suppose we have a priori probabilities of the populations, $q_1, \dots, q_m$. Then the expected loss is

$$(2) \qquad \sum_{i=1}^{m} q_i \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{m} C(j|i) P(j|i, R) \right\}.$$

We should like to choose $R_1, \dots, R_m$ to make this a minimum.

Since we have a priori probabilities for the populations, we can define the conditional probability of an observation coming from a population given the values of the components of the vector x. The conditional probability of the observation coming from π_i is

$$(3) \qquad \frac{q_i p_i(x)}{\sum_{k=1}^{m} q_k p_k(x)}.$$

If we classify the observation as from π_j, the expected loss is

$$(4) \qquad \sum_{\substack{i=1 \\ i \neq j}}^{m} \frac{q_i p_i(x)}{\sum_{k=1}^{m} q_k p_k(x)} C(j|i).$$

We minimize the expected loss at this point if we choose j so as to minimize (4); that is, we consider

$$(5) \qquad \sum_{\substack{i=1 \\ i \neq j}}^{m} q_i p_i(x) C(j|i)$$

for all j and select that j that gives the minimum. (If two different indices give the minimum, it is irrelevant which index is selected.) This procedure assigns the point x to one of the R_j. Following this procedure for each x, we define our regions $R_1, \dots, R_m$. The classification procedure, then, is to classify an observation as coming from π_j if it falls in R_j.

Theorem 6.7.1. *If q_i is the a priori probability of drawing an observation from population π_i with density $p_i(x)$, $i = 1, \dots, m$, and if the cost of misclassifying an observation from π_i as from π_j is $C(j|i)$, then the regions of classification, $R_1, \dots, R_m$, that minimize the expected cost are defined by assigning x to R_k if*

$$\sum_{\substack{i=1\\i\neq k}}^{m} q_i p_i(x) C(k|i) < \sum_{\substack{i=1\\i\neq j}}^{m} q_i p_i(x) C(j|i), \qquad j = 1, \dots, m, \quad j \neq k. \tag{6}$$

[If (6) holds for all j ($j \neq k$) except for h indices and the inequality is replaced by equality for those indices, then this point can be assigned to any of the $h+1$ π's.] If the probability of equality between the right-hand and left-hand sides of (6) is zero for each k and j under π_i (each i), then the minimizing procedure is unique except for sets of probability zero.

Proof. We now verify this result. Let

$$h_j(x) = \sum_{\substack{i=1\\i\neq j}}^{m} q_i p_i(x) C(j|i). \tag{7}$$

Then the expected loss of a procedure R is

$$\sum_{j=1}^{m} \int_{R_j} h_j(x)\, dx = \int h(x|R)\, dx, \tag{8}$$

where $h(x|R) = h_j(x)$ for x in R_j. For the Bayes procedure R^* described in the theorem, $h(x|R)$ is $h(x|R^*) = \min_i h_i(x)$. Thus the difference between the expected loss for any procedure R and for R^* is

$$\begin{aligned} \int [h(x|R) - h(x|R^*)]\, dx &= \sum_j \int_{R_j} [h_j(x) - \min h_i(x)]\, dx \\ &\geq 0. \end{aligned} \tag{9}$$

Equality can hold only if $h_j(x) = \min_i h_i(x)$ for x in R_j except for sets of probability zero. ■

Let us see how this method applies when $C(j|i) = 1$ for all i and j, $i \neq j$. Then in R_k

$$\sum_{\substack{i=1 \\ i \neq k}}^{m} q_i p_i(x) < \sum_{\substack{i=1 \\ i \neq j}}^{m} q_i p_i(x), \qquad j \neq k. \tag{10}$$

Subtracting $\sum_{i=1,\, i \neq k, j}^{m} q_i p_i(x)$ from both sides of (10), we obtain

$$q_j p_j(x) < q_k p_k(x), \qquad j \neq k. \tag{11}$$

In this case the point x is in R_k if k is the index for which $q_i p_i(x)$ is a maximum; that is, π_k is the most probable population.

Now suppose that we do not have a priori probabilities. Then we cannot define an unconditional expected loss for a classification procedure. However, we can define an expected loss on the condition that the observation comes from a given population. The conditional expected loss if the observation is from π_i is

$$\sum_{\substack{j=1 \\ j \neq i}}^{m} C(j|i) P(j|i, R) = r(i, R). \tag{12}$$

A procedure R is *at least as good* as R^* if $r(i, R) \leq r(i, R^*)$, $i = 1, \ldots, m$; R is *better* if at least one inequality is strict. R is *admissible* if there is no procedure R^* that is better. A class of procedures is *complete* if for every procedure R outside the class there is a procedure R^* in the class that is better.

Now let us show that a Bayes procedure is admissible. Let R be a Bayes procedure; let R^* be another procedure. Since R is Bayes,

$$\sum_{i=1}^{m} q_i r(i, R) \leq \sum_{i=1}^{m} q_i r(i, R^*). \tag{13}$$

Suppose $q_1 > 0$, $q_2 > 0$, $r(2, R^*) < r(2, R)$, and $r(i, R^*) \leq r(i, R)$, $i = 3, \ldots, m$. Then

$$q_1[r(1, R) - r(1, R^*)] \leq \sum_{i=2}^{m} q_i[r(i, R^*) - r(i, R)] < 0, \tag{14}$$

and $r(1, R) < r(1, R^*)$. Thus R^* is not better than R.

Theorem 6.7.2. *If $q_i > 0$, $i = 1, \ldots, m$, then a Bayes procedure is admissible.*

We shall now assume that $C(i|j)=1$, $i \neq j$, and $\Pr\{p_i(x)=0|\pi_j\}=0$. The latter condition implies that all $p_i(x)$ are positive on the same set (except for a set of measure 0). Suppose $q_i=0$ for $i=1,\dots,t$, and $q_i>0$ for $i=t+1,\dots,m$. Then for the Bayes solution R_i, $i=1,\dots,t$, is empty (except for a set of probability 0), as seen from (11) [that is, $p_m(x)=0$ for x in R_i]. It follows that $r(i,R)=\sum_{j\neq i}P(j|i,R)=1-P(i|i,R)=1$ for $i=1,\dots,t$. Then $(R_{t+1},\dots,R_m)$ is a Bayes solution for the problem involving $p_{t+1}(x),\dots,p_m(x)$ and $q_{t+1},\dots,q_m$. It follows from Theorem 6.7.2 that no procedure R^* for which $P(i|i,R^*)=0$, $i=1,\dots,t$, can be better than the Bayes procedure. Now consider a procedure R^* such that R_1^* includes a set of positive probability so that $P(1|1,R^*)>0$. For R^* to be better than R,

$$
\begin{aligned}
P(i|i,R) &= \int_{R_i} p_i(x)\,dx \\
&\leq P(i|i,R^*) = \int_{R_i^*} p_i(x)\,dx, \qquad i=2,\dots,m.
\end{aligned}
\tag{15}
$$

In such a case a procedure R^{**} where R_i^{**} is empty, $i=1,\dots,t$, $R_i^{**}=R_i^*$, $i=t+1,\dots,m-1$, and $R_m^{**}=R_m^*\cup R_1^*\cup\cdots\cup R_t^*$ would give risks such that

$$
\begin{aligned}
&P(i|i,R^{**})=0, && i=1,\dots,t,\\
&P(i|i,R^{**})=P(i|i,R^*)\geq P(i|i,R), && i=t+1,\dots,m-1,\\
&P(m|m,R^{**})>P(m|m,R^*)\geq P(m|m,R).
\end{aligned}
\tag{16}
$$

Then $R_{t+1}^{**},\dots,R_m^{**})$ would be better than $(R_{t+1},\dots,R_m)$ for the $(m-t)$-decision problem, which contradicts the preceding discussion.

Theorem 6.7.3. *If* $C(i|j)=1$, $i\neq j$, *and* $\Pr\{p_i(x)=0|\pi_j\}=0$, *then a Bayes procedure is admissible.*

The converse is true without conditions (except that the parameter space is finite).

Theorem 6.7.4. *Every admissible procedure is a Bayes procedure.*

We shall not prove this theorem. It is Theorem 1 of Section 2.10 of Ferguson (1967), for example. The class of Bayes procedures is minimal complete if each Bayes procedure is unique (for the specified probabilities).

The minimax procedure is the Bayes procedure for which the risks are equal.

There are available general treatments of statistical decision procedures by Wald (1950), Blackwell and Girshick (1954), Ferguson (1967), De Groot (1970), Berger (1980b), and others.

6.8. CLASSIFICATION INTO ONE OF SEVERAL MULTIVARIATE NORMAL POPULATIONS

We shall now apply the theory of Section 6.7 to the case in which each population has a normal distribution. [See von Mises (1945).] We assume that the means are different and the covariance matrices are alike. Let $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$ be the distribution of π_i. The density is given by (1) of Section 6.4. At the outset the parameters are assumed known. For general costs with known a priori probabilities we can form the m functions (5) of Section 6.7 and define the region R_j as consisting of points x such that the jth function is minimum.

In the remainder of our discussion we shall assume that the costs of misclassification are equal. Then we use the functions

$$(1) \qquad u_{jk}(x) = \log\frac{p_j(x)}{p_k(x)} = \left[x - \tfrac{1}{2}(\boldsymbol{\mu}^{(j)} + \boldsymbol{\mu}^{(k)})\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(k)}).$$

If a priori probabilities are known, the region R_j is defined by those x satisfying

$$(2) \qquad R_j\colon u_{jk}(x) > \log\frac{q_k}{q_j}, \qquad k = 1, \dots, m, \quad k \neq j.$$

Theorem 6.8.1. *If q_i is the a priori probability of drawing an observation from $\pi_i = N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i = 1, \dots, m$, and if the costs of misclassification are equal, then the regions of classification, $R_1, \dots, R_m$, that minimize the expected cost are defined by* (2), *where $u_{jk}(x)$ is given by* (1).

It should be noted that each $u_{jk}(x)$ is the classification function related to the jth and kth populations, and $u_{jk}(x) = -u_{kj}(x)$. Since these are linear functions, the region R_i is bounded by hyperplanes. If the means span an $(m-1)$-dimensional hyperplane (for example, if the vectors $\boldsymbol{\mu}^{(i)}$ are linearly independent and $p \geq m-1$), then R_i is bounded by $m-1$ hyperplanes.

In the case of no set of a priori probabilities known, the region R_j is defined by inequalities

$$(3) \qquad u_{jk}(x) \geq c_j - c_k, \qquad k = 1, \dots, m, \quad k \neq j.$$

The constants c_k can be taken nonnegative. These sets of regions form the class of admissible procedures. For the minimax procedure these constants are determined so all $P(i|i, R)$ are equal.

We now show how to evaluate the probabilities of correct classification. If X is a random observation, we consider the random variables

$$U_{ji} = \left[X - \tfrac{1}{2}\left(\boldsymbol{\mu}^{(i)} + \boldsymbol{\mu}^{(j)}\right)\right]'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}\right). \tag{4}$$

Here $U_{ji} = -U_{ij}$. Thus we use $m(m-1)/2$ classification functions if the means span an $(m-1)$-dimensional hyperplane. If X is from π_j, then U_{ji} is distributed according to $N(\frac{1}{2}\Delta_{ji}^2, \Delta_{ji}^2)$, where

$$\Delta_{ji}^2 = \left(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}\right). \tag{5}$$

The covariance of U_{ji} and U_{jk} is

$$\Delta_{jk.ji} = \left(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(k)}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(i)}\right). \tag{6}$$

To determine the constants c_j we consider the integrals

$$P(j|j.R) = \int_{c_j - c_m}^{\infty} \cdots \int_{c_j - c_1}^{\infty} f_j \, du_{j1} \cdots du_{j.j-1} \, du_{j,j+1} \cdots du_{jm}, \tag{7}$$

where f_j is the density of U_{ji}, $i = 1, 2, \ldots, m$, $i \neq j$.

Theorem 6.8.2. *If π_i is $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$ and the costs of misclassification are equal, then the regions of classification, $R_1, \ldots, R_m$, that minimize the maximum conditional expected loss are defined by* (3), *where $u_{jk}(x)$ is given by* (1). *The constants c_j are determined so that the integrals* (7) *are equal.*

As an example consider the case of $m = 3$. There is no loss of generality in taking $p = 2$, for the density for higher p can be projected on the two-dimensional plane determined by the means of the three populations if they are not collinear (i.e., we can transform the vector x into u_{12}, u_{13}, and $p-2$ other coordinates, where these last $p-2$ components are distributed independently of u_{12} and u_{13} and with zero means). The regions R_j are determined by three half lines as shown in Figure 6.2. If this procedure is minimax, we cannot move the line between R_1 and R_2 nearer $(\mu_1^{(1)}, \mu_2^{(1)})$, the line between R_2 and R_3 nearer $(\mu_1^{(2)}, \mu_2^{(2)})$, and the line between R_3 and R_1 nearer $(\mu_1^{(3)}, \mu_2^{(3)})$ and still retain the equality $P(1|1, R) = P(2|2, R) = P(3|3, R)$ without leaving a triangle that is not included in any region. Thus, since the regions must exhaust the space, the lines must meet in a point, and the equality of probabilities determines $c_i - c_j$ uniquely.

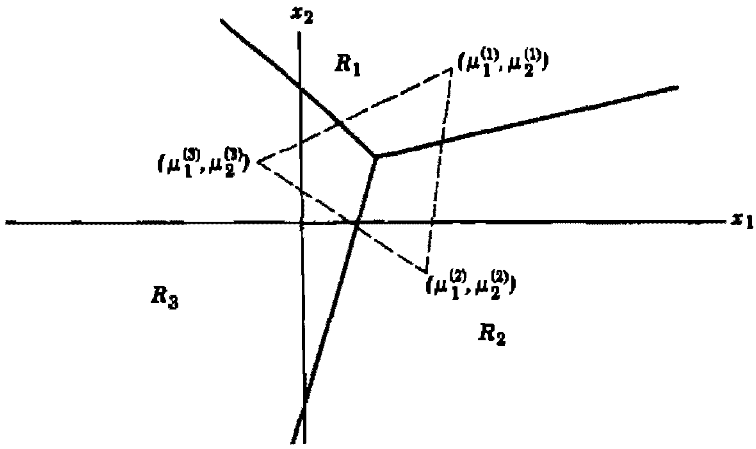

Figure 6.2. Classification regions.

To do this in a specific case in which we have numerical values for the components of the vectors $\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \boldsymbol{\mu}^{(3)}$, and the matrix $\boldsymbol{\Sigma}$, we would consider the three ($\leq p+1$) joint distributions, each of two U_{ij}'s ($j \neq i$). We could try the values of $c_i = 0$ and, using tables [Pearson (1931)] of the bivariate normal distribution, compute $P(i|i, R)$. By a trial-and-error method we could obtain c_i to approximate the above condition.

The preceding theory has been given on the assumption that the parameters are known. If they are not known and if a sample from each population is available, the estimators of the parameters can be substituted in the definition of $u_{ij}(x)$. Let the observations be $x_1^{(i)}, \dots, x_{N_i}^{(i)}$ from $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i = 1, \dots, m$. We estimate $\boldsymbol{\mu}^{(i)}$ by

$$\bar{x}^{(i)} = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} x_\alpha^{(i)} \tag{8}$$

and $\boldsymbol{\Sigma}$ by S defined by

$$\left(\sum_{i=1}^{m} N_i - m \right) S = \sum_{i=1}^{m} \sum_{\alpha=1}^{N_i} \left(x_\alpha^{(i)} - \bar{x}^{(i)} \right) \left(x_\alpha^{(i)} - \bar{x}^{(i)} \right)'. \tag{9}$$

Then, the analog of $u_{ij}(x)$ is

$$w_{ij}(x) = \left[x - \tfrac{1}{2} \left(\bar{x}^{(i)} + \bar{x}^{(j)} \right) \right]' S^{-1} \left(\bar{x}^{(i)} - \bar{x}^{(j)} \right). \tag{10}$$

If the variables above are random, the distributions are different from those of U_{ij}. However, as $N_i \to \infty$, the joint distributions approach those of U_{ij}. Hence, for sufficiently large samples one can use the theory given above.

Table 6.2

	Mean		
Measurement	Brahmin (π_1)	Artisan (π_2)	Korwa (π_3)
Stature (x_1)	164.51	160.53	158.17
Sitting height (x_2)	86.43	81.47	81.16
Nasal depth (x_3)	25.49	23.84	21.44
Nasal height (x_4)	51.24	48.62	46.72

6.9. AN EXAMPLE OF CLASSIFICATION INTO ONE OF SEVERAL MULTIVARIATE NORMAL POPULATIONS

Rao (1948a) considers three populations consisting of the Brahmin caste (π_1), the Artisan caste (π_2), and the Korwa caste (π_3) of India. The measurements for each individual of a caste are stature (x_1), sitting height (x_2), nasal depth (x_3), and nasal height (x_4). The means of these variables in the three populations are given in Table 6.2. The matrix of correlations for all the populations is

$$(1) \qquad \begin{bmatrix} 1.0000 & 0.5849 & 0.1774 & 0.1974 \\ 0.5849 & 1.0000 & 0.2094 & 0.2170 \\ 0.1774 & 0.2094 & 1.0000 & 0.2910 \\ 0.1974 & 0.2170 & 0.2910 & 1.0000 \end{bmatrix}.$$

The standard deviations are $\sigma_1 = 5.74$, $\sigma_2 = 3.20$, $\sigma_3 = 1.75$, $\sigma_4 = 3.50$. We assume that each population is normal. Our problem is to divide the space of the four variables x_1, x_2, x_3, x_4 into three regions of classification. We assume that the costs of misclassification are equal. We shall find (i) a set of regions under the assumption that drawing a new observation from each population is equally likely ($q_1 = q_2 = q_3 = \frac{1}{3}$), and (ii) a set of regions such that the largest probability of misclassification is minimized (the minimax solution).

We first compute the coefficients of $\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$ and $\Sigma^{-1}(\mu^{(1)} - \mu^{(3)})$. Then $\Sigma^{-1}(\mu^{(2)} - \mu^{(3)}) = \Sigma^{-1}(\mu^{(1)} - \mu^{(3)}) - \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$. Then we calculate $\frac{1}{2}(\mu^{(i)} + \mu^{(j)})'\Sigma^{-1}(\mu^{(i)} - \mu^{(j)})$. We obtain the discriminant functions†

$$(2) \qquad \begin{aligned} u_{12}(x) &= -0.0708x_1 + 0.4990x_2 + 0.3373x_3 + 0.0887x_4 - 43.13, \\ u_{13}(x) &= 0.0003x_1 + 0.3550x_2 + 1.1063x_3 + 0.1375x_4 - 62.49, \\ u_{23}(x) &= 0.0711x_1 - 0.1440x_2 + 0.7690x_3 + 0.0488x_4 - 19.36. \end{aligned}$$

†Due to an error in computations, Rao's discriminant functions are incorrect. I am indebted to Mr. Peter Frank for assistance in the computations.

Table 6.3

Population of x	u	Means	Standard Deviation	Correlation
π_1	u_{12}	1.491	1.727	0.8658
	u_{13}	3.487	2.641	
π_2	u_{21}	1.491	1.727	-0.3894
	u_{23}	1.031	1.436	
π_3	u_{31}	3.487	2.641	0.7983
	u_{32}	1.031	1.436	

The other three functions are $u_{21}(x) = -u_{12}(x)$, $u_{31}(x) = -u_{13}(x)$, and $u_{32}(x) = -u_{23}(x)$. If there are a priori probabilities and they are equal, the best set of regions of classification are R_1: $u_{12}(x) \geq 0$, $u_{13}(x) \geq 0$; R_2: $u_{21}(x) \geq 0$, $u_{23}(x) \geq 0$; and R_3: $u_{31}(x) \geq 0$, $u_{32}(x) \geq 0$. For example, if we obtain an individual with measurements x such that $u_{12}(x) \geq 0$ and $u_{13}(x) \geq 0$, we classify him as a Brahmin.

To find the probabilities of misclassification when an individual is drawn from population π_g we need the means, variances, and covariances of the proper pairs of u's. They are given in Table 6.3.†

The probabilities of misclassification are then obtained by use of the tables for the bivariate normal distribution. These probabilities are 0.21 for π_1, 0.42 for π_2, and 0.25 for π_3. For example, if measurements are made on a Brahmin, the probability that he is classified as an Artisan or Korwa is 0.21.

The minimax solution is obtained by finding the constants c_1, c_2, and c_3 for (3) of Section 6.8 so that the probabilities of misclassification are equal. The regions of classification are

$$\begin{aligned} &R_1' : u_{12}(x) \geq 0.54, && u_{13}(x) \geq 0.29; \\ (3) \qquad &R_2' : u_{21}(x) \geq -0.54, && u_{23}(x) \geq -0.25; \\ &R_3' : u_{31}(x) \geq -0.29, && u_{32}(x) \geq 0.25. \end{aligned}$$

The common probability of misclassification (to two decimal places) is 0.30. Thus the maximum probability of misclassification has been reduced from 0.42 to 0.30.

†Some numerical errors in Anderson (1951a) are corrected in Table 6.3 and (3).

6.10. CLASSIFICATION INTO ONE OF TWO KNOWN MULTIVARIATE NORMAL POPULATIONS WITH UNEQUAL COVARIANCE MATRICES

6.10.1. Likelihood Procedures

Let π_1 and π_2 be $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\mu}^{(1)} \neq \boldsymbol{\mu}^{(2)}$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. When the parameters are known, the likelihood ratio is

$$
\begin{aligned}
(1) \qquad \frac{p_1(x)}{p_2(x)} &= \frac{|\boldsymbol{\Sigma}_2|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}_1^{-1}(x-\boldsymbol{\mu}^{(1)})\right]}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}_2^{-1}(x-\boldsymbol{\mu}^{(2)})\right]} \\
&= |\boldsymbol{\Sigma}_2|^{\frac{1}{2}} |\boldsymbol{\Sigma}_1|^{-\frac{1}{2}} \exp\left[\tfrac{1}{2}(x-\boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}_2^{-1}(x-\boldsymbol{\mu}^{(2)}) \right. \\
&\qquad \left. -\tfrac{1}{2}(x-\boldsymbol{\mu}^{(1)})'\boldsymbol{\Sigma}_1^{-1}(x-\boldsymbol{\mu}^{(1)})\right].
\end{aligned}
$$

The logarithm of (1) is quadratic in x. The probabilities of misclassification are difficult to compute. [One can make a linear transformation of x so that its covariance matrix is I and the matrix of the quadratic form is diagonal; then the logarithm of (1) has the distribution of a linear combination of noncentral χ^2-variables plus a constant.]

When the parameters are unknown, we consider the problem as testing the hypothesis that $x, x_1^{(1)}, \dots, x_{N_1}^{(1)}$ are observations from $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_1)$ and $x_1^{(2)}, \dots, x_{N_2}^{(2)}$ are observations from $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_2)$ against the alternative that $x_1^{(1)}, \dots, x_{N_1}^{(1)}$ are observations from $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_1)$ and $x, x_1^{(2)}, \dots, x_{N_2}^{(2)}$ are observations from $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_2)$. Under the first hypothesis the maximum likelihood estimators are $\hat{\boldsymbol{\mu}}_1^{(1)} = (N_1\bar{x}^{(1)} + x)/(N_1+1)$, $\hat{\boldsymbol{\mu}}_1^{(2)} = \bar{x}^{(2)}$,

$$
(2) \qquad
\begin{aligned}
\hat{\boldsymbol{\Sigma}}_1(1) &= \frac{1}{N_1+1}\left[A_1 + \frac{N_1}{N_1+1}(x-\bar{x}^{(1)})(x-x^{(1)})'\right], \\
\hat{\boldsymbol{\Sigma}}_2(1) &= \frac{1}{N_2}A_2,
\end{aligned}
$$

where $A_i = \sum_{\alpha=1}^{N_i}(x_\alpha^{(i)} - \bar{x}^{(i)})(x_\alpha^{(i)} - \bar{x}^{(i)})'$, $i = 1, 2$. (See Section 6.5.5.) Under the second hypothesis the maximum likelihood estimators are $\hat{\boldsymbol{\mu}}_2^{(1)} = \bar{x}^{(1)}$, $\hat{\boldsymbol{\mu}}_2^{(2)} = (N_2\bar{x}^{(2)} + x)/(N_2+1)$,

$$
(3) \qquad
\begin{aligned}
\hat{\boldsymbol{\Sigma}}_1(2) &= \frac{1}{N_1}A_1, \\
\hat{\boldsymbol{\Sigma}}_2(2) &= \frac{1}{N_2+1}\left[A_2 + \frac{N_2}{N_2+1}(x-\bar{x}^{(2)})(x-\bar{x}^{(2)})'\right].
\end{aligned}
$$

The likelihood ratio criterion is

$$\text{(4)} \quad \frac{|\hat{\Sigma}_1(2)|^{\frac{1}{2}N_1}|\hat{\Sigma}_2(2)|^{\frac{1}{2}(N_2+1)}}{|\hat{\Sigma}_1(1)|^{\frac{1}{2}(N_1+1)}|\hat{\Sigma}_2(1)|^{\frac{1}{2}N_2}} = \frac{\left[1+(x-\bar{x}^{(2)})'A_2^{-1}(x-\bar{x}^{(2)})\right]^{\frac{1}{2}(N_2+1)}}{\left[1+(x-\bar{x}^{(1)})'A_1^{-1}(x-\bar{x}^{(1)})\right]^{\frac{1}{2}(N_1+1)}} \cdot \frac{(N_1+1)^{\frac{1}{2}(N_1+1)p}N_2^{\frac{1}{2}N_2 p}|A_2|^{\frac{1}{2}}}{N_1^{\frac{1}{2}N_1 p}(N_2+1)^{\frac{1}{2}(N_2+1)p}|A_1|^{\frac{1}{2}}}.$$

The observation x is classified into π_1 if (4) is greater than 1 and into π_2 if (4) is less than 1.

An alternative criterion is to plug estimates into the logarithm of (1). Use

$$\text{(5)} \qquad \tfrac{1}{2}\left[(x-\bar{x}^{(2)})'S_2^{-1}(x-\bar{x}^{(2)}) - (x-\bar{x}^{(1)})'S_1^{-1}(x-\bar{x}^{(1)})\right]$$

to classify into π_1 if (5) is large and into π_2 if (5) is small. Again it is difficult to evaluate the probabilities of misclassification.

6.10.2. Linear Procedures

The best procedures when $\Sigma_1 \neq \Sigma_2$ are not linear; when the parameters are known, the best procedures are based on a quadratic function of the vector observation x. The procedure depends very much on the assumed normality. For example, in the case of $p = 1$, the region for classification with one population is an interval and for the other is the complement of the interval —that is, where the observation is sufficiently large or sufficiently small. In the bivariate case the regions are defined by conic sections; for example, the region of classification into one population might be the interior of an ellipse or the region between two hyperbolas. In general, the regions are defined by means of a quadratic function of the observations which is not necessarily a positive definite quadratic form. These procedures depend very much on the assumption of normality and especially on the shape of the normal distribution far from its center. For instance, in the univariate case cited above the region of classification into the first population is a finite interval because the density of the first population falls off in either direction more rapidly than the density of the second because its standard deviation is smaller.

One may want to use a classification procedure in a situation where the two populations are centered around different points and have different patterns of scatter, and where one considers multivariate normal distributions to be reasonably good approximations for these two populations near their centers and between their two centers (though not far from the centers, where the densities are small). In such a case one may want to divide the sample space into the two regions of classification by some simple curve or surface. The simplest is a line or hyperplane; the procedure may then be termed linear.

Let b ($\neq 0$) be a vector (of p components) and c a scalar. An observation x is classified as from the first population if $b'x \geq c$ and as from the second if $b'x < c$. We are primarily interested in situations where the important difference between the two populations is the difference between the centers; we assume $\mu^{(1)} \neq \mu^{(2)}$ as well as $\Sigma_1 \neq \Sigma_2$, and that Σ_1 and Σ_2 are nonsingular.

When sampling from the ith population, $b'x$ has a univariate normal distribution with mean $\mathscr{E}(b'x|\pi_i) = b'\mu^{(i)}$ and variance

$$\mathscr{E}\left(b'x - b'\mu^{(i)}\right)^2 | \pi_i = \mathscr{E}b'\left(x - \mu^{(i)}\right)\left(x - \mu^{(i)}\right)'b | \pi_i = b'\Sigma_i b. \tag{6}$$

The probability of misclassifying an observation when it comes from the first population is

$$\begin{aligned} P(2|1) &= \Pr\{b'x < c | \pi_1\} = \Pr\left\{ \frac{b'x - b'\mu^{(1)}}{(b'\Sigma_1 b)^{\frac{1}{2}}} < \frac{c - b'\mu^{(1)}}{(b'\Sigma_1 b)^{\frac{1}{2}}} \middle| \pi_1 \right\} \\ &= \Phi\left(\frac{c - b'\mu^{(1)}}{(b'\Sigma_1 b)^{\frac{1}{2}}} \right) = 1 - \Phi\left(\frac{b'\mu^{(1)} - c}{(b'\Sigma_1 b)^{\frac{1}{2}}} \right). \end{aligned} \tag{7}$$

The probability of misclassifying an observation when it comes from the second population is

$$\begin{aligned} P(1|2) &= \Pr\{b'x \geq c | \pi_2\} = \Pr\left\{ \frac{b'x - b'\mu^{(2)}}{(b'\Sigma_2 b)^{\frac{1}{2}}} \geq \frac{c - b'\mu^{(2)}}{(b'\Sigma_2 b)^{\frac{1}{2}}} \middle| \pi_2 \right\} \\ &= 1 - \Phi\left(\frac{c - b'\mu^{(2)}}{(b'\Sigma_2 b)^{\frac{1}{2}}} \right). \end{aligned} \tag{8}$$

It is desired to make these probabilities small or, equivalently, to make the arguments

$$y_1 = \frac{b'\mu^{(1)} - c}{(b'\Sigma_1 b)^{\frac{1}{2}}}, \qquad y_2 = \frac{c - b'\mu^{(2)}}{(b'\Sigma_2 b)^{\frac{1}{2}}} \tag{9}$$

large. We shall consider making y_1 large for given y_2.

When we eliminate c from (9), we obtain

$$y_1 = \left[b'\gamma - y_2(b'\Sigma_2 b)^{\frac{1}{2}}\right](b'\Sigma_1 b)^{-\frac{1}{2}}, \tag{10}$$

where $\gamma = \mu^{(1)} - \mu^{(2)}$. To maximize y_1 for given y_2 we differentiate y_1 with respect to b to obtain

$$\frac{\partial y_1}{\partial b} = \left[\gamma - y_2(b'\Sigma_2 b)^{-\frac{1}{2}}\Sigma_2 b\right](b'\Sigma_1 b)^{-\frac{1}{2}} - \left[b'\gamma - y_2(b'\Sigma_2 b)^{\frac{1}{2}}\right](b'\Sigma_1 b)^{-\frac{3}{2}}\Sigma_1 b. \tag{11}$$

If we let

$$t_1 = \frac{b'\gamma - y_2(b'\Sigma_2 b)^{\frac{1}{2}}}{b'\Sigma_1 b}, \tag{12}$$

$$t_2 = \frac{y_2}{\sqrt{b'\Sigma_2 b'}}, \tag{13}$$

then (11) set equal to $\mathbf{0}$ is

$$(t_1\Sigma_1 + t_2\Sigma_2)b = \gamma. \tag{14}$$

Note that (13) and (14) imply (12). If there is a pair t_1, t_2, and a vector b satisfying (12) and (13), then c is obtained from (9) as

$$c = y_2\sqrt{b'\Sigma_2 b} + b'\mu^{(2)} = t_2 b'\Sigma_2 b + b'\mu^{(2)}. \tag{15}$$

Then from (9), (12), and (13)

$$y_1 = \frac{b'\mu^{(1)} - \left(t_2 b'\Sigma_2 b + b'\mu^{(2)}\right)}{\sqrt{b'\Sigma_1 b}} = t_1\sqrt{b'\Sigma_1 b}. \tag{16}$$

Now consider (14) as a function of t $(0 \le t \le 1)$. Let $t_1 = t$ and $t_2 = 1 - t$; then $b = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}\gamma$. Define $v_1 = t_1\sqrt{b'\Sigma_1 b}$ and $v_2 = t_2\sqrt{b'\Sigma_2 b}$. The derivative of v_1^2 with respect to t is

$$\begin{aligned}
&\frac{d}{dt}t^2\gamma'[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\Sigma_1[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\gamma \\
&= 2t\gamma'[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\Sigma_1[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\gamma \\
&\quad - t^2\gamma'[t\Sigma_1 + (1-t)\Sigma_2]^{-1}(\Sigma_1 - \Sigma_2)[t\Sigma_1 + (1-t)\Sigma_2]^{-1} \\
&\qquad \cdot\Sigma_1[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\gamma \\
&\quad - t^2\gamma'[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\Sigma_1[t\Sigma_1 + (1-t)\Sigma_2]^{-1} \\
&\qquad \cdot(\Sigma_1 - \Sigma_2)[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\gamma \\
&= t\gamma'[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\left\{\Sigma_2[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\Sigma_1\right. \\
&\qquad \left. + \Sigma_1[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\Sigma_2\right\}[t\Sigma_1 + (1-t)\Sigma_2]^{-1}\gamma
\end{aligned} \tag{17}$$

by the following lemma.

Lemma 6.10.1. *If Σ_1 and Σ_2 are positive definite and $t_1 > 0, t_2 > 0$, then*

$$\Sigma_2[t_1\Sigma_1 + t_2\Sigma_2]^{-1}\Sigma_1 \tag{18}$$

is positive definite.

Proof. The matrix (18) is

$$\Sigma_2\big[\Sigma_1(t_1\Sigma_2^{-1} + t_2\Sigma_1^{-1})\Sigma_2\big]^{-1}\Sigma_1 = \big(t_1\Sigma_2^{-1} + t_2\Sigma_1^{-1}\big)^{-1}. \tag{19}$$ ■

Similarly $dv_2^2/dt < 0$. Since $v_1 \geq 0$, $v_2 \geq 0$, we see that v_1 increases with t from 0 at $t = 0$ to $\sqrt{\gamma'\Sigma_1^{-1}\gamma}$ at $t = 1$ and v_2 decreases from $\sqrt{\gamma'\Sigma_2^{-1}\gamma}$ at $t = 0$ to 0 at $t = 1$. The coordinates v_1 and v_2 are continuous functions of t. For given y_2, $0 \leq y_2 \leq \sqrt{\gamma'\Sigma_2^{-1}\gamma}$, there is a t such that $y_2 = v_2 = t_2\sqrt{b'\Sigma_2 b}$ and b satisfies (14) for $t_1 = t$ and $t_2 = 1 - t$. Then $y_1 = v_1 = t_1\sqrt{b'\Sigma_1 b}$ maximizes y_1 for that value of y_2. Similarly given y_1, $0 \leq y_1 \leq \sqrt{\gamma'\Sigma_1^{-1}\gamma}$, there is a t such that $y_1 = v_1 = t_1\sqrt{b'\Sigma_1 b}$ and b satisfies (14) for $t_1 = t$ and $t_2 = 1 - t$, and $y_2 = v_2 = t_2\sqrt{b'\Sigma_2 b}$ maximizes y_2. Note that $y_1 \geq 0$, $y_2 \geq 0$ implies the errors of misclassification are not greater than $\frac{1}{2}$.

We now argue that the set of y_1, y_2 defined this way correspond to admissible linear procedures. Let x_1, x_2 be in this set, and suppose another procedure defined by z_1, z_2 were better than x_1, x_2, that is, $x_1 \leq z_1$, $x_2 \leq z_2$ with at least one strict inequality. For $y_1 = z_1$ let y_2^* be the maximum y_2 among linear procedures; then $z_1 = y_1$, $z_2 \leq y_2^*$ and hence $x_1 \leq y_1$, $x_2 \leq y_2^*$. However, this is possible only if $x_1 = y_1$, $x_2 = y_2^*$, because $dy_1/dy_2 < 0$. Now we have a contradiction to the assumption that z_1, z_2 was better than x_1, x_2. Thus x_1, x_2 corresponds to an admissible linear procedure.

Use of Admissible Linear Procedures

Given t_1 and t_2 such that $t_1\Sigma_1 + t_2\Sigma_2$ is positive definite, one would compute the optimum b by solving the linear equations (15) and then compute c by one of (9). Usually t_1 and t_2 are not given, but a desired solution is specified in another way. We consider three ways.

Minimization of One Probability of Misclassification for a Specified Probability of the Other

Suppose we are given y_2 (or, equivalently, the probability of misclassification when sampling from the second distribution) and we want to maximize y_1 (or, equivalently, minimize the probability of misclassification when sampling from the first distribution). Suppose $y_2 > 0$ (i.e., the given probability of misclassification is less than $\frac{1}{2}$). Then if the maximum $y_1 \geq 0$, we want to find $t_2 = 1 - t_1$ such that $y_2 = t_2(b'\Sigma_2 b)^{\frac{1}{2}}$, where $b = [t_1\Sigma_1 + t_2\Sigma_2]^{-1}\gamma$. The solution can be approximated by trial and error, since y_2 is an increasing function of t_2. For $t_2 = 0$, $y_2 = 0$; and for $t_2 = 1$, $y_2 = (b'\Sigma_2 b)^{\frac{1}{2}} = (b'\gamma)^{\frac{1}{2}} = (\gamma'\Sigma_2^{-1}\gamma)$, where $\Sigma_2 b = \gamma$. One could try other values of t_2 successively by solving (14) and inserting in $b'\Sigma_2 b$ until $t_2(b'\Sigma_2 b)^{\frac{1}{2}}$ agreed closely enough with the desired y_2. [$y_1 > 0$ if the specified $y_2 < (\gamma'\Sigma_2^{-1}\gamma)^{\frac{1}{2}}$.]

The Minimax Procedure

The minimax procedure is the admissible procedure for which $y_1 = y_2$. Since for this procedure both probabilities of correct classification are greater than $\frac{1}{2}$, we have $y_1 = y_2 > 0$ and $t_1 > 0$, $t_2 > 0$. We want to find t $(= t_1 = 1 - t_2)$ so that

$$\begin{aligned} (20) \qquad 0 = y_1^2 - y_2^2 &= t^2 b'\Sigma_1 b - (1-t)^2 b'\Sigma_2 b \\ &= b'\left[t^2\Sigma_1 - (1-t)^2\Sigma_2\right]b. \end{aligned}$$

Since y_1^2 increases with t and y_2^2 decreases with increasing t, there is one and only one solution to (20), and this can be approximated by trial and error by guessing a value of t $(0 < t < 1)$, solving (14) for b, and computing the quadratic form on the right of (20). Then another t can be tried.

An alternative approach is to set $y_1 = y_2$ in (9) and solve for c. Then the common value of $y_1 = y_2$ is

$$(21) \qquad \frac{b'\gamma}{(b'\Sigma_1 b)^{\frac{1}{2}} + (b'\Sigma_2 b)^{\frac{1}{2}}},$$

and we want to find b to maximize this, where b is of the form

$$(22) \qquad [t\Sigma_1 + (1-t)\Sigma_2]^{-1}\gamma$$

with $0 < t < 1$.

When $\Sigma_1 = \Sigma_2$, twice the maximum of (21) is the squared Mahalanobis distance between the populations. This suggests that when Σ_1 may be unequal to Σ_2, twice the maximum of (21) might be called the distance between the populations.

Welch and Wimpress (1961) have programmed the minimax procedure and applied it to the recognition of spoken sounds.

Case of A Priori Probabilities

Suppose we are given a priori probabilities, q_1 and q_2, of the first and second populations, respectively. Then the probability of a misclassification is

$$(23) \quad q_1[1 - \Phi(y_1)] + q_2[1 - \Phi(y_2)] = 1 - [q_1\Phi(y_1) + q_2\Phi(y_2)],$$

which we want to minimize. The solution will be an admissible linear procedure. If we know it involves $y_1 \geq 0$ and $y_2 \geq 0$, we can substitute $y_1 = t(b'\Sigma_1 b)^{\frac{1}{2}}$ and $y_2 = (1-t)(b'\Sigma_2 b)^{\frac{1}{2}}$, where $b = [t\Sigma_1 + (1-t)\Sigma_2]^{-1}\gamma$, into (23) and set the derivative of (23) with respect to t equal to 0, obtaining

$$(24) \qquad q_1\phi(y_1)\frac{dy_1}{dt} + q_2\phi(y_2)\frac{dy_2}{dt} = 0,$$

where $\phi(u) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}u^2}$. There does not seem to be any easy or direct way of solving (24) for t. The left-hand side of (24) is not necessarily monotonic. In fact, there may be several roots to (24). If there are, the absolute minimum will be found by putting the solution into (23). (We remind the reader that the curve of admissible error probabilities is not necessary convex.)

Anderson and Bahadur (1962) studied these linear procedures in general, including $y_1 < 0$ and $y_2 < 0$. Clunies-Ross and Riffenburgh (1960) approached the problem from a more geometric point of view.

PROBLEMS

6.1. (Sec. 6.3) Let π_i be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. Find the form of the admissible classification procedures.

6.2. (Sec. 6.3) Prove that every complete class of procedures includes the class of admissible procedures.

6.3. (Sec. 6.3) Prove that if the class of admissible procedures is complete, it is minimal complete.

6.4. (Sec. 6.3) The *Neyman–Pearson fundamental lemma* states that of all tests at a given significance level of the null hypothesis that x is drawn from $p_1(x)$ against alternative that it is drawn from $p_2(x)$ the most powerful test has the critical region $p_1(x)/p_2(x) < k$. Show that the discussion in Section 6.3 proves this result.

6.5. (Sec. 6.3) When $p(x) = n(x|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ find the best test of $\boldsymbol{\mu} = \mathbf{0}$ against $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ at significance level ε. Show that this test is uniformly most powerful against all alternatives $\boldsymbol{\mu} = c\boldsymbol{\mu}^*$, $c > 0$. Prove that there is no uniformly most powerful test against $\boldsymbol{\mu} = \boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}^{(2)}$ unless $\boldsymbol{\mu}^{(1)} = c\boldsymbol{\mu}^{(2)}$ for some $c > 0$.

6.6. (Sec. 6.4) Let $P(2|1)$ and $P(1|2)$ be defined by (14) and (15). Prove if $-\frac{1}{2}\Delta^2 < c < \frac{1}{2}\Delta^2$, then $P(2|1)$ and $P(1|2)$ are decreasing functions of Δ.

6.7. (Sec. 6.4) Let $x' = (x^{(1)\prime}, x^{(2)\prime})$. Using Problem 5.23 and Problem 6.6, prove that the class of classification procedures based on x is uniformly as good as the class of procedures based on $x^{(1)}$.

6.8. (Sec. 6.5.1) Find the criterion for classifying irises as *Iris setosa* or *Iris versicolor* on the basis of data given in Section 5.3.4. Classify a random sample of 5 *Iris virginica* in Table 3.4.

6.9. (Sec. 6.5.1) Let $W(x)$ be the classification criterion given by (2). Show that the T^2-criterion for testing $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}) = N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$ is proportional to $W(\bar{\boldsymbol{x}}^{(1)})$ and $W(\bar{\boldsymbol{x}}^{(2)})$.

6.10. (Sec. 6.5.1) Show that the probabilities of misclassification of $x_1, \dots, x_N$ (all assumed to be from either π_1 or π_2) decrease as N increases.

6.11. (Sec. 6.5) Show that the elements of M are invariant under the transformation (34) and that any function of the sufficient statistics that is invariant is a function of M.

6.12. (Sec. 6.5) Consider $d'x^{(i)}$. Prove that the ratio

$$\frac{\left(d'\bar{x}^{(1)} - d'\bar{x}^{(2)}\right)^2}{\sum_{\alpha=1}^{N_1}\left(d'x_\alpha^{(1)} - d'\bar{x}^{(1)}\right)^2 + \sum_{\alpha=1}^{N_2}\left(d'x_\alpha^{(2)} - d'\bar{x}^{(2)}\right)^2}$$

6.13. (Sec. 6.6) Show that the derivative of (2) to terms of order n^{-1} is

$$-\phi\left(\tfrac{1}{2}\Delta\right)\left\{\frac{1}{2} + \frac{1}{n}\left[\frac{p-1}{\Delta^2} + \frac{p-2}{4} - \frac{p}{8}\Delta^2\right]\right\}.$$

6.14. (Sec. 6.6) Show $\mathscr{E}D^2$ is (4). [*Hint:* Let $\boldsymbol{\Sigma} = \boldsymbol{I}$ and show that $\mathscr{E}(\boldsymbol{S}^{-1}|\boldsymbol{\Sigma} = \boldsymbol{I}) = [n/(n-p-1)]\boldsymbol{I}$.]

6.15. (Sec. 6.6.2) Show

$$\begin{aligned}
&\Pr\left\{\frac{Z - \frac{1}{2}D^2}{D} \le u \,\middle|\, \pi_1\right\} - \Pr\left\{\frac{Z - \frac{1}{2}\Delta^2}{\Delta} \le u \,\middle|\, \pi_1\right\} \\
&\quad = \phi(u)\left\{\frac{1}{2N_1\Delta^2}\left[u^3 + (p-3)u - \Delta^2 u + p\Delta\right]\right. \\
&\qquad + \frac{1}{2N_2\Delta^2}\left[u^3 + 2\Delta u^2 + (p - 3 + \Delta^2)u - \Delta^3 + p\Delta\right] \\
&\qquad \left. + \frac{1}{4n}\left[3u^3 + 4\Delta u^2 + (2p - 3 + \Delta^2)u + 2(p-1)\Delta\right]\right\} + O(n^{-2}).
\end{aligned}$$

6.16. (Sec. 6.8) Let π_i be $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$, $i = 1, \dots, m$. If the $\boldsymbol{\mu}^{(i)}$ are on a line (i.e., $\boldsymbol{\mu}^{(i)} = \boldsymbol{\mu} + \nu_i \boldsymbol{\beta}$), show that for admissible procedures the R_i are defined by parallel planes. Thus show that only one discriminant function $u_{jk}(x)$ need be used.

6.17. (Sec. 6.8) In Section 8.8 data are given on samples from four populations of skulls. Consider the first two measurements and the first three samples. Construct the classification functions $u_{ij}(x)$. Find the procedure for $q_i = N_i/(N_1 + N_2 + N_3)$. Find the minimax procedure.

6.18. (Sec. 6.10) Show that $b'x = c$ is the equation of a plane that is tangent to an ellipsoid of constant density of π_1 and to an ellipsoid of constant density of π_2 at a common point.

6.19. (Sec. 6.8) Let $x_1^{(i)}, \dots, x_{N_i}^{(i)}$ be observations from $N(\mu^{(i)}, \Sigma)$, $i = 1, 2, 3$, and let x be an observation to be classified. Give explicitly the maximum likelihood rule.

6.20. (Sec. 6.5) Verify (33).

CHAPTER 7

The Distribution of the Sample Covariance Matrix and the Sample Generalized Variance

7.1. INTRODUCTION

The sample covariance matrix, $S = [1/(N-1)]\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$, is an unbiased estimator of the population covariance matrix Σ. In Section 4.2 we found the density of $A = (N-1)S$ in the case of a 2×2 matrix. In Section 7.2 this result will be generalized to the case of a matrix A of any order. When $\Sigma = I$, this distribution is in a sense a generalization of the χ^2-distribution. The distribution of A (or S), often called the Wishart distribution, is fundamental to multivariate statistical analysis. In Sections 7.3 and 7.4 we discuss some properties of the Wishart distribution.

The generalized variance of the sample is defined as $|S|$ in Section 7.5; it is a measure of the scatter of the sample. Its distribution is characterized. The density of the set of all correlation coefficients when the components of the observed vector are independent is obtained in Section 7.6.

The inverted Wishart distribution is introduced in Section 7.7 and is used as an a priori distribution of Σ to obtain a Bayes estimator of the covariance matrix. In Section 7.8 we consider improving on S as an estimator of Σ with respect to two loss functions. Section 7.9 treats the distributions for sampling from elliptically contoured distributions.

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

7.2. THE WISHART DISTRIBUTION

We shall obtain the distribution of $A=\sum_{\alpha=1}^{N}(X_\alpha-\bar{X})(X_\alpha-\bar{X})'$, where $X_1,\dots,X_N$ $(N>p)$ are independent, each with the distribution $N(\boldsymbol{\mu},\Sigma)$. As was shown in Section 3.3, A is distributed as $\sum_{\alpha=1}^{n}Z_\alpha Z_\alpha'$, where $n=N-1$ and $Z_1,\dots,Z_n$ are independent, each with the distribution $N(\mathbf{0},\Sigma)$. We shall show that the density of A for A positive definite is

$$(1)\qquad \frac{|A|^{\frac{1}{2}(n-p-1)}\exp\left(-\frac{1}{2}\operatorname{tr}\Sigma^{-1}A\right)}{2^{\frac{1}{2}np}\pi^{p(p-1)/4}|\Sigma|^{\frac{1}{2}n}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]}.$$

We shall first consider the case of $\Sigma=I$. Let

$$(2)\qquad (Z_1,\dots,Z_n)=\begin{pmatrix} v_1' \\ \vdots \\ v_p' \end{pmatrix}.$$

Then the elements of $A=(a_{ij})$ are inner products of these n-component vectors, $a_{ij}=v_i'v_j$. The vectors $v_1,\dots,v_p$ are independently distributed, each according to $N(\mathbf{0},I_n)$. It will be convenient to transform to new coordinates according to the Gram–Schmidt orthogonalization. Let $w_1=v_1$,

$$(3)\qquad w_i=v_i-\sum_{j=1}^{i-1}w_j\frac{w_j'v_i}{v_j'w_j},\qquad i=2,\dots,p.$$

We prove by induction that w_k is orthogonal to w_i, $k<i$. Assume $w_k'w_h=0$, $k\neq h$, $k,h=1,\dots,i-1$; then take the inner product of w_k and (3) to obtain $w_k'w_i=0$, $k=1,\dots,i-1$. (Note that $\Pr\{\|w_i\|=0\}=0$.)

Define $t_{ii}=\|w_i\|=\sqrt{w_i'w_i}$, $i=1,\dots,p$, and $t_{ij}=v_i'w_j/\|w_j\|$, $j=1,\dots,i-1$, $i=2,\dots,p$. Since $v_i=\sum_{j=1}^{i}(t_{ij}/\|w_j\|)w_j$,

$$(4)\qquad a_{hi}=v_h'v_i=\sum_{j=1}^{\min(h,i)}t_{hj}t_{ij}.$$

If we define the lower triangular matrix $T=(t_{ij})$ with $t_{ii}>0$, $i=1,\dots,p$, and $t_{ij}=0$, $i<j$, then

$$(5)\qquad A=TT'.$$

Note that t_{ij}, $j=1,\dots,i-1$, are the first $i-1$ coordinates of v_i in the coordinate system with $w_1,\dots,w_{i-1}$ as the first $i-1$ coordinate axes. (See Figure 7.1.) The sum of the other $n-i+1$ coordinates squared is $\|v_i\|^2-\sum_{j=1}^{i-1}t_{ij}^2=t_{ii}^2=\|w_i\|^2$; w_i is the vector from v_i to its projection on $w_1,\dots,w_{i-1}$ (or equivalently on $v_1,\dots,v_{i-1}$).

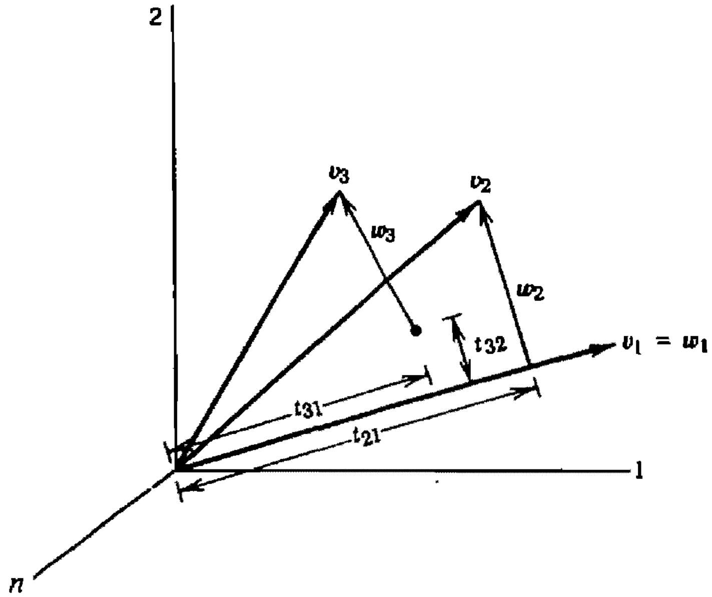

Figure 7.1. Transformation of coordinates.

Lemma 7.2.1. *Conditional on $w_1, \dots, w_{i-1}$ (or equivalently on $v_1, \dots, v_{i-1}$), $t_{i1}, \dots, t_{i,i-1}$ and t_{ii}^2 are independently distributed; t_{ij} is distributed according to $N(0,1)$, $i > j$; and t_{ii}^2 has the χ^2-distribution with $n - i + 1$ degrees of freedom.*

Proof. The coordinates of v_i referred to the new orthogonal coordinates with $v_1, \dots, v_{i-1}$ defining the first coordinate axes are independently normally distributed with means 0 and variances 1 (Theorem 3.3.1). t_{ii}^2 is the sum of the coordinates squared omitting the first $i - 1$. ■

Since the conditional distribution of $t_{i1}, \dots, t_{ii}$ does not depend on $v_1, \dots, v_{i-1}$, they are distributed independently of $t_{11}, t_{21}, t_{22}, \dots, t_{i-1,i-1}$.

Corollary 7.2.1. *Let $Z_1, \dots, Z_n$ $(n \geq p)$ be independently distributed, each according to $N(\mathbf{0}, \boldsymbol{I})$; let $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha' = \boldsymbol{T}\boldsymbol{T}'$, where $t_{ij} = 0$, $i < j$, and $t_{ii} > 0$, $i = 1, \dots, p$. Then $t_{11}, t_{21}, \dots, t_{pp}$ are independently distributed; t_{ij} is distributed according to $N(0,1)$, $i > j$; and t_{ii}^2 has the χ^2-distribution with $n - i + 1$ degrees of freedom.*

Since t_{ii} has density $2^{-\frac{1}{2}(n-i-1)} t^{n-i} e^{-\frac{1}{2}t^2} / \Gamma[\frac{1}{2}(n+1-i)]$, the joint density of t_{ij}, $j = 1, \dots, i$, $i = 1, \dots, p$, is

$$\prod_{i=1}^{p} \frac{t_{ii}^{n-i} \exp\left(-\frac{1}{2}\sum_{j=1}^{i} t_{ij}^2\right)}{\pi^{\frac{1}{2}(i-1)} 2^{\frac{1}{2}n-1} \Gamma\left[\frac{1}{2}(n+1-i)\right]} \tag{6}$$

$$= \frac{\prod_{i=1}^{p} t_{ii}^{n-i} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{i} t_{ij}^2\right)}{2^{\frac{1}{2}p(n-2)} \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left[\frac{1}{2}(n+1-i)\right]}.$$

Let C be a lower triangular matrix ($c_{ij} = 0$, $i < j$) such that $\Sigma = CC'$ and $c_{ii} > 0$. The linear transformation $T^* = CT$, that is,

$$t_{ij}^* = \sum_{k=j}^{i} c_{ik} t_{kj}, \qquad i \geq j, \tag{7}$$
$$= 0, \qquad i < j,$$

can be written

$$\begin{bmatrix} t_{11}^* \\ t_{21}^* \\ t_{22}^* \\ \vdots \\ t_{p1}^* \\ \vdots \\ t_{pp}^* \end{bmatrix} = \begin{bmatrix} c_{11} & 0 & 0 & \cdots & 0 & \cdots & 0 \\ x & c_{22} & 0 & \cdots & 0 & \cdots & 0 \\ x & x & c_{22} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ x & x & x & \cdots & c_{pp} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ x & x & x & \cdots & x & \cdots & c_{pp} \end{bmatrix} \begin{bmatrix} t_{11} \\ t_{21} \\ t_{22} \\ \vdots \\ t_{p1} \\ \vdots \\ t_{pp} \end{bmatrix}, \tag{8}$$

where x denotes an element, possibly nonzero. Since the matrix of the transformation is triangular, its determinant is the product of the diagonal elements, namely, $\prod_{i=1}^p c_{ii}^i$. The Jacobian of the transformation from T to T^* is the reciprocal of the determinant. The density of T^* is obtained by substituting into (6) $t_{ii} = t_{ii}^*/c_{ii}$ and

$$\sum_{i=1}^{p} \sum_{j=1}^{i} t_{ij}^2 = \operatorname{tr} TT' \tag{9}$$
$$= \operatorname{tr} C^{-1} T^* T^{*\prime} (C^{-1})'$$
$$= \operatorname{tr} T^* T^{*\prime} C'^{-1} C^{-1}$$
$$= \operatorname{tr} T^* T^{*\prime} \Sigma^{-1} = \operatorname{tr} T^{*\prime} \Sigma^{-1} T^*,$$

and using $\prod_{i=1}^p c_{ii}^2 = |C||C'| = |\Sigma|$.

Theorem 7.2.1. *Let $Z_1, \dots, Z_n$ $(n \geq p)$ be independently distributed, each according to $N(\mathbf{0}, \Sigma)$; let $A = \sum_{\alpha=1}^n Z_\alpha Z_\alpha'$; and let $A = T^* T^{*\prime}$, where $t_{ij}^* = 0$, $i < j$, and $t_{ii}^* > 0$. Then the density of T^* is*

$$\frac{\prod_{i=1}^p t_{ii}^{*\,n-i} e^{-\frac{1}{2}\operatorname{tr} \Sigma^{-1} T^* T^{*\prime}}}{2^{\frac{1}{2}p(n-2)} \pi^{p(p-1)/4} |\Sigma|^{\frac{1}{2}n} \prod_{i=1}^p \Gamma\left[\frac{1}{2}(n+1-i)\right]}. \tag{10}$$

We can write (4) as $a_{hi} = \sum_{j=1}^{i} t_{hj}^* t_{ij}^*$ for $h \geq i$. Then

$$\frac{\partial a_{hi}}{\partial t_{kl}^*} = 0, \qquad k > h, \tag{11}$$

$$= 0, \qquad k = h, \quad l > i;$$

that is, $\partial a_{hi} / \partial t_{kl}^* = 0$ if k, l is beyond h, i in the lexicographic ordering. The Jacobian of the transformation from A to T^* is the determinant of the lower triangular matrix with diagonal elements

$$\frac{\partial a_{hh}}{\partial t_{hh}^*} = 2t_{hh}^*, \tag{12}$$

$$\frac{\partial a_{hi}}{\partial t_{hi}^*} = t_{ii}^*, \qquad h > i, \tag{13}$$

The Jacobian is therefore $2^p \prod_{i=1}^{p} t_{ii}^{*\,p+1-i}$. The Jacobian of the transformation from T^* to A is the reciprocal.

Theorem 7.2.2. *Let $Z_1, \ldots, Z_n$ be independently distributed, each according to $N(\mathbf{0}, \Sigma)$. The density of $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ is*

$$\frac{|A|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\operatorname{tr} \Sigma^{-1} A}}{2^{\frac{1}{2}pn} \pi^{p(p-1)/4} |\Sigma|^{\frac{1}{2}n} \prod_{i=1}^{p} \Gamma\left[\frac{1}{2}(n+1-i)\right]} \tag{14}$$

for A positive definite, and 0 otherwise.

Corollary 7.2.2. *Let $X_1, \ldots, X_N$ $(N > p)$ be independently distributed, each according to $N(\boldsymbol{\mu}, \Sigma)$. Then the density of $A = \sum_{\alpha=1}^{N} (X_\alpha - \bar{X})(X_\alpha - \bar{X})'$ is (14) for $n = N - 1$.*

The density (14) will be denoted by $w(A|\Sigma, n)$, and the associated distribution will be termed $W(\Sigma, n)$. If $n < p$, then A does not have a density, but its distribution is nevertheless defined, and we shall refer to it as $W(\Sigma, n)$.

Corollary 7.2.3. *Let $X_1, \ldots, X_N$ $(N > p)$ be independently distributed, each according to $N(\boldsymbol{\mu}, \Sigma)$. The distribution of $S = (1/n)\sum_{\alpha=1}^{N} (X_\alpha - \bar{X})(X_\alpha - \bar{X})'$ is $W[(1/n)\Sigma, n]$, where $n = N - 1$.*

Proof. S has the distribution of $\sum_{\alpha=1}^{n} [(1/\sqrt{n}) Z_\alpha][(1/\sqrt{n}) Z_\alpha]'$, where $(1/\sqrt{n}) Z_1, \ldots, (1/\sqrt{n}) Z_N$ are independently distributed, each according to $N(\mathbf{0}, (1/n)\Sigma)$. Theorem 7.2.2 implies this corollary. ∎

The Wishart distribution for $p = 2$ as given in Section 4.2.1 was derived by Fisher (1915). The distribution for arbitrary p was obtained by Wishart (1928) by a geometric argument using $v_1, \dots, v_p$ defined above. As noted in Section 3.2, the ith diagonal element of A is the squared length of the ith vector, $a_{ii} = v_i'v_i = \|v_i\|^2$, and the i, jth off-diagonal element of A is the product of the lengths of v_i and v_j and the cosine of the angle between them. The matrix A specifies the lengths and configuration of the vectors.

We shall give a geometric interpretation[†] of the derivation of the density of the rectangular coordinates t_{ij}, $i \geq j$, when $\Sigma = I$. The probability element of t_{11} is approximately the probability that $\|v_1\|$ lies in the interval $t_{11} < \|v_1\| < t_{11} + dt_{11}$. This is the probability that v_1 falls in a spherical shell in n dimensions with inner radius t_{11} and thickness dt_{11}. In this region, the density $(2\pi)^{-\frac{1}{2}n} \exp(-\frac{1}{2}v_1'v_1)$ is approximately constant, namely, $(2\pi)^{-\frac{1}{2}n} \exp(-\frac{1}{2}t_{11}^2)$. The surface area of the unit sphere in n dimensions is $C(n) = 2\pi^{\frac{1}{2}n}/\Gamma(\frac{1}{2}n)$ (Problems 7.1–7.3), and the volume of the spherical shell is approximately $C(n)t_{11}^{n-1}\,dt_{11}$. The probability element is the product of the volume and approximate density, namely,

$$2^{-(\frac{1}{2}n-1)} t_{11}^{n-1} \exp\left(-\tfrac{1}{2}t_{11}^2\right) dt_{11} / \Gamma\left(\tfrac{1}{2}n\right). \tag{15}$$

The probability element of $t_{i1}, \dots, t_{i,i-1}, t_{ii}$ given $v_1, \dots, v_{i-1}$ (i.e., given $w_1, \dots, w_{i-1}$) is approximately the probability that v_i falls in the region for which $t_{i1} < v_i'w_1/\|w_1\| < t_{i1} + dt_{i1}, \dots, t_{i,i-1} < v_i'w_{i-1}/\|w_{i-1}\| < t_{i,i-1} + dt_{i,i-1}$, and $t_{ii} < \|w_i\| < t_{ii} - dt_{ii}$, where w_i is the projection of v_i on the $(n-i+1)$-dimensional space orthogonal to $w_1, \dots, w_{i-1}$. Each of the first $i-1$ pairs of inequalities defines the region between two hyperplanes (the different pairs being orthogonal). The last pair of inequalities defines a cylindrical shell whose intersection with the $(i-1)$-dimensional hyperplane spanned by $v_1, \dots, v_{i-1}$ is a spherical shell in $n-i+1$ dimensions with inner radius t_{ii}. In this region the density $(2\pi)^{-\frac{1}{2}n} \exp(-\frac{1}{2}v_i'v_i)$ is approximately constant, namely, $(2\pi)^{-\frac{1}{2}n} \exp(-\frac{1}{2}\Sigma_{j=1}^{i} t_{ij}^2)$. The volume of the region is approximately $dt_{i1} \cdots dt_{i,i-1}\, C(n-i+1)t_{ii}^{n-i}\,dt_{ii}$. The probability element is

$$\frac{2^{-(\frac{1}{2}n-1)}\pi^{-\frac{1}{2}(i-1)} t_{ii}^{n-i} \exp\left(-\frac{1}{2}\Sigma_{j=1}^{i} t_{ij}^2\right)}{\Gamma\left[\frac{1}{2}(n+1-i)\right]} dt_{i1} \cdots dt_{ii}. \tag{16}$$

Then the product of (15) and (16) for $i = 2, \dots, p$ is (6) times $dt_{11} \cdots dt_{pp}$.

This analysis, which exactly parallels the geometric derivation by Wishart [and later by Mahalanobis, Bose, and Roy (1937)], was given by Sverdrup

[†] In the first edition of this book, the derivation of the Wishart distribution and its geometric interpretation were in terms of the nonorthogonal vectors $v_1, \dots, v_p$.

(1947) [and by Fog (1948) for $p = 3$]. Another method was used by Madow (1938), who drew on the distribution of correlation coefficients (for $\Sigma = I$) obtained by Hotelling by considering certain partial correlation coefficients. Hsu (1939b) gave an inductive proof, and Rasch (1948) gave a method involving the use of a functional equation. A different method is to obtain the characteristic function and invert it, as was done by Ingham (1933) and by Wishart and Bartlett (1933).

Cramér (1946) verified that the Wishart distribution has the characteristic function of A. By means of alternative matrix transformations Elfving (1947), Mauldon (1955), and Olkin and Roy (1954) derived the Wishart distribution via the *Bartlett decomposition*; Kshirsagar (1959) based his derivation on random orthogonal transformations. Narain (1948), (1950) and Ogawa (1953) used a regression approach. James (1954), Khatri and Ramachandran (1958), and Khatri (1963) applied different methods. Giri (1977) used invariance. Wishart (1948) surveyed the derivations up to that date. Some of these methods are indicated in the problems.

The relation $A = TT'$ is known as the *Bartlett decomposition* [Bartlett (1939)], and the (nonzero) elements of T were termed *rectangular coordinates* by Mahalanobis, Bose, and Roy (1937).

Corollary 7.2.4

$$\int_{B>0}\cdots\int |B|^{t-\frac{1}{2}(p+1)}e^{-\operatorname{tr} B}\,dB = \pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left[t-\tfrac{1}{2}(i-1)\right]. \tag{17}$$

Proof. Here $B > \mathbf{0}$ denotes B positive definite. Since (14) is a density, its integral for $A > \mathbf{0}$ is 1. Let $\Sigma = I$, $A = 2B$ ($dA = 2\,dB$), and $n = 2t$. Then the fact that the integral is 1 is identical to (17) for t a half integer. However, if we derive (14) from (6), we can let n be any real number greater than $p-1$. In fact (17) holds for complex t such that $\mathscr{R}t > p-1$. ($\mathscr{R}t$ means the real part of t.) ∎

Definition 7.2.1. *The multivariate gamma function is*

$$\Gamma_p(t) = \pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left[t-\tfrac{1}{2}(i-1)\right]. \tag{18}$$

The Wishart density can be written

$$w(A|\Sigma, n) = \frac{|A|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}\operatorname{tr}\Sigma^{-1}A}}{2^{\frac{1}{2}pn}|\Sigma|^{\frac{1}{2}n}\Gamma_p\left(\frac{1}{2}n\right)}. \tag{19}$$

7.3. SOME PROPERTIES OF THE WISHART DISTRIBUTION

7.3.1. The Characteristic Function

The characteristic function of the Wishart distribution can be obtained directly from the distribution of the observations. Suppose $Z_1, \ldots, Z_n$ are distributed independently, each with density

$$\frac{1}{(2\pi)^{\frac{1}{2}p}|\Sigma|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}z'\Sigma^{-1}z\right). \tag{1}$$

Let

$$A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'. \tag{2}$$

Introduce the $p \times p$ matrix $\Theta = (\theta_{ij})$ with $\theta_{ij} = \theta_{ji}$. The characteristic function of $A_{11}, A_{22}, \ldots, A_{pp}, 2A_{12} 2A_{13}, \ldots, 2A_{p-1,p}$ is

$$\begin{aligned} \mathscr{E} \exp[i \operatorname{tr}(A\Theta)] &= \mathscr{E} \exp\left(i \operatorname{tr} \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha' \Theta\right) \\ &= \mathscr{E} \exp\left(i \operatorname{tr} \sum_{\alpha=1}^{n} Z_\alpha' \Theta Z_\alpha\right) \\ &= \mathscr{E} \exp\left(i \sum_{\alpha=1}^{n} Z_\alpha' \Theta Z_\alpha\right). \end{aligned} \tag{3}$$

It follows from Lemma 2.6.1 that

$$\mathscr{E} \exp\left(i \sum_{\alpha=1}^{n} Z_\alpha' \Theta Z_\alpha\right) = \prod_{\alpha=1}^{n} \mathscr{E} \exp(iZ_\alpha' \Theta Z_\alpha) = \left[\mathscr{E} \exp(iZ'\Theta Z)\right]^n, \tag{4}$$

where Z has the density (1). For Θ real, there is a real nonsingular matrix B such that

$$B'\Sigma^{-1}B = I, \tag{5}$$

$$B'\Theta B = D, \tag{6}$$

where D is a real diagonal matrix (Theorem A.2.2 of the Appendix). If we let $z = By$, then

$$\begin{aligned} \mathscr{E} \exp(iZ'\Theta Z) &= \mathscr{E} \exp(iY'DY) \\ &= \mathscr{E} \prod_{j=1}^{p} \exp\left(id_{jj} Y_j^2\right) \\ &= \prod_{j=1}^{p} \mathscr{E} \exp\left(id_{jj} Y_j^2\right) \end{aligned} \tag{7}$$

by Lemma 2.6.2. The jth factor in the second product is $\mathscr{E}\exp(id_{jj}Y_j^2)$, where Y_j has the distribution $N(0,1)$; this is the characteristic function of the χ^2-distribution with one degree of freedom, namely $(1-2id_{jj})^{-\frac{1}{2}}$ [as can be proved by expanding $\exp(id_{jj}y_j^2)$ in a power series and integrating term by term]. Thus

$$(8)\qquad \mathscr{E}\exp(iZ'\Theta Z)=\prod_{j=1}^{p}(1-2id_{jj})^{-\frac{1}{2}}=|I-2iD|^{-\frac{1}{2}}$$

since $I-2iD$ is a diagonal matrix. From (5) and (6) we see that

$$(9)\qquad \begin{aligned}|I-2iD| &= |B'\Sigma^{-1}B-2iB'\Theta B|\\ &= |B'(\Sigma^{-1}-2i\Theta)B|\\ &= |B'|\cdot|\Sigma^{-1}-2i\Theta|\cdot|B|\\ &= |B|^2\cdot|\Sigma^{-1}-2i\Theta|,\end{aligned}$$

$|B'|\cdot|\Sigma^{-1}|\cdot|B|=|I|=1$, and $|B|^2=1/|\Sigma^{-1}|$. Combining the above results, we obtain

$$(10)\qquad \mathscr{E}\exp[i\operatorname{tr}(A\Theta)]=\frac{|\Sigma^{-1}|^{\frac{1}{2}n}}{|\Sigma^{-1}-2i\Theta|^{\frac{1}{2}n}}=|I-2i\Theta\Sigma|^{-\frac{1}{2}n}.$$

It can be shown that the result is valid provided $(\mathscr{R}(\sigma^{jk}-2i\theta_{jk}))$ is positive definite. In particular, it is true for all real Θ. It also holds for Σ singular.

Theorem 7.3.1. *If $Z_1,\dots,Z_n$ are independent, each with distribution $N(\mathbf{0},\Sigma)$, then the characteristic function of $A_{11},\dots,A_{pp},2A_{12},\dots,2A_{p-1,p}$, where $(A_{ij})=A=\sum_{\alpha=1}^{n}Z_\alpha Z'_\alpha$, is given by* (10).

7.3.2. The Sum of Wishart Matrices

Suppose the A_i, $i=1,2$, are distributed independently according to $W(\Sigma,n_i)$, respectively. Then A_1 is distributed as $\sum_{\alpha=1}^{n_1}Z_\alpha Z'_\alpha$, and A_2 is distributed as $\sum_{\alpha=n_1+1}^{n_1+n_2}Z_\alpha Z'_\alpha$, where $Z_1,\dots,Z_{n_1+n_2}$ are independent, each with distribution $N(\mathbf{0},\Sigma)$. Then $A=A_1+A_2$ is distributed as $\sum_{\alpha=1}^{n}Z_\alpha Z'_\alpha$, where $n=n_1+n_2$. Thus A is distributed according to $W(\Sigma,n)$. Similarly, the sum of q matrices distributed independently, each according to a Wishart distribution with covariance Σ, has a Wishart distribution with covariance matrix Σ and number of degrees of freedom equal to the sum of the numbers of degrees of freedom of the component matrices.

Theorem 7.3.2. *If $A_1, \dots, A_q$ are independently distributed with A_i distributed according to $W(\Sigma, n_i)$, then*

$$A = \sum_{i=1}^{q} A_i \tag{11}$$

is distributed according to $W(\Sigma, \sum_{i=1}^{q} n_i)$.

7.3.3. A Certain Linear Transformation

We shall frequently make the transformation

$$A = CBC', \tag{12}$$

where C is a nonsingular $p \times p$ matrix. If A is distributed according to $W(\Sigma, n)$, then B is distributed according to $W(\Phi, n)$ where

$$\Phi = C^{-1} \Sigma C'^{-1}. \tag{13}$$

This is proved by the following argument: Let $A = \sum_{\alpha=1}^{n} Z_\alpha Z'_\alpha$, where $Z_1, \dots, Z_n$ are independently distributed, each according to $N(\mathbf{0}, \Sigma)$. Then $Y_\alpha = C^{-1} Z_\alpha$ is distributed according to $N(\mathbf{0}, \Phi)$. However,

$$B = \sum_{\alpha=1}^{n} Y_\alpha Y'_\alpha = C^{-1} \sum_{\alpha=1}^{n} Z_\alpha Z'_\alpha C'^{-1} = C^{-1} A C'^{-1} \tag{14}$$

is distributed according to $W(\Phi, n)$. Finally, $|\partial(A)/\partial(B)|$, the Jacobian of the transformation (12), is

$$\left| \frac{\partial(A)}{\partial(B)} \right| = \frac{w(B, \Phi, n)}{w(A, \Sigma, n)} = \frac{|B|^{\frac{1}{2}(n-p-1)} |\Sigma|^{\frac{1}{2}n}}{|A|^{\frac{1}{2}(n-p-1)} |\Phi|^{\frac{1}{2}n}} = \operatorname{mod} |C|^{p+1}. \tag{15}$$

Theorem 7.3.3. *The Jacobian of the transformation* (12) *from A to B, where A and B are symmetric, is* $\operatorname{mod}|C|^{p+1}$.

7.3.4. Marginal Distribntions

If A is distributed according to $W(\Sigma, n)$, the marginal distribution of any arbitrary set of the elements of A may be awkward to obtain. However, the marginal distribution of some sets of elements can be found easily. We give some of these in the following two theorems.

Theorem 7.3.4. *Let A and Σ be partitioned into q and $p-q$ rows and columns,*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{16}$$

If A is distributed according to $W(\Sigma, n)$, then A_{11} is distributed according to $W(\Sigma_{11}, n)$.

Proof. A is distributed as $\sum_{\alpha=1}^{n} Z_\alpha Z'_\alpha$, where the Z_α are independent, each with the distribution $N(\mathbf{0}, \Sigma)$. Partition Z_α into subvectors of q and $p-q$ components, $Z_\alpha = (Z_\alpha^{(1)\prime}, Z_\alpha^{(2)\prime})'$. Then $Z_1^{(1)}, \dots, Z_n^{(1)}$ are independent, each with the distribution $N(\mathbf{0}, \Sigma_{11})$, and A_{11} is distributed as $\sum_{\alpha=1}^{n} Z_\alpha^{(1)} Z_\alpha^{(1)\prime}$, which has the distribution $W(\Sigma_{11}, n)$. ∎

Theorem 7.3.5. *Let A and Σ be partitioned into $p_1, p_2, \dots, p_q$ rows and columns $(p_1 + \cdots + p_q = p)$,*

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1q} \\ \vdots & & \vdots \\ A_{q1} & \cdots & A_{qq} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1q} \\ \vdots & & \vdots \\ \Sigma_{q1} & \cdots & \Sigma_{qq} \end{pmatrix}. \tag{17}$$

If $\Sigma_{ij} = \mathbf{0}$ for $i \neq j$ and if A is distributed according to $W(\Sigma, n)$, then $A_{11}, A_{22}, \dots, A_{qq}$ are independently distributed and A_{jj} is distributed according to $W(\Sigma_{jj}, n)$.

Proof. A is distributed as $\sum_{\alpha=1}^{n} Z_\alpha Z'_\alpha$, where $Z_1, \dots, Z_n$ are independently distributed, each according to $N(\mathbf{0}, \Sigma)$. Let Z_α be partitioned

$$Z_\alpha = \begin{pmatrix} Z_\alpha^{(1)} \\ \vdots \\ Z_\alpha^{(q)} \end{pmatrix} \tag{18}$$

as A and Σ have been partitioned. Since $\Sigma_{ij} = \mathbf{0}$, the sets $Z_1^{(1)}, \dots, Z_n^{(1)}, \dots, Z_1^{(q)}, \dots, Z_n^{(q)}$ are independent. Then $A_{11} = \sum_{\alpha=1}^{n} Z_\alpha^{(1)} Z_\alpha^{(1)\prime}, \dots, A_{qq} = \sum_{\alpha=1}^{n} Z_\alpha^{(q)} Z_\alpha^{(q)\prime}$ are independent. The rest of Theorem 7.3.5 follows from Theorem 7.3.4. ∎

7.3.5. Conditional Distributions

In Section 4.3 we considered estimation of the parameters of the conditional distribution of $X^{(1)}$ given $X^{(2)} = x^{(2)}$. Application of Theorem 7.2.2 to Theorem 4.3.3 yields the following theorem:

Theorem 7.3.6. *Let A and Σ be partitioned into q and $p-q$ rows and columns as in* (16). *If A is distributed according to $W(\Sigma, n)$, the distribution of $A_{11\cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ is $W(\Sigma_{11\cdot 2}, n-p+q)$, $n \geq p-q$.*

Note that Theorem 7.3.6 implies that $A_{11\cdot 2}$ is independent of A_{22} and $A_{12}A_{22}^{-1}$ regardless of Σ.

7.4. COCHRAN'S THEOREM

Cochran's theorem [Cochran (1934)] is useful in proving that certain *vector quadratic forms* are distributed as sums of *vector squares*. It is a statistical statement of an algebraic theorem, which we shall give as a lemma.

Lemma 7.4.1. *If the $N \times N$ symmetric matrix C_i has rank r_i, $i = 1, \ldots, m$, and*

$$\sum_{i=1}^{m} C_i = I_N, \tag{1}$$

then

$$\sum_{i=1}^{m} r_i = N \tag{2}$$

is a necessary and sufficient condition for there to exist an $N \times N$ orthogonal matrix P such that for $i = 1, \ldots, m$

$$PC_iP' = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{3}$$

where I is of order r_i, the upper left-hand $\mathbf{0}$ is square of order $\sum_{j=1}^{i-1} r_j$ (which is vacuous for $i = 1$), and the lower-right hand $\mathbf{0}$ is square of order $\sum_{j=i+1}^{m} r_j$ (which is vacuous for $i = m$).

Proof. The necessity follows from the fact that (1) implies that the sum of (3) over $i = 1, \ldots, m$ is I_N. Now let us prove the sufficiency; we assume (2).

There exists an orthogonal matrix P_j such that $P_j C_j P_j'$ is diagonal with diagonal elements the characteristic roots of C_j. The number of nonzero roots is r_j, the rank of C_j, and the number of 0 roots is $N - r_j$. We write

$$P_j C_j P_j' = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \Delta_j & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{4}$$

where the partitioning is according to (3), and Δ_j is diagonal of order r_j. This is possible in view of (2). Then

$$P_j \sum_{\substack{i=1 \\ i \neq j}}^{m} C_i P_j' = P_j(I - C_j)P_j' = \begin{pmatrix} I & 0 & 0 \\ 0 & I - \Delta_j & 0 \\ 0 & 0 & I \end{pmatrix}. \tag{5}$$

Since the rank of (5) is not greater than $\sum_{i=1}^{m} r_i - r_j = N - r_j$, which is the sum of the orders of the upper left-hand and lower right-hand I's in (5), the rank of $I - \Delta_j$ is 0 and $\Delta_j = I$. (Thus the r_j nonzero roots of C_j are 1, and C_j is positive semidefinite.) From (4) we obtain

$$C_j = P_j' \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix} P_j = B_j B_j', \tag{6}$$

where B_j consists of the r_j columns of P_j' corresponding to I in (6). From (1) we obtain

$$I = \sum_{j=1}^{m} B_j B_j' = (B_1, B_2, \ldots, B_m) \begin{pmatrix} B_1' \\ B_2' \\ \vdots \\ B_m' \end{pmatrix} = P'P, \tag{7}$$

where $P = (B_1, B_2, \ldots, B_m)'$. ■

We now state a multivariate analog to Cochran's theorem.

Theorem 7.4.1. *Suppose $Y_1, \ldots, Y_N$ are independently distributed, each according to $N(\mathbf{0}, \Sigma)$. Suppose the matrix $(c_{\alpha\beta}^i) = C_i$ used in forming*

$$Q_i = \sum_{\alpha, \beta = 1}^{N} c_{\alpha\beta}^i Y_\alpha Y_\beta', \qquad i = 1, \ldots, m, \tag{8}$$

is of rank r_i, and suppose

$$\sum_{i=1}^{m} Q_i = \sum_{\alpha=1}^{N} Y_\alpha Y_\alpha'. \tag{9}$$

Then (2) is a necessary and sufficient condition for $Q_1, \dots, Q_m$ to be independently distributed with Q_i having the distribution $W(\Sigma, r_i)$.

It follows from (3) that C_i is idempotent. See Section A.2 of the Appendix.

This theorem is useful in generalizing results from the univariate analysis of variance. (See Chapter 8.) As an example of the use of this theorem, let us prove that the mean of a sample of size N times its transpose and a multiple of the sample covariance matrix are independently distributed with a singular and a nonsingular Wishart distribution, respectively. Let $Y_1, \dots, Y_N$ be independently distributed, each according to $N(\mathbf{0}, \Sigma)$. We shall use the matrices $C_1 = (c_{\alpha\beta}^{(1)}) = (1/N)$ and $C_2 = (c_{\alpha\beta}^{(2)}) = [\delta_{\alpha\beta} - (1/N)]$. Then

$$Q_1 = \sum_{\alpha,\beta=1}^{N} \frac{1}{N} Y_\alpha Y_\beta' = N\bar{Y}\bar{Y}', \tag{10}$$

$$\begin{aligned} Q_2 &= \sum_{\alpha,\beta=1}^{N} \left(\delta_{\alpha\beta} - \frac{1}{N}\right) Y_\alpha Y_\beta' \\ &= \sum_{\alpha=1}^{N} Y_\alpha Y_\alpha' - N\bar{Y}\bar{Y}' \\ &= \sum_{\alpha=1}^{N} (Y_\alpha - \bar{Y})(Y_\alpha - \bar{Y})', \end{aligned} \tag{11}$$

and (9) is satisfied. The matrix C_1 is of rank 1; the matrix C_2 is of rank $N-1$ (since the rank of the sum of two matrices is less than or equal to the sum of the ranks of the matrices and the rank of the second matrix is less than N). The conditions of the theorem are satisfied; therefore Q_1 is distributed as ZZ', where Z is distributed according to $N(\mathbf{0}, \Sigma)$, and Q_2 is distributed independently according to $W(\Sigma, N-1)$.

Anderson and Styan (1982) have given a survey of proofs and extensions of Cochran's theorem.

7.5. THE GENERALIZED VARIANCE

7.5.1. Definition of the Generalized Variance

One multivariate analog of the variance σ^2 of a univariate distribution is the covariance matrix Σ. Another multivariate analog is the scalar $|\Sigma|$, which is called the *generalized variance* of the multivariate distribution [Wilks (1932); see also Frisch (1929)]. Similarly, the generalized variance of the sample of vectors $x_1, \dots, x_N$ is

$$|S| = \left| \frac{1}{N-1} \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' \right|. \tag{1}$$

In some sense each of these is a measure of spread. We consider them here because the sample generalized variance will recur in many likelihood ratio criteria for testing hypotheses.

A geometric interpretation of the sample generalized variance comes from considering the p rows of $X = (x_1, \dots, x_N)$ as p vectors in N-dimensional space. In Section 3.2 it was shown that the rows of

$$(x_1 - \bar{x}, \dots, x_N - \bar{x}) = X - \bar{x}\varepsilon', \tag{2}$$

where $\varepsilon = (1, \dots, 1)'$, are orthogonal to the equiangular line (through the origin and ε); see Figure 3.2. Then the entries of

$$A = (X - \bar{x}\varepsilon')(X - \bar{x}\varepsilon')' \tag{3}$$

are the inner products of rows of $X - \bar{x}\varepsilon'$.

We now define a *parallelotope* determined by p vectors $v_1, \dots, v_p$ in an n-dimensional space $(n \geq p)$. If $p = 1$, the parallelotope is the line segment v_1. If $p = 2$, the parallelotope is the parallelogram with v_1 and v_2 as principal edges; that is, its sides are v_1, v_2, v_1 translated so its initial endpoint is at v_2, and v_2 translated so its initial endpoint is at v_1. See Figure 7.2. If $p = 3$, the parallelotope is the conventional parallelepided with v_1, v_2, and v_3 as

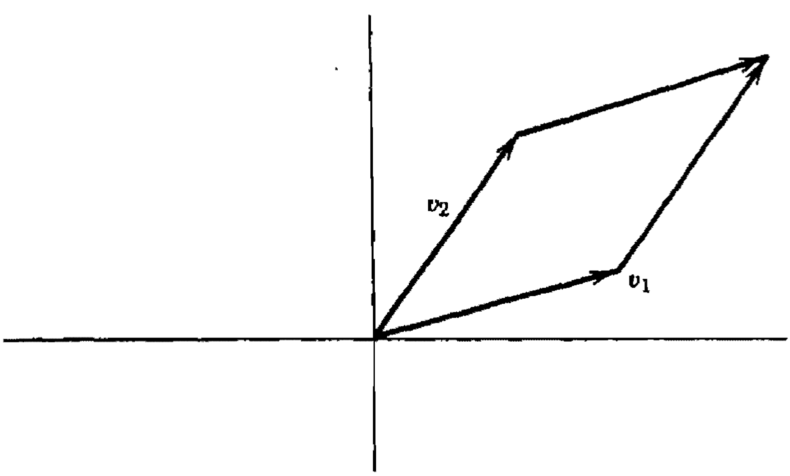

Figure 7.2. A parallelogram.

principal edges. In general, the parallelotope is the figure defined by the principal edges $v_1, \ldots, v_p$. It is cut out by p pairs of parallel $(p-1)$-dimensional hyperplanes, one hyperplane of a pair being spanned by $p-1$ of $v_1, \ldots, v_p$ and the other hyperplane going through the endpoint of the remaining vector.

Theorem 7.5.1. *If $V = (v_1, \ldots, v_p)$, then the square of the p-dimensional volume of the parallelotope with $v_1, \ldots, v_p$ as principal edges is $|V'V|$.*

Proof. If $p = 1$, then $|V'V| = v_1'v_1 = \|v_1\|^2$, which is the square of the one-dimensional volume of v_1. If two k-dimensional parallelotopes have bases consisting of $(k-1)$-dimensional parallelotopes of equal $(k-1)$-dimensional volumes and equal altitudes, their k-dimensional volumes are equal [since the k-dimensional volume is the integral of the $(k-1)$-dimensional volumes]. In particular, the volume of a k-dimensional parallelotope is equal to the volume of a parallelotope with the same base (in $k-1$ dimensions) and same altitude with sides in the kth direction orthogonal to the first $k-1$ directions. Thus the volume of the parallelotope with principal edges $v_1, \ldots, v_k$, say P_k, is equal to the volume of the parallelotope with principal edges $v_1, \ldots, v_{k-1}$, say P_{k-1}, times the altitude of P_k over P_{k-1}; that is,

$$\mathrm{Vol}(P_k) = \mathrm{Vol}(P_{k-1}) \times \mathrm{Alt}(P_k | P_{k-1}). \tag{4}$$

It follows (by induction) that

$$\mathrm{Vol}(P_p) = \mathrm{Vol}(P_1) \times \mathrm{Alt}(P_2|P_1) \times \cdots \times \mathrm{Alt}(P_p|P_{p-1}). \tag{5}$$

By the construction in Section 7.2 the altitude of P_k over P_{k-1} is $t_{kk} = \|w_k\|$; that is, t_{kk} is the distance of v_k from the $(k-1)$-dimensional space spanned by $v_1, \ldots, v_{k-1}$ (or $w_1, \ldots, w_{k-1}$). Hence $\mathrm{Vol}(P_p) = \prod_{k=1}^p t_{kk}$. Since $|V'V| = |TT'| = \prod_{i=1}^p t_{ii}^2$, the theorem is proved. ■

We now apply this theorem to the parallelotope having the rows of (2) as principal edges. The dimensionality in Theorem 7.5.1 is arbitrary (but at least p).

Corollary 7.5.1. *The square of the p-dimensional volume of the parallelotope with the rows of* (2) *as principal edges is $|A|$, where A is given by* (3).

We shall see later that many multivariate statistics can be given an interpretation in terms of these volumes. These volumes are analogous to distances that arise in special cases when $p = 1$.

We now consider a geometric interpretation of $|A|$ in terms of N points in p-space. Let the columns of the matrix (2) be $y_1, \dots, y_N$, representing N points in p-space. When $p = 1$, $|A| = \sum_\alpha y_{1\alpha}^2$, which is the sum of squares of the distances from the points to the origin. In general $|A|$ is the sum of squares of the volumes of all parallelotopes formed by taking as principal edges p vectors from the set $y_1, \dots, y_N$.

We see that

$$
(6) \qquad |A| = \begin{vmatrix} \sum_\alpha y_{1\alpha}^2 & \cdots & \sum_\alpha y_{1\alpha} y_{p-1,\alpha} & \sum_\beta y_{1\beta} y_{p\beta} \\ \vdots & & \vdots & \vdots \\ \sum_\alpha y_{p-1,\alpha} y_{1\alpha} & \cdots & \sum_\alpha y_{p-1,\alpha}^2 & \sum_\beta y_{p-1,\beta} y_{p\beta} \\ \sum_\alpha y_{p\alpha} y_{1\alpha} & \cdots & \sum_\alpha y_{p\alpha} y_{p-1,\alpha} & \sum_\beta y_{p\beta}^2 \end{vmatrix}
$$

$$
= \sum_\beta \begin{vmatrix} \sum_\alpha y_{1\alpha}^2 & \cdots & \sum_\alpha y_{1\alpha} y_{p-1,\alpha} & y_{1\beta} y_{p\beta} \\ \vdots & & \vdots & \vdots \\ \sum_\alpha y_{p-1,\alpha} y_{1\alpha} & \cdots & \sum_\alpha y_{p-1,\alpha}^2 & y_{p-1,\beta} y_{p\beta} \\ \sum_\alpha y_{p\alpha} y_{1\alpha} & \cdots & \sum_\alpha y_{p\alpha} y_{p-1,\alpha} & y_{p\beta}^2 \end{vmatrix}
$$

by the rule for expanding determinants. [See (24) of Section A.1 of the Appendix.] In (6) the matrix A has been partitioned into $p - 1$ and 1 columns. Applying the rule successively to the columns, we find

$$
(7) \qquad |A| = \sum_{\alpha_1, \dots, \alpha_p = 1}^{N} |y_{i\alpha_j} y_{j\alpha_j}|.
$$

By Theorem 7.5.1 the square of the volume of the parallelotope with $y_{\gamma_1}, \dots, y_{\gamma_p}$, $\gamma_1 < \cdots < \gamma_p$, as principal edges is

$$
(8) \qquad V_{\gamma_1, \dots, \gamma_p}^2 = \left| \sum_\beta y_{i\beta} y_{j\beta} \right|,
$$

where the sum on β is over $(\gamma_1, \dots, \gamma_p)$. If we now expand this determinant in the manner used for $|A|$, we obtain

$$
(9) \qquad V_{\gamma_1, \dots, \gamma_p}^2 = \sum |y_{i\beta_j} y_{j\beta_j}|,
$$

where the sum is for each β_i over the range $(\gamma_1, \ldots, \gamma_p)$. Summing (9) over all different sets $(\gamma_1 < \cdots \gamma_p)$, we obtain (7). ($|y_{i\beta_j} y_{l\beta_j}| = 0$ if two or more β_i are equal.) Thus $|A|$ is the sum of volumes squared of all different parallelotopes formed by sets of p of the vectors y_α as principal edges. If we replace y_α by $x_\alpha - \bar{x}$, we can state the following theorem:

Theorem 7.5.2. *Let $|S|$ be defined by* (1), *where $x_1, \ldots, x_N$ are the N vectors of a sample. Then $|S|$ is proportional to the sum of squares of the volumes of all the different parallelotopes formed by using as principal edges p vectors with p of $x_1, \ldots, x_N$ as one set of endpoints and $\bar{x}$ as the other, and the factor of proportionality is $1/(N-1)^p$.*

The population analog of $|S|$ is $|\Sigma|$, which can also be given a geometric interpretation. From Section 3.3 we know that

$$\Pr\{X'\Sigma^{-1}X \le \chi_p^2(\alpha)\} = 1 - \alpha \tag{10}$$

if X is distributed according to $N(\mathbf{0}, \Sigma)$; that is, the probability is $1 - \alpha$ that X fall inside the ellipsoid

$$x'\Sigma^{-1}x = \chi_p^2(\alpha). \tag{11}$$

The volume of this ellipsoid is $C(p)|\Sigma|^{\frac{1}{2}}[\chi_p^2(\alpha)]^{\frac{1}{2}p}/p$, where $C(p)$ is defined in Problem 7.3.

7.5.2. Distribution of the Sample Generalized Variance

The distribution of $|S|$ is the same as the distribution of $|A|/(N-1)^p$, where $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ and $Z_1, \ldots, Z_n$ are distributed independently, each according to $N(\mathbf{0}, \Sigma)$, and $n = N - 1$. Let $Z_\alpha = CY_\alpha$, $\alpha = 1, \ldots, n$, where $CC' = \Sigma$. Then $Y_1, \ldots, Y_n$ are independently distributed, each with distribution $N(\mathbf{0}, I)$. Let

$$B = \sum_{\alpha=1}^{n} Y_\alpha Y_\alpha' = \sum_{\alpha=1}^{n} C^{-1} Z_\alpha Z_\alpha' (C^{-1})' = C^{-1}A(C^{-1});\tag{12}$$

then $|A| = |C| \cdot |B| \cdot |C'| = |B| \cdot |\Sigma|$. By the development in Section 7.2 we see that $|B|$ has the distribution of $\prod_{i=1}^{p} t_{ii}^2$ and that $t_{11}^2, \ldots, t_{pp}^2$ are independently distributed with χ^2-distributions.

Theorem 7.5.3. *The distribution of the generalized variance $|S|$ of a sample $X_1, \ldots, X_N$ from $N(\boldsymbol{\mu}, \Sigma)$ is the same as the distribution of $|\Sigma|/(N-1)^p$ times the product of p independent factors, the distribution of the ith factor being the χ^2-distribution with $N - i$ degrees of freedom.*

If $p = 1$, $|S|$ has the distribution of $|\Sigma| \cdot \chi^2_{N-1}/(N-1)$. If $p = 2$, $|S|$ has the distribution of $|\Sigma| \chi^2_{N-1} \cdot \chi^2_{N-2}/(N-1)^2$. It follows from Problem 7.15 or 7.37 that when $p = 2$, $|S|$ has the distribution of $|\Sigma|(\chi^2_{2N-4})^2/(2N-2)^2$. We can write

$$|A| = |\Sigma| \times \chi^2_{N-1} \times \chi^2_{N-2} \times \cdots \times \chi^2_{N-p}. \tag{13}$$

If $p = 2r$, then $|A|$ is distributed as

$$|\Sigma| \left(\chi^2_{2N-4} \times \chi^2_{2N-8} \times \cdots \times \chi^2_{2N-4r}\right)^2 / 2^{2r}. \tag{14}$$

Since the hth moment of a χ^2-variable with m degrees of freedom is $2^h\Gamma(\frac{1}{2}m + h)/\Gamma(\frac{1}{2}m)$ and the moment of a product of independent variables is the product of the moments of the variables, the hth moment of $|A|$ is

$$\begin{aligned} |\Sigma|^h \prod_{i=1}^{p} \left\{ 2^h \frac{\Gamma[\frac{1}{2}(N-i)+h]}{\Gamma[\frac{1}{2}(N-i)]} \right\} &= 2^{hp}|\Sigma|^h \frac{\prod_{i=1}^{p}\Gamma[\frac{1}{2}(N-i)+h]}{\prod_{i=1}^{p}\Gamma[\frac{1}{2}(N-i)]} \\ &= 2^{hp}|\Sigma|^h \frac{\Gamma_p[\frac{1}{2}(N-1)+h]}{\Gamma_p[\frac{1}{2}(N-1)]}. \end{aligned} \tag{15}$$

Thus

$$\mathscr{E}|A| = |\Sigma| \prod_{i=1}^{p} (N-i). \tag{16}$$

$$\mathscr{V}(|A|) = |\Sigma|^2 \prod_{i=1}^{p}(N-i) \left[\prod_{j=1}^{p} \Gamma(N-j+2) - \prod_{j=1}^{p} \Gamma(N-j) \right], \tag{17}$$

where $\mathscr{V}(|A|)$ is the variance of $|A|$.

7.5.3. The Asymptotic Distribution of the Sample Generalized Variance

Let $|B|/n^p = V_1(n) \times V_2(n) \times \cdots \times V_p(n)$, where the V's are independently distributed and $nV_i(n) = \chi^2_{n-p+i}$. Since χ^2_{n-p+i} is distributed as $\sum_{\alpha=1}^{n-p+i} W_\alpha^2$, where the W_α are independent, each with distribution $N(0,1)$, the central limit theorem (applied to W_α^2) states that

$$\frac{nV_i(n) - (n-p+i)}{\sqrt{2(n-p+i)}} = \sqrt{n}\,\frac{V_i(n) - 1 + \frac{p-i}{n}}{\sqrt{2}\sqrt{1 - \frac{p-i}{n}}} \tag{18}$$

is asymptotically distributed according to $N(0,1)$. Then $\sqrt{n}[V_i(n) - 1]$ is asymptotically distributed according to $N(0,2)$. We now apply Theorem 4.2.3.

We have

$$(19)\qquad U(n) = \begin{pmatrix} V_1(n) \\ \vdots \\ V_p(n) \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

$|B|/n^p = w = f(u_1, \dots, u_p) = u_1 u_2 \cdots u_p$, $\mathbf{T} = 2I$, $\partial f/\partial u_i|_{u=b} = 1$, *and* $\phi_b' \mathbf{T} \phi_b = 2p$. *Thus*

$$(20)\qquad \sqrt{n}\left(\frac{|B|}{n^p} - 1\right)$$

is asymptotically distributed according to $N(0, 2p)$.

Theorem 7.5.4. *Let S be a $p \times p$ sample covariance matrix with n degrees of freedom. Then $\sqrt{n}\,(|S|/|\Sigma| - 1)$ is asymptotically normally distributed with mean 0 and variance $2p$.*

7.6. DISTRIBUTION OF THE SET OF CORRELATION COEFFICIENTS WHEN THE POPULATION COVARIANCE MATRIX IS DIAGONAL

In Section 4.2.1 we found the distribution of a single sample correlation when the corresponding population correlation was zero. Here we shall find the density of the set r_{ij}, $i<j$, $i,j = 1, \dots, p$, when $\rho_{ij} = 0$, $i<j$.

We start with the distribution of A when Σ is diagonal; that is, $W[(\sigma_{ii}\delta_{ij}), n]$. The density of A is

$$(1)\qquad \frac{|a_{ij}|^{\frac{1}{2}(n-p-1)} \exp\left(-\frac{1}{2}\sum_{i=1}^p a_{ii}/\sigma_{ii}\right)}{2^{\frac{1}{2}np} \prod_{i=1}^p \sigma_{ii}^{\frac{1}{2}n} \Gamma_p\left(\frac{1}{2}n\right)},$$

since

$$(2)\qquad |\Sigma| = \begin{vmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{vmatrix} = \prod_{i=1}^p \sigma_{ii}.$$

We make the transformation

$$(3)\qquad a_{ij} = \sqrt{a_{ii}}\sqrt{a_{jj}}\, r_{ij}, \qquad i \neq j,$$

$$(4)\qquad a_{ii} = a_{ii}.$$

The Jacobian is the product of the Jacobian of (4) and that of (3) for a_{ii} fixed. The Jacobian of (3) is the determinant of a $p(p-1)/2$-order diagonal matrix with diagonal elements $\sqrt{a_{ii}}\sqrt{a_{jj}}$. Since each particular subscript k, say, appears in the set r_{ij} $(i<j)$ $p-1$ times, the Jacobian is

$$(5) \qquad J=\prod_{i=1}^{p} a_{ii}^{\frac{1}{2}(p-1)}.$$

If we substitute from (3) and (4) into $w[A|(\sigma_{ii}\delta_{ij}),n]$ and multiply by (5), we obtain as the joint density of $\{a_{ii}\}$ and $\{r_{ij}\}$

$$(6) \qquad \frac{\left|\sqrt{a_{ii}}\sqrt{a_{jj}}\,r_{ij}\right|^{\frac{1}{2}(n-p-1)}\exp\left(-\frac{1}{2}\sum_{i=1}^{p}a_{ii}/\sigma_{ii}\right)}{2^{\frac{1}{2}np}\prod_{i=1}^{p}\sigma_{ii}^{\frac{1}{2}n}\Gamma_p\left(\frac{1}{2}n\right)}\prod_{i=1}^{p}a_{ii}^{\frac{1}{2}(p-1)}$$

$$=\frac{|r_{ij}|^{\frac{1}{2}(n-p-1)}}{\Gamma_p\left(\frac{1}{2}n\right)}\prod_{i=1}^{p}\left\{\frac{a_{ii}^{\frac{1}{2}n-1}\exp\left(-\frac{1}{2}a_{ii}/\sigma_{ii}\right)}{2^{\frac{1}{2}n}\sigma_{ii}^{\frac{1}{2}n}}\right\},$$

since

$$(7) \qquad \left|\sqrt{a_{ii}}\sqrt{a_{jj}}\,r_{ij}\right|=\left(\prod_{i=1}^{p}a_{ii}\right)|r_{ij}|,$$

where $r_{ii}=1$. In the ith term of the product on the right-hand side of (6), let $a_{ii}/(2\sigma_{ii})=u_i$; then the integral of this term is

$$(8) \qquad \int_0^\infty \frac{a_{ii}^{\frac{1}{2}n-1}\exp\left(-\frac{1}{2}a_{ii}/\sigma_{ii}\right)}{2^{\frac{1}{2}n}\sigma_{ii}^{\frac{1}{2}n}}\,da_{ii}=\int_0^\infty u_i^{\frac{1}{2}n-1}e^{-u_i}\,du_i=\Gamma\left(\tfrac{1}{2}n\right)$$

by definition of the gamma function (or by the fact that a_{ii}/σ_{ii} has the χ^2-density with n degrees of freedom). Hence the density of r_{ij} is

$$(9) \qquad \frac{\Gamma^p\left(\frac{1}{2}n\right)|r_{ij}|^{\frac{1}{2}(n-p-1)}}{\Gamma_p\left(\frac{1}{2}n\right)}.$$

Theorem 7.6.1. *If $X_1,\dots,X_N$ are independent, each with distribution $N[\mu,(\sigma_{ii}\delta_{ij})]$, then the density of the sample correlation coefficients is given by (9) where $n=N-1$.*

7.7. THE INVERTED WISHART DISTRIBUTION AND BAYES ESTIMATION OF THE COVARIANCE MATRIX

7.7.1. The Inverted Wishart Distribution

As indicated in Section 3.4.2, Bayes estimators are usually admissible. The calculation of Bayes estimators is facilitated when the prior distributions of the parameter is chosen conveniently. When there is a sufficient statistic, there will exist a family of prior distributions for the parameter such that the posterior distribution is a member of this family; such a family is called a *conjugate family of distributions*. In Section 3.4.2 we saw that the normal family of priors is conjugate to the normal family of distributions when the covariance matrix is given. In this section we shall consider Bayesian estimation of the covariance matrix and estimation of the mean vector and the covariance matrix.

Theorem 7.7.1. *If A has the distribution $W(\Sigma, m)$, then $B = A^{-1}$ has the density*

$$\frac{|\boldsymbol{\Psi}|^{\frac{1}{2}m}|B|^{-\frac{1}{2}(m+p+1)}e^{-\frac{1}{2}\operatorname{tr}\boldsymbol{\Psi}B^{-1}}}{2^{\frac{1}{2}mp}\Gamma_p(\frac{1}{2}m)} \tag{1}$$

for B positive definite and 0 elsewhere, where $\boldsymbol{\Psi} = \Sigma^{-1}$.

Proof. By Theorem A.4.6 of the Appendix, the Jacobian of the transformation $A = B^{-1}$ is $|B|^{-(p+1)}$. Substitution of B^{-1} for A in (16) of Section 7.2 and multiplication by $|B|^{-(p+1)}$ yields (1). ∎

We shall call (1) the density of the *inverted Wishart distribution with m degrees of freedom*† and denote the distribution by $W^{-1}(\boldsymbol{\Psi}, m)$ and the density by $w^{-1}(B|\boldsymbol{\Psi}, m)$. We shall call $\boldsymbol{\Psi}$ the *precision matrix* or *concentration matrix*.

7.7.2. Bayes Estimation of the Covariance Matrix

The covariance matrix of a sample of size N from $N(\boldsymbol{\mu}, \Sigma)$ has the distribution of $(1/n)A$, where A has the distribution $W(\Sigma, n)$ and $n = N - 1$. We shall now show that if Σ is assigned an inverted Wishart distribution, then the conditional distribution of Σ given A is an inverted Wishart distribution. In other words, the family of inverted Wishart distributions for Σ is conjugate to the family of Wishart distributions.

†The definition of the number of degrees of freedom differs from that of Giri (1977), p. 104, and Muirhead (1982), p. 113.

Theorem 7.7.2. *If A has the distribution $W(\Sigma, n)$ and Σ has the a priori distribution $W^{-1}(\Psi, m)$, then the conditional distribution of Σ is $W^{-1}(A + \Psi, n+m)$.*

Proof. The joint density of A and Σ is

$$\frac{|\Psi|^{\frac{1}{2}m}|\Sigma|^{-\frac{1}{2}(n+m+p+1)}|A|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}\operatorname{tr}(A+\Psi)\Sigma^{-1}}}{2^{\frac{1}{2}(n+m)p}\Gamma_p(\frac{1}{2}n)\Gamma_p(\frac{1}{2}m)} \tag{2}$$

for A and Σ positive definite. The marginal density of A is the integral of (2) over the set of Σ positive definite. Since the integral of (1) with respect to B is 1 identically in Ψ, the integral of (2) with respect to Σ is

$$\frac{\Gamma_p[\frac{1}{2}(n+m)]|\Psi|^{\frac{1}{2}m}|A|^{\frac{1}{2}(n-p-1)}|A+\Psi|^{-\frac{1}{2}(n+m)}}{\Gamma_p(\frac{1}{2}n)\Gamma_p(\frac{1}{2}m)} \tag{3}$$

for A positive definite. The conditional density of Σ given A is the ratio of (2) to (3), namely,

$$\frac{|A+\Psi|^{\frac{1}{2}(n+m)}|\Sigma|^{-\frac{1}{2}(n+m+p+1)}e^{-\frac{1}{2}\operatorname{tr}(A+\Psi)\Sigma^{-1}}}{2^{\frac{1}{2}(n+m)p}\Gamma_p[\frac{1}{2}(n+m)]}, \tag{4}$$

which is $w^{-1}(\Sigma|A+\Psi, n+m)$. ■

Corollary 7.7.1. *If nS has the distribution $W(\Sigma, n)$ and Σ has the a priori distribution $W^{-1}(\Psi, m)$, then the conditional distribution of Σ given S is $W^{-1}(nS + \Psi, n+m)$.*

Corollary 7.7.2. *If nS has the distribution $W(\Sigma, n)$, Σ has the a priori distribution $W^{-1}(\Psi, m)$, and the loss function is $\operatorname{tr}(D-\Sigma)G(D-\Sigma)H$, where G and H are positive definite, then the Bayes estimator for Σ is*

$$\frac{1}{n+m-p-1}(nS+\Psi). \tag{5}$$

Proof. It follows from Section 3.4.2 that the Bayes estimator for Σ is $\mathscr{E}(\Sigma|S)$. From Theorem 7.7.2 we see that Σ^{-1} has the a posteriori distribution $W[(nS+\Psi)^{-1}, n+m]$. The theorem results from the following lemma. ■

Lemma 7.7.1. *If A has the distribution $W(\Sigma, n)$, then*

$$\mathscr{E}A^{-1} = \frac{1}{n-p-1}\Sigma^{-1}. \tag{6}$$

Proof. If C is a nonsingular matrix such that $\Sigma = CC'$, then A has the distribution of CBC', where B has the distribution $W(I, n)$, and $\mathscr{E}A^{-1} = (C')^{-1}(\mathscr{E}B^{-1})C^{-1}$. By symmetry the diagonal elements of $\mathscr{E}B^{-1}$ are the same and the off-diagonal elements are the same; that is, $\mathscr{E}B^{-1} = k_1 I + k_2 \varepsilon\varepsilon'$. For every orthogonal matrix Q, QBQ' has the distribution $W(I, n)$ and hence $\mathscr{E}(QBQ')^{-1} = Q\mathscr{E}B^{-1}Q' = \mathscr{E}B^{-1}$. Thus $k_2 = 0$. A diagonal element of B^{-1} has the distribution of $(\chi^2_{n-p+1})^{-1}$. (See, e.g., the proof of Theorem 5.2.2.) Since $\mathscr{E}(\chi^2_{n-p+1})^{-1} = (n-p-1)^{-1}$, $\mathscr{E}B^{-1} = (n-p-1)^{-1}I$. Then (6) follows. ■

We note that $(n-p-1)A^{-1} = [(n-p-1)/(n-1)]S^{-1}$ is an unbiased estimator of the precision Σ^{-1}.

If μ is known, the unbiased estimator of Σ is $(1/N)\Sigma_{\alpha=1}^{N}(x_\alpha - \mu)(x_\alpha - \mu)'$. The above can be applied with n replaced by N. Note that if n (or N) is large. (5) is approximately S.

Theorem 7.7.3. *Let $x_1, \dots, x_N$ be observations from $N(\mu, \Sigma)$. Suppose μ and Σ have the a priori density $n(\mu|\nu, (1/K)\Sigma) \times w^{-1}(\Sigma|\Psi, m)$. Then the a posteriori density of μ and Σ given $\bar{x} = (1/N)\Sigma_{\alpha=1}^{N} x_\alpha$, and $S = (1/n)\Sigma_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$ is*

$$n\left(\mu \left| \frac{1}{N+K}(N\bar{x} + K\nu), \frac{1}{N+K}\Sigma\right.\right) \cdot w^{-1}\left(\Sigma \left| \Psi + nS + \frac{NK}{N+K}(\bar{x}-\nu)(\bar{x}-\nu)', N+m\right.\right). \tag{7}$$

Proof. Since $\bar{x}$ and $nS = A$ are a sufficient set of statistics, we can consider the joint density of $\bar{x}$, A, μ, and Σ, which is

$$\frac{K^{\frac{1}{2}p}N^{\frac{1}{2}p}|\Psi|^{\frac{1}{2}m}|\Sigma|^{-\frac{1}{2}(N+m+p+2)}|A|^{\frac{1}{2}(N-p-2)}}{2^{\frac{1}{2}(N+m+1)p}\pi^p\Gamma_p[\frac{1}{2}(N-1)]\Gamma_p(\frac{1}{2}m)} \cdot \exp\left\{-\tfrac{1}{2}\left[N(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) + \operatorname{tr} A\Sigma^{-1} + K(\mu-\nu)'\Sigma^{-1}(\mu-\nu) + \operatorname{tr}\Psi\Sigma^{-1}\right]\right\}. \tag{8}$$

The marginal density of $\bar{x}$ and A is the integral of (8) with respect to μ and Σ. The exponential in (8) is $-\frac{1}{2}$ times

$$\begin{aligned}(N+K)\mu'\Sigma^{-1}\mu - 2(N\bar{x}+K\nu)\Sigma^{-1}\mu &+ N\bar{x}'\Sigma^{-1}\bar{x} + K\nu'\Sigma^{-1}\nu + \operatorname{tr}(A+\Psi)\Sigma^{-1} \\ = (N+K)\left[\mu - \frac{1}{N+K}(N\bar{x}+K\nu)\right]'\Sigma^{-1}&\left[\mu - \frac{1}{N+K}(N\bar{x}+K\nu)\right] \\ + \frac{NK}{N+K}(\bar{x}-\nu)'\Sigma^{-1}(\bar{x}-\nu) &+ \operatorname{tr}(A+\Psi)\Sigma^{-1}.\end{aligned} \tag{9}$$

The integral of (8) with respect to $\boldsymbol{\mu}$ is

$$(10)\qquad \frac{K^{\frac{1}{2}p}N^{\frac{1}{2}p}|\boldsymbol{\Psi}|^{\frac{1}{2}m}|\boldsymbol{\Sigma}|^{-\frac{1}{2}(N+m+p+1)}|A|^{\frac{1}{2}(N-p-2)}}{(N+K)^{\frac{1}{2}p}2^{\frac{1}{2}(N+m)p}\pi^{\frac{1}{2}p}\Gamma_p[\frac{1}{2}(N-1)]\Gamma_p(\frac{1}{2}m)}$$

$$\cdot\exp\left\{-\frac{1}{2}\left[\operatorname{tr} A\boldsymbol{\Sigma}^{-1}+\frac{NK}{N+K}(\bar{\boldsymbol{x}}-\boldsymbol{\nu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\nu})+\operatorname{tr}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right]\right\}.$$

In turn, the integral of (10) with respect to $\boldsymbol{\Sigma}$ is

$$(11)\qquad \frac{K^{\frac{1}{2}p}N^{\frac{1}{2}}\Gamma_p[\frac{1}{2}(N+m)]}{\pi^{\frac{1}{2}p}\Gamma_p[\frac{1}{2}(N-1)]\Gamma_p(\frac{1}{2}m)(N+K)^{\frac{1}{2}p}}$$

$$\cdot|A|^{\frac{1}{2}(N-p-2)}|\boldsymbol{\Psi}|^{\frac{1}{2}m}\left|\boldsymbol{\Psi}+A+\frac{NK}{N+K}(\bar{\boldsymbol{x}}-\boldsymbol{\nu})(\bar{\boldsymbol{x}}-\boldsymbol{\nu})'\right|^{-\frac{1}{2}(N+m)}.$$

The conditional density of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given $\bar{\boldsymbol{x}}$ and A is the ratio of (8) to (11), namely,

$$(12)\qquad \frac{(N+K)^{\frac{1}{2}p}|\boldsymbol{\Sigma}|^{-\frac{1}{2}(N+m+p+2)}\left|\boldsymbol{\Psi}+A+\frac{NK}{N+K}(\bar{\boldsymbol{x}}-\boldsymbol{\nu})(\bar{\boldsymbol{x}}-\boldsymbol{\nu})'\right|^{\frac{1}{2}(N+m)}}{2^{\frac{1}{2}(N+m+1)p}\pi^{\frac{1}{2}p}\Gamma_p[\frac{1}{2}(N+m)]}$$

$$\cdot\exp\left\{-\frac{1}{2}\left[(N+K)\left[\boldsymbol{\mu}-\frac{1}{N+K}(N\bar{\boldsymbol{x}}+K\boldsymbol{\nu})\right]'\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\mu}-\frac{1}{N+K}(N\bar{\boldsymbol{x}}+K\boldsymbol{\nu})\right]\right.\right.$$

$$\left.\left.+\operatorname{tr}\left[\boldsymbol{\Psi}+A+\frac{NK}{N+K}(\bar{\boldsymbol{x}}-\boldsymbol{\nu})(\bar{\boldsymbol{x}}-\boldsymbol{\nu})'\right]\boldsymbol{\Sigma}^{-1}\right]\right\}.$$

Then (12) can be written as (7). ■

Corollary 7.7.3. *If $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ are observations from $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have the a priori density $n[\boldsymbol{\mu}|\boldsymbol{\nu},(1/K)\boldsymbol{\Sigma}]\times w^{-1}(\boldsymbol{\Sigma}|\boldsymbol{\Psi},m)$, and if the loss function is $(\boldsymbol{d}-\boldsymbol{\mu})'\boldsymbol{J}(\boldsymbol{d}-\boldsymbol{\mu})-\operatorname{tr}(\boldsymbol{D}-\boldsymbol{\Sigma})\boldsymbol{G}(\boldsymbol{D}-\boldsymbol{\Sigma})\boldsymbol{H}$, then the Bayes estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are*

$$(13)\qquad \frac{1}{N+K}(N\bar{\boldsymbol{x}}+K\boldsymbol{\nu})$$

and

$$(14)\qquad \frac{1}{N+m-p-1}\left[nS+\boldsymbol{\Psi}+\frac{NK}{N+K}(\bar{\boldsymbol{x}}-\boldsymbol{\nu})(\bar{\boldsymbol{x}}-\boldsymbol{\nu})'\right],$$

respectively.

The estimator of $\boldsymbol{\mu}$ is a weighted average of the sample mean $\bar{\boldsymbol{x}}$ and the a priori mean $\boldsymbol{\nu}$. If N is large, the a priori mean has relatively little weight.

The estimator of Σ is a weighted average of the sample covariances S, Ψ, and a term deriving from the difference between the sample mean and the a priori mean. If N is large, the estimator is close to the sample covariance matrix.

Theorem 7.7.4. *If $x_1, \dots, x_N$ are observations from $N(\mu, \Sigma)$ and if μ and Σ have the a priori density $n[\mu|\nu, (1/K)\Sigma] \times w^{-1}(\Sigma|\Psi, m)$, then the marginal a posteriori density of μ given $\bar{x}$ and S is*

$$\frac{(N+K)^{\frac{1}{2}p}\Gamma[\frac{1}{2}(N+m+1)]|B|^{-\frac{1}{2}}}{\pi^{\frac{1}{2}p}\Gamma[\frac{1}{2}(N+m+1-p)][1+(N+K)(\mu-\mu^*)'B^{-1}(\mu-\mu^*)]^{\frac{1}{2}(N+m-1)}}, \tag{15}$$

where μ^* is (13) *and B is $N+m-p-1$ times* (14).

Proof. The exponent in (12) is $-\frac{1}{2}$ times

$$\operatorname{tr}[B+(N+K)(\mu-\mu^*)(\mu-\mu^*)']\Sigma^{-1}. \tag{16}$$

Then the integral of (12) with respect to Σ is

$$\frac{(N+K)^{\frac{1}{2}p}\Gamma_p[\frac{1}{2}(N+m+1)]|B|^{\frac{1}{2}(N+m)}}{\pi^{\frac{1}{2}p}\Gamma_p[\frac{1}{2}(N+m)]|B+(N+K)(\mu-\mu^*)(\mu-\mu^*)'|^{\frac{1}{2}(N+m+1)}}. \tag{17}$$

Since $|B+xx'| = |B|(1+x'B^{-1}x)$ (Corollary A.3.1), (15) follows. ■

The density (15) is the multivariate t-distribution with $N+m+1-p$ degrees of freedom. See Section 2.7.5, Examples.

7.8. IMPROVED ESTIMATION OF THE COVARIANCE MATRIX

Just as the sample mean $\bar{x}$ can be improved on as an estimator of the population mean μ when the loss function is quadratic, so can the sample covariance S be improved on as an estimator of the population covariance Σ for certain loss functions. The loss function for estimation of the location parameter μ was invariant with respect to translation ($x \to x+a$, $\mu \to \mu+a$), and the risk of the sample mean (which is the unique unbiased function of the sufficient statistic when Σ is known) does not depend on the parameter value. The natural group of transformations of covariance matrices is multiplication on the left by a nonsingular matrix and on the right by its transpose

($x \to Cx$, $S \to CSC'$, $\Sigma \to C\Sigma C'$). We consider two loss functions which are invariant with respect to such transformations.

One loss function is quadratic:

$$L_q(\Sigma, G) = \operatorname{tr}(G-\Sigma)\Sigma^{-1}(G-\Sigma)\Sigma^{-1} \tag{1}$$
$$= \operatorname{tr}(G\Sigma^{-1} - I)^2,$$

where G is a positive definite matrix. The other is based on the form of the likelihood function:

$$L_l(\Sigma, G) = \operatorname{tr} G\Sigma^{-1} - \log|G\Sigma^{-1}| - p. \tag{2}$$

(See Lemma 3.2.2 and alternative proofs in Problems 3.4, 3.8, and 3.12.) Each of these is 0 when $G = \Sigma$ and is positive when $G \neq \Sigma$. The second loss function approaches ∞ as G approaches a singular matrix or when one or more elements (or one or more characteristic roots) of G approaches ∞. (See proof of Lemma 3.2.2.) Each is invariant with respect to transformations $G^* = CGC'$, $\Sigma^* = C\Sigma C'$. We can see some properties of the loss functions from $L_q(I, D) = \sum_{i=1}^p (d_{ii} - 1)^2$ and $L_l(I, D) = \sum_{i=1}^p (d_{ii} - \log d_{ii} - 1)$, where D is diagonal. (By Theorem A.2.2 of the Appendix for arbitrary positive definite Σ and symmetric G, there exists a nonsingular C such that $C\Sigma C' = I$ and $CGC' = D$.) If we let $g = (g_{11}, \dots, g_{pp}, g_{12}, \dots, g_{p-1,p})'$, $s = (s_{11}, \dots, s_{pp}, s_{12}, \dots, s_{p-1,p})'$, $\sigma = (\sigma_{11}, \dots, \sigma_{pp}, \sigma_{12}, \dots, \sigma_{p-1,p})'$, and $\Phi = \mathscr{E}(s-\sigma)(s-\sigma)'$, then $L_q(\Sigma G)$ is a constant multiple of $(g-\sigma)'\Phi^{-1}(g-\sigma)$. (See Problem 7.33.)

The maximum likelihood estimator $\hat{\Sigma}$ and the unbiased estimator S are of the form aA, where A has the distribution $W(\Sigma, n)$ and $n = N-1$.

Theorem 7.8.1. *The quadratic risk of aA is minimized at $a = 1/(n+p+1)$, and its value is $p(p+1)/(n+p+1)$. The likelihood risk of aA is minimized at $a = 1/n$ (i.e., $aA = S$), and its value of $p \log n - \sum_{i=1}^p \mathscr{E} \log \chi^2_{n+1-i}$.*

Proof. By the invariance of the loss function

$$\mathscr{E}_\Sigma L_q(\Sigma, aA) = \mathscr{E}_I L_q(I, aA^*) \tag{3}$$
$$= \mathscr{E}_I \operatorname{tr}(aA^* - I)^2$$
$$= \mathscr{E}_I \left(a^2 \sum_{i,j=1}^p a_{ij}^{*2} - 2a \sum_{i=1}^p a_{ii}^* + p \right)$$
$$= a^2\left[(2n+n^2)p + np(p-1)\right] - 2anp + p$$
$$= p\left[n(n+p+1)a^2 - 2na + 1\right].$$

which has its minimum at $a = 1/(n+p+1)$. Similarly

$$\begin{aligned}(4) \qquad \mathscr{E}_\Sigma L_l(\Sigma, aA) &= \mathscr{E}_I L_l(I, aA^*)\\ &= \mathscr{E}_I\{a \operatorname{tr} A^* - \log|A^*| - p\log a - p\}\\ &= p[na - \log a - 1] - \mathscr{E}_I \log|A^*|,\end{aligned}$$

which is minimized at $a = 1/n$. ∎

Although the minimum risk of the estimator of the form aA is constant for its loss function, the estimator is not minimax. We shall now consider estimators $G(A)$ such that

$$(5) \qquad G(HAH') = HG(A)H'$$

for lower triangular matrices H. The two loss functions are invariant with respect to transformations $G^* = HGH'$, $\Sigma^* = H\Sigma H'$.

Let $A = I$ and H be the diagonal matrix D_i with -1 as the ith diagonal element and 1 as each other diagonal element. Then $HAH' = I$, and the i,jth component of (5) is

$$(6) \qquad g_{ij}(I) = -g_{ij}(I), \qquad j \neq i.$$

Hence, $g_{ij}(I) = 0$, $i \neq j$, and $G(I)$ is diagonal, say D. Since $A = TT'$ for T lower triangular, we have

$$\begin{aligned}(7) \qquad G(A) &= G(TIT')\\ &= TG(I)T'\\ &= TDT',\end{aligned}$$

where D is a diagonal matrix not depending on A. We note in passing that if (5) holds for *all* nonsingular H, then $D = aI$ for some a. (H can be taken as a permutation matrix.)

If $\Sigma = KK'$, where K is lower triangular, then

(8)

$$\begin{aligned}\mathscr{E}_\Sigma L[\Sigma, G(A)] &= \int L[\Sigma, G(A)]C(p,n)|\Sigma|^{-\frac{1}{2}n}|A|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}\operatorname{tr}\Sigma^{-1}A}\,dA\\ &= \int L[KK', G(A)]C(p,n)|KK'|^{-\frac{1}{2}n}|A|^{\frac{1}{2}(n-p-1)}\\ &\qquad \cdot e^{-\frac{1}{2}\operatorname{tr}K'^{-1}K^{-1}}\,dA\end{aligned}$$

$$= \int L[KK', G(KA^*K')]C(p,n)|A^*|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\operatorname{tr} A^*} dA^*$$

$$= \mathscr{E}_I L[KK', KG(A^*)K']$$

$$= \mathscr{E}_I L[I, G(A^*)]$$

by invariance of the loss function. The risk does not depend on Σ.

For the quadratic loss function we calculate

$$(9) \quad \mathscr{E}_I L_q[I, G(A)] = \mathscr{E}_I L_q[I, TDT']$$

$$= \mathscr{E}_I \operatorname{tr}(TDT' - I)^2$$

$$= \mathscr{E}_I \operatorname{tr}(TDT'TDT' - 2TDT' + I)$$

$$= \mathscr{E}_I \sum_{i,j,k,l=1}^{p} t_{ij} d_j t_{kj} t_{kl} d_l t_{il} - 2\mathscr{E}_I \sum_{i,j=1}^{p} t_{ij}^2 d_j + p.$$

The expectations can be evaluated by using the fact that the (nonzero) elements of T are independent, t_{ii}^2 has the χ^2-distribution with $n+1-i$ degrees of freedom, and t_{ij}, $i > j$, has the distribution $N(0,1)$. Then

$$(10) \qquad \mathscr{E}_I L_q[I, G(A)] = d'Fd - 2f'd + p,$$

where $F = (f_{ij})$, $f = (f_i)$,

$$(11) \qquad f_{ii} = (n+p-2i+1)(n+p-2i+3),$$

$$f_{ij} = n+p-2j+1, \qquad i < j,$$

$$f_i = n+p+2i+1,$$

and $d = (d_1, \ldots, d_p)'$. Since $d'Fd = \mathscr{E}\operatorname{tr}(TDT')^2 > 0$, F is positive definite and (10) has a unique minimum. It is attained at $d = F^{-1}f$, and the minimum is $p - f'F^{-1}f$.

Theorem 7.8.2. *With respect to the quadratic loss function the best estimator invariant with respect to linear transformations $\Sigma \to H\Sigma H'$, $A \to HAH'$, where H is lower triangular, is $G(A) = TDT'$, where D is the diagonal matrix whose diagonal elements compose $d = F^{-1}f$, F and f are defined by* (11), *and $A = TT'$ with T lower triangular.*

Since $d = F^{-1}f$ is not proportional to $\varepsilon = (1, \dots, 1)'$, that is, $F\varepsilon$ is not proportional to f (see Problem 7.28), this estimator has a smaller (quadratic) loss than any estimator of the form aA (which is the only type of estimator invariant under the full linear group). Kiefer (1957) showed that if an estimator is minimax in the class of estimators invariant with respect to a group of transformations satisfying certain conditions,† then it is minimax with respect to all estimators. In this problem the group of triangular linear transformations satisfies the conditions, while the group of all linear transformations does not.

The definition of this estimator depends on the coordinate system and on the numbering of the coordinates. These properties are intuitively unappealing.

Theorem 7.8.3. *The estimator $G(A)$ defined in Theorem 7.8.2 is minimax with respect to the quadratic loss function.*

In the case of $p = 2$

$$d_1 = \frac{(n+1)^2 - (n-1)}{(n+1)^2(n+3) - (n-1)}, \qquad d_2 = \frac{(n+1)(n+2)}{(n+1)^2(n+3) - (n-1)}. \tag{12}$$

The risk is

$$2\frac{3n^2 + 5n + 4}{n^3 + 5n^2 + 6n + 4}. \tag{13}$$

The difference between the risks of the best estimator aA and the best estimator TDT' is

$$\frac{6}{n+3} - \frac{6n^2 + 10n + 8}{n^3 + 5n^2 + 6n + 4} = \frac{2n(n-1)}{(n+3)(n^3 + 5n^2 + 6n + 4)}. \tag{14}$$

The difference is $\frac{1}{55}$ for $n = 2$ (relative to $\frac{6}{5}$), and $\frac{1}{47}$ for $n = 3$ (relative to 1); it is of the order $2/n^2$; the improvement due to using the estimator TDT' is not great, at least for $p = 2$.

For the likelihood loss function we calculate

$$\begin{aligned} &\mathscr{E}_I L_l[I, G(A)] \\ &\quad = \mathscr{E}_I L_l[I, TDT'] \\ &\quad = \mathscr{E}_I[\operatorname{tr} TDT' - \log|TDT'| - p] \end{aligned} \tag{15}$$

†The essential condition is that the group is solvable. See Kiefer (1966) and Kudo (1955).

$$= \mathscr{E}_I\left[\sum_{i,j=1}^{p} t_{ij}^2 d_j - \sum_{i=1}^{p} \log t_{ii}^2\right] - \sum_{i=1}^{p} \log d_i - p$$

$$= \sum_{j=1}^{p} (n+p-2j+1)\, d_j - \sum_{j=1}^{p} \log d_j - \sum_{j=1}^{p} \mathscr{E} \log \chi_{n+1-j}^2 - p.$$

The minimum of (15) occurs at $d_j = 1/(n+p-2j+1)$, $j = 1,\ldots,p$.

Theorem 7.8.4. *With respect to the likelihood loss function, the best estimator invariant with respect to linear transformations $\Sigma \to H\Sigma H'$, $A \to HAH'$, where H is lower triangular, is $G(A) = TDT'$, where the jth diagonal element of the diagonal matrix D is $1/(n+p-2j+1)$, $j=1,\ldots,p$, and $A = TT'$, with T lower triangular. The minimum risk is*

$$\mathscr{E}_\Sigma L[\Sigma, G(A)] = \sum_{j=1}^{p} \log(n+p-2j+1) - \sum_{j=1}^{p} \mathscr{E} \log \chi_{n+1-j}^2. \tag{16}$$

Theorem 7.8.5. *The estimator $G(A)$ defined in Theorem 7.8.4 is minimax with respect to the likelihood loss function.*

James and Stein (1961) gave this estimator. Note that the reciprocals of the weights $1/(n+p-1), 1/(n+p-3),\ldots, 1/(n-p+1)$ are symmetrically distributed about the reciprocal of $1/n$.

If $p = 2$,

$$G(A) = \frac{1}{n+1}A + \begin{pmatrix} 0 & 0 \\ 0 & \dfrac{2}{n^2-1} \end{pmatrix} \frac{|A|}{a_{11}}, \tag{17}$$

$$\mathscr{E}G(A) = \frac{n}{n+1}\Sigma + \frac{2}{n+1}\begin{pmatrix} 0 & 0 \\ 0 & \dfrac{|\Sigma|}{\sigma_{11}} \end{pmatrix}. \tag{18}$$

The difference between the risks of the best estimator aA and the best estimator TDT' is

$$p\log n - \sum_{j=1}^{p} \log(n+p-2j+1) = -\sum_{j=1}^{p} \log\left(1 + \frac{p-2j+1}{n}\right). \tag{19}$$

If $p = 2$, the improvement is

$$-\log\left(1+\frac{1}{n}\right) - \log\left(1-\frac{1}{n}\right) = -\log\left(1-\frac{1}{n^2}\right) \tag{20}$$

$$= \frac{1}{n^2} + \frac{1}{2n^4} + \frac{1}{3n^6} + \cdots,$$

which is 0.288 for $n = 2$, 0.118 for $n = 3$, 0.065 for $n = 4$, etc. The risk (19) is $O(1/n^2)$ for any p. (See Problem 7.31.)

An obvious disadvantage of these estimators is that they depend on the coordinate system. Let P_i be the ith permutation matrix, $i = 1, \dots, p!$, and let $P_i A P_i' = T_i T_i'$, where T_i is lower triangular and $t_{jj} > 0$, $j = 1, \dots, p$. Then a randomized estimator that does not depend on the numbering of coordinates is to let the estimator be $P_i' T_i D T_i' P_i$ with probability $1/p!$; this estimator has the same risk as the estimator for the original numbering of coordinates. Since the loss functions are convex, $(1/p!)\Sigma_i P_i' T_i D T_i' P_i$ will have at least as good a risk function; in this case the risk will depend on Σ.

Haff (1980) has shown that $G(A) = [1/(n+p+1)](A + \gamma u C)$, where γ is constant, $0 \le \gamma \le 2(p-1)/(n-p+3)$, $u = 1/\operatorname{tr}(A^{-1}C)$ and C is an arbitrary positive definite matrix, has a smaller quadratic risk than $[1/(n+p+1)]A$. The estimator $G(A) = (1/n)[A + ut(u)C]$, where $t(u)$ is an absolutely continuous, nonincreasing function, $0 \le t(u) \le 2(p-1)/n$, has a smaller likelihood risk than S.

7.9. ELLIPTICALLY CONTOURED DISTRIBUTIONS

7.9.1. Observations Elliptically Contoured

Consider $x_1, \dots, x_N$ observations on a random vector X with density

$$|\Lambda|^{-\frac{1}{2}} g\left[(x-\nu)'\Lambda^{-1}(x-\nu)\right]. \tag{1}$$

Let $A = \sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$, $n = N-1$, $S = (1/n)A$. Then $S \xrightarrow{p} \Sigma$ as $N \to \infty$. The limiting normal distribution of $\sqrt{N}\operatorname{vec}(S - \Sigma)$ was given in Theorem 3.6.2.

The lower triangular matrix T, satisfying $A = TT'$, was used in Section 7.2 in deriving the distribution of A and hence of S. Define the lower triangular matrix $\tilde{T}$ by $S = \tilde{T}\tilde{T}'$, $\tilde{t}_{ii} \ge 0$, $i = 1, \dots, p$. Then $\tilde{T} = (1/\sqrt{n})T$. If $\Sigma = I$, then

$S \xrightarrow{p} I$ and $\tilde{T} \xrightarrow{p} I$, $\sqrt{N}(S-I)$ and $\sqrt{N}(\tilde{T}-I)$ have limiting normal distributions, and

$$\sqrt{N}(S-I) = \sqrt{N}(\tilde{T}-I) + \sqrt{N}(\tilde{T}-I)' + O_p(1). \tag{2}$$

That is, $\sqrt{N}(s_{ii}-1) = 2\sqrt{N}(\tilde{t}_{ii}-1) + O_p(1)$, and $\sqrt{N}s_{ij} = \sqrt{N}\tilde{t}_{ij} + O_p(1)$, $i>j$. When $\Sigma = I$, the set $\sqrt{N}(s_{11}-1),\dots,\sqrt{N}(s_{pp}-1)$ and the set $\sqrt{N}s_{ij}$, $i>j$, are asymptotically independent; $\sqrt{N}s_{12},\dots,\sqrt{N}s_{p-1,p}$ are mutually asymptotically independent, each with variance $1+\kappa$; the limiting variance of $\sqrt{N}(s_{ii}-1)$ is $3\kappa+2$; and the limiting covariance of $\sqrt{N}(s_{ii}-1)$ and $\sqrt{N}(s_{jj}-1)$, $i \neq j$, is κ.

Theorem 7.9.1. *If $\Sigma = I_p$, the limiting distribution of $\sqrt{N}(\tilde{T}-I_p)$ is normal with mean $\mathbf{0}$. The variance of a diagonal element is $(3\kappa+2)/4$; the covariance of two diagonal elements is $\kappa/4$; the variance of an off-diagonal element is $\kappa+1$; the off-diagonal elements are uncorrelated and are uncorrelated with the diagonal elements.*

Let $X = \boldsymbol{\nu} + CY$, where Y has the density $g(y'y)$, $\Lambda = CC'$, and $\Sigma = \mathscr{E}(X-\boldsymbol{\nu})(X-\boldsymbol{\nu})' = (\mathscr{E}R^2/p)\Lambda = \Gamma\Gamma'$, and C and Γ are lower triangular. Let S be the sample covariance of a sample of N on X. Let $S = \tilde{T}\tilde{T}'$. Then $S \xrightarrow{p} \Sigma$, $\tilde{T} \xrightarrow{p} \Gamma$, and

$$\sqrt{N}(S-\Sigma) = \sqrt{N}(\tilde{T}-\Gamma)\Gamma' + \Gamma\sqrt{N}(\tilde{T}-\Gamma)' + O_p(1). \tag{3}$$

The limiting distribution of $\sqrt{T}(\tilde{T}-\Gamma)$ is normal, and the covariance can be calculated from (3) and the covariances of the elements of $\sqrt{N}(S-\Sigma)$. Since the primary interest in T is to find the distribution of S, we do not pursue this further here.

7.9.2. Elliptically Contoured Matrix Distributions

Let X $(N \times p)$ have the density

$$|C|^{-N} g\left[C^{-1}(X-\boldsymbol{\varepsilon}_N\boldsymbol{\nu}')'(X-\boldsymbol{\varepsilon}_N\boldsymbol{\nu}')(C')^{-1}\right] \tag{4}$$

based on the left spherical density $g(Y'Y)$.

Theorem 7.9.2. *Define $T=(t_{ij})$ by $Y'Y = TT'$, $t_{ij}=0$, $i<j$, and $t_{ii} \geq 0$. If the density of Y is $g(Y'Y)$, then the density of T is*

$$\prod_{i=1}^{p}\left\{\frac{2\pi^{\frac{1}{2}(N+1-i)}}{\Gamma\left[\frac{1}{2}(N+1-i)\right]}t_{ii}^{N-i}\right\}g(TT') = \frac{2^p\pi^{\frac{1}{2}Np}}{\Gamma_p\left(\frac{1}{2}N\right)}\prod_{i=1}^{p}t_{ii}^{N-i}g(TT'). \tag{5}$$

Proof. Let $Y = (v_1, \dots, v_p)$. Define w_i and w_i recursively by $w_1 = v_1$, $u_1 = w_1/\|w_1\|$,

$$w_i = v_i - \sum_{j=1}^{i-1} w_j \frac{w_j' v_i}{\|w_j\|^2} = v_i - \sum_{j=1}^{i-1} u_j u_j' v_i, \tag{6}$$

and $u_i = w_i/\|w_i\|$. Then $w_i' w_j = 0$, $u_i' u_j = 0$, $i \neq j$, and $u_i' u_i = 1$. Conditional on $v_1, \dots, v_{i-1}$ (that is, $w_1, \dots, w_{i-1}$), let Q_i be an orthogonal matrix with $u_1', \dots, u_{i-1}'$ as the first $i - 1$ rows; that is,

$$Q_i' = (u_1, \dots, u_{i-1}, Q_i^{*\prime}). \tag{7}$$

(See Lemma A.4.2.) Define

$$z_i = Q_i v_i = \begin{bmatrix} t_{i1} \\ \vdots \\ t_{i,i-1} \\ z_i^* \end{bmatrix}. \tag{8}$$

This transformation of v_i is linear and has Jacobian 1. The vector z_i^* has $N + 1 - i$ components. Note that $\|z_i^*\|^2 = \|w_i\|^2$,

$$v_i = \sum_{j=1}^{i-1} t_{ij} u_j + w_i = \sum_{i=1}^{i-1} t_{ij} u_j + Q_i^{*\prime} z_i^*, \tag{9}$$

$$v_i' v_i = \sum_{j=1}^{i-1} t_{ij}^2 + z_i^{*\prime} z_i^* = \sum_{j=1}^{i} t_{ij}^2, \tag{10}$$

$$v_j' v_i = \sum_{k=1}^{j} t_{jk} t_{ik}, \qquad j < i. \tag{11}$$

The transformation from $Y = (v_1, \dots, v_p)$ to $z_1, \dots, z_p$ has Jacobian 1.

To obtain the density of T convert z_i^* to polar coordinates and integrate with respect to the angular coordinates. (See Section 2.7.1.) ∎

The above proof follows the lines of the proof of (6) in Section 7.2, but does not use information about the normal distribution, such as $t_{ii}^2 \overset{d}{=} \chi_{N+1-i}^2$. See also Fang and Zhang (1990), Theorem 3.4.1.

Let C be a lower triangular matrix such that $\Lambda = CC'$. Define $X = YC'$.

Theorem 7.9.3. *If X $(N \times p)$ has the density*

$$|C|^{-N} g\left[C^{-1} X' X (C')^{-1}\right], \tag{12}$$

then the lower triangular matrix T^* satisfying $X'X = T^*T^{*\prime}$ and $t_{ii} \geq 0$ has the density

$$\frac{2^p \pi^{\frac{1}{2}Np}}{\Gamma_p(\frac{1}{2}N)|\Lambda|^{\frac{1}{2}N}} \prod_{i=1}^{p} t_{ii}^{*\,N-1} g\left[C^{-1}T^*T^{*\prime}(C')^{-1}\right]. \tag{13}$$

Let $A = X'X = T^*T^{*\prime}$.

Theorem 7.9.4. *If X has the density* (12), *then $A = X'X$ has the density*

$$\frac{\pi^{\frac{1}{2}p[N-\frac{1}{2}(p-1)]}}{\Gamma_p(\frac{1}{2}N)|\Lambda|^{\frac{1}{2}N}} |A|^{\frac{1}{2}(N-p+1)} g\left[C^{-1}A(C')^{-1}\right]. \tag{14}$$

The class of densities $g(\operatorname{tr} Y'Y)$ is a subclass of densities $g(Y'Y)$. Let $X = \boldsymbol{\varepsilon}_N \boldsymbol{\nu}' + YC'$. Then the density of X is

$$|\Lambda|^{-\frac{1}{2}N} g\left[\operatorname{tr}(X - \boldsymbol{\varepsilon}_N \boldsymbol{\nu}')\Lambda^{-1}(X - \boldsymbol{\varepsilon}_N \boldsymbol{\mu}')'\right]. \tag{15}$$

A stochastic representation of X is $\operatorname{vec} X \overset{d}{=} R(C \otimes I_N)\operatorname{vec} U + \boldsymbol{\nu} \otimes \boldsymbol{\varepsilon}_N$. Theorems 7.9.3 and 7.9.4 can be specialized to this form. Then Theorem 3.6.5 holds.

Theorem 7.9.5. *Let X have the density* (12) *where Λ is diagonal. Let $S = (N-1)^{-1}(X - \boldsymbol{\varepsilon}_N \bar{\mathbf{x}}')'(X - \boldsymbol{\varepsilon}_N \bar{\mathbf{x}}')$ and $R = (\operatorname{diag} S)^{-\frac{1}{2}} S (\operatorname{diag} S)^{-\frac{1}{2}}$. Then the density of R is* (9) *of Section* 7.6.

PROBLEMS

7.1. (Sec. 7.2) A transformation from rectangular to *polar coordinates* is

$$\begin{aligned}
y_1 &= w \sin \theta_1, \\
y_2 &= w \cos \theta_1 \sin \theta_2, \\
y_3 &= w \cos \theta_1 \cos \theta_2 \sin \theta_3, \\
&\ \ \vdots \\
y_{n-1} &= w \cos \theta_1 \cos \theta_2 \cdots \cos \theta_{n-2} \sin \theta_{n-1}, \\
y_n &= w \cos \theta_1 \cos \theta_2 \cdots \cos \theta_{n-2} \cos \theta_{n-1},
\end{aligned}$$

where $-\frac{1}{2}\pi < \theta_i \leq \frac{1}{2}\pi$, $i = 1, \ldots, n-2$, $-\pi < \theta_{n-1} \leq \pi$, and $0 \leq w < \infty$.

(a) Prove $w^2 = \sum y_\alpha^2$. [*Hint:* Compute in turn $y_n^2 + y_{n-1}^2, (y_n^2 + y_{n-1}^2) + y_{n-2}^2$, and so forth.]

(b) Show that the Jacobian is $w^{n-1} \cos^{n-2}\theta_1 \cos^{n-3}\theta_2 \cdots \cos\theta_{n-2}$. [*Hint:* Prove

$$\left| \frac{\partial(y_1, \dots, y_n)}{\partial(\theta_1, \dots, \theta_{n-1}, w)} \right| \cdot \begin{vmatrix} \cos\theta_1 & 0 & \cdots & 0 & 0 \\ 0 & \cos\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \cos\theta_{n-1} & 0 \\ w\sin\theta_1 & w\sin\theta_2 & \cdots & w\sin\theta_{n-1} & 1 \end{vmatrix}$$

$$= \begin{vmatrix} w & x & \cdots & x & & x & & \\ 0 & w\cos\theta_1 & \cdots & x & & x & & \\ \vdots & \vdots & & \vdots & & \vdots & & \\ 0 & 0 & \cdots & w\cos\theta_1 & \cdots \cos\theta_{n-2} & x & & \\ 0 & 0 & \cdots & 0 & & \cos\theta_1 & \cdots & \cos\theta_{n-1} \end{vmatrix},$$

where x denotes elements whose explicit values are not needed.]

7.2. (Sec. 7.2) Prove that

$$\int_{-\pi/2}^{\pi/2} \cos^{h-1}\theta\, d\theta = \frac{\Gamma(\frac{1}{2}h)\Gamma(\frac{1}{2})}{\Gamma[\frac{1}{2}(h+1)]}.$$

[*Hint:* Let $\cos^2\theta = u$, and use the definition of $B(p,q)$.]

7.3. (Sec. 7.2) Use Problems 7.1 and 7.2 to prove that the *surface area of a sphere* of unit radius in n dimensions is

$$C(n) = \frac{2\pi^{\frac{1}{2}n}}{\Gamma(\frac{1}{2}n)}.$$

7.4. (Sec. 7.2) Use Problems 7.1, 7.2, and 7.3 to prove that if the density of $y' = y_1, \dots, y_n)$ is $f(y'y)$, then the density of $u = y'y$ is $\frac{1}{2}C(n)f(u)u^{\frac{1}{2}n-1}$.

7.5. (Sec. 7.2) *χ^2-distribution.* Use Problem 7.4 to show that if $y_1, \dots, y_n$ are independently distributed, each according to $N(0,1)$, then $U = \sum_{\alpha=1}^n y_\alpha^2$ has the density $u^{\frac{1}{2}n-1}e^{-\frac{1}{2}u}/[2^{\frac{1}{2}n}\Gamma(\frac{1}{2}n)]$, which is the χ^2-density with n degrees of freedom.

7.6. (Sec. 7.2) Use (9) of Section 7.6 to derive the distribution of A.

7.7. (Sec. 7.2) Use the proof of Theorem 7.2.1 to demonstrate $\Pr\{|A| = 0\} = 0$.

7.8. (Sec. 7.2) *Independence of estimators of the parameters of the complex normal distribution.* Let $z_1, \dots, z_N$ be N observations from the complex normal distribution with mean θ and covariance matrix P. (See Problem 2.64.) Show that $\bar{Z}$ and $A = \sum_{\alpha=1}^{N}(Z_\alpha - \bar{Z})(Z_\alpha - \bar{Z})^*$ are independently distributed, and show that A has the distribution of $\sum_{\alpha=1}^{n} W_\alpha W_\alpha^*$, where $W_1, \dots, W_n$ are independently distributed, each according to the complex normal distribution with mean $\mathbf{0}$ and covariance matrix P.

7.9. (Sec. 7.2) *The complex Wishart distribution.* Let $W_1, \dots, W_n$ be independently distributed, each according to the complex normal distribution with mean $\mathbf{0}$ and covariance matrix P. (See Problem 2.64.) Show that the density of $B = \sum_{\alpha=1}^{n} W_\alpha W_\alpha^*$ is

$$\frac{|B|^{n-p} e^{-\frac{1}{2}\operatorname{tr} BP^{-1}}}{|P|^n \pi^{\frac{1}{2}p(p-1)} \prod_{i=1}^{p} \Gamma(n+1-i)}.$$

7.10. (Sec. 7.3) Find the characteristic function of A from $W(\Sigma, n)$. [*Hint:* From $\int w(A|\Sigma, n)\,dA = 1$, one derives

$$\int \frac{|A|^{\frac{1}{2}(n-p-1)} \exp\left(-\frac{1}{2}\operatorname{tr} \Phi^{-1} A\right) dA}{2^{\frac{1}{2}pn}\Gamma_p\left(\frac{1}{2}n\right)} = |\Phi|^{\frac{1}{2}n}$$

as an identity in Φ.] Note that comparison of this result with that of Section 7.3.1 is a proof of the Wishart distribution.

7.11. (Sec. 7.3.2) Prove Theorem 7.3.2 by use of characteristic functions.

7.12. (Sec. 7.3.1) Find the first two moments of the elements of A by differentiating the characteristic function (11).

7.13. (Sec. 7.3) Let $Z_1, \dots, Z_n$ be independently distributed, each according to $N(\mathbf{0}, I)$. Let $W = \sum_{\alpha,\beta=1}^{n} b_{\alpha\beta} Z_\alpha Z_\beta'$. Prove that if $a'Wa = \chi_m^2$ for all a such that $a'a = 1$, then W is distributed according to $W(I, m)$. [*Hint:* Use the characteristic function of $a'Wa$.]

7.14. (Sec. 7.4) Let x_α be an observation from $N(\beta z_\alpha, \Sigma)$, $\alpha = 1, \dots, N$, where z_α is a scalar. Let $b = \sum_\alpha z_\alpha x_\alpha / \sum_\alpha z_\alpha^2$. Use Theorem 7.4.1 to show that $\sum_\alpha x_\alpha x_\alpha' - bb'\sum_\alpha z_\alpha^2$ and bb' are independent.

7.15. (Sec. 7.4) Show that

$$\mathscr{E}\left(\chi_{N-1}^2 \chi_{N-2}^2\right)^h = \mathscr{E}\left(\chi_{2N-4}^2/4\right)^h, \qquad h \geq 0,$$

by use of the duplication formula for the gamma function; χ_{N-1}^2 and χ_{N-2}^2 are independent. Hence show that the distribution of $\chi_{N-1}^2 \chi_{N-2}^2$ is the distribution of $\chi_{2N-4}^2/4$.

7.16. (Sec. 7.4) Verify that Theorem 7.4.1 follows from Lemma 7.4.1. [*Hint:* Prove that Q_i having the distribution $W(\Sigma, r_i)$ implies the existence of (6) where I is of order r_i and that the independence of the Q_i's implies that the I's in (6) do not overlap.]

7.17. (Sec. 7.5) Find $\mathscr{E}|A|^h$ directly from $W(\Sigma, n)$. [*Hint:* The fact that

$$\int w(A|\Sigma, n)\, dA \equiv 1$$

shows

$$\int |A|^{\frac{1}{2}(n-p-1)} \exp\left(-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}A\right) dA = 2^{\frac{1}{2}np}|\Sigma|^{\frac{1}{2}n}\Gamma_p\left(\tfrac{1}{2}n\right)$$

as an identity in n.]

7.18. (Sec. 7.5) Consider the confidence region for μ given by

$$N(\bar{x} - \mu^*)'S^{-1}(\bar{x} - \mu^*) \le \frac{(N-1)p}{N-p}F_{p, N-p}(\varepsilon),$$

where $\bar{x}$ and S are based on a sample of N from $N(\mu, \Sigma)$. Find the expected value of the volume of the confidence region.

7.19. (Sec. 7.6) Prove that if $\Sigma = I$, the joint density of $r_{ij\cdot p}$, $i, j = 1, \dots, p-1$, and $r_{1p}, \dots, r_{p-1,p}$ is

$$\frac{\Gamma^{p-1}\left[\frac{1}{2}(n-1)\right]|R_{11\cdot p}|^{\frac{1}{2}(n-p-1)}}{\pi^{(p-1)(p-2)/4}\prod_{i=1}^{p-1}\Gamma\left[\frac{1}{2}(n-i)\right]} \prod_{i=1}^{p-1} \frac{\Gamma\left(\frac{1}{2}n\right)}{\pi^{\frac{1}{2}}\Gamma\left[\frac{1}{2}(n-1)\right]}\left(1 - r_{ip}^2\right)^{\frac{1}{2}(n-3)},$$

where $R_{11\cdot p} = (r_{ij\cdot p})$. [*Hint:* $r_{ij\cdot p} = (r_{ij} - r_{ip}r_{jp})/(\sqrt{1-r_{ip}^2}\sqrt{1-r_{jp}^2})$ and $|r_{ij}| = |\sqrt{1-r_{ip}^2}\sqrt{1-r_{jp}^2}\, r_{ij\cdot p}|$. Use (9).]

7.20. (Sec. 7.6) Prove that the joint density of $r_{12\cdot 3,\dots,p}, r_{13\cdot 4,\dots,p}, r_{23\cdot 4,\dots,p}, \dots, r_{1p}, \dots, r_{p-1,p}$ is

$$\frac{\Gamma\left\{\frac{1}{2}[n-(p-2)]\right\}}{\pi^{\frac{1}{2}}\Gamma\left\{\frac{1}{2}[n-(p-1)]\right\}}\left(1 - r_{12\cdot 3,\dots,p}^2\right)^{\frac{1}{2}[n-(p+1)]}$$

$$\cdot \prod_{i=1}^{2} \frac{\Gamma\left\{\frac{1}{2}[n-(p-3)]\right\}}{\pi^{\frac{1}{2}}\Gamma\left\{\frac{1}{2}[n-(p-2)]\right\}}\left(1 - r_{i3\cdot 4,\dots,p}^2\right)^{\frac{1}{2}(n-p)}$$

$$\cdots \prod_{i=1}^{p-2} \frac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\pi^{\frac{1}{2}}\Gamma\left[\frac{1}{2}(n-2)\right]}\left(1 - r_{i,p-1\cdot p}^2\right)^{\frac{1}{2}(n-4)}$$

$$\cdot \prod_{i=1}^{p-1} \frac{\Gamma\left(\frac{1}{2}n\right)}{\pi^{\frac{1}{2}}\Gamma\left[\frac{1}{2}(n-1)\right]}\left(1 - r_{ip}^2\right)^{\frac{1}{2}(n-3)}.$$

[*Hint:* Use the result of Problem 7.19 inductivity.]

7.21. (Sec. 7.6) Prove (without the use of Problem 7.20) that if $\Sigma = I$, then $r_{1p}, \ldots, r_{p-1,p}$ are independently distributed. [*Hint:* $r_{ip} = a_{ip}/(\sqrt{a_{ii}}\sqrt{a_{pp}})$. Prove that the pairs $(a_{1p}, a_{11}), \ldots, (a_{p-1,p}, a_{p-1,p-1})$ are independent when $(z_{1p}, \ldots, z_{np})$ are fixed, and note from Section 4.2.1 that the marginal distribution of r_{ip}, conditional on $z_{\alpha p}$, does not depend on $z_{\alpha p}$.]

7.22. (Sec. 7.6) Prove (without the use of Problems 7.19 and 7.20) that if $\Sigma = I$, then the set $r_{1p}, \ldots, r_{p-1,p}$ is independent of the set $r_{ij\cdot p}$, $i, j = 1, \ldots, p-1$. [*Hint:* From Section 4.3.2 a_{pp}, and (a_{ip}) are independent of $(a_{ij\cdot p})$. Prove that $a_{pp}, (a_{ip})$, and a_{ii}, $i = 1, \ldots, p-1$, are independent of $(r_{ij\cdot p})$ by proving that $a_{ii\cdot p}$ are independent of $(r_{ij\cdot p})$. See Problem 4.21.]

7.23. (Sec. 7.6) Prove the conclusion of Problem 7.20 by using Problems 7.21 and 7.22.

7.24. (Sec. 7.6) Reverse the steps in Problem 7.20 to derive (9) of Section 7.6.

7.25. (Sec. 7.6) Show that when $p = 3$ and Σ is diagonal r_{12}, r_{13}, r_{23} are not mutually independent.

7.26. (Sec. 7.6) Show that when Σ is diagonal the set r_{ij} are pairwise independent.

7.27. (Sec. 7.7) *Multivariate t-distribution.* Let y and u be independently distributed according to $N(\mathbf{0}, \Sigma)$ and the χ_n^2-distribution, respectively, and let $\sqrt{n/u}\, y = x - \boldsymbol{\mu}$.

(a) Show that the density of x is

$$\frac{\Gamma[\frac{1}{2}(n+p)]}{\Gamma(\frac{1}{2}n) n^{\frac{1}{2}p} \pi^{\frac{1}{2}p} |\Sigma|^{\frac{1}{2}} \left[1 - \frac{1}{n}(x - \boldsymbol{\mu})'\Sigma^{-1}(x - \boldsymbol{\mu})\right]^{\frac{1}{2}(n+p)}}.$$

(b) Show that $\mathscr{E}x = \boldsymbol{\mu}$ and

$$\mathscr{E}(x - \boldsymbol{\mu})(x - \boldsymbol{\mu})' = \frac{n}{n-2}\Sigma.$$

7.28. (Sec. 7.8) Prove that $F\varepsilon$ is not proportional to f by calculating $F\varepsilon$.

7.29. (Sec. 7.8) Prove for $p = 2$

$$TDT' = d_1 A + (d_2 - d_1)\begin{pmatrix} 0 & 0 \\ 0 & \dfrac{|A|}{a_{11}} \end{pmatrix}.$$

7.30. (Sec. 7.8) Verify (17) and (18). [*Hint:* To verify (18) let $\Sigma = KK'$, $A = KA^*K'$, and $A^* = T^*T^*$, where K and T^* are lower triangular.]

7.31. (Sec. 7.8) Prove for optimal D

$$\mathscr{E}_l L_l(I,S) - \mathscr{E}_l L_l(I,TDT') = -\sum_{i=1}^{\frac{1}{2}p} \log\left[1-\left(\frac{p-2i+1}{n}\right)^2\right], \qquad p \text{ even},$$

$$= -\sum_{i=1}^{\frac{1}{2}(p-1)} \log\left[1-\left(\frac{p-2i+1}{n}\right)^2\right], \qquad p \text{ odd}.$$

7.32. (Sec. 7.8) Prove $L_q(\Sigma, G)$ and $L_l(\Sigma, G)$ are invariant with respect to transformations $G^* = CGC'$, $\Sigma^* = C\Sigma C'$ for C nonsingular.

7.33. (Sec. 7.8) Prove $L_q(\Sigma, G)$ is a multiple of $(g-\sigma)'\Phi^{-1}(g-\sigma)$. *Hint:* Transform so $\Sigma = I$. Then show

$$\Phi = \frac{1}{n}\begin{pmatrix} 2I & 0 \\ 0 & I \end{pmatrix}.$$

7.34. (Sec. 7.8) Verify (11).

7.35. Let the density of Y be $f(y) = K$ for $y'y \le p+2$ and 0 elsewhere. Prove that $K = \Gamma(\frac{1}{2}p+1)/[(p+2)\pi]^{\frac{1}{2}p}$, and show that $\mathscr{E}Y = 0$ and $\mathscr{E}YY' = I$.

7.36. (Sec. 7.2) *Dirichlet distribution.* Let $Y_1, \dots, Y_m$ be independently distributed as χ^2-variables with $p_1, \dots, p_m$ degrees of freedom, respectively. Define $Z_i = Y_i / \sum_{j=1}^m Y_j$, $i = 1, \dots, m$. Show that the density of $Z_1, \dots, Z_{m-1}$ is

$$\frac{\Gamma\left(\frac{1}{2}\sum_{i=1}^m p_i\right)}{\prod_{i=1}^m \Gamma\left(\frac{1}{2}p_i\right)} z_1^{\frac{1}{2}p_1 - 1} \cdots z_m^{\frac{1}{2}p_m - 1}, \quad z_m \equiv 1 - \sum_{i=1}^{m-1} z_i,$$

for $z_i \ge 0$, $i = 1, \dots, m$.

7.37. (Sec. 7.5) Show that if χ^2_{N-1} and χ^2_{N-2} are independently distributed, then $\chi^2_{N-1}\chi^2_{N-2}$ is distributed as $(\chi^2_{2N-4})^2/4$. [*Hint:* In the joint density of $x = \chi^2_{N-1}$ and $y = \chi^2_{N-2}$ substitute $z = 2\sqrt{xy}$, $x = x$, and express the marginal density of z as $z^{N-3}h(z)$, where $h(z)$ is an integral with respect to x. Find $h'(z)$, and solve the differential equation. See Srivastava and Khatri (1979), Chapter 3.]

CHAPTER 8

Testing the General Linear Hypothesis; Multivariate Analysis of Variance

8.1. INTRODUCTION

In this chapter we generalize the univariate least squares theory (i.e., regression analysis) and the analysis of variance to vector variates. The algebra of the multivariate case is essentially the same as that of the univariate case. This leads to distribution theory that is analogous to that of the univariate case and to test criteria that are analogs of F-statistics. In fact, given a univariate test, we shall be able to write down immediately a corresponding multivariate test. Since the analysis of variance based on the model of fixed effects can be obtained from least squares theory, we obtain directly a theory of multivariate analysis of variance. However, in the multivariate case there is more latitude in the choice of tests of significance.

In univariate least squares we consider scalar dependent variates $x_1, \dots, x_N$ drawn from populations with expected values $\boldsymbol{\beta}' z_1, \dots, \boldsymbol{\beta}' z_N$, respectively, where $\boldsymbol{\beta}$ is a column vector of q components and each of the z_α is a column vector of q known components. Under the assumption that the variances in the populations are the same, the least squares estimator of $\boldsymbol{\beta}'$ is

$$b' = \left(\sum_{\alpha=1}^{N} x_\alpha z_\alpha' \right) \left(\sum_{\alpha=1}^{N} z_\alpha z_\alpha' \right)^{-1}. \tag{1}$$

An Introduction to Multivariate Statistical Analysis, *Third Edition*. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

If the populations are normal, the vector is the maximum likelihood estimator of $\boldsymbol{\beta}$. The unbiased estimator of the common variance σ^2 is

$$s^2 = \sum_{\alpha=1}^{N} (x_\alpha - b'z_\alpha)^2/(N-q), \tag{2}$$

and under the assumption of normality, the maximum likelihood estimator of σ^2 is $\hat{\sigma}^2 = (N-q)s^2/N$.

In the multivariate case x_α is a vector, $\boldsymbol{\beta}'$ is replaced by a matrix $\boldsymbol{\beta}$, and σ^2 is replaced by a covariance matrix Σ. The estimators of $\boldsymbol{\beta}$ and Σ, given in Section 8.2, are matric analogs of (1) and (2).

To test a hypothesis concerning $\boldsymbol{\beta}$, say the hypothesis $\boldsymbol{\beta} = \mathbf{0}$, we use an F-test. A criterion equivalent to the F-ratio is

$$\frac{1}{[q/(N-q)]F+1} = \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}, \tag{3}$$

where $\hat{\sigma}_0^2$ is the maximum likelihood estimator of σ^2 under the null hypothesis. We shall find that the likelihood ratio criterion for the corresponding multivariate hypothesis, say $\boldsymbol{\beta} = \mathbf{0}$, is the above with the variances replaced by generalized variances. The distribution of the likelihood ratic criterion under the null hypothesis is characterized, the moments are found, and some specific distributions obtained. Satisfactory approximations are given as well as tables of significance points (Appendix B).

The hypothesis testing problem is invariant under several groups of linear transformations. Other invariant criteria are treated, including the Lawley–Hotelling trace, the Bartlett–Nanda–Pillai trace, and the Roy maximum root criteria. Some comparison of power is made.

Confidence regions or simultaneous confidence intervals for elements of $\boldsymbol{\beta}$ can be based on the likelihood ratio test, the Lawley–Hotelling trace test, and the Roy maximum root test. Procedures are given explicitly for several problems of the analysis of variance. Optimal properties of admissibility, unbiasedness, and monotonicity of power functions are studied. Finally, the theory and methods are extended to elliptically contoured distributions.

8.2. ESTIMATORS OF PARAMETERS IN MULTIVARIATE LINEAR REGRESSION

8.2.1. Maximum Likelihood Estimators; Least Squares Estimators

Suppose $x_1, \dots, x_N$ are a set of N independent observations, x_α being drawn from $N(\boldsymbol{\beta} z_\alpha, \Sigma)$. Ordinarily the vectors z_α (with q components) are known vectors, and the $p \times p$ matrix Σ and the $p \times q$ matrix $\boldsymbol{\beta}$ are unknown. We assume $N \geq p + q$ and the rank of

$$Z = (z_1, \dots, z_N) \tag{1}$$

is q. We shall estimate Σ and $\boldsymbol{\beta}$ by the method of maximum likelihood. The likelihood function is

$$L = (2\pi)^{-\frac{1}{2}Np} |\Sigma^*|^{-\frac{1}{2}N} \exp\left[-\tfrac{1}{2} \sum_{\alpha=1}^{N} (x_\alpha - \boldsymbol{\beta}^* z_\alpha)' \Sigma^{*-1} (x_\alpha - \boldsymbol{\beta}^* z_\alpha) \right]. \tag{2}$$

In (2) the elements of Σ^* and $\boldsymbol{\beta}^*$ are indeterminates. The method of maximum likelihood specifies the estimators of Σ and $\boldsymbol{\beta}$ based on the given sample $x_1, z_1, \dots, x_N, z_N$ as the Σ^* and $\boldsymbol{\beta}^*$ that maximize (2). It is convenient to use the following lemma.

Lemma 8.2.1. *Let*

$$B = \sum_{\alpha=1}^{N} x_\alpha z'_\alpha \left(\sum_{\alpha=1}^{N} z_\alpha z'_\alpha \right)^{-1}. \tag{3}$$

Then for any $p \times q$ matrix F

$$\sum_{\alpha=1}^{N} (x_\alpha - F z_\alpha)(x_\alpha - F z_\alpha)' = \sum_{\alpha=1}^{N} (x_\alpha - B z_\alpha)(x_\alpha - B z_\alpha)' + (B - F) \sum_{\alpha=1}^{N} z_\alpha z'_\alpha (B - F)'. \tag{4}$$

Proof. The left-hand side of (4) is

$$\sum_{\alpha=1}^{N} [(x_\alpha - B z_\alpha) + (B - F) z_\alpha][(x_\alpha - B z_\alpha) + (B - F) z_\alpha]', \tag{5}$$

which is equal to the right-hand side of (4) because

$$\sum_{\alpha=1}^{N} z_\alpha (x_\alpha - B z_\alpha)' = \mathbf{0} \tag{6}$$

by virtue of (3). ∎

The exponential in L is $-\frac{1}{2}$ times

(7)

$$\operatorname{tr}\Sigma^{*-1}\sum_{\alpha=1}^{N}(x_\alpha-\beta^* z_\alpha)(x_\alpha-\beta^* z_\alpha)' = \operatorname{tr}\Sigma^{*-1}\sum_{\alpha=1}^{N}(x_\alpha - Bz_\alpha)(x_\alpha - Bz_\alpha)' + \operatorname{tr}\Sigma^{*-1}(B-\beta^*)A(B-\beta^*)',$$

where

$$A=\sum_{\alpha=1}^{N} z_\alpha z'_\alpha. \tag{8}$$

The likelihood is maximized with respect to β^* by minimizing the last term in (7).

Lemma 8.2.2. *If A and G are positive definite, $\operatorname{tr} FAF'G>0$ for $F\neq 0$.*

Proof. Let $A=HH'$, $G=KK'$. Then

$$\begin{aligned}\operatorname{tr} FAF'G &= \operatorname{tr} FHH'F'KK' = \operatorname{tr} K'FHH'F'K \\ &= \operatorname{tr}(K'FH)(K'FH)'>0\end{aligned} \tag{9}$$

for $F\neq 0$ because then $K'FH\neq 0$ since H and K are nonsingular. ■

It follows from (7) and the lemma that L is maximized with respect to β^* by $\beta^* = B$, that is,

$$\hat{\beta}=CA^{-1}, \tag{10}$$

where

$$C=\sum_{\alpha=1}^{N} x_\alpha z'_\alpha. \tag{11}$$

Then by Lemma 3.2.2, L is maximized with respect to Σ^* at

$$\hat{\Sigma}=\frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha-\hat{\beta}z_\alpha)(x_\alpha-\hat{\beta}z_\alpha)'. \tag{12}$$

This is the multivariate analog of $\hat{\sigma}^2=(N-q)s^2/N$ defined by (2) of Section 8.1.

Theorem 8.2.1. *If x_α is an observation from $N(\beta z_\alpha, \Sigma)$, $\alpha=1,\dots,N$, with $(z_1,\dots,z_N)$ of rank q, the maximum likelihood estimator of β is given by* (10), *where $C=\sum_\alpha x_\alpha z'_\alpha$ and $A=\sum_\alpha z_\alpha z'_\alpha$. The maximum likelihood estimator of Σ is given by* (12).

A useful algebraic result follows from (12) and (4) with $F = 0$:

$$N\hat{\Sigma} = \sum_{\alpha=1}^{N} x_\alpha x'_\alpha - \hat{\boldsymbol{\beta}} A \hat{\boldsymbol{\beta}}' = \sum_{\alpha=1}^{N} x_\alpha x'_\alpha - CA^{-1}C'. \tag{13}$$

Now let us consider a geometric interpretation of the estimation procedure. Let the ith row of $(x_1, \dots, x_N)$ be x_i^* (with N components) and the ith row of $(z_1, \dots, z_N)$ be z_i^* (with N components). Then $\sum_j \hat{\beta}_{ij} z_j^*$, being a linear combination of the vectors $z_1^*, \dots, z_q^*$, is a vector in the q-space spanned by $z_1^*, \dots, z_q^*$, and is in fact, of all such vectors, the one nearest to x_i^*; hence, it is the projection of x_i^* on the q-space. Thus $x_i^* - \sum_j \hat{\beta}_{ij} z_j^*$ is the vector orthogonal to the q-space going from the projection of x_i^* on the q-space to x_i^*. Translate this vector so that one endpoint is at the origin. Then the set of p vectors $x_1^* - \sum_j \hat{\beta}_{1j} z_j^*, \dots, x_p^* - \sum_j \hat{\beta}_{pj} z_j^*$ is a set of vectors emanating from the origin. $N\hat{\sigma}_{ii} = (x_i^* - \sum_j \hat{\beta}_{ij} z_j^*)(x_i^* - \sum_j \hat{\beta}_{ij} z_j^*)'$ is the square of the length of the ith such vector, and $N\hat{\sigma}_{ij} = (x_i^* - \sum_h \hat{\beta}_{ih} z_h^*)(x_j^* - \sum_g \hat{\beta}_{jg} z_g^*)'$ is the product of the length of the ith vector, the length of the jth vector, and the cosine of the angle between them.

The equations defining the maximum likelihood estimator of $\boldsymbol{\beta}$, namely, $AB' = C'$, consist of p sets of q linear equations in q unknowns. Each set can be solved by the method of pivotal condensation or successive elimination (Section A.5 of the Appendix). The forward solutions are the same (except the right-hand sides) for all sets. Use of (13) to compute $N\hat{\Sigma}$ involves an efficient computation of $\hat{\boldsymbol{\beta}} A \hat{\boldsymbol{\beta}}'$.

Let $X_\alpha = (x_{1\alpha}, \dots, x_{p\alpha})'$, $\mathbf{B} = (b_1, \dots, b_p)'$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)'$. Then $\mathscr{E} x_{i\alpha} = \boldsymbol{\beta}'_i z_\alpha$ and b_i is the least squares estimator of $\boldsymbol{\beta}_i$. If G is a positive definite matrix, then $\operatorname{tr} G \sum_{\alpha=1}^{n} (x_\alpha - F z_\alpha)(x_\alpha - F z_\alpha)'$ is minimized by $F = \mathbf{B}$. This is another sense in which $\mathbf{B}$ is the least squares estimator.

8.2.2. Distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$

Now let us find the joint distribution of $\hat{\beta}_{ig}$ $(i = 1, \dots, p,\ g = 1, \dots, q)$. The joint distribution is normal since the $\hat{\beta}_{ig}$ are linear combinations of the $X_{i\alpha}$. From (10) we see that

$$\begin{aligned} \mathscr{E}\hat{\boldsymbol{\beta}} &= \mathscr{E} \sum_{\alpha=1}^{N} X_\alpha z'_\alpha A^{-1} \\ &= \sum_{\alpha=1}^{N} \boldsymbol{\beta} z_\alpha z'_\alpha A^{-1} = \boldsymbol{\beta} A A^{-1} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{14}$$

Thus $\hat{\boldsymbol{\beta}}$ is an *unbiased estimator* of $\boldsymbol{\beta}$. The covariance between $\hat{\boldsymbol{\beta}}'_i$ and $\hat{\boldsymbol{\beta}}'_j$, two rows of $\hat{\boldsymbol{\beta}}$, is

(15)

$$\begin{aligned}
\mathscr{E}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)' &= A^{-1}\mathscr{E}\sum_{\alpha=1}^{N}(X_{i\alpha} - \mathscr{E}X_{i\alpha})z_\alpha \sum_{\gamma=1}^{N}(X_{j\gamma} - \mathscr{E}X_{j\gamma})z'_\gamma A^{-1} \\
&= A^{-1}\sum_{\alpha,\gamma=1}^{N}\mathscr{E}(X_{i\alpha} - \mathscr{E}X_{i\alpha})(X_{j\gamma} - \mathscr{E}X_{j\gamma})z_\alpha z'_\gamma A^{-1} \\
&= A^{-1}\sum_{\alpha,\gamma=1}^{N}\delta_{\alpha\gamma}\sigma_{ij}z_\alpha z'_\gamma A^{-1} \\
&= A^{-1}\sum_{\alpha=1}^{N}\sigma_{ij}z_\alpha z'_\alpha A^{-1} \\
&= \sigma_{ij}A^{-1}AA^{-1} \\
&= \sigma_{ij}A^{-1}.
\end{aligned}$$

To summarize, the vector of pq components $(\hat{\boldsymbol{\beta}}'_1, \dots, \hat{\boldsymbol{\beta}}'_p)' = \operatorname{vec}\hat{\boldsymbol{\beta}}'$ is normally distributed with mean $(\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_p)' = \operatorname{vec}\boldsymbol{\beta}'$ and covariance matrix

$$\begin{pmatrix} \sigma_{11}A^{-1} & \sigma_{12}A^{-1} & \cdots & \sigma_{1p}A^{-1} \\ \sigma_{21}A^{-1} & \sigma_{22}A^{-1} & \cdots & \sigma_{2p}A^{-1} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1}A^{-1} & \sigma_{p2}A^{-1} & \cdots & \sigma_{pp}A^{-1} \end{pmatrix}. \tag{16}$$

The matrix (16) is the Kronecker (or direct) product of the matrices $\boldsymbol{\Sigma}$ and A^{-1}, denoted by $\boldsymbol{\Sigma} \otimes A^{-1}$.

From Theorem 4.3.3 it follows that $N\hat{\boldsymbol{\Sigma}} = \sum_{\alpha=1}^{N} x_\alpha x'_\alpha - \hat{\boldsymbol{\beta}}A\hat{\boldsymbol{\beta}}'$ is distributed according to $W(\boldsymbol{\Sigma}, N-q)$. From this we see that an unbiased estimator of $\boldsymbol{\Sigma}$ is $S = [N/(N-q)]\hat{\boldsymbol{\Sigma}}$.

Theorem 8.2.2. *The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ based on a set of N observations, the αth from $N(\boldsymbol{\beta}z_\alpha, \boldsymbol{\Sigma})$, is normally distributed with mean $\boldsymbol{\beta}$, and the covariance matrix of the ith and jth rows of $\hat{\boldsymbol{\beta}}$ is $\sigma_{ij}A^{-1}$, where $A = \sum_\alpha z_\alpha z'_\alpha$. The maximum likelihood estimator $\hat{\boldsymbol{\Sigma}}$ multiplied by N is independently distributed according to $W(\boldsymbol{\Sigma}, N-q)$, where q is the number of components of z_α.*

The density then can be written [by virtue of (4)]

$$(17)\qquad \frac{1}{(2\pi)^{\frac{1}{2}pN}|\Sigma|^{\frac{1}{2}N}}\exp\left(-\tfrac{1}{2}\operatorname{tr}\left\{\Sigma^{-1}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})A(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'+N\hat{\Sigma}\right]\right\}\right).$$

This proves the following:

Corollary 8.2.1. *$\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$ form a sufficient set of statistics for $\boldsymbol{\beta}$ and Σ.*

A useful theorem is the following.

Theorem 8.2.3. *Let X_α be distributed according to $N(\boldsymbol{\beta} z_\alpha, \Sigma)$, $\alpha = 1, \dots, N$, and suppose $X_1, \dots, X_N$ are independent.*

(a) *If $w_\alpha = Hz_\alpha$ and $\Gamma = \boldsymbol{\beta} H^{-1}$, then X_α is distributed according to $N(\Gamma w_\alpha, \Sigma)$.*

(b) *The maximum likelihood estimator of Γ based on observations x_α on X_α, $\alpha = 1, \dots, N$, is $\hat{\Gamma} = \hat{\boldsymbol{\beta}} H^{-1}$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$.*

(c) *$\hat{\Gamma}(\sum_\alpha w_\alpha w_\alpha')\hat{\Gamma}' = \hat{\boldsymbol{\beta}} A \hat{\boldsymbol{\beta}}'$, where $A = \sum_\alpha z_\alpha z_\alpha'$ and the maximum likelihood estimator of $N\Sigma$ is $N\hat{\Sigma} = \sum_\alpha x_\alpha x_\alpha' - \hat{\Gamma}(\sum_\alpha w_\alpha w_\alpha')\hat{\Gamma}' = \sum_\alpha x_\alpha x_\alpha' - \hat{\boldsymbol{\beta}} A \hat{\boldsymbol{\beta}}'$.*

(d) *$\hat{\Gamma}$ and $\hat{\Sigma}$ are independently distributed.*

(e) *$\hat{\Gamma}$ is normally distributed with mean Γ and the covariance matrix of the ith and jth rows of $\hat{\Gamma}$ is $\sigma_{ij}(HAH')^{-1} = \sigma_{ij}H'^{-1}A^{-1}H^{-1}$.*

The proof is left to the reader.

An estimator F is a *linear estimator* of β_{ig} if $F = \sum_{\alpha=1}^N f_\alpha' x_\alpha$. It is a *linear unbiased estimator* of β_{ig} if

$$(18)\qquad \beta_{ig} = \mathscr{E}F = \mathscr{E}\sum_{\alpha=1}^N f_\alpha' x_\alpha = \sum_{\alpha=1}^N f_\alpha' \boldsymbol{\beta} z_\alpha = \sum_{\alpha=1}^N \sum_{j=1}^p \sum_{h=1}^q f_{j\alpha}\beta_{jh} z_{h\alpha}$$

is an identity in $\boldsymbol{\beta}$, that is, if

$$(19)\qquad \begin{aligned} \sum_{\alpha=1}^N f_{j\alpha} z_{h\alpha} &= 1, \qquad & j = i,\quad h = g, \\ &= 0, & \text{otherwise}. \end{aligned}$$

A linear unbiased estimator is *best* if it has minimum variance over all linear unbiased estimators; that is, if $\mathscr{E}(F - \beta_{ig})^2 \le \mathscr{E}(G - \beta_{ig})^2$ for $G = \sum_{\alpha=1}^N g_\alpha' x_\alpha$ and $\mathscr{E}G = \beta_{ig}$.

Theorem 8.2.4. *The least squares estimator is the best linear unbiased estimator of β_{ig}.*

Proof. Let $\tilde{\beta}_{ig} = \sum_{\alpha=1}^{N} \sum_{j=1}^{p} f_{j\alpha} x_{j\alpha}$ be an arbitrary unbiased estimator of β_{ig}, and let $\hat{\beta}_{ig} = \sum_{\alpha=1}^{N} \sum_{h=1}^{q} x_{i\alpha} z_{h\alpha} a^{hg}$ be the least squares estimator, where $A = \sum_{\alpha=1}^{N} z_\alpha z'_\alpha$. Then

(20)
$$\mathscr{E}\left(\tilde{\beta}_{ig} - \beta_{ig}\right)^2 = \mathscr{E}\left[\hat{\beta}_{ig} - \beta_{ig} + \left(\tilde{\beta}_{ig} - \hat{\beta}_{ig}\right)\right]^2$$
$$= \mathscr{E}\left(\hat{\beta}_{ig} - \beta_{ig}\right)^2 + 2\,\mathscr{E}\left(\hat{\beta}_{ig} - \beta_{ig}\right)\left(\tilde{\beta}_{ig} - \hat{\beta}_{ig}\right) + \mathscr{E}\left(\tilde{\beta}_{ig} - \hat{\beta}_{ig}\right)^2.$$

Because $\tilde{\beta}_{ig}$ and $\hat{\beta}_{ig}$ are unbiased, $\tilde{\beta}_{ig} - \beta_{ig} = \sum_{\alpha=1}^{N} \sum_{j=1}^{p} f_{j\alpha} u_{j\alpha}$, $\hat{\beta}_{ig} - \beta_{ig} = \sum_{\alpha=1}^{N} \sum_{h=1}^{q} u_{i\alpha} z_{h\alpha} a^{hg}$, and

$$\tilde{\beta}_{ig} - \hat{\beta}_{ig} = \sum_{\alpha=1}^{N} \sum_{j=1}^{p} \left(f_{j\alpha}^{ig} - \delta_{ij} \sum_{h=1}^{q} z_{h\alpha} a^{hg} \right) u_{j\alpha}, \tag{21}$$

where $\delta_{ii} = 1$ and $\delta_{ij} = 0$, $i \neq j$. Then

(22)
$$\mathscr{E}\left(\hat{\beta}_{ig} - \beta_{ig}\right)\left(\tilde{\beta}_{ig} - \hat{\beta}_{ig}\right)$$
$$= \mathscr{E} \sum_{\alpha,\gamma=1}^{N} \sum_{h=1}^{q} z_{h\alpha} a^{hg} u_{i\alpha} \sum_{j=1}^{p} \left(f_{j\gamma} - \delta_{ij} \sum_{h'=1}^{q} z_{h'\gamma} a^{h'g} \right) u_{j\gamma}$$
$$= \sum_{\alpha=1}^{N} \sum_{h=1}^{q} \sum_{j=1}^{p} z_{h\alpha} a^{hg} \left(f_{j\alpha} - \delta_{ij} \sum_{h'=1}^{q} z_{h'\alpha} a^{h'g} \right) \sigma_{ij}$$
$$= \sigma_{ii} a^{gg} - \sigma_{ii} \sum_{h=1}^{q} \sum_{h'=1}^{q} a_{hh'} a^{hg} a^{h'g}$$
$$= 0.$$

Then (20) implies $\mathscr{E}(\tilde{\beta}_{ig} - \beta_{ig})^2 \geq \mathscr{E}(\hat{\beta}_{ig} - \beta_{ig})^2$. ■

8.3. LIKELIHOOD RATIO CRITERIA FOR TESTING LINEAR HYPOTHESES ABOUT REGRESSION COEFFICIENTS

8.3.1. Likelihood Ratio Criteria

Suppose we partition

$$\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2\right) \tag{1}$$

so that $\boldsymbol{\beta}_1$ has q_1 columns and $\boldsymbol{\beta}_2$ has q_2 columns. We shall derive the likelihood ratio criterion for testing the hypothesis

$$H: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*, \tag{2}$$

where $\boldsymbol{\beta}_1^*$ is a given matrix. The maximum of the likelihood function L for the sample $x_1, \dots, x_N$ is

$$\max_{\boldsymbol{\beta}, \boldsymbol{\Sigma}} L = (2\pi)^{-\frac{1}{2}pN} |\hat{\boldsymbol{\Sigma}}_\Omega|^{-\frac{1}{2}N} e^{-\frac{1}{2}pN}, \tag{3}$$

where $\hat{\boldsymbol{\Sigma}}_\Omega$ is given by (12) or (13) of Section 8.2.

To find the maximum of the likelihood function for the parameters restricted to ω defined by (2) we let

$$y_\alpha = x_\alpha - \boldsymbol{\beta}_1^* z_\alpha^{(1)}, \qquad \alpha = 1, \dots, N, \tag{4}$$

where

$$z_\alpha = \begin{pmatrix} z_\alpha^{(1)} \\ z_\alpha^{(2)} \end{pmatrix}, \qquad \alpha = 1, \dots, N, \tag{5}$$

is partitioned in a manner corresponding to the partitioning of $\boldsymbol{\beta}$. Then y_α can be considered as an observation from $N(\boldsymbol{\beta}_2 z_\alpha^{(2)}, \boldsymbol{\Sigma})$. The estimator of $\boldsymbol{\beta}_2$ is obtained by the procedure of Section 8.2 as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{2\omega} &= \sum_{\alpha=1}^{N} y_\alpha z_\alpha^{(2)\prime} A_{22}^{-1} = \sum_{\alpha=1}^{N} \left(x_\alpha - \boldsymbol{\beta}_1^* z_\alpha^{(1)}\right) z_\alpha^{(2)\prime} A_{22}^{-1} \\ &= (C_2 - \boldsymbol{\beta}_1^* A_{12}) A_{22}^{-1} \end{aligned} \tag{6}$$

with C and A partitioned in the manner corresponding to the partitioning of $\boldsymbol{\beta}$ and z_α,

$$C = \begin{pmatrix} C_1 & C_2 \end{pmatrix}, \tag{7}$$

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}. \tag{8}$$

The estimator of $\boldsymbol{\Sigma}$ is given by

$$\begin{aligned} N\hat{\boldsymbol{\Sigma}}_\omega &= \sum_{\alpha=1}^{N} \left(y_\alpha - \hat{\boldsymbol{\beta}}_{2\omega} z_\alpha^{(2)}\right)\left(y_\alpha - \hat{\boldsymbol{\beta}}_{2\omega} z_\alpha^{(2)}\right)' \\ &= \sum_{\alpha=1}^{N} y_\alpha y_\alpha' - \hat{\boldsymbol{\beta}}_{2\omega} A_{22} \hat{\boldsymbol{\beta}}_{2\omega}' \\ &= \sum_{\alpha=1}^{N} \left(x_\alpha - \boldsymbol{\beta}_1^* z_\alpha^{(1)}\right)\left(x_\alpha - \boldsymbol{\beta}_1^* z_\alpha^{(1)}\right)' - \hat{\boldsymbol{\beta}}_{2\omega} A_{22} \hat{\boldsymbol{\beta}}_{2\omega}'. \end{aligned} \tag{9}$$

Thus the maximum of the likelihood function over ω is

$$(10)\qquad \max_{\boldsymbol{\beta}_2,\boldsymbol{\Sigma}} L = (2\pi)^{-\frac{1}{2}pN}|\hat{\boldsymbol{\Sigma}}_\omega|^{-\frac{1}{2}N}e^{-\frac{1}{2}pN}.$$

The likelihood ratio criterion for testing H is (10) divided by (3), namely,

$$(11)\qquad \lambda = \frac{|\hat{\boldsymbol{\Sigma}}_\Omega|^{\frac{1}{2}N}}{|\hat{\boldsymbol{\Sigma}}_\omega|^{\frac{1}{2}N}}.$$

In testing H, one rejects the hypothesis if $\lambda < \lambda_0$, where λ_0 is a suitably chosen number.

A special case of this problem led to Hotelling's T^2-criterion. If $q = q_1 = 1$ $(q_2 = 0)$, $z_\alpha = 1$, $\alpha = 1,\dots,N$, and $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = \boldsymbol{\mu}$, then the T^2-criterion for testing the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is a monotonic function of (11) for $\boldsymbol{\beta}_1^* = \boldsymbol{\mu}_0$.

The hypothesis $\boldsymbol{\mu} = \mathbf{0}$ and the T^2-statistic are invariant with respect to the transformations $X^* = DX$ and $x_\alpha^* = Dx_\alpha$, $\alpha = 1,\dots,N$, for nonsingular D. Similarly, in this problem the null hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$ and the likelihood ratio criterion for testing it are invariant with respect to nonsingular linear transformations.

Theorem 8.3.1. *The likelihood ratio criterion* (11) *for testing the null hypothesis* $\boldsymbol{\beta}_1 = \mathbf{0}$ *is invariant with respect to transformations* $x_\alpha^* = Dx_\alpha$, $\alpha = 1,\dots,N$, *for nonsingular* D.

Proof. The estimators in terms of x_α^* are

$$(12)\qquad \hat{\boldsymbol{\beta}}^* = DCA^{-1} = D\hat{\boldsymbol{\beta}},$$

$$(13)\qquad \hat{\boldsymbol{\Sigma}}_\Omega^* = \frac{1}{N}\sum_{\alpha=1}^N\left(Dx_\alpha - D\hat{\boldsymbol{\beta}}z_\alpha\right)\left(Dx_\alpha - D\hat{\boldsymbol{\beta}}z_\alpha\right)' = D\hat{\boldsymbol{\Sigma}}_\Omega D',$$

$$(14)\qquad \hat{\boldsymbol{\beta}}_{2\omega}^* = DC_2A_{22}^{-1} = D\hat{\boldsymbol{\beta}}_{2\omega},$$

$$(15)\qquad \hat{\boldsymbol{\Sigma}}_\omega^* = \frac{1}{N}\sum_{\alpha=1}^N\left(Dx_\alpha - D\hat{\boldsymbol{\beta}}_{2\omega}z_\alpha^{(2)}\right)\left(Dx_\alpha - D\hat{\boldsymbol{\beta}}_{2\omega}z_\alpha^{(2)}\right)' = D\hat{\boldsymbol{\Sigma}}_\omega D'. \qquad \blacksquare$$

8.3.2. Geometric Interpretation

An insight into the algebra developed here can be given in terms of a geometric interpretation. It will be convenient to use the following lemma:

Lemma 8.3.1.

$$(16)\qquad \hat{\boldsymbol{\beta}}_{2\omega} - \hat{\boldsymbol{\beta}}_{2\Omega} = \left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)A_{12}A_{22}^{-1}.$$

Proof. The normal equation $\hat{\boldsymbol{\beta}}_\Omega A = C$ is written in partitioned form

$$(17)\qquad \left(\hat{\boldsymbol{\beta}}_{1\Omega} A_{11} + \hat{\boldsymbol{\beta}}_{2\Omega} A_{21}, \hat{\boldsymbol{\beta}}_{1\Omega} A_{12} + \hat{\boldsymbol{\beta}}_{2\Omega} A_{22}\right) = (C_1, C_2).$$

Thus $\hat{\boldsymbol{\beta}}_{2\Omega} = C_2 A_{22}^{-1} - \hat{\boldsymbol{\beta}}_{1\Omega} A_{12} A_{22}^{-1}$. The lemma follows by comparison with (6). ■

We can now write

$$\begin{aligned}(18)\qquad X - \boldsymbol{\beta} Z &= \left(X - \hat{\boldsymbol{\beta}}_\Omega Z\right) + \left(\hat{\boldsymbol{\beta}}_{2\Omega} - \boldsymbol{\beta}_2\right) Z_2 + \left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right) Z_1 \\ &= \left(X - \hat{\boldsymbol{\beta}}_\Omega Z\right) + \left(\hat{\boldsymbol{\beta}}_{2\omega} - \boldsymbol{\beta}_2\right) Z_2 \\ &\quad - \left(\hat{\boldsymbol{\beta}}_{2\omega} - \hat{\boldsymbol{\beta}}_{2\Omega}\right) Z_2 + \left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right) Z_1 \\ &= \left(X - \hat{\boldsymbol{\beta}}_\Omega Z\right) + \left(\hat{\boldsymbol{\beta}}_{2\omega} - \boldsymbol{\beta}_2\right) Z_2 \\ &\quad + \left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)\left(Z_1 - A_{12} A_{22}^{-1} Z_2\right)\end{aligned}$$

as an identity; here $X = (x_1, \dots, x_N)$, $Z_1 = (z_1^{(1)}, \dots, z_N^{(1)})$, and $Z_2 = (z_1^{(2)}, \dots, z_N^{(2)})$. The rows of $Z = (Z_1', Z_2')'$ span a q-dimensional subspace in N-space. Each row of $\boldsymbol{\beta} Z$ is a vector in the q-space, and hence each row of $X - \boldsymbol{\beta} Z$ is a vector from a vector in the q-space to the corresponding row vector of X. Each row vector of $X - \boldsymbol{\beta} Z$ is expressed above as the sum of three row vectors. The first matrix on the right of (18) has as its ith row a vector orthogonal to the q-space and leading to the ith row vector of X (as shown in the preceding section). The row vectors of $(\hat{\boldsymbol{\beta}}_{2\omega} - \boldsymbol{\beta}_2) Z_2$ are vectors in the q_2-space spanned by the rows of Z_2 (since they are linear combinations of the rows of Z_2). The row vectors of $(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)(Z_1 - A_{12} A_{22}^{-1} Z_2)$ are vectors in the q_1-space of $Z_1 - A_{12} A_{22}^{-1} Z_2$, and this space is in the q-space of Z, but orthogonal to the q_2-space of Z_2 [since $(Z_1 - A_{12} A_{22}^{-1} Z_2) Z_2' = \mathbf{0}$]. Thus each row of $X - \boldsymbol{\beta} Z$ is indicated in Figure 8.1 as the sum of three orthogonal vectors: one vector is in the space orthogonal to Z, one is in the space of Z_2, and one is in the subspace of Z that is orthogonal to Z_2.

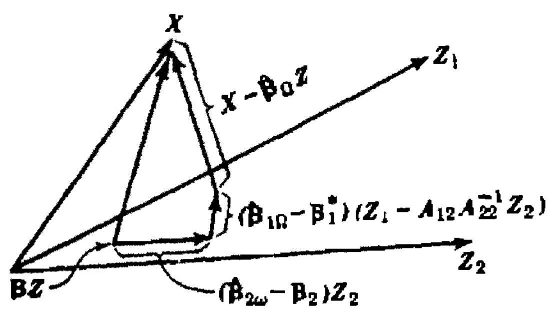

Figure 8.1

From the orthogonality relations we have

$$(19)\quad \begin{aligned}(X-\boldsymbol{\beta}Z)(X-\boldsymbol{\beta}Z)' &= (X-\hat{\boldsymbol{\beta}}_\Omega Z)(X-\hat{\boldsymbol{\beta}}_\Omega Z)' + (\hat{\boldsymbol{\beta}}_{2\omega}-\boldsymbol{\beta}_2)Z_2Z_2'(\hat{\boldsymbol{\beta}}_{2\omega}-\boldsymbol{\beta}_2)' \\ &\quad + (\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)(Z_1-A_{12}A_{22}^{-1}Z_2)(Z_1-A_{12}A_{22}^{-1}Z_2)'(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)' \\ &= N\hat{\boldsymbol{\Sigma}}_\Omega + (\hat{\boldsymbol{\beta}}_{2\omega}-\boldsymbol{\beta}_2)A_{22}(\hat{\boldsymbol{\beta}}_{2\omega}-\boldsymbol{\beta}_2)' \\ &\quad + (\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)(A_{11}-A_{12}A_{22}^{-1}A_{21})(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)'.\end{aligned}$$

If we subtract $(\hat{\boldsymbol{\beta}}_{2\omega}-\boldsymbol{\beta}_2)Z_2$ from both sides of (18), we have

$$(20)\quad X-\boldsymbol{\beta}_1^*Z_1-\hat{\boldsymbol{\beta}}_{2\omega}Z_2 = (X-\hat{\boldsymbol{\beta}}_\Omega Z) + (\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)(Z_1-A_{12}A_{22}^{-1}Z_2).$$

From this we obtain

$$(21)\quad \begin{aligned}N\hat{\boldsymbol{\Sigma}}_\omega &= (X-\boldsymbol{\beta}_1^*Z_1-\hat{\boldsymbol{\beta}}_{2\omega}Z_2)(X-\boldsymbol{\beta}_1^*Z_1-\hat{\boldsymbol{\beta}}_{2\omega}Z_2)' \\ &= (X-\hat{\boldsymbol{\beta}}_\Omega Z)(X-\hat{\boldsymbol{\beta}}_\Omega Z)' \\ &\quad + (\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)(Z_1-A_{12}A_{22}^{-1}Z_2)(Z_1-A_{12}A_{22}^{-1}Z_2)'(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)' \\ &= N\hat{\boldsymbol{\Sigma}}_\Omega + (\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)(A_{11}-A_{12}A_{22}^{-1}A_{21})(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)'.\end{aligned}$$

The determinant $|\hat{\boldsymbol{\Sigma}}_\Omega| = (1/N^p)|(X-\hat{\boldsymbol{\beta}}_\Omega Z)(X-\hat{\boldsymbol{\beta}}_\Omega Z)'|$ is proportional to the volume squared of the parallelotope spanned by the row vectors of $X-\hat{\boldsymbol{\beta}}_\Omega Z$ (translated to the origin). The determinant $|\hat{\boldsymbol{\Sigma}}_\omega| = (1/N^p)|(X-\boldsymbol{\beta}_1^*Z_1-\hat{\boldsymbol{\beta}}_{2\omega}Z_2)(X-\boldsymbol{\beta}_1^*Z_1-\hat{\boldsymbol{\beta}}_{2\omega}Z_2)'|$ is proportional to the volume squared of the parallelotope spanned by the row vectors of $X-\boldsymbol{\beta}_1^*Z_1-\hat{\boldsymbol{\beta}}_{2\omega}Z_2$ (translated to the origin); each of these vectors is the part of the vector of $X-\boldsymbol{\beta}_1^*Z_1$ that is orthogonal to Z_2. Thus the test based on the likelihood ratio criterion depends on the ratio of volumes of parallelotopes. One parallelotope involves vectors orthogonal to Z, and the other involves vectors orthogonal to Z_2.

From (15) we see that the density of $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ can be written as

$$(22)\quad \frac{1}{(2\pi)^{\frac{1}{2}pN}|\boldsymbol{\Sigma}|^{\frac{1}{2}N}}\exp\Big(-\tfrac{1}{2}\operatorname{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big[N\hat{\boldsymbol{\Sigma}} + (\hat{\boldsymbol{\beta}}_{2\omega}-\boldsymbol{\beta}_2)A_{22}(\hat{\boldsymbol{\beta}}_{2\omega}-\boldsymbol{\beta}_2)' + (\hat{\boldsymbol{\beta}}_{1\omega}-\boldsymbol{\beta}_1^*)(A_{11}-A_{12}A_{22}^{-1}A_{21})(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1^*)'\Big]\Big\}\Big).$$

Thus, $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\beta}}_{1\Omega}$, and $\hat{\boldsymbol{\beta}}_{2\omega}$ form a sufficient set of statistics for $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$.

Wilks (1932) first gave the likelihood ratio criterion for testing the equality of mean vectors from several populations (Section 8.8). Wilks (1934) and Bartlett (1934) extended its use to regression coefficients.

8.3.3. The Canonical Form

In studying the distributions of criteria it will be convenient to put the distribution of the observations in canonical form. This amounts to picking a coordinate system in the N-dimensional space so that the first q_1 coordinate axes are in the space of Z that is orthogonal to Z_2, the next q_2 coordinate axes are in the space of Z_2, and the last n $(=N-q)$ coordinate axes are orthogonal to the Z-space.

Let P_2 be a $q_2 \times q_2$ matrix such that

$$I = P_2 A_{22} P_2' = (P_2 Z_2)(P_2 Z_2)', \tag{23}$$

and let P_1 be a $q_1 \times q_1$ matrix such that $(A_{11\cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21})$

$$I = P_1 A_{11\cdot 2} P_1' = \left[P_1(Z_1 - A_{12}A_{22}^{-1}Z_2)\right]\left[P_1(Z_1 - A_{12}A_{22}^{-1}Z_2)\right]'. \tag{24}$$

Then define the $N \times N$ orthogonal matrix Q as

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \end{pmatrix} = \begin{pmatrix} P_1(Z_1 - A_{12}A_{22}^{-1}Z_2) \\ P_2 Z_2 \\ Q_3 \end{pmatrix}, \tag{25}$$

where Q_3 is any $n \times N$ matrix making Q orthogonal. Then the columns of

$$W = \begin{pmatrix} W_1 & W_2 & W_3 \end{pmatrix} = XQ' = X\begin{pmatrix} Q_1' & Q_2' & Q_3' \end{pmatrix} \tag{26}$$

are independently normally distributed with covariance matrix $\boldsymbol{\Sigma}$ (Theorem 3.3.1). Then

$$\begin{aligned} \mathscr{E}W_1 = \mathscr{E}XQ_1' &= (\boldsymbol{\beta}_1 Z_1 + \boldsymbol{\beta}_2 Z_2)(Z_1 - A_{12}A_{22}^{-1}Z_2)'P_1' \\ &= \boldsymbol{\beta}_1 A_{11\cdot 2} P_1' = \boldsymbol{\beta}_1 P_1^{-1}, \end{aligned} \tag{27}$$

$$\begin{aligned} \mathscr{E}W_2 = \mathscr{E}XQ_2' &= (\boldsymbol{\beta}_1 Z_1 + \boldsymbol{\beta}_2 Z_2)Z_2'P_2' \\ &= (\boldsymbol{\beta}_1 A_{12} + \boldsymbol{\beta}_2 A_{22})P_2', \end{aligned} \tag{28}$$

$$\mathscr{E}W_3 = \mathscr{E}XQ_3' = \boldsymbol{\beta} Z Q_3' = \mathbf{0}. \tag{29}$$

Let

(30) $\Gamma_1 = (\gamma_1, \dots, \gamma_{q_1}) = \beta_1 A_{11\cdot 2} P_1' = \beta_1 P_1^{-1}$,

(31) $\Gamma_2 = (\gamma_{q_1+1}, \dots, \gamma_q) = (\beta_1 A_{12} + \beta_2 A_{22}) P_2'$,

(32) $W = \begin{pmatrix} W_1 & W_2 & W_3 \end{pmatrix} = (w_1, \dots, w_{q_1}, w_{q_1+1}, \dots, w_q, w_{q+1}, \dots, w_N)$.

Then $w_1, \dots, w_N$ are independently normally distributed with covariance matrix Σ and $\mathscr{E} w_\alpha = \gamma_\alpha$, $\alpha = 1, \dots, q$, and $\mathscr{E} w_\alpha = \mathbf{0}$, $\alpha = q+1, \dots, N$.

The hypothesis $\beta_1 = \beta_1^*$ can be transformed to $\beta_1 = \mathbf{0}$ by subtraction, that is, by letting $x_\alpha - \beta_1^* z_\alpha^{(1)} = y_\alpha$, as in Section 8.3.1. In canonical form then, the hypothesis is $\Gamma_1 = \mathbf{0}$. We can study problems in the canonical form, if we wish, and transform solutions back to terms of X and Z.

In (17), which is the partitioned form of $\hat{\beta}_\Omega A = C$, eliminate $\hat{\beta}_{2\Omega}$ to obtain

$$\begin{aligned} \hat{\beta}_{1\Omega}\left(A_{11} - A_{12} A_{22}^{-1} A_{21}\right) &= C_1 - C_2 A_{22}^{-1} A_{21} \\ &= X\left(Z_1' - Z_2' A_{22}^{-1} A_{21}\right) \\ &= W_1 P_1'^{-1}; \end{aligned} \tag{33}$$

that is, $W_1 = \hat{\beta}_{1\Omega} A_{11\cdot 2} P_1' = \hat{\beta}_{1\Omega} P_1^{-1}$ and $\Gamma_1 = \beta_1 P_1^{-1}$. Similarly, from (6) we obtain

$$\hat{\beta}_{2\omega} A_{22} + \beta_1^* A_{12} = C_2 = X Z_2' = W_2 P_2'^{-1}; \tag{34}$$

that is, $W_2 = (\hat{\beta}_{2\omega} A_{22} + \beta_1^* A_{12}) P_2' = \beta_{2\omega} P_2^{-1} + \beta_1^* A_{12} P_2^{-1}$ and $\Gamma_2 = \beta_2 P_2^{-1} + \beta_1 A_{12} P_2^{-1}$.

8.4. THE DISTRIBUTION OF THE LIKELIHOOD RATIO CRITERION WHEN THE HYPOTHESIS IS TRUE

8.4.1. Characterization of the Distribution

The likelihood ratio criterion is the $\frac{1}{2}N$th power of

$$U = \lambda^{2/N} = \frac{|\hat{\Sigma}_\Omega|}{|\hat{\Sigma}_\omega|} = \frac{|N\hat{\Sigma}_\Omega|}{\left|N\hat{\Sigma}_\Omega + (\hat{\beta}_{1\Omega} - \beta_1^*) A_{11\cdot 2} (\hat{\beta}_{1\Omega} - \beta_1^*)'\right|}, \tag{1}$$

where $A_{11\cdot 2} = A_{11} - A_{12} A_{22}^{-1} A_{21}$. We shall study the distribution and the moments of U when $\beta_1 = \beta_1^*$. It has been shown in Section 8.2 that $N\hat{\Sigma}_\Omega$ is distributed according to $W(\Sigma, n)$, where $n = N - q$, and the elements of $\hat{\beta}_\Omega - \beta$ have a joint normal distribution independent of $N\hat{\Sigma}_\Omega$.

From (33) of Section 8.3, we have

$$(2) \qquad (\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)A_{11\cdot 2}(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)' = (W_1 - \boldsymbol{\Gamma}_1)P_1 A_{11\cdot 2} P_1'(W_1 - \boldsymbol{\Gamma}_1)' = (W_1 - \boldsymbol{\Gamma}_1)(W_1 - \boldsymbol{\Gamma}_1)',$$

by (24) of Section 8.3; the columns of $W_1 - \boldsymbol{\Gamma}_1$ are independently distributed, each according to $N(\mathbf{0}, \boldsymbol{\Sigma})$.

Lemma 8.4.1. $(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)A_{11\cdot 2}(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)'$ *is distributed according to* $W(\boldsymbol{\Sigma}, q_1)$.

Lemma 8.4.2. *The criterion U has the distribution of*

$$(3) \qquad U = \frac{|G|}{|G+H|},$$

where G is distributed according to $W(\boldsymbol{\Sigma}, n)$, H is distributed according to $W(\boldsymbol{\Sigma}, m)$, where $m = q_1$, and G and H are independent.

Let

$$(4) \qquad G = N\hat{\boldsymbol{\Sigma}}_\Omega = XX' - XZ'(ZZ')^{-1}ZX',$$

$$(5) \qquad G + H = N\hat{\boldsymbol{\Sigma}}_\Omega + (\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)A_{11\cdot 2}(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)' = N\hat{\boldsymbol{\Sigma}}_\omega = YY' - YZ_2'(Z_2Z_2')^{-1}Z_2Y',$$

where $Y = X - \boldsymbol{\beta}_1^* Z_1 = X - (\boldsymbol{\beta}_1^* \;\; \mathbf{0})Z$. Then

$$(6) \qquad G = YY' - YZ'(ZZ')^{-1}ZY'.$$

We shall denote this criterion as $U_{p,m,n}$, where p is the dimensionality, $m = q_1$ is the number of columns of $\boldsymbol{\beta}_1$, and $n = N - q$ is the number of degrees of freedom of G.

We now proceed to characterize the distribution of U as the product of beta variables (Section 5.2). Write the criterion U as

$$(7) \qquad U = V_1 V_2 \cdots V_p,$$

where $V_1 = g_{11}/(g_{11} + h_{11})$,

$$(8) \qquad V_i = \frac{|G_i|}{|G_{i-1}|} \bigg/ \frac{|G_i + H_i|}{|G_{i-1} + H_{i-1}|}, \qquad i = 2, \ldots, p,$$

and G_i and H_i are the submatrices of G and H, respectively, of the first i rows and columns. Correspondingly, let $y_\alpha^{(i)}$ consist of the first i components of $y_\alpha = x_\alpha - \boldsymbol{\beta}_1^* z_\alpha^{(1)}$, $\alpha = 1, \dots, N$. We shall show that V_i is the length squared of the vector from $y_i^* = (y_{i1}, \dots, y_{iN})$ to its projection on Z and $Y_{i-1} = (y_1^{(i-1)}, \dots, y_N^{(i-1)})$ divided by the length squared of the vector from y_i^* to its projection on Z_2 and Y_{i-1}.

Lemma 8.4.3. *Let y be an N-component row vector and U an $r \times N$ matrix. Then the sum of squares of the residuals of y from its regression on U is*

$$\frac{\begin{vmatrix} yy' & yU' \\ Uy' & UU' \end{vmatrix}}{|UU'|}. \tag{9}$$

Proof. By Corollary A.3.1 of the Appendix, (9) is $yy' - yU'(UU')^{-1}Uy'$, which is the sum of squares of residuals as indicated in (13) of Section 8.2. ∎

Lemma 8.4.4. *V_i defined by (8) is the ratio of the sum of squares of the residuals of $y_{i1}, \dots, y_{iN}$ from their regression on $y_1^{(i-1)}, \dots, y_N^{(i-1)}$ and Z to the sum of squares of residuals of $y_{i1}, \dots, y_{iN}$ from their regression on $y_1^{(i-1)}, \dots, y_N^{(i-1)}$ and Z_2.*

Proof. The numerator of V_i can be written [from (13) of Section 8.2]

$$\frac{|G_i|}{|G_{i-1}|} = \frac{|Y_iY_i' - Y_iZ'(ZZ')^{-1}ZY_i'|}{|Y_{i-1}Y_{i-1}' - Y_{i-1}Z'(ZZ')^{-1}ZY_{i-1}'|} \tag{10}$$

$$= \frac{\begin{vmatrix} Y_iY_i' & Y_iZ' \\ ZY_i' & ZZ' \end{vmatrix} \Big/ |ZZ'|}{\begin{vmatrix} Y_{i-1}Y_{i-1}' & Y_{i-1}Z' \\ ZY_{i-1}' & ZZ' \end{vmatrix} \Big/ |ZZ'|}$$

$$= \frac{\begin{vmatrix} Y_{i-1}Y_{i-1}' & Y_iy_i^{*\prime} & Y_{i-1}Z' \\ y_i^*Y_{i-1}' & y_i^*y_i^{*\prime} & y_i^*Z' \\ ZY_{i-1}' & Zy_i^{*\prime} & ZZ' \end{vmatrix}}{\begin{vmatrix} Y_{i-1}Y_{i-1}' & Y_{i-1}Z' \\ ZY_{i-1}' & ZZ' \end{vmatrix}}$$

$$= \frac{\begin{vmatrix} y_i^* y_i^{*\prime} & y_i^* \begin{bmatrix} Y_{i-1}' & Z' \end{bmatrix} \\ \begin{bmatrix} Y_{i-1} \\ Z \end{bmatrix} y_i^{*\prime} & \begin{bmatrix} Y_{i-1} \\ Z \end{bmatrix} \begin{bmatrix} Y_{i-1}' & Z' \end{bmatrix} \end{vmatrix}}{\left| \begin{bmatrix} Y_{i-1} \\ Z \end{bmatrix} \begin{bmatrix} Y_{i-1}' & Z' \end{bmatrix} \right|}$$

$$= y_i^* y_i^{*\prime} - y_i^* \begin{pmatrix} Y_{i-1}' & Z' \end{pmatrix} \begin{bmatrix} Y_{i-1} Y_{i-1}' & Y_{i-1} Z' \\ Z Y_{i-1}' & Z Z' \end{bmatrix}^{-1} \begin{pmatrix} Y_{i-1} \\ Z \end{pmatrix} y_i^{*\prime}$$

by Corollary A.3.1. Application of Lemma 8.4.3 shows that the right-hand side of (10) is the sum of squares of the residuals of y_i^* on Y_{i-1} and Z. The denominator is evaluated similarly with Z replaced by Z_2. ■

The ratio V_i is the $2/N$th power of the likelihood ratio criterion for testing the hypothesis that the regression of $y_i^* = x_i^* - \boldsymbol{\beta}_{i1}^* Z_1$ on Z_1 is $\mathbf{0}$ (in the presence of regression on Y_{i-1} and Z_2); here $\boldsymbol{\beta}_{i1}^*$ is the ith row of $\boldsymbol{\beta}_1^*$. For $i=1$, g_{11} is the sum of squares of the residuals of $y_1^* = (y_{11}, \dots, y_{1N})$ from its regression on Z, and $g_{11} + h_{11}$ is the sum of squares of the residuals from Z_2. The ratio $V_1 = g_{11}/(g_{11} + h_{11})$, which is approximate to test the hypothesis that regression of y_1^* on Z_1 is $\mathbf{0}$, is distributed as $\chi_n^2/(\chi_n^2 + \chi_m^2)$ (by Lemma 8.4.2) and has the beta distribution $\beta(v; \frac{1}{2}n, \frac{1}{2}m)$. (See Section 5.2, for example.) Thus V_i has the beta density

$$\beta\left[v; \tfrac{1}{2}(n+1-i), \tfrac{1}{2}m\right] = \frac{\Gamma\left[\frac{1}{2}(n+m+1-i)\right]}{\Gamma\left[\frac{1}{2}(n+1-i)\right]\Gamma\left(\frac{1}{2}m\right)} v^{\frac{1}{2}(n+1-i)-1}(1-v)^{\frac{1}{2}m-1}, \tag{11}$$

for $0 \le v \le 1$ and 0 for v outside this interval. Since this distribution does not depend on Y_{i-1}, we see that the ratio V_i is independent of Y_{i-1}, and hence independent of $|V_1, \dots, V_{i-1}$. Then $V_1, \dots, V_p$ are independent.

Theorem 8.4.1. *The distribution of U defined by* (3) *is the distribution of the product $\prod_{i=1}^p V_i$, where $V_1, \dots, V_p$ are independent and V_i has the density* (11).

The cdf of U can be found by integrating the joint density of $V_1, \dots, V_p$ over the range

$$\prod_{i=1}^p V_i \le u. \tag{12}$$

We shall now show that for given $N-q_2$ the indices p and q_1 can be interchanged; that is, the distributions of $U_{p,q_1,N-q_2-q_1}=U_{p,m,n}$ and of $U_{q_1,p,N-q_2-p}=U_{m,p,n+m-p}$ are the same. The joint density of G and W_1 defined in Section 8.2 when $\Sigma=I$ and $\boldsymbol{\beta}_1=\mathbf{0}$ is

$$(13)\qquad \frac{|G|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}\operatorname{tr}G-\frac{1}{2}\operatorname{tr}W_1W_1'}}{2^{\frac{1}{2}np}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right](2\pi)^{\frac{1}{2}mp}}.$$

Let $G+W_1W_1'=J=CC'$ and let $W_1=CU$. Then

$$(14)\qquad U_{p,m,n}=\frac{|G|}{|G+W_1W_1'|}=\frac{|CC'-CUU'C'|}{|CC'|}=|I_p-UU'|$$

$$=\begin{vmatrix} I_p & U \\ U' & I_m \end{vmatrix}=\begin{vmatrix} I_m & U' \\ U & I_p \end{vmatrix}=|I_m-U'U|;$$

the fourth and sixth equalities follow from Theorem A.3.2 of the Appendix, and the fifth from permutation of rows and columns. Since the Jacobian of $W_1=CU$ is mod$|C|^m=|J|^{\frac{1}{2}m}$, the joint density of J and U is

$$(15)\qquad \frac{|J|^{\frac{1}{2}(n+m-p-1)}e^{-\frac{1}{2}\operatorname{tr}J}}{2^{\frac{1}{2}(n+m)p}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+m+1-i)\right]}$$

$$\cdot\prod_{i=1}^{p}\left\{\frac{\Gamma\left[\frac{1}{2}(n+m+1-i)\right]}{\Gamma\left[\frac{1}{2}(n+1-i)\right]}\right\}\frac{|I_p-UU'|^{\frac{1}{2}(n-p-1)}}{\pi^{\frac{1}{2}mp}}$$

for J and I_p-UU' positive definite, and 0 otherwise. Thus J and U are independently distributed; the density of J is the first term in (15), namely, $w(J|I_p,n+m)$, and the density of U is the second term, namely, of the form

$$(16)\qquad K|I_p-UU'|^{\frac{1}{2}(n-p-1)}$$

for I_p-UU' positive definite, and 0 otherwise. Let $U_*=U'$, $p^*=m$, $m^*=p$, and $n^*=n+m-p$. Then the density of U_* is

$$(17)\qquad K|I_p-U_*'U_*|^{\frac{1}{2}(n-p-1)}$$

for $I_p-U_*'U_*$ positive definite, and 0 otherwise. By (14), $|I_p-U_*'U_*|=|I_m-U_*U_*'|$, and hence the density of U_* is

$$(18)\qquad K|I_{p^*}-U_*U_*'|^{\frac{1}{2}(n^*-p^*-1)},$$

which is of the form of (16) with p replaced by $p^* = m$, m replaced by $m^* = p$, and $n - p - 1$ replaced by $n^* - p^* - 1 = n - p - 1$. Finally we note that $U_{p,m,n}$ given by (14) is $|I_m - U_* U_*'| = U_{m,p,n+m-p}$.

Theorem 8.4.2. *When the hypothesis is true, the distribution of $U_{p,q_1,N-q_1-q_2}$ is the same as that of $U_{q_1,p,N-p-q_2}$ (i.e., that of $U_{p,m,n}$ is that of $U_{m,p,n+m-p}$).*

8.4.2. Moments

Since (11) is a density and hence integrates to 1, by change of notation

$$\int_0^1 v^{a-1}(1-v)^{b-1}\,dv = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{19}$$

From this fact we see that the hth moment of V_i is

$$\begin{aligned} \mathscr{E}V_i^h &= \int_0^1 \frac{\Gamma[\frac{1}{2}(n+m+1-i)]}{\Gamma[\frac{1}{2}(n+1-i)]\Gamma(\frac{1}{2}m)} v^{\frac{1}{2}(n+1-i)+h-1}(1-v)^{\frac{1}{2}m-1}\,dv \\ &= \frac{\Gamma[\frac{1}{2}(n+1-i)+h]\Gamma[\frac{1}{2}(n+m+1-i)]}{\Gamma[\frac{1}{2}(n+1-i)]\Gamma[\frac{1}{2}(n+m+1-i)+h]}. \end{aligned} \tag{20}$$

Since $V_1, \ldots, V_p$ are independent, $\mathscr{E}U^h = \mathscr{E}\prod_{i=1}^p V_i^h = \prod_{i=1}^p \mathscr{E}V_i^h$. We obtain the following theorem:

Theorem 8.4.3. *The hth moment of U [if $h > -\frac{1}{2}(n+1-p)$] is*

$$\begin{aligned} \mathscr{E}U^h &= \prod_{i=1}^p \frac{\Gamma[\frac{1}{2}(n+1-i)+h]\Gamma[\frac{1}{2}(n+m+1-i)]}{\Gamma[\frac{1}{2}(n+1-i)]\Gamma[\frac{1}{2}(n+m+1-i)+h]} \\ &= \prod_{i=1}^p \frac{\Gamma[\frac{1}{2}(N-q_1-q_2+1-i)+h]\Gamma[\frac{1}{2}(N-q_2+1-i)]}{\Gamma[\frac{1}{2}(N-q_1-q_2+1-i)]\Gamma[\frac{1}{2}(N-q_2+1-i)+h]}. \end{aligned} \tag{21}$$

In the first expression p can be replaced by m, m by p, and n by $n+m-p$.

Suppose p is even, that is, $p = 2r$. We use the duplication formula

$$\Gamma(\alpha+\tfrac{1}{2})\Gamma(\alpha+1) = \frac{\sqrt{\pi}\,\Gamma(2\alpha+1)}{2^{2\alpha}}. \tag{22}$$

Then the hth moment of $U_{2r,m,n}$ is

(23)

$$\begin{aligned}\mathscr{E}U_{2r,m,n}^h = \prod_{j=1}^r \Bigg\{ & \frac{\Gamma[\frac{1}{2}(m+n+2)-j]}{\Gamma[\frac{1}{2}(m+n+2)-j+h]} \frac{\Gamma[\frac{1}{2}(m+n+1)-j]}{\Gamma[\frac{1}{2}(m+n+1)-j+h]} \\ & \cdot \frac{\Gamma[\frac{1}{2}(n+2)-j+h]\Gamma[\frac{1}{2}(n+1)-j+h]}{\Gamma[\frac{1}{2}(n+2)-j]\Gamma[\frac{1}{2}(n+1)-j]} \Bigg\} \\ = \prod_{j=1}^r & \left\{ \frac{\Gamma(m+n+1-2j)\Gamma(n+1-2j+2h)}{\Gamma(m+n-1-2j+2h)\Gamma(n+1-2j)} \right\}.\end{aligned}$$

It is clear from the definition of the beta function that (23) is

$$\begin{aligned}(24)\qquad & \prod_{j=1}^r \left\{ \int_0^1 \frac{\Gamma(m+n+1-2j)}{\Gamma(n+1-2j)\Gamma(m)} y^{(n+1-2j)+2h-1}(1-y)^{m-1}\,dy \right\} \\ & = \prod_{j=1}^r \mathscr{E}Y_j^{2h} = \mathscr{E}\left(\prod_{j=1}^r Y_j^2 \right)^h,\end{aligned}$$

where the Y_j are independent and Y_j has density $\beta(y; n+1-2j, m)$.

Suppose p is odd; that is, $p = 2s+1$. Then

$$(25)\qquad \mathscr{E}U_{2s+1,m,n}^h = \mathscr{E}\left(\prod_{i=1}^s Z_i^2 Z_{s+1} \right)^h,$$

where the Z_i are independent and Z_i has density $\beta(z; n+1-2i, m)$ for $i = 1, \ldots, s$ and Z_{s+1} is distributed with density $\beta[z; (n+1-p)/2, m/2]$.

Theorem 8.4.4. *$U_{2r,m,n}$ is distributed as $\prod_{i=1}^r Y_i^2$, where $Y_1, \ldots, Y_r$ are independent and Y_i has density $\beta(y; n+1-2i, m)$; $U_{2s+1,m,n}$ is distributed as $\prod_{i=1}^s Z_i^2 Z_{s+1}$, where the Z_i, $i = 1, \ldots, s$, are independent and Z_i has density $\beta(z; n+1-2i, m)$, and Z_{s+1} is independently distributed with density $\beta[z; \frac{1}{2}(n+1-p), \frac{1}{2}m]$.*

8.4.3. Some Special Distributions

$p = 1$

From the preceding characterization we see that the density of $U_{1,m,n}$ is

$$(26)\qquad \frac{\Gamma[\frac{1}{2}(n+m)]}{\Gamma(\frac{1}{2}n)\Gamma(\frac{1}{2}m)} u^{\frac{1}{2}n-1}(1-u)^{\frac{1}{2}m-1}.$$

Another way of writing $U_{1,m,n}$ is

$$U_{1,m,n} = \frac{1}{1+\sum_{i=1}^m Y_i^2/g_{11}} = \frac{1}{1+(m/n)F_{m,n}}, \tag{27}$$

where g_{11} is the one element of $G = N\hat{\Sigma}_\Omega$ and $F_{m,n}$ is an F-statistic. Thus

$$\frac{1-U_{1,m,n}}{U_{1,m,n}} \cdot \frac{n}{m} = F_{m,n}. \tag{28}$$

Theorem 8.4.5. *The distribution of* $[(1-U_{1,m,n})/U_{1,m,n}]\cdot n/m$ *is the F-distribution with m and n degrees of freedom; the distribution of* $[(1-U_{p,1,n}/U_{p,1,n}]\cdot(n+1-p)/p$ *is the F-distribution with p and* $n+1-p$ *degrees of freedom.*

$p = 2$

From Theorem 8.4.4, we see that the density of $\sqrt{U_{2,m,n}}$ is

$$\frac{\Gamma(n+m-1)}{\Gamma(n-1)\Gamma(m)} x^{n-2}(1-x)^{m-1}, \tag{29}$$

and thus the density of $U_{2,m,n}$ is

$$\frac{\Gamma(n+m-1)}{2\Gamma(n-1)\Gamma(m)} u^{\frac{1}{2}(n-3)}(1-\sqrt{u})^{m-1}. \tag{30}$$

From (29) it follows that

$$\frac{1-\sqrt{U_{2,m,n}}}{\sqrt{U_{2,m,n}}} \cdot \frac{n-1}{m} = F_{2m,2(n-1)}. \tag{31}$$

Theorem 8.4.6. *The distribution of* $[(1-\sqrt{U_{2,m,n}})/\sqrt{U_{2,m,n}}]\cdot(n-1)/m$ *is the F-distribution with* $2m$ *and* $2(n-1)$ *degrees of freedom; the distribution of* $[(1-\sqrt{U_{p,2,n}})/\sqrt{U_{p,2,n}}]\cdot(n+1-p)/p$ *is the F-distribution with* $2p$ *and* $2(n+1-p)$ *degrees of freedom.*

p Even

Wald and Brookner (1941) gave a method for finding the distribution of $U_{p,m,n}$ for p or m even. We shall present the method of Schatzoff (1966a). It will be convenient first to consider $U_{p,m,n}$ for $m = 2r$. We can write the event $\prod_{i=1}^p V_i \le u$ as

$$Y_1 + \cdots + Y_p \ge -\log u, \tag{32}$$

where $Y_1, \ldots, Y_p$ are independent and $Y_i = -\log V_i$ has the density

$$(33)\quad K_i e^{-\frac{1}{2}(n+1-i)y}(1-e^{-y})^{r-1} = K_i \sum_{j=0}^{r-1}(-1)^j \binom{r-1}{j} e^{-[\frac{1}{2}(n+1-i)+j]y}$$

for $0 \le y < \infty$ and 0 otherwise, and

$$(34)\qquad K_i = \frac{\Gamma[\frac{1}{2}(n+1-i)+r]}{\Gamma[\frac{1}{2}(n+1-i)]\Gamma(r)} = \frac{1}{(r-1)!}\prod_{j=0}^{r-1}\frac{n+1-i+2j}{2}.$$

The joint density of $Y_1, \ldots, Y_p$ is then a linear combination of terms $\exp[-\sum_{i=1}^p a_i y_i]$. The density of $W_j = \sum_{i=1}^j Y_i$ can be obtained inductively from the density of $W_{j-1} = \sum_{i=1}^{j-1} Y_i$ and Y_j, $j = 2, \ldots, p$, which is a linear combination of terms $w_{j-1}^k e^{cw_{j-1}+a_j y_j}$. The density of W_j consists of linear combinations of

$$(35)\qquad e^{a_j w_j}\int_0^{w_j} w^k e^{(c-a_j)w}\,dw = e^{a_j w_j}\cdot\frac{w_j^{k+1}}{k+1} \qquad \text{if } a_j = c,$$

$$= e^{cw_j}\sum_{h=0}^{k}(-1)^h \frac{k!}{(k-h)!}\frac{w_j^{k-h}}{(c-a_j)^{h+1}}$$

$$+(-1)^{k+1}e^{a_j w_j}\frac{k!}{(c-a_j)^{k+1}} \qquad \text{if } a_j \neq c.$$

The evaluation involves integration by parts.

Theorem 8.4.7. *If p is even or if m is even, the density of $U_{p,m,n}$ can be expressed as a linear combination of terms $(-\log u)^k u^l$, where k is an integer and l is a half integer.*

From (35) we see that the cumulative distribution function of $-\log U$ is a linear combination of terms $w^k e^{-lw}$ and hence the cumulative distribution function of U is a linear combination of terms $(-\log u)^k u^l$. The values of k and l and the coefficients depend on p, m, and n. They can be obtained by inductively carrying out the procedure leading to Theorem 8.4.7. Pillai and Gupta (1969) used Theorem 8.4.3 for obtaining distributions.

An alternative approach is to use Theorem 8.4.4. The complement to the cumulative distribution function $U_{2r,m,n}$ is

(36)

$$\begin{aligned}\Pr\{U_{2r,m,n} \geq u\} &= \Pr\left\{\prod_{i=1}^{r} Y_i > \sqrt{u}\right\} \\ &= \int_{\sqrt{u}}^{1}\int_{\frac{\sqrt{u}}{y_1}}^{1}\cdots\int_{\frac{\sqrt{u}}{\prod_{i=1}^{r-1} y_i}}^{1} \prod_{i=1}^{r} \beta(y_i|n+1-2i, m)\, dy_r \cdots dy_2\, dy_1.\end{aligned}$$

In the density, $(1-y_i)^{m-1}$ can be expanded by the binomial theorem. Then all integrations are expressed as integrations of powers of the variables.

As an example, consider $r=2$. The density of Y_1 and Y_2 is

$$\begin{aligned}(37) \qquad & C y_1^{n-2} y_2^{n-4}(1-y_1)^{m-1}(1-y_2)^{m-1} \\ &= C\sum_{i,j=0}^{m-1} \frac{[(m-1)!]^2(-1)^{i+j}}{(m-i-1)!(m-j-1)!i!j!} y_1^{n-2+i} y_2^{n-4+j},\end{aligned}$$

where

$$(38) \qquad C = \frac{\Gamma(n+m-1)\Gamma(n+m-3)}{\Gamma(n-1)\Gamma(n-3)\Gamma^2(m)}.$$

The complement to the cdf of $U_{4,m,n}$ is

$$\begin{aligned}(39) \quad \Pr\{U_{4,m,n} \geq u\} &= C\sum_{i,j=0}^{m-1} \frac{[(m-1)!]^2(-1)^{i+j}}{(m-i-1)!(m-j-1)!i!j!} \\ &\quad \cdot \int_{\sqrt{u}}^{1}\int_{\frac{\sqrt{u}}{y_1}}^{1} y_1^{n-2+i} y_2^{n-4+j}\, dy_2\, dy_1 \\ &= C\sum_{i,j=0}^{m-1} \frac{[(m-1)!]^2(-1)^{i+j}}{(m-i-1)!(m-j-1)!i!j!(n-3+j)} \\ &\quad \cdot \int_{\sqrt{u}}^{1}\left[y_1^{n-2+i} - u^{\frac{1}{2}(n-3+j)} y_1^{1+i-j}\right] dy_1.\end{aligned}$$

The last step of the integration yields powers of $\sqrt{u}$ and products of powers of $\sqrt{u}$ and $\log u$ (for $1+i-j=-1$).

Particular Values

Wilks (1935) gives explicitly the distributions of U for $p = 1$, $p = 2$, $p = 3$ with $m = 3$; $p = 3$ with $m = 4$; and $p = 4$ with $m = 4$. Wilks's formula for $p = 3$ with $m = 4$ appears to be incorrect; see the first edition of this book. Consul (1966) gives many distributions for special cases. See also Mathai (1971).

8.4.4. The Likelihood Ratio Procedure

Let $u_{p,m,n}(\alpha)$ be the α significance point for $U_{p,m,n}$; that is,

$$\Pr\{U_{p,m,n} \le u_{p,m,n}(\alpha) | H \text{ true}\} = \alpha. \tag{40}$$

It is shown in Section 8.5 that $-[n - \frac{1}{2}(p - m + 1)] \log U_{p,m,n}$ has a limiting χ^2-distribution with pm degrees of freedom. Let $\chi^2_{pm}(\alpha)$ denote the α significance point of χ^2_{pm}, and let

$$C_{p,m,n-p+1}(\alpha) = \frac{-[n - \frac{1}{2}(p - m + 1)] \log u_{p,m,n}(\alpha)}{\chi^2_{pm}(\alpha)}. \tag{41}$$

Table B.1 [from Pearson and Hartley (1972)] gives value of $C_{p,m,M}(\alpha)$ for $\alpha = 0.10$ and 0.05, $p = 1(1)10$, various even values of m, and $M = n - p + 1 = 1(1)10(2)20, 24, 30, 40, 60, 120$.

To test a null hypothesis one computes $U_{p,m,n}$ and rejects the null hypothesis at significance level α if

$$-[n - \tfrac{1}{2}(p - m + 1)] \log U_{p,m,n} > C_{p,m,n-p+1}(\alpha) \chi^2_{pm}(\alpha). \tag{42}$$

Since $C_{p,m,n}(\alpha) > 1$, the hypothesis is accepted if the left-hand side of (42) is less than $\chi^2_{pm}(\alpha)$.

The purpose of tabulating $C_{p,m,M}(\alpha)$ is that linear interpolation is reasonably accurate because the entries decrease monotonically and smoothly to 1 as M increases. Schatzoff (1966a) has recommended interpolation for odd p by using adjacent even values of p and displays some examples. The table also indicates how accurate the χ^2-approximation is. The table has been extended by Pillai and Gupta (1969).

8.4.5. A Step-down Procedure

The criterion U has been expressed in (7) as the product of independent beta variables $V_1, V_2, \dots, V_p$. The ratio V_i is a least squares criterion for testing the null hypothesis that in the regression of $x_i^* - \beta_{i1}^* Z_1$ on $Z = (Z_1' \; Z_2')'$ and

X_{i-1} the coefficient of Z_1 is $\mathbf{0}$. The null hypothesis that the regression of X on Z_1 is $\boldsymbol{\beta}_1^*$, which is equivalent to the hypothesis that the regression of $X - \boldsymbol{\beta}_1^* Z_1$ on Z_1 is $\mathbf{0}$, is composed of the hypotheses that the regression of $x_i^* - \boldsymbol{\beta}_{i1}^* Z_1$ on Z_1 is $\mathbf{0}$, $i = 1, \dots, p$. Hence the null hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ can be tested by use of $V_1, \dots, V_p$.

Since V_i has the beta density (11) under the hypothesis $\boldsymbol{\beta}_{i1} = \boldsymbol{\beta}_{i1}^*$,

$$\frac{1 - V_i}{V_i} \frac{n - i + 1}{m} \tag{43}$$

has the F-distribution with m and $n - i + 1$ degrees of freedom. The step-down testing procedure is to compare (43) for $i = 1$ with the significance point $F_{m,n}(\varepsilon_1)$; if (43) for $i = 1$ is larger, reject the null hypothesis that the regression of $x_1^* - \boldsymbol{\beta}_{11}^* Z_1$ on Z_1 is $\mathbf{0}$ and hence reject the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$. If this first component null hypothesis is accepted, compare (43) for $i = 2$ with $F_{m,n-1}(\varepsilon_2)$. In sequence, the component null hypotheses are tested. If one is rejected, the sequence is stopped and the hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ is rejected. If all component null hypotheses are accepted, the composite hypothesis is accepted. When the hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ is true, the probability of accepting it is $\prod_{i=1}^p (1 - \varepsilon_i)$. Hence the significance level of the step-down test is $1 - \prod_{i=1}^p (1 - \varepsilon_i)$.

In the step-down procedure the investigator usually has a choice of the ordering of the variables[†] (i.e., the numbering of the components of X) and a selection of component significance levels. It seems reasonable to order the variables in descending order of importance. The choice of significance levels will affect the power. If ε_i is a very small number, it will take a correspondingly large deviation from the ith null hypothesis to lead to rejection. In the absence of any other reason, the component significance levels can be taken equal. This procedure, of course, is not invariant with respect to linear transformation of the dependent vector variable. However, before carrying out a step-down procedure, a linear transformation can be used to determine the p variables.

The factors can be grouped. For example, group $x_1, \dots, x_k$ into one set and $x_{k+1}, \dots, x_p$ into another set. Then $U_{k,m,n} = \prod_{i=1}^k V_i$ can be used to test the null hypothesis that the first k rows of $\boldsymbol{\beta}_1$ are the first k rows of $\boldsymbol{\beta}_1^*$. Subsequently $\prod_{i=k+1}^p V_i$ is used to test the hypothesis that the last $p - k$ rows of $\boldsymbol{\beta}_1$ are those of $\boldsymbol{\beta}_1^*$; this latter criterion has the distribution under the null hypothesis of $U_{p-k,m,n-k}$.

[†] In some cases the ordering of variables may be imposed; for example, x_1 might be an observation at the first time point, x_2 at the second time point, and so on.

The investigator may test the null hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ by the likelihood ratio procedure. If the hypothesis is rejected, he may look at the factors $V_1, \dots, V_p$ to try to determine which rows of $\boldsymbol{\beta}_1$ might be different from $\boldsymbol{\beta}_1^*$.

The factors can also be used to obtain confidence regions for $\boldsymbol{\beta}_{11}, \dots, \boldsymbol{\beta}_{p1}$. Let $v_i(\varepsilon_i)$ be defined by

$$\frac{1 - v_i(\varepsilon_i)}{v_i(\varepsilon_i)} \cdot \frac{n - i + 1}{m} = F_{m, n-i+1}(\varepsilon_i). \tag{44}$$

Then a confidence region for $\boldsymbol{\beta}_{i1}$ of confidence $1 - \varepsilon_i$ is

$$\frac{\begin{vmatrix} x_i^* x_i^{*\prime} & x_i^* X_{i-1}' & x_i^* Z' \\ X_{i-1} x_i^{*\prime} & X_{i-1} X_{i-1}' & X_{i-1} Z' \\ Z x_i^{*\prime} & Z X_{i-1}' & ZZ' \end{vmatrix}}{\begin{vmatrix} (x_i^* - \overline{\boldsymbol{\beta}}_{i1} Z_1)(x_i^* - \overline{\boldsymbol{\beta}}_{i1} Z_1)' & (x_i^* - \overline{\boldsymbol{\beta}}_{i1} Z_1) X_{i-1}' & (x_i^* - \overline{\boldsymbol{\beta}}_{i1} Z_1) Z_2' \\ X_{i-1}(x_i^* - \overline{\boldsymbol{\beta}}_{i1} Z_1)' & X_{i-1} X_{i-1}' & X_{i-1} Z_2' \\ Z_2 (x_i^* - \overline{\boldsymbol{\beta}}_{i1} Z_1)' & Z_2 X_{i-1}' & Z_2 Z_2' \end{vmatrix}} \tag{45}$$

$$\cdot \frac{\begin{vmatrix} X_{i-1} X_{i-1}' & Z_{i-1} Z_2^* \\ Z_2 X_{i-1}' & Z_2 X_2' \end{vmatrix}}{\begin{vmatrix} X_{i-1} X_{i-1}' & X_{i-1} Z' \\ Z X_{i-1}' & ZZ' \end{vmatrix}} \geq v_i(\varepsilon_i).$$

8.5. AN ASYMPTOTIC EXPANSION OF THE DISTRIBUTION OF THE LIKELIHOOD RATIO CRITERION

8.5.1. General Theory of Asymptotic Expansions

In this section we develop a large-sample distribution theory for the criterion studied in this chapter. First we develop a general asymptotic expansion of the distribution of a random variable whose moments are certain functions of gamma functions [Box (1949)]. Then we apply it to the case of the likelihood ratio criterion for the linear hypothesis.

We consider a random variable W $(0 \leq W \leq 1)$ with hth moment†

$$\mathscr{E} W^h = K \left(\frac{\prod_{j=1}^b y_j^{y_j}}{\prod_{k=1}^a x_k^{x_k}} \right)^h \frac{\prod_{k=1}^a \Gamma[x_k(1+h) + \xi_k]}{\prod_{j=1}^b \Gamma[y_j(1+h) + \eta_j]}, \qquad h = 0, 1, \dots, \tag{1}$$

† In all cases where we apply this result, the parameters x_k, ξ_k, y_j, and η_j will be such that there is a distribution with such moments.

where K is a constant such that $\mathscr{E}W^0 = 1$ and

$$\sum_{k=1}^{a} x_k = \sum_{j=1}^{b} y_j. \tag{2}$$

It will be observed that the hth moment of $\lambda = U_{p,q_1,n}^{\frac{1}{2}N}$ is of this form where $x_k = \frac{1}{2}N = y_j$, $\xi_k = \frac{1}{2}(-q+1-k)$, $\eta_j = \frac{1}{2}(-q_2+1-j)$, $a = b = p$. We treat a more general case here because applications later in this book require it.

If we let

$$M = -2\log W, \tag{3}$$

the characteristic function of ρM $(0 \le \rho < 1)$ is

$$\begin{aligned} \phi(t) &= \mathscr{E}e^{it\rho M} \\ &= \mathscr{E}W^{-2it\rho} \\ &= K\left(\frac{\prod_{j=1}^{b} y_j^{y_j}}{\prod_{k=1}^{a} x_k^{x_k}}\right)^{-2it\rho} \frac{\prod_{k=1}^{a}\Gamma[x_k(1-2it\rho)+\xi_k]}{\prod_{j=1}^{b}\Gamma[y_j(1-2it\rho)+\eta_j]}. \end{aligned} \tag{4}$$

Here ρ is arbitrary; later it will depend on N. If $a = b$, $x_k = y_k$, $\xi_k \le \eta_k$, then (1) is the hth moment of the product of powers of variables with beta distributions, and then (1) holds for all h for which the gamma functions exist. In this case (4) is valid for all real t. We shall assume here that (4) holds for all real t, and in each case where we apply the result we shall verify this assumption.

Let

$$\Phi(t) = \log\phi(t) = g(t) - g(0), \tag{5}$$

where

$$\begin{aligned} g(t) = 2it\rho&\left(\sum_{k=1}^{a} x_k \log x_k - \sum_{j=1}^{b} y_j \log y_j\right) \\ &+ \sum_{k=1}^{a}\log\Gamma[\rho x_k(1-2it)+\beta_k+\xi_k] \\ &- \sum_{j=1}^{b}\log\Gamma[\rho y_j(1-2it)+\varepsilon_j+\eta_j], \end{aligned}$$

where $\beta_k = (1-\rho)x_k$ and $\varepsilon_j = (1-\rho)y_j$. The form $g(t) - g(0)$ makes $\Phi(0) = 0$, which agrees with the fact that K is such that $\phi(0) = 1$. We make use of an expansion formula for the gamma function [Barnes (1899), p. 64] which is asymptotic in x for bounded h:

$$\log \Gamma(x+h) = \log\sqrt{2\pi} + \left(x+h-\tfrac{1}{2}\right)\log x - x - \sum_{r=1}^{m}(-1)^r \frac{B_{r+1}(h)}{r(r+1)x^r} + R_{m+1}(x), \tag{6}$$

where† $R_{m+1}(x) = O(x^{-(m+1)})$ and $B_r(h)$ is the Bernoulli polynomial of degree r and order unity defined by‡

$$\frac{\tau e^{h\tau}}{e^{\tau}-1} = \sum_{r=0}^{\infty} \frac{\tau^r}{r!} B_r(h). \tag{7}$$

The first three polynomials are $[B_0(h) = 1]$

$$\begin{aligned} B_1(h) &= h - \tfrac{1}{2}, \\ B_2(h) &= h^2 - h + \tfrac{1}{6}, \\ B_3(h) &= h^3 - \tfrac{3}{2}h^2 + \tfrac{1}{2}h. \end{aligned} \tag{8}$$

Taking $x = \rho x_k(1-2it)$, $\rho y_j(1-2it)$ and $h = \beta_k + \xi_k$, $\varepsilon_j + \eta_j$ in turn, we obtain

$$\begin{aligned} \Phi(t) &= Q - g(0) - \tfrac{1}{2}f\log(1-2it) \\ &\quad + \sum_{r=1}^{m} \omega_r (1-2it)^{-r} + \sum_{k=1}^{a} O\left(x_k^{-(m+1)}\right) + \sum_{j=1}^{b} O\left(y_j^{-(m+1)}\right), \end{aligned} \tag{9}$$

where

$$f = -2\left\{\sum_k \xi_k - \sum_j \eta_j - \tfrac{1}{2}(a-b)\right\}, \tag{10}$$

$$\omega_r = \frac{(-1)^{r+1}}{r(r+1)}\left\{\sum_k \frac{B_{r+1}(\beta_k+\xi_k)}{(\rho x_k)^r} - \sum_j \frac{B_{r+1}(\varepsilon_j+\eta_j)}{(\rho y_j)^r}\right\}, \tag{11}$$

$$\begin{aligned} Q &= \tfrac{1}{2}(a-b)\log 2\pi - \tfrac{1}{2}f\log\rho \\ &\quad + \sum_k \left(x_k + \xi_k - \tfrac{1}{2}\right)\log x_k - \sum_j \left(y_j + \eta_j - \tfrac{1}{2}\right)\log y_j. \end{aligned} \tag{12}$$

†$R_{m+1}(x) = O(x^{-(m+1)})$ means $|x^{m+1}R_{m+1}(x)|$ is bounded as $|x| \to \infty$.

‡This definition differs slightly from that of Whittaker and Watson [(1943), p. 126], who expand $\tau(e^{h\tau}-1)/(e^{\tau}-1)$. If $B_r^*(h)$ is this second type of polynomial, $B_1(h) = B_1^*(h) - \frac{1}{2}$, $B_{2r}(h) = B_{2r}^*(h) + (-1)^{r+1}B_r$, where B_r is the rth Bernoulli number, and $B_{2r+1}(h) = B_{2r+1}^*(h)$.

One resulting form for $\phi(t)$ (which we shall not use here) is

$$\phi(t) = e^{\Phi(t)} = e^{Q-g(0)}(1-2it)^{-\frac{1}{2}f} \sum_{v=0}^{m} a_v (1-2it)^{-v} + R_{m+1}^{*}, \tag{13}$$

where $\sum_{v=0}^{m} a_v z^{-v}$ is the sum of the first $m+1$ terms in the series expansion of $\exp(-\sum_{r=0}^{m} \omega_r z^{-r})$, and R_{m+1}^{*} is a remainder term. Alternatively,

$$\Phi(t) = -\tfrac{1}{2}f \log(1-2it) + \sum_{r=1}^{m} \omega_r \left[(1-2it)^{-r} - 1 \right] + R'_{m+1}, \tag{14}$$

where

$$R'_{m+1} = \sum_k O\left(x_k^{-(m+1)}\right) + \sum_j O\left(y_j^{-(m+1)}\right). \tag{15}$$

In (14) we have expanded $g(0)$ in the same way we expanded $g(t)$ and have collected similar terms.

Then

$$\begin{aligned}
\phi(t) &= e^{\Phi(t)} \\
&= (1-2it)^{-\frac{1}{2}f} \exp\left[\sum_{r=1}^{m} \omega_r (1-2it)^{-r} - \sum_{r=1}^{m} \omega_r + R'_{m+1} \right] \\
&= (1-2it)^{-\frac{1}{2}f} \left\{ \prod_{r=1}^{m} \left[1 + \omega_r (1-2it)^{-r} + \frac{1}{2!} \omega_r^2 (1-2it)^{-2r} \cdots \right] \right. \\
&\qquad \left. \times \prod_{r=1}^{m} \left(1 - \omega_r + \frac{1}{2!} \omega_r^2 - \cdots \right) + R''_{m+1} \right\} \\
&= (1-2it)^{-\frac{1}{2}f} \left[1 + T_1(t) + T_2(t) + \cdots + T_m(t) + R'''_{m+1} \right],
\end{aligned} \tag{16}$$

where $T_r(t)$ is the term in the expansion with terms $\omega_1^{s_1} \cdots \omega_r^{s_r}$, $\sum i s_i = r$; for example,

$$T_1(t) = \omega_1 \left[(1-2it)^{-1} - 1 \right], \tag{17}$$

$$T_2(t) = \omega_2 \left[(1-2it)^{-2} - 1 \right] + \tfrac{1}{2} \omega_1^2 \left[(1-2it)^{-2} - 2(1-2it)^{-1} + 1 \right]. \tag{18}$$

In most applications, we will have $x_k = c_k \theta$ and $y_j = d_j \theta$, where c_k and d_j will be constant and θ will vary (i.e., will grow with the sample size). In this case if ρ is chosen so $(1-\rho)x_k$ and $(1-\rho)y_j$ have limits, then R'''_{m+1} is $O(\theta^{-(m+1)})$. We collect in (16) all terms $\omega_1^{s_1} \cdots \omega_r^{s_r}$, $\sum i s_i = r$, because these terms are $O(\theta^{-r})$.

It will be observed that $T_r(t)$ is a polynomial of degree r in $(1-2it)^{-1}$ and each term of $(1-2it)^{-\frac{1}{2}f}T_r(t)$ is a constant times $(1-2it)^{-\frac{1}{2}v}$ for an integral v. We know that $(1-2it)^{-\frac{1}{2}v}$ is the characteristic function of the χ^2-density with v degrees of freedom; that is,

$$
\begin{aligned}
(19)\qquad g_v(z) &= \frac{1}{2^{\frac{1}{2}v}\Gamma(\frac{1}{2}v)} z^{\frac{1}{2}v-1} e^{-\frac{1}{2}z} \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi}(1-2it)^{-\frac{1}{2}v} e^{-itz}\,dt.
\end{aligned}
$$

Let

$$
(20)\qquad
\begin{aligned}
S_r(z) &= \int_{-\infty}^{\infty} \frac{1}{2\pi}(1-2it)^{-\frac{1}{2}f} T_r(t) e^{-itz}\,dt, \\
R^{\text{iv}}_{m+1} &= \int_{-\infty}^{\infty} \frac{1}{2\pi}(1-2it)^{-\frac{1}{2}f} R'''_{m+1}\, e^{-itz}\,dt.
\end{aligned}
$$

Then the density of ρM is

$$
\begin{aligned}
(21)\quad \int_{-\infty}^{\infty} \frac{1}{2\pi}\phi(t) e^{-itz}\,dt &= \sum_{r=0}^{m} S_r(z) + R^{\text{iv}}_{m+1} \\
&= g_f(z) + \omega_1\left[g_{f+2}(z) - g_f(z)\right] \\
&\quad + \Big\{\omega_2\left[g_{f+4}(z) - g_f(z)\right] \\
&\qquad + \frac{\omega_1^2}{2}\left[g_{f+4}(z) - 2g_{f+2}(z) + g_f(z)\right]\Big\} \\
&\quad + \cdots + S_m(z) + R^{\text{iv}}_{m+1}.
\end{aligned}
$$

Let

$$
(22)\qquad
\begin{aligned}
U_r(z_0) &= \int_0^{z_0} S_r(z)\,dz, \\
R^{\text{v}}_{m+1} &= \int_0^{z_0} R^{\text{iv}}_{m+1}\,dz.
\end{aligned}
$$

The cdf of M is written in terms of the cdf of ρM, which is the integral of the density, namely,

$$
\begin{aligned}
(23)\quad \Pr\{M \le M_0\} &= \Pr\{\rho M \le \rho M_0\} \\
&= \sum_{r=0}^{m} U_r(\rho M_0) + R_{m+1}^{v} \\
&= \Pr\{\chi_f^2 \le \rho M_0\} + \omega_0\left(\Pr\{\chi_{f+2}^2 \le \rho M_0\} - \Pr\{\chi_f^2 \le \rho M_0\}\right) \\
&\quad + \Big[\omega_2\left(\Pr\{\chi_{f+4}^2 \le \rho M_0\} - \Pr\{\chi_f^2 \le \rho M_0\}\right) + \frac{\omega_1^2}{2}\left(\Pr\{\chi_{f+4}^2 \le \rho M_0\}\right. \\
&\qquad \left. - 2\Pr\{\chi_{f+2}^2 \le \rho M_0\} + \Pr\{\chi_f^2 \le \rho M_0\}\right)\Big] \\
&\quad + \cdots + U_m(\rho M_0) + R_{m+1}^{v}.
\end{aligned}
$$

The remainder R_{m+1}^{v} is $O(\theta^{-(m+1)})$; this last statement can be verified by following the remainder terms along. (In fact, to make the proof rigorous one needs to verify that each remainder is of the proper order in a uniform sense.)

In many cases it is desirable to choose ρ so that $\omega_1 = 0$. In such a case using only the first term of (23) gives an error of order θ^{-2}.

Further details of the expansion can be found in Box's paper (1949).

Theorem 8.5.1. *Suppose that $\mathscr{E}W^h$ is given by* (1) *for all purely imaginary h, with* (2) *holding. Then the* cdf *of $-2\rho\log W$ is given by* (23). *The error, R_{m+1}^{v}, is $O(\theta^{-(m+1)})$ if $x_k \ge c_k\theta$, $y_j \ge d_j\theta$ ($c_k > 0$, $d_j > 0$), and if $(1-\rho)x_k$, $(1-\rho)y_j$ have limits, where ρ may depend on θ.*

Box also considers approximating the distribution of $-2\rho\log W$ by an F-distribution. He finds that the error in this approximation can be made to be of order θ^{-3}.

8.5.2. Asymptotic Distribution of the Likelihood Ratio Criterion

We now apply Theorem 8.5.1 to the distribution of $-2\log\lambda$, the likelihood ratio criterion developed in Section 8.3. We let $W = \lambda$. The hth moment of λ is

$$
(24)\qquad \mathscr{E}\lambda^h = K\frac{\prod_{k=1}^{p}\Gamma\left[\frac{1}{2}(N-q+1-k+Nh)\right]}{\prod_{j=1}^{p}\Gamma\left[\frac{1}{2}(N-q_2+1-j+Nh)\right]},
$$

and this holds for all h for which the gamma functions exist, including purely imaginary h. We let $a = b = p$,

$$
(25)\qquad
\begin{aligned}
x_k &= \tfrac{1}{2}N, & \xi_k &= \tfrac{1}{2}(-q+1-k), & \beta_k &= \tfrac{1}{2}(1-\rho)N,\\
y_j &= \tfrac{1}{2}N, & \eta_j &= \tfrac{1}{2}(-q_2+1-j), & \varepsilon_j &= \tfrac{1}{2}(1-\rho)N.
\end{aligned}
$$

We observe that

$$
\begin{aligned}
(26)\quad 2\omega_1 = \sum_{k=1}^{p} &\left\{ \frac{\left\{\frac{1}{2}[(1-\rho)N-q+1-k]\right\}^2 - \frac{1}{2}[(1-\rho)N-q+1-k]}{\frac{1}{2}\rho N} \right.\\
&\left. - \frac{\left\{\frac{1}{2}[(1-\rho)N-q_2+1-k]\right\}^2 - \frac{1}{2}[(1-\rho)N-q_2+1-k]}{\frac{1}{2}\rho N} \right\}\\
&= \frac{2}{\rho N}\sum_{k=1}^{p}\left[\frac{-2[(1-\rho)N-q_2+1-k]q_1+q_1^2}{4}+\frac{q_1}{2}\right]\\
&= \frac{pq_1}{2\rho N}[-2(1-\rho)N+2q_2-2+(p+1)+q_1+2].
\end{aligned}
$$

To make this zero, we require that

$$
(27)\qquad \rho = \frac{N-q_2-\frac{1}{2}(p+q_1+1)}{N}.
$$

Then

$$
\begin{aligned}
(28)\qquad \Pr&\left\{-2\frac{k}{N}\log\lambda \le z\right\}\\
&= \Pr\{-k\log U_{p,q_1,N-c} \le z\}\\
&= \Pr\{\chi^2_{pq_1} \le z\}\\
&\quad + \frac{\gamma_2}{k^2}\left(\Pr\{\chi^2_{pq_1+4} \le z\} - \Pr\{\chi^2_{pq_1} \le z\}\right)\\
&\quad + \frac{1}{k^4}\left[\gamma_4\left(\Pr\{\chi^2_{pq_1+8} \le z\} - \Pr\{\chi^2_{pq_1} \le z\}\right)\right.\\
&\qquad\left. - \gamma_2^2\left(\Pr\{\chi^2_{pq_1+4} \le z\} - \Pr\{\chi^2_{pq_1} \le z\}\right)\right] + R_5^{v},
\end{aligned}
$$

where

$$k = \rho N = N - q_2 - \tfrac{1}{2}(p + q_1 + 1) = n - \tfrac{1}{2}(p - q_1 + 1), \tag{29}$$

$$\gamma_2 = \frac{pq_1(p^2 + q_1^2 - 5)}{48}, \tag{30}$$

$$\gamma_4 = \frac{\gamma_2^2}{2} + \frac{pq_1}{1920}\left[3p^4 + 3q_1^4 + 10p^2q_1^2 - 50(p^2 + q_1^2) + 159\right]. \tag{31}$$

Since $\lambda = U_{p,q_1,n}^{\frac{1}{2}N}$, where $n = N - q$, (28) gives $\Pr\{-k \log U_{p,q_1,n} \le z\}$.

Theorem 8.5.2. *The* cdf *of* $-k \log U_{p,q_1,n}$ *is given by* (28) *with* $k = n - \frac{1}{2}(p - q_1 + 1)$, *and* γ_2 *and* γ_4 *given by* (30) *and* (31), *respectively. The remainder term is* $O(N^{-6})$.

The coefficient $k = n - \frac{1}{2}(p - q_1 + 1)$ is known as the *Bartlett correction.* If the first term of (28) is used, the error is of the order N^{-2}; if the second, N^{-4}; and if the third,† N^{-6}. The second term is always negative and is numerically maximum for $z = \sqrt{(pq_1 + 2)(pq_1)}$ $(= pq_1 + 1$, approximately). For $p \ge 3, q_1 \ge 3$, we have $\gamma_2/k^2 \le [(p^2 + q_1^2)/k]^2/96$, and the contribution of the second term lies between $-0.005[(p^2 + q_1^2)/k]^2$ and 0. For $p \ge 3$, $q_1 \ge 3$, we have $\gamma_4 \le \gamma_2^2$, and the contribution of the third term is numerically less than $(\gamma_2/k^2)^2$. A rough rule that may be followed is that use of the first term is accurate to three decimal places if $p^2 + q_1^2 \le k/3$.

As an example of the calculation, consider the case of $p = 3$, $q_1 = 6$, $N - q_2 = 24$, and $z = 26.0$ (the 10% significance point χ_{18}^2). In this case $\gamma_2/k^2 = 0.048$ and the second term is -0.007: $\gamma_4/k^4 = 0.0015$ and the third term is -0.0001. Thus the probability of $-19 \log U_{3,6,18} \le 26.0$ is 0.893 to three decimal places.

Since

$$-\left[n - \tfrac{1}{2}(p - m + 1)\right] \log u_{p,m,n}(\alpha) = C_{p,m,n-p+1}(\alpha)\chi_{pm}^2(\alpha), \tag{32}$$

the proportional error in approximating the left-hand side by $\chi_{pm}^2(\alpha)$ is $C_{p,m,n-p+1} - 1$. The proportional error increases slowly with p and m.

8.5.3. A Normal Approximation

Mudholkar and Trivedi (1980), (1981) developed a normal approximation to the distribution of $-\log U_{p,m,n}$ which is asymptotic as p and/or $m \to \infty$. It is related to the Wilson–Hilferty normal approximation for the χ^2-distribution.

†Box has shown that the term of order N^{-5} is 0 and gives the coefficients to be used in the term of order N^{-6}.

First, we give the background of the approximation. Suppose $\{Y_k\}$ is a sequence of nonnegative random variables such that $(Y_k - \mu_k)/\sigma_k \xrightarrow{\mathcal{L}} N(0,1)$ as $k \to \infty$, where $\mathscr{E}Y_k = \mu_k$ and $\mathscr{V}(Y_k) = \sigma_k^2$. Suppose also that $\mu_k \to \infty$ and σ_k^2/μ_k is bounded as $k \to \infty$. Let $Z_k = (Y_k/\mu_k)^h$. Then

$$\text{(33)} \qquad \frac{Z_k - 1}{h(\sigma_k/\mu_k)} = \frac{\mu_k(Z_k - 1)}{h\sigma_k} \xrightarrow{\mathcal{L}} N(0,1)$$

by Theorem 4.2.3. The approach to normality may be accelerated by choosing h to make the distribution of Z_k nearly symmetric as measured by its third cumulant. The normal distribution is to be used as an approximation and is justified by its accuracy in practice. However, it will be convenient to develop the ideas in terms of limits, although rigor is not necessary.

By a Taylor expansion we express the hth moment of Y_k/μ_k as

$$\begin{aligned}\text{(34)} \quad \mathscr{E}Z_k &= \mathscr{E}\left(\frac{Y_k}{\mu_k}\right)^h \\ &= 1 + \frac{h(h-1)}{2}\frac{\sigma_k^2}{\mu_k} \\ &\quad + \frac{h(h-1)(h-2)}{24}\frac{4\phi_k - 3(h-3)\left(\sigma_k^2/\mu_k\right)^2}{\mu_k^2} + O(\mu_k^{-3}),\end{aligned}$$

where $\phi_k = \mathscr{E}(Y_k - \mu_k)^3/\mu_k$, assumed bounded. The rth moment of Z_k is expressed by replacment of h by rh in (34). The central moments of Z_k are

$$\text{(35)} \quad \mathscr{E}(Z_k - 1)^2 = h^2\frac{\sigma_k^2/\mu_k}{\mu_k} + \frac{h^2(h-1)}{2}\frac{2\phi_k + (3h-5)\left(\sigma_l^2/\mu_k\right)^2}{\mu_k^2} + O(\mu_k^{-3}),$$

$$\text{(36)} \qquad \mathscr{E}(Z_k - 1)^3 = h^3\frac{\phi_k + 3(h-1)\left(\sigma_k^2/\mu_k\right)^2}{\mu_k^2} + O(\mu_k^{-3}).$$

To make the third moment approximately 0 we take h to be

$$\text{(37)} \qquad h_0 = 1 - \frac{\mathscr{E}(Y_k - \mu_k)^3\mu_k}{3\sigma_k^4}.$$

Then $Z_k = (Y_k/\mu_k)^{h_0}$ is treated as normally distributed with mean and variance given by (34) and (35), respectively, with $h = h_0$.

Now we consider $-\log U_{p,m,n} = -\sum_{i=1}^{p} \log V_i$, where $V_1, \dots, V_p$ are independent and V_i has the density $\beta(x; (n+1-i)/2, m/2)$, $i = 1, \dots, p$. As $n \to \infty$ and $m \to \infty$, $-\log V_i$ tends to normality. If V has the density $\beta(x; a/2, b/2)$, the moment generating function of $-\log V$ is

$$(38) \qquad \mathscr{E} e^{-t \log V} = \frac{\Gamma[(a+b)/2]\Gamma(a/2-t)}{\Gamma(a/2)\Gamma[(a+b)/2-t]}.$$

Its logarithm is the cumulant generating function. Differentiation of the last yields as the rth cumulant of V

$$(39) \qquad C_r = (-1)^r \left[\psi^{(r-1)}\left(\frac{a}{2}\right) - \psi^{(r-1)}\left(\frac{a+b}{2}\right) \right], \qquad r = 1, 2, \dots .$$

where $\psi(w) = d \log \Gamma(w)/dw$. [See Abramovitz and Stegun (1972), p. 258, for example.] From $\Gamma(w+1) = w\Gamma(w)$ we obtain the recursion relation $\psi(w+1) = \psi(w) + 1/w$. This yields for $s = 0$ and l an integer

$$(40) \qquad \psi^{(s)}(w+l) - \psi^{(s)}(w) = (-1)^{s+1} s! \sum_{j=0}^{l-1} \frac{1}{(w+j)^{s+1}}.$$

The validity of (40) for $s = 1, 2, \dots$ is verified by differentiation. [The expression for $\psi'(Z)$ in the first line of page 223 of Mudholkar and Trivedi (1981) is incorrect.] Thus for $b = 2l$

$$(41) \qquad C_r = (r-1)! \sum_{j=0}^{l-1} \frac{1}{(a/2+j)^r}.$$

From these results we obtain as the rth cumulant of $-\log U_{p,2l,n}$

$$(42) \qquad \kappa_r(-\log U_{p,2l,n}) = 2^r (r-1)! \sum_{i=1}^{p} \sum_{j=0}^{l-1} \frac{1}{(n-i+1-2j)^r}.$$

As $l \to \infty$ the series diverges for $r = 1$ and converges for $r = 2, 3$, and hence $\kappa_r/\kappa_1 \to 0$, $r = 2, 3$. The same is true as $p \to \infty$ (if n/p approaches a positive constant).

Given n, p, and l, the first three cumulants are calculated from (42). Then h_0 is determined from (37), and $(-\log U_{p,2l,n})^{h_0}$ is treated as approximately normally distributed with mean and variance calculated from (34) and (35) for $h = h_0$.

Mudholkar and Trivedi (1980) calculated the error of approximation for significance levels of 0.01 and 0.05 for n from 4 to 66, $p = 3, 7$, and

$q = 2, 6, 10$. The maximum error is less than 0.0007; in most cases the error is considerably less. The error for the χ^2-approximation is much larger, especially for small values of n.

In case of m odd the rth cumulant can be approximated by

$$(43) \qquad 2^r(r-1)!\sum_{i=1}^{p}\left[\sum_{j=0}^{\frac{1}{2}(m-3)}\frac{1}{(n-i+1-2j)^r}+\frac{1}{2}\frac{1}{(n-i+m)^r}\right].$$

Davis (1933, 1935) gave tables of $\psi(w)$ and its derivatives.

8.5.4. An F-Approximation

Rao (1951) has used the expansion of Section 8.5.2 to develop an expansion of the distribution of another function of $U_{p,m,n}$ in terms of beta distributions. The constants can be adjusted so that the term after the leading one is of order m^{-4}. A good approximation is to consider

$$(44) \qquad \frac{1-U^{1/s}}{U^{1/s}}\cdot\frac{ks-r}{pm}$$

as F with pm and $ks-r$ degrees of freedom, where

$$(45) \qquad s=\sqrt{\frac{p^2m^2-4}{p^2+m^2-5}}, \qquad r=\frac{pm}{2}-1,$$

and k is $n-\frac{1}{2}(p-m-1)$. For $p = 1$ or 2 or $m = 1$ or 2 the F-distribution is exactly as given in Section 8.4. If $ks - r$ is not an integer, interpolation between two integer values can be used. For smaller values of m this approximation is more accurate than the χ^2-approximation.

8.6. OTHER CRITERIA FOR TESTING THE LINEAR HYPOTHESIS

8.6.1. Functions of Roots

Thus far the only test of the linear hypothesis we have considered is the likelihood ratio test. In this section we consider other test procedures.

Let $\hat{\boldsymbol{\Sigma}}_\Omega$, $\hat{\boldsymbol{\beta}}_{1\Omega}$, and $\hat{\boldsymbol{\beta}}_{2\omega}$ be the estimates of the parameters in $N(\boldsymbol{\beta}\mathbf{z}, \boldsymbol{\Sigma})$, based on a sample of N observations. These are a sufficient set of statistics, and we shall base test procedures on them. As was shown in Section 8.3, if the hypothesis is $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$, one can reformulate the hypothesis as $\boldsymbol{\beta}_1 = \mathbf{0}$ (by replacing x_α by $x_\alpha - \beta_1^* z_\alpha^{(1)}$). Moreover,

$$
\begin{aligned}
(1)\qquad \beta z_\alpha &= \beta_1 z_\alpha^{(1)} + \beta_2 z_\alpha^{(2)} \\
&= \beta_1\left(z_\alpha^{(1)} - A_{12}A_{22}^{-1}z_\alpha^{(2)}\right) + \left(\beta_2 + \beta_1 A_{12}A_{22}^{-1}\right)z_\alpha^{(2)} \\
&= \beta_1 z_\alpha^{*(1)} + \beta_2^* z_\alpha^{(2)},
\end{aligned}
$$

where $\sum_\alpha z_\alpha^{*(1)} z_\alpha^{(2)\prime} = 0$ and $\sum_\alpha z_\alpha^{*(1)} z_\alpha^{*(1)\prime} = A_{11\cdot 2}$. Then $\hat{\beta}_1 = \hat{\beta}_{1\Omega}$ and $\hat{\beta}_2^* = \hat{\beta}_{2\omega}$.

We shall use the principle of invariance to reduce the set of tests to be considered. First, if we make the transformation $X_\alpha^* = X_\alpha + \Gamma z_\alpha^{(2)}$, we leave the null hypothesis invariant, since $\mathscr{E} X_\alpha^* = \beta_1 z_\alpha^{*(1)} + (\beta_2^* + \Gamma)z_\alpha^{(2)}$ and $\beta_2^* + \Gamma$ is unspecified. The only invariants of the sufficient statistics are $\hat{\Sigma}$ and $\hat{\beta}_1$ (since for each $\hat{\beta}_2^*$, there is a Γ that transforms it to 0, that is, $-\hat{\beta}_2^*$).

Second, the null hypothesis is invariant under the transformation $z_\alpha^{**(1)} = Cz_\alpha^{*(1)}$ (C nonsingular); the transformation carries β_1 to $\beta_1 C^{-1}$. Under this transformation $\hat{\Sigma}$ and $\hat{\beta}_1 A_{11\cdot 2}\hat{\beta}_1'$ are invariant; we consider $A_{11\cdot 2}$ as information relevant to inference. However, these are the only invariants. For consider a function of $\hat{\beta}_1$ and $A_{11\cdot 2}$, say $f(\hat{\beta}_1, A_{11\cdot 2})$. Then there is a C^* that carries this into $f(\hat{\beta}_1 C^{*-1}, I)$, and a further orthogonal transformation carries this into $f(T, I)$, where $t_{iv} = 0$, $i < v$, $t_{ii} \geq 0$. (If each row of T is considered a vector in q_1-space, the rotation of coordinate axes can be done so the first vector is along the first coordinate axis, the second vector is in the plane determined by the first two coordinate axes, and so forth). But T is a function of $TT' = \hat{\beta}_1 A_{11\cdot 2}\hat{\beta}_1'$; that is, the elements of T are uniquely determined by this equation and the preceding restrictions. Thus our tests will depend on $\hat{\Sigma}$ and $\hat{\beta}_1 A_{11\cdot 2}\hat{\beta}_1'$. Let $N\hat{\Sigma} = G$ and $\hat{\beta}_1 A_{11\cdot 2}\hat{\beta}_1' = H$.

Third, the null hypothesis is invariant when x_α is replaced by Kx_α, for Σ and β_2^* are unspecified. This transforms G to KGK' and H to KHK'. The only invariants of G and H under such transformations are the roots of

$$
(2)\qquad |H - lG| = 0.
$$

It is clear the roots are invariant, for

$$
\begin{aligned}
(3)\qquad 0 &= |KHK' - lKGK'| \\
&= |K(H - lG)K'| \\
&= |K| \cdot |H - lG| \cdot |K'|.
\end{aligned}
$$

On the other hand, these are the only invariants, for given G and H there is a K such that $KGK' = I$ and

$$\text{(4)} \qquad KHK' = L = \begin{pmatrix} l_1 & 0 & \cdots & 0 \\ 0 & l_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & l_p \end{pmatrix},$$

where $l_1 \geq \cdots l_p$ are the roots of (2). (See Theorem A.2.2 of the Appendix.)

Theorem 8.6.1. *Let x_α be an observation from $N(\beta_1 z_\alpha^{*(1)} + \beta_2^* z_\alpha^{(2)}, \Sigma)$, where $\sum_\alpha z_\alpha^{*(1)} z_\alpha^{(2)\prime} = 0$ and $\sum_\alpha z_\alpha^{*(1)} z_\alpha^{*(1)\prime} = A_{11\cdot 2}$. The only functions of the sufficient statistics and $A_{11\cdot 2}$ invariant under the transformations $x_\alpha^* = x_\alpha + \Gamma z_\alpha^{(2)}$, $z_\alpha^{**(1)} = C z_\alpha^{*(1)}$, and $x_\alpha^* = K x_\alpha$ are the roots of (2), where $G = N\hat{\Sigma}$ and $H = \hat{\beta}_1 A_{11\cdot 2} \hat{\beta}_1'$.*

The likelihood ratio criterion is a function of

$$\text{(5)} \qquad U = \frac{|G|}{|G+H|} = \frac{|KGK'|}{|KGK' + KHK'|} = \frac{|I|}{|I+L|} = \prod_{i=1}^{p} (1 + l_i)^{-1},$$

which is clearly invariant under the transformations.

Intuitively it would appear that good tests should reject the null hypothesis when the roots in some sense are large, for if β_1 is very different from $\mathbf{0}$, then $\hat{\beta}_1$ will tend to be large and so will H. Some other criteria that have been suggested are (a) $\sum l_i$, (b) $\sum l_i/(1 + l_i)$, (c) $\max l_i$, and (d) $\min l_i$. In each case we reject the null hypothesis if the criterion exceeds some specified number.

8.6.2. The Lawley–Hotelling Trace Criterion

Let K be the matrix such that $KGK' = I$ [$G = K^{-1}(K')^{-1}$, or $G^{-1} = K'K$] and so (4) holds. Then the sum of the roots can be written

$$\text{(6)} \qquad \sum_{i=1}^{p} l_i = \operatorname{tr} L = \operatorname{tr} KHK' = \operatorname{tr} HK'K = \operatorname{tr} HG^{-1}.$$

This criterion was suggested by Lawley (1938), Bartlett (1939), and Hotelling (1947), (1951). The test procedure is to reject the hypothesis if (6) is greater than a constant depending on p, m, and n.

The general distribution† of $\operatorname{tr} HG^{-1}$ cannot be characterized as easily as that of $U_{p,m,n}$. In the case of $p=2$, Hotelling (1951) obtained an explicit expression for the distribution of $\operatorname{tr} HG^{-1} = l_1 + l_2$. A slightly different form of this distribution is obtained from the density of the two roots l_1 and l_2 in Chapter 13. It is

$$(7)\quad \Pr\{\operatorname{tr} HG^{-1} \le w\} = I_{w/(2+w)}(m-1, n-1)$$
$$-\frac{\sqrt{\pi}\,\Gamma\left[\frac{1}{2}(m+n-1)\right]}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)}(1+w)^{-\frac{1}{2}(n-1)} I_{w^2/(2+w)^2}\left[\tfrac{1}{2}(m-1), \tfrac{1}{2}(n-1)\right],$$

where $I_x(a,b)$ is the *incomplete beta function*, that is, the integral of $\beta(y; a, b)$ from 0 to x.

Constantine (1966) expressed the density of $\operatorname{tr} HG^{-1}$ as an infinite series in generalized Laguerre polynomials and as an infinite series in zonal polynomials; these series, however, converge only for $\operatorname{tr} HG^{-1} < 1$. Davis (1968) showed that the analytic continuation of these series satisfies a system of linear homogeneous differential equations of order p. Davis (1970a, 1970b) used a solution to compute tables as given in Appendix B.

Under the null hypothesis, G is distributed as $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ $(n = N-q)$ and H is distributed as $\sum_{\upsilon=1}^{q_1} Y_\upsilon Y_\upsilon'$, where the Z_α and Y_υ are independent, each with distribution $N(\mathbf{0}, \mathbf{\Sigma})$. Since the roots are invariant under the previously specified linear transformation, we can choose K so that $K\mathbf{\Sigma}K' = I$ and let $G^* = KGK'$ $[= \sum(KZ_\alpha)(KZ_\alpha)']$ and $H^* = KHK'$. This is equivalent to assuming at the outset that $\mathbf{\Sigma} = I$.

Now

$$(8)\qquad \underset{N\to\infty}{\operatorname{plim}} \frac{1}{N}G = \underset{n\to\infty}{\operatorname{plim}} \frac{n}{n+q}\frac{1}{n}\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha' = I.$$

This result follows applying the (weak) law of large numbers to each element of $(1/n)G$,

$$(9)\qquad \underset{n\to\infty}{\operatorname{plim}} \frac{1}{n}\sum_{\alpha=1}^{n} Z_{i\alpha}Z_{j\alpha} = \mathscr{E} Z_{i\alpha}Z_{j\alpha} = \delta_{ij}.$$

Theorem 8.6.2. *Let $f(H)$ be a function whose discontinuities form a set of probability zero when H is distributed as $\sum_{\upsilon=1}^{q_1} Y_\upsilon Y_\upsilon'$ with the Y_υ independent, each with distribution $N(\mathbf{0}, I)$. Then the limiting distribution of $f(NHG^{-1})$ is the distribution of $f(H)$.*

†Lawley (1938) purported to derive the exact distribution, but the result is in error

Proof. This is a straightforward application of a general theorem [for example, Theorem 2 of Chernoff (1956)] to the effect that if the cdf of X_n converges to that of X (at every continuity point of the latter) and if $g(x)$ is a function whose discontinuities form a set of probability 0 according to the distribution of X, then the cdf of $g(X_n)$ converges to that of $g(X)$. In our case X_n consists of the components of H and G, and X consists of the components of H and I. ■

Corollary 8.6.1. *The limiting distribution of N $\operatorname{tr} HG^{-1}$ or $n \operatorname{tr} HG^{-1}$ is the χ^2-distribution with pq_1 degrees of freedom.*

This follows from Theorem 8.6.2, because

$$\operatorname{tr} H = \sum_{i=1}^{p} h_{ii} = \sum_{i=1}^{p} \sum_{v=1}^{q_1} Y_{iv}^2. \tag{10}$$

Ito (1956), (1960) developed asymptotic formulas, and Fujikoshi (1973) extended them. Let $w_{p,m,n}(\alpha)$ be the α significance point of $\operatorname{tr} HG^{-1}$; that is,

$$\Pr\left\{\operatorname{tr} HG^{-1} \geq w_{p,m,n}(\alpha)\right\} = \alpha, \tag{11}$$

and let $\chi_k^2(\alpha)$ be the α-significance point of the χ^2-distribution with k degrees of freedom. Then

$$nw_{p,m,n}(\alpha) = \chi_{pm}^2(\alpha) + \frac{1}{2n}\left[\frac{p+m+1}{pm+2}\chi_{pm}^4(\alpha) + (p-m+1)\chi_{pm}^2(\alpha)\right] + O(n^{-2}). \tag{12}$$

Ito also gives the term of order n^{-2}. See also Muirhead (1970). Davis (1970a), (1970b) evaluated the accuracy of the approximation (12). Ito also found

$$\Pr\{n \operatorname{tr} HG^{-1} \leq z\} = G_{pm}(z) - \frac{1}{2n}\left[\frac{p+m+1}{pm+2}z^2 + (p-m+1)g_{pm}(z)\right] + O(n^{-2}), \tag{13}$$

where $G_k(z) = \Pr\{\chi_k^2 \leq z\}$ and $g_k(z) = (d/dz)G_k(z)$. Pillai (1956) suggested another approximation to $nw_{p,m,n}(\alpha)$, and Pillai and Samson (1959) gave moments of $\operatorname{tr} HG^{-1}$. Pillai and Young (1971) and Krishnaiah and Chang (1972) evaluated the Laplace transform of $\operatorname{tr} HG^{-1}$ and showed how to invert the transform. Khatri and Pillai (1966) suggest an approximate distribution based on moments. Pillai and Young (1971) suggest approximate distributions based on the first three moments.

Tables of the significance points are given by Grubbs (1954) for $p = 2$ and by Davis (1970a) for $p = 3$ and 4, Davis (1970b) for $p = 5$, and Davis (1980) for $p = 6(1)10$; approximate significance points have been given by Pillai (1960). Davis's tables are reproduced in Table B.2.

8.6.3. The Bartlett–Nanda–Pillai Trace Criterion

Another criterion, proposed by Bartlett (1939), Nanda (1950), and Pillai (1955), is

$$\begin{aligned} V &= \sum_{i=1}^{p} \frac{l_i}{1+l_i} = \operatorname{tr} L(I+L)^{-1} \\ &= \operatorname{tr} KHK'(KGK' + KHK')^{-1} \\ &= \operatorname{tr} HK'[K(G+H)K']^{-1}K \\ &= \operatorname{tr} H(G+H)^{-1}, \end{aligned} \tag{14}$$

where as before K is such that $KGK' = I$ and (4) holds. In terms of the roots $f_i = l_i/(1+l_i)$, $i = 1, \dots, p$, of

$$|H - f(H+G)| = 0, \tag{15}$$

the criterion is $\sum_{i=1}^{p} f_i$. In principle, the cdf, density, and moments under the null hypothesis can be found from the density of the roots (Sec. 13.2.3),

$$C \prod_{i=1}^{p} f_i^{\frac{1}{2}(m-p-1)} \prod_{i=1}^{p} (1-f_i)^{\frac{1}{2}(n-p-1)} \prod_{i<j} (f_i - f_j), \tag{16}$$

where

$$C = \frac{\pi^{\frac{1}{2}p^2} \Gamma_p\left[\frac{1}{2}(m+n)\right]}{\Gamma_p\left(\frac{1}{2}n\right) \Gamma_p\left(\frac{1}{2}m\right) \Gamma_p\left(\frac{1}{2}p\right)} \tag{17}$$

for $1 > f_1 > \cdots > f_p > 0$, and 0 otherwise. If $m - p$ and $n - p$ are odd, the density is a polynomial in $f_1, \dots, f_p$. Then the density and cdf of the sum of the roots are polynomials.

Many authors have written about the moments, Laplace transforms, densities, and cdfs, using various approaches. Nanda (1950) derived the distribution for $p = 2, 3, 4$ and $m = p + 1$. Pillai (1954), (1956), (1960) and Pillai and

Mijares (1959) calculated the first four moments of V and proposed approximating the distribution by a beta distribution based on the first four moments. Pillai and Jayachandran (1970) show how to evaluate the moment generating function as a weighted sum of determinants whose elements are incomplete gamma functions; they derive exact densities for some special cases and use them for a table of significance points. Krishnaiah and Chang (1972) express the distributions as linear combinations of inverse Laplace transforms of the products of certain double integrals and further develop this technique for finding the distribution. Davis (1972b) showed that the distribution satisfies a differential equation and showed the nature of the solution. Khatri and Pillai (1968) obtained the (nonnull) distributions in series forms. The characteristic function (under the null hypothesis) was given by James (1964). Pillai and Jayachandran (1967) found the nonnull distribution for $p = 2$ and computed power functions. For an extensive bibliography see Krishnaiah (1978).

We now turn to the asymptotic theory. It follows from Theorem 8.6.2 that nV or NV has a limiting χ^2-distribution with pm degrees of freedom.

Let $v_{p,m,n}(\alpha)$ be defined by

$$\Pr\left\{\operatorname{tr} \boldsymbol{H}(\boldsymbol{H}+\boldsymbol{G})^{-1} \geq v_{p,m,n}(\alpha)\right\} = \alpha. \tag{18}$$

Then Davis (1970a), (1970b), Fujikoshi (1973), and Rothenberg (1977) have shown that

$$\begin{aligned} nv_{p,m,n}(\alpha) = \chi^2_{pm}(\alpha) + \frac{1}{2n}\bigg[&-\frac{p+m+1}{pm+2}\chi^4_{pm}(\alpha) \\ &+(p-m+1)\chi^2_{pm}(\alpha)\bigg] + O(n^{-2}). \end{aligned} \tag{19}$$

Since we can write (for the likelihood ratio test)

$$nu_{p,m,n}(\alpha) = \chi^2_{pm}(\alpha) + \frac{1}{2n}(p-m+1)\chi^2_{pm}(\alpha) + O(n^{-2}), \tag{20}$$

we have the comparison

$$nw_{p,m,n}(\alpha) = nu_{p,m,n}(\alpha) + \frac{1}{2n}\cdot\frac{p+m+1}{pm+2}\chi^4_{pm}(\alpha) + O(n^{-2}), \tag{21}$$

$$nv_{p,m,n}(\alpha) = nu_{p,m,n}(\alpha) + \frac{1}{2n}\cdot\frac{p+m+1}{pm+2}\chi^4_{pm}(\alpha) + O(n^{-2}). \tag{22}$$

An asymptotic expansion [Muirhead (1970), Fujikoshi (1973)] is

$$(23)\quad \Pr\{nV \le z\} = G_{pn}(z) + \frac{pm}{4n}\big[(m-p-1)G_{pm}(z) + 2(p+1)G_{pm+2}(z) - (p+m+1)G_{pm+4}(z)\big] + O(n^{-2}).$$

Higher-order terms are given by Muirhead and Fujikoshi.

Tables. Pillai (1960) tabulated 1% and 5% significance points of V for $p = 2(1)8$ based on fitting Pearson curves (i.e., beta distributions with adjusted ranges) to the first four moments. Mijares (1964) extended the tables to $p = 50$. Table B.3 of some significance points of $(n+m)V/m = \operatorname{tr}(1/m)\boldsymbol{H}\{[1/(n+m)](\boldsymbol{G}+\boldsymbol{H})\}^{-1}$ is from *Concise Statistical Tables*, and was computed on the same basis as Pillai's. Schuurman, Krishnaiah, and Chattopodhyay (1975) gave exact significance points of V for $p = 2(1)5$; a more extensive table is in their technical report (ARL 73-0008). A comparison of some values with those of *Concise Statistical Tables* (Appendix B) shows a maximum difference of 3 in the third decimal place.

8.6.4. The Roy Maximum Root Criterion

Any characteristic root of $\boldsymbol{H}\boldsymbol{G}^{-1}$ can be used as a test criterion. Roy (1953) proposed l_1, the maximum characteristic root of $\boldsymbol{H}\boldsymbol{G}^{-1}$, on the basis of his union–intersection principle. The test procedure is to reject the null hypothesis if l_1 is greater than a certain number, or equivalently, if $f_1 = l_1/(1+l_1) = R$ is greater than a number $r_{p,m,n}(\alpha)$ which satisfies

$$(24)\qquad \Pr\{R \ge r_{p,m,n}(\alpha)\} = \alpha.$$

The density of the roots $f_1, \dots, f_p$ for $p \le m$ under the null hypothesis is given in (16). The cdf of $R = f_1$, $\Pr\{f_1 \le f^*\}$, can be obtained from the joint density by integration over the range $0 \le f_p \le \cdots \le f_1 \le f^*$. If $m-p$ and $n-p$ are both odd, the density of $f_1, \dots, f_p$ is a polynomial; then the cdf of f_1 is a polynomial in f^* and the density of f_1 is a polynomial. The only difficulty in carrying out the integration is keeping track of the different terms.

Roy [(1945), (1957), Appendix 9] developed a method of integration that results in a cdf that is a linear combination of products of univariate beta densities and beta cdfs. The cdf of f_1 for $p = 2$ is

$$(25)\quad \Pr\{f_1 \le f\} = I_f(m-1, n-1) - \frac{\sqrt{\pi}\,\Gamma\big[\frac12(m+n-1)\big]}{\Gamma(\frac12 m)\Gamma(\frac12 n)} f^{\frac12(m-1)}(1-f)^{\frac12(n-1)} I_f\big[\tfrac12(m-1), \tfrac12(n-1)\big].$$

This is derived in Section 13.5. Roy (1957), Chapter 8, gives the cdfs for $p = 3$ and 4 also.

By Theorem 8.6.2 the limiting distribution of the largest characteristic root of nHG^{-1}, NHG^{-1}, $nH(H+G)^{-1}$, or $NH(H+G)^{-1}$ is the distribution of the largest characteristic root of H having the distribution $W(I, m)$. The densities of the roots of H are given in Section 13.3. In principle, the marginal density of the largest root can be obtained from the joint density by integration, but in actual fact the integration is more difficult than that for the density of the roots of HG^{-1} or $H(H+G)^{-1}$.

The literature on this subject is too extensive to summarize here. Nanda (1948) obtained the distribution for $p = 2$, 3, 4, and 5. Pillai (1954), (1956), (1965), (1967) treated the distribution under the null hypothesis. Other results were obtained by Sugiyama and Fukutomi (1966) and Sugiyama (1967). Pillai (1967) derived an appropriate distribution as a linear combination of incomplete beta functions. Davis (1972a) showed that the density of a single ordered root satisfies a differential equation and (1972b) derived a recurrence relation for it. Hayakawa (1967), Khatri and Pillai (1968), Pillai and Sugiyama (1969), and Khatri (1972) treated the noncentral case. See Krishnaiah (1978) for more references.

Tables. Tables of the percentage points have been calculated by Nanda (1951) and Foster and Rees (1957) for $p = 2$, Foster (1957) for $p = 3$, Foster (1958) for $p = 4$, and Pillai (1960) for $p = 2(1)6$ on the basis of an approximation. [See also Pillai (1956), (1960), (1964), (1965), (1967).] Heck (1960) presented charts of the significance points for $p = 2(1)6$. Table B.4 of significance points of nl_1/m is from *Concise Statistical Tables*, based on the approximation by Pillai (1967).

8.6.5. Comparison of Powers

The four tests that have been given most consideration are those based on Wilks's U, the Lawley–Hotelling W, the Bartlett–Nanda–Pillai V, and Roy's R. To guide in the choice of one of these four, we would like to compare power functions. The first three have been compared by Rothenberg on the basis of the asymptotic expansions of their distributions in the nonnull case.

Let $\nu_1^N, \dots, \nu_p^N$ be the roots of

$$|(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)A_{11\cdot2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)' - \nu\boldsymbol{\Sigma}| = 0. \tag{26}$$

The distribution of

$$\operatorname{tr}(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)A_{11\cdot2}(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)'\boldsymbol{\Sigma}^{-1} \tag{27}$$

is the noncentral χ^2-distribution with pm degrees of freedom and noncentrality parameter $\sum_{i=1}^{p} \nu_i^N$. As $N \to \infty$, the quantity $(1/n)G$ or $(1/N)G$ approaches $\boldsymbol{\Sigma}$ with probability one. If we let $N \to \infty$ and $A_{11 \cdot 2}$ is unbounded, the noncentrality parameter grows indefinitely and the power approaches 1. It is more informative to consider a sequence of alternatives such that the powers of the different tests are different. Suppose $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^N$ is a sequence of matrices such that as $N \to \infty$, $(\boldsymbol{\beta}_1^N - \boldsymbol{\beta}_1^*)A_{11 \cdot 2}(\boldsymbol{\beta}_1^N - \boldsymbol{\beta}_1^*)'$ approaches a limit and hence $\nu_1^N, \dots, \nu_p^N$ approach some limiting values $\nu_1, \dots, \nu_p$, respectively. Then the limiting distribution of $N \operatorname{tr} HG^{-1}$, $n \operatorname{tr} HG^{-1}$, $N \operatorname{tr} H(H+G)^{-1}$, and $n \operatorname{tr} H(H+G)^{-1}$ is the noncentral χ^2-distribution with pm degrees of freedom and noncentrality parameter $\sum_{i=1}^{p} \nu_i$. Similarly for $-N \log U$ and $-n \log U$.

Rothenberg (1977) has shown under the above conditions that

$$
(28) \quad \Pr\{U \le u_{p,m,n}(\alpha)\} = 1 - G_{pm}\left[\chi^2_{pm}(\alpha) \,\middle|\, \sum_{i=1}^{p} \nu_i\right]
- \frac{1}{2n}\left\{(p+m+1)\sum_{i=1}^{p} \nu_i g_{pm+4}\left[\chi^2_{pm}(\alpha)\right]
+ \sum_{i=1}^{p} \nu_i^2 g_{pm+6}\left[\chi^2_{pm}(\alpha)\right]\right\} + o\left(\frac{1}{n}\right),
$$

$$
(29) \quad \Pr\{\operatorname{tr} HG^{-1} \ge w_{p,m,n}(\alpha)\}
= 1 - G_{pm}\left[\chi^2_{pm}(\alpha) \,\middle|\, \sum_{i=1}^{p} \nu_i\right]
- \frac{1}{2n}\left\{(p+m+1)\sum_{i=1}^{p} \nu_i g_{pm+4}\left[\chi^2_{pm}(\alpha)\right]
+ \sum_{i=1}^{p} \nu_i^2 g_{pm+6}\left[\chi^2_{pm}(\alpha)\right]
- \left[\sum_{i=1}^{p} \nu_i^2 - \frac{p+m+1}{pm+2}\left(\sum_{i=1}^{p} \nu_i\right)^2\right] g_{pm+8}\left[\chi^2_{pm}(\alpha)\right]\right\} + o\left(\frac{1}{n}\right),
$$

$$
\begin{aligned}
(30)\quad \Pr\left\{\operatorname{tr} H(H+G)^{-1} \geq v_{p,m,n}(\alpha)\right\} &= 1 - G_{pm}\left[\chi^2_{pm}(\alpha)\,\middle|\,\sum_{i=1}^{p} \nu_i\right] \\
&\quad - \frac{1}{2n}\Bigg\{(p+m+1)\sum_{i=1}^{p} \nu_i g_{pm+4}\left[\chi^2_{pm}(\alpha)\right] \\
&\quad + \sum_{i=1}^{p} \nu_i^2 g_{pm+6}\left[\chi^2_{pm}(\alpha)\right] \\
&\quad + \left[\sum_{i=1}^{p} \nu_i^2 - \frac{p+m+1}{pm+2}\left(\sum_{i=1}^{p} \nu_i\right)^2\right] g_{pm+8}\left[\chi^2_{pm}(\alpha)\right]\Bigg\} + o\left(\frac{1}{n}\right),
\end{aligned}
$$

where $G_f(x|y)$ is the noncentral χ^2-distribution with f degrees of freedom and noncentrality parameter y, and $g_f(x)$ is the (central) χ^2-density with f degrees of freedom. The leading terms are the noncentral χ^2-distribution; the power functions of the three tests agree to this order. The power functions of the two trace tests differ from that of the likelihood ratio test by $\pm g_{pm+8}[\chi^2_{pm}(\alpha)]/(2n)$ times

$$
(31)\quad \sum_{i=1}^{p} \nu_i^2 - \frac{p+m+1}{pm+2}\left(\sum_{i=1}^{p} \nu_i\right)^2 = \sum_{i=1}^{p} (\nu_i - \bar{\nu})^2 - \frac{p(p-1)(p+2)}{pm+2}\bar{\nu}^2,
$$

where $\bar{\nu} = \sum_{i=1}^{p} \nu_i/p$. This is positive if

$$
(32)\quad \frac{\sigma_\nu}{\bar{\nu}} > \sqrt{\frac{(p-1)(p+2)}{pm+2}},
$$

where $\sigma_\nu^2 = \sum_{i=1}^{p}(\nu_i - \bar{\nu})^2/p$ is the (population) variance of $\nu_1, \dots, \nu_p$; the left-hand side of (32) is the coefficient of variation. If the ν_i's are relatively variable in the sense that (32) holds, the power of the Lawley–Hotelling trace test is greater than that of the likelihood ratio test, which in turn is greater than that of the Bartlett–Nanda–Pillai trace test (to order $1/n$); if the inequality (32) is reversed, the ordering of power is reversed.

The differences between the powers decrease as n increases for fixed $\nu_1, \dots, \nu_p$. (However, this comparison is not very meaningful, because increasing n decreases $\boldsymbol{\beta}_1^N - \boldsymbol{\beta}_1^*$ and increases $Z'Z$.)

A number of numerical comparisons have been made. Schatzoff (1966b) and Olson (1974) have used Monte Carlo methods; Mikhail (1965), Pillai and Jayachandran (1967), and Lee (1971a) have used asymptotic expansions of distributions. All of these results agree with Rothenberg's. Among these three procedures, the Bartlett–Nanda–Pillai trace test is to be preferred if the roots are roughly equal in the alternative, and the Lawley–Hotelling trace is more powerful when the roots are substantially unequal. Wilks's likelihood ratio test seems to come in second best; in a sense it is maximin.

As noted in Section 8.6.4, the Roy largest root has a limiting distribution which is not a χ^2-distribution under the null hypothesis and is not a noncentral χ^2-distribution under a sequence of alternative hypotheses. Hence the comparison of Rothenberg cannot be extended to this case. In fact, the distributions in the nonnull case are difficult to evaluate. However, the Monte Carlo results of Schatzoff (1966b) and Olson (1974) are clear-cut. The maximum root test has greatest power if the alternative is one-dimensional, that is, if $\nu_2 = \cdots = \nu_p = 0$. On the other hand, if the alternative is not one-dimensional, then the maximum root test is inferior.

These test procedures tend to be robust. Under the null hypothesis the limiting distribution of $\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*$ suitably normalized is normal with mean $\mathbf{0}$ and covariances the same as if X were normal, as long as its distribution satisfies some condition such as bounded fourth-order moments. Then $\hat{\boldsymbol{\Sigma}}_\Omega = (1/N)G$ converges with probability one. The limiting distribution of each criterion suitably normalized is the same as if X were normal. Olson (1974) studied the robustness under departures from covariance homogeneity as well as departures from normality. His conclusion was that the two trace tests and the likelihood ratio test were rather robust, and the maximum root test least robust. See also Pillai and Hsu (1979).

Berndt and Savin (1977) have noted that

$$\operatorname{tr} H(H+G)^{-1} \le \log U^{-1} \le \operatorname{tr} HG^{-1}. \tag{33}$$

(See Problem 8.19.) If the χ^2 significance point is used, then a larger criterion may lead to rejection while a smaller one may not.

8.7. TESTS OF HYPOTHESES ABOUT MATRICES OF REGRESSION COEFFICIENTS AND CONFIDENCE REGIONS

8.7.1. Testing Hypotheses

Suppose we are given a set of vector observations $x_1, \ldots, x_N$ with accompanying fixed vectors $z_1, \ldots, z_N$, where x_α is an observation from $N(\boldsymbol{\beta} z_\alpha, \boldsymbol{\Sigma})$. We let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 \; \boldsymbol{\beta}_2)$ and $z'_\alpha = (z_\alpha^{(1)\prime}, z_\alpha^{(2)\prime})$, where $\boldsymbol{\beta}_1$ and $z_\alpha^{(1)\prime}$ have q_1 $(= q - q_2)$ columns. The null hypothesis is

$$H: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*, \tag{1}$$

where $\boldsymbol{\beta}_1^*$ is a specified matrix. Suppose the desired significance level is α. A test procedure is to compute

$$U = \frac{|N\hat{\Sigma}_\Omega|}{|N\hat{\Sigma}_\omega|} \tag{2}$$

and compare this number with $u_{p,q_1,n}(\alpha)$, the α significance point of the $U_{p,q_1,n}$-distribution. For $p = 2, \dots, 10$ and even m, Table 1 in Appendix B can be used. For $m = 2, \dots, 10$ and even p the same table can be used with m replaced by p and p replaced by m. (M as given in the table remains unchanged.) For p and m both odd, interpolation between even values of either p or m will give sufficient accuracy for most purposes. For reasonably large n, the asymptotic theory can be used. An equivalent procedure is to calculate $\Pr\{U_{p,m,n} \le U\}$; if this is less than α, the null hypothesis is rejected.

Alternatively one can use the Lawley–Hotelling trace criterion

$$\begin{aligned} W &= \operatorname{tr}\left(N\hat{\Sigma}_\omega - N\hat{\Sigma}_\Omega\right)\left(N\hat{\Sigma}_\Omega\right)^{-1} \\ &= \operatorname{tr}\left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)A_{11\cdot 2}\left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)'\left(N\hat{\Sigma}_\Omega\right)^{-1}, \end{aligned} \tag{3}$$

the Pillai trace criterion

$$\begin{aligned} V &= \operatorname{tr}\left(N\hat{\Sigma}_\omega - N\hat{\Sigma}_\Omega\right)\left(N\hat{\Sigma}_\omega\right)^{-1} \\ &= \operatorname{tr}\left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)A_{11\cdot 2}\left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)'\left(N\hat{\Sigma}_\omega\right)^{-1}, \end{aligned} \tag{4}$$

or the Roy maximum root criterion R, where R is the maximum root of

$$|N\hat{\Sigma}_\omega - N\hat{\Sigma}_\Omega - rN\hat{\Sigma}_\Omega| = \left|\left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)A_{11\cdot 2}\left(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)' - rN\hat{\Sigma}_\Omega\right| = 0. \tag{5}$$

These criteria can be referred to the appropriate tables in Appendix B.

We outline an approach to computing the criterion. If we let $\boldsymbol{y}_\alpha = \boldsymbol{x}_\alpha - \boldsymbol{\beta}_1^* z_\alpha^{(1)}$, then $\boldsymbol{y}_\alpha$ can be considered as an observation from $N(\boldsymbol{\Delta} z_\alpha, \Sigma)$, where $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1 \;\; \boldsymbol{\Delta}_2) = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^* \;\; \boldsymbol{\beta}_2)$. Then the null hypothesis is $H: \boldsymbol{\Delta}_1 = \mathbf{0}$, and

$$\sum y_\alpha y_\alpha' = \sum x_\alpha x_\alpha' - \boldsymbol{\beta}_1^* C_1' - C_1 \boldsymbol{\beta}_1^{*\prime} + \boldsymbol{\beta}_1^* A_{11} \boldsymbol{\beta}_1^{*\prime}, \tag{6}$$

$$\sum y_\alpha z_\alpha' = C - \boldsymbol{\beta}_1^*\left(A_{11} \quad A_{12}\right). \tag{7}$$

Thus the problem of testing the hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ is equivalent to testing the hypothesis $\boldsymbol{\Delta}_1 = \mathbf{0}$, where $\mathscr{E} y_\alpha = \boldsymbol{\Delta} z_\alpha$. Hence let us suppose the problem is testing the hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$. Then $N\hat{\Sigma}_\omega = \sum x_\alpha x_\alpha' - \hat{\boldsymbol{\beta}}_{2\omega} A_{22} \hat{\boldsymbol{\beta}}_{2\omega}'$ and

$N\hat{\boldsymbol{\Sigma}}_\Omega = \Sigma \boldsymbol{x}_\alpha \boldsymbol{x}'_\alpha - \hat{\boldsymbol{\beta}}_\Omega A \hat{\boldsymbol{\beta}}'_\Omega$. We have discussed in Section 8.2.2 the computation of $\hat{\boldsymbol{\beta}}_\Omega A \hat{\boldsymbol{\beta}}'_\Omega$ and hence $N\hat{\boldsymbol{\Sigma}}_\Omega$. Then $\hat{\boldsymbol{\beta}}_{2\omega} A_{22} \hat{\boldsymbol{\beta}}'_{2\omega}$ can be computed in a similar manner If the method is laid out as

$$\begin{pmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}}'_{2\Omega} \\ \hat{\boldsymbol{\beta}}'_{1\Omega} \end{pmatrix} = \begin{pmatrix} C'_2 \\ C'_1 \end{pmatrix}, \tag{8}$$

the first q_2 rows and columns of A^* and of A^{**} are the same as the result of applying the forward solution to the left-hand side of

$$A_{22} \hat{\boldsymbol{\beta}}'_{2\omega} = C'_2, \tag{9}$$

and the first q_2 rows of $\bar{C}^*$ and $\bar{C}^{**}$ are the same as the result of applying the forward solution to the right-hand side of (9). Thus $\hat{\boldsymbol{\beta}}_{2\omega} A_{22} \hat{\boldsymbol{\beta}}'_{2\omega} = \bar{C}_2^* \bar{C}_2^{**\prime}$, where $\bar{C}^{*\prime} = (\bar{C}_2^{*\prime} \ \bar{C}_1^{*\prime})$ and $\bar{C}^{**\prime} = (\bar{C}_2^{**\prime} \ \bar{C}_1^{**\prime})$.

The method implies a method for computing a determinant. In Section A.5 of the Appendix it is shown that the result of the forward solution is $\boldsymbol{F}A = A^*$. Thus $|\boldsymbol{F}| \cdot |A| = |A^*|$. Since the determinant of a triangular matrix is the product of its diagonal elements, $|\boldsymbol{F}| = 1$ and $|A| = |A^*| = \prod_{i=1}^q a_{ii}^*$. This result holds for any positive definite matrix in place of A (with a suitable modification of $\boldsymbol{F}$) and hence can be used to compute $|N\hat{\boldsymbol{\Sigma}}_\Omega|$ and $|N\hat{\boldsymbol{\Sigma}}_\omega|$.

8.7.2. Confidence Regions Based on U

We have considered tests of hypotheses $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$, where $\boldsymbol{\beta}_1^*$ is specified. In the usual way we can deduce from the family of tests a confidence region for $\boldsymbol{\beta}_1$. From the theory given before, we know that the probability is $1-\alpha$ of drawing a sample so that

$$\frac{|N\hat{\boldsymbol{\Sigma}}_\Omega|}{|N\hat{\boldsymbol{\Sigma}}_\Omega + (\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1) A_{11\cdot 2} (\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1)'|} \ge u_{p,q_1,n}(\alpha). \tag{10}$$

Thus if we make the confidence-region statement that $\boldsymbol{\beta}_1$ satisfies

$$\frac{|N\hat{\boldsymbol{\Sigma}}_\Omega|}{|N\hat{\boldsymbol{\Sigma}}_\Omega + (\hat{\boldsymbol{\beta}}_{1\Omega} - \bar{\boldsymbol{\beta}}_1) A_{11\cdot 2} (\hat{\boldsymbol{\beta}}_{1\Omega} - \bar{\boldsymbol{\beta}}'_1)'|} \ge u_{p,q_1,n}(\alpha), \tag{11}$$

where (11) is interpreted as an inequality on $\boldsymbol{\beta}_1 = \bar{\boldsymbol{\beta}}_1$, then the probability is $1-\alpha$ of drawing a sample such that the statement is true.

Theorem 8.7.1. *The region* (11) *in the* $\bar{\boldsymbol{\beta}}_1$*-space is a confidence region for* $\boldsymbol{\beta}_1$ *with confidence coefficient* $1-\alpha$.

Usually the set of $\overline{\boldsymbol{\beta}}_1$ satisfying (11) is difficult to visualize. However, the inequality can be used to determine whether trial matrices are included in the region.

8.7.3. Simultaneous Confidence Intervals Based on the Lawley–Hotelling Trace

Each test procedure implies a set of confidence regions. The Lawley–Hotelling trace criterion can be used to develop simultaneous confidence intervals for linear combinations of elements of $\boldsymbol{\beta}_1$. A confidence region with confidence coefficient $1-\alpha$ is

$$\operatorname{tr}\left(\hat{\boldsymbol{\beta}}_{1\Omega}-\overline{\boldsymbol{\beta}}_1'\right)A_{11\cdot2}\left(\hat{\boldsymbol{\beta}}_{1\Omega}-\overline{\boldsymbol{\beta}}_1'\right)'\left(N\hat{\boldsymbol{\Sigma}}_\Omega\right)^{-1}\le w_{p,m,n}(\alpha). \tag{12}$$

To derive the confidence bounds we generalize Lemma 5.3.2.

Lemma 8.7.1. *For positive definite matrices A and G,*

$$|\operatorname{tr}\Phi'Y|\le\sqrt{\operatorname{tr}A^{-1}\Phi'G\Phi}\,\sqrt{\operatorname{tr}AY'G^{-1}Y}. \tag{13}$$

Proof. Let $b=\operatorname{tr}\Phi'Y/\operatorname{tr}A^{-1}\Phi'G\Phi$. Then

$$\begin{aligned} 0&\le\operatorname{tr}A(Y-bG\Phi A^{-1})'G^{-1}(Y-bG\Phi A^{-1})\\ &=\operatorname{tr}AY'G^{-1}Y-b\operatorname{tr}\Phi'Y-b\operatorname{tr}Y'\Phi+b^2\operatorname{tr}\Phi'G\Phi A^{-1}\\ &=\operatorname{tr}AY'G^{-1}Y-\frac{(\operatorname{tr}\Phi'Y)^2}{\operatorname{tr}A^{-1}\Phi'G\Phi}, \end{aligned} \tag{14}$$

which yields (13). ∎

Now (12) and (13) imply that

$$\begin{aligned} \left|\operatorname{tr}\Phi'\hat{\boldsymbol{\beta}}_{1\Omega}-\operatorname{tr}\Phi'\overline{\boldsymbol{\beta}}_1\right|&=\left|\operatorname{tr}\Phi'\left(\hat{\boldsymbol{\beta}}_{1\Omega}-\overline{\boldsymbol{\beta}}_1\right)\right|\\ &\le\sqrt{\operatorname{tr}A_{11\cdot2}^{-1}\Phi'N\hat{\boldsymbol{\Sigma}}_\Omega\Phi\cdot\operatorname{tr}A_{11\cdot2}\left(\hat{\boldsymbol{\beta}}_{1\Omega}-\overline{\boldsymbol{\beta}}_1\right)'\left(N\hat{\boldsymbol{\Sigma}}_\Omega\right)^{-1}\left(\hat{\boldsymbol{\beta}}_{1\Omega}-\overline{\boldsymbol{\beta}}_1\right)}\\ &\le\sqrt{\operatorname{tr}A_{11\cdot2}^{-1}\Phi'N\hat{\boldsymbol{\Sigma}}_\Omega\Phi}\,\sqrt{w_{p,m,n}(\alpha)} \end{aligned} \tag{15}$$

holds for all $p\times m$ matrices Φ. We assert that

$$\begin{aligned} \operatorname{tr}\Phi'\hat{\boldsymbol{\beta}}_{1\Omega}-\sqrt{N\operatorname{tr}A_{11\cdot2}^{-1}\Phi'\hat{\boldsymbol{\Sigma}}_\Omega\Phi}\,\sqrt{w_{p,m,n}(\alpha)}&\le\operatorname{tr}\Phi'\overline{\boldsymbol{\beta}}_1\\ &\le\operatorname{tr}\Phi'\hat{\boldsymbol{\beta}}_{1\Omega}+\sqrt{N\operatorname{tr}A_{11\cdot2}^{-1}\Phi'\hat{\boldsymbol{\Sigma}}_\Omega\Phi}\,\sqrt{w_{p,m,n}(\alpha)} \end{aligned} \tag{16}$$

holds for all Φ with confidence $1-\alpha$.

The confidence region (12) can be explored by use of (16) for various Φ. If $\phi_{ik}=1$ for some pair (I,K) and 0 for other elements, then (16) gives an interval for β_{IK}. If $\phi_{ik}=1$ for a pair (I,K), -1 for (I,L), and 0 otherwise, the interval pertains to $\beta_{IK}-\beta_{IL}$, the difference of coefficients of two independent variables. If $\phi_{ik}=1$ for a pair (I,K), -1 for (J,K), and 0 otherwise, one obtains an interval for $\beta_{IK}-\beta_{JK}$, the difference of coefficients for two dependent variables.

8.7.4. Simultaneous Confidence Intervals Based on the Roy Maximum Root Criterion

A confidence region with confidence $1-\alpha$ based on the maximum root criterion is

$$\text{(17)} \qquad \operatorname{ch}_1\big(\hat{\boldsymbol{\beta}}_{1\Omega}-\overline{\boldsymbol{\beta}}_1\big)A_{11\cdot 2}\big(\hat{\boldsymbol{\beta}}_{1\Omega}-\overline{\boldsymbol{\beta}}_1\big)'\big(N\hat{\boldsymbol{\Sigma}}_\Omega\big)^{-1}\le r_{p,m,n}(\alpha),$$

where $\operatorname{ch}_1(C)$ denotes the largest characteristic root of C. We can derive simultaneous confidence bounds from (17). From Lemma 5.3.2, we find for any vectors a and b

(18)

$$\begin{aligned}
\big[a'(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)b\big]^2 &= \big\{\big[(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)'a\big]'b\big\}^2\\
&\le \big[(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)'a\big]'A_{11\cdot 2}\big[(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)'a\big]\cdot b'A_{11\cdot 2}^{-1}b\\
&= \frac{a'(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)A_{11\cdot 2}(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)'a}{a'Ga}\cdot a'Ga\cdot b'A_{11\cdot 2}^{-1}b\\
&\le \operatorname{ch}_1\big[(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)A_{11\cdot 2}(\hat{\boldsymbol{\beta}}_{1\Omega}-\boldsymbol{\beta}_1)'G^{-1}\big]\cdot a'Ga\cdot b'A_{11\cdot 2}^{-1}b\\
&\le r_{p,m,n}(\alpha)\cdot a'Ga\cdot b'A_{11\cdot 2}^{-1}b
\end{aligned}$$

with probability $1-\alpha$; the second inequality follows from Theorem A.2.4 of the Appendix. Then a set of confidence intervals on all linear combinations $a'\boldsymbol{\beta}_1 b$ holding with confidence $1-\alpha$ is

$$\text{(19)} \quad a'\hat{\boldsymbol{\beta}}_{1\Omega}b-\sqrt{r_{p,m,n}(\alpha)\cdot a'Ga\cdot b'A_{11\cdot 2}^{-1}b}\le a'\overline{\boldsymbol{\beta}}_1 b$$
$$\le a'\hat{\boldsymbol{\beta}}_{1\Omega}b+\sqrt{r_{p,m,n}(\alpha)\cdot a'Ga\cdot b'A_{11\cdot 2}^{-1}b}.$$

The linear combinations are $a'\boldsymbol{\beta}_1 b=\sum_{i=1}^{p}\sum_{h=1}^{m}a_i\beta_{ih}b_h$. If $a_1=1$, $a_i=0$, $i\ne 1$, and $b_1=1$, $b_h=0$, $h\ne 1$, the linear combination is simply β_{11}. If $a_1=1$, $a_i=0$, $i\ne 1$, and $b_1=1$, $b_2=-1$, $b_h=0$, $h\ne 1,2$, the linear combination is $\beta_{11}-\beta_{12}$.

We can compare these intervals with (16) for $\Phi = ab'$, which is of rank 1. The term subtracted from and added to tr $\Phi'\hat{\beta}_{1\Omega} = a'\hat{\beta}_{1\Omega}b$ is the square root of

(20) $$w_{p,m,n}(\alpha) \cdot \operatorname{tr} A_{11\cdot 2}^{-1} ba'Gab' = w_{p,m,n}(\alpha) \cdot a'Ga \cdot b'A_{11\cdot 2}^{-1} b.$$

This is greater than the term subtracted and added to $a'\hat{\beta}_{1\Omega}b$ in (19) because $w_{p,m,n}(\alpha)$, pertaining to the sum of the roots, is greater than $r_{p,m,n}(\alpha)$, relating to one root. The bounds (16) hold for all $p \times m$ matrices Φ, while (19) holds only for matrices ab' of rank 1.

Mudholkar (1966) gives a very general method of constructing simultaneous confidence intervals based on symmetric gauge functions. Gabriel (1969) relates confidence bounds to simultaneous test procedures. Wijsman (1979) showed that under certain conditions the confidence sets based on the maximum root are smallest. [See also Wijsman (1980).]

8.8. TESTING EQUALITY OF MEANS OF SEVERAL NORMAL DISTRIBUTIONS WITH COMMON COVARIANCE MATRIX

In univariate analysis it is well known that many hypotheses can be put in the form of hypotheses concerning regression coefficients. The same is true for the corresponding multivariate cases. As an example we consider testing the hypothesis that the means of, say, q normal distributions with a common covariance matrix are equal.

Let $y_\alpha^{(i)}$ be an observation from $N(\mu^{(i)}, \Sigma)$, $\alpha = 1, \dots, N_i$, $i = 1, \dots, q$. The null hypothesis is

(1) $$H: \mu^{(1)} = \cdots = \mu^{(q)}.$$

To put the problem in the form considered earlier in this chapter, let

(2) $$X = (x_1 \; x_2 \cdots x_{N_1} \; x_{N_1+1} \cdots x_N) = \left(y_1^{(1)} \; y_2^{(2)} \cdots y_{N_1}^{(1)} \; y_1^{(1)} \cdots y_{N_q}^{(q)}\right)$$

with $N = N_1 + \cdots + N_q$. Let

(3) $$Z = (z_1 \quad z_2 \quad \cdots \quad z_{N_1} \quad z_{N_1+1} \quad \cdots \quad z_N)$$
$$= \begin{pmatrix} 1 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \end{pmatrix};$$

that is, $z_{i\alpha}=1$ if $N_1+\cdots+N_{i-1}<\alpha\le N_1+\cdots+N_i$, and $z_{i\alpha}=0$ otherwise, for $i=1,\dots,q-1$, and $z_{q\alpha}=1$ (all α). Let $\boldsymbol{\beta}=(\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2)$, where

$$\begin{aligned}\boldsymbol{\beta}_1&=\left(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(q)},\dots,\boldsymbol{\mu}^{(q-1)}-\boldsymbol{\mu}^{(q)}\right),\\ \boldsymbol{\beta}_2&=\boldsymbol{\mu}^{(q)}.\end{aligned}\tag{4}$$

Then $\boldsymbol{x}_\alpha$ is an observation from $N(\boldsymbol{\beta}z_\alpha,\boldsymbol{\Sigma})$, and the null hypothesis is $\boldsymbol{\beta}_1=\mathbf{0}$. Thus we can use the above theory for finding the criterion for testing the hypothesis.

We have

$$A=\sum_{\alpha=1}^{N}z_\alpha z_\alpha'=\begin{pmatrix}N_1&0&\cdots&0&N_1\\0&N_2&\cdots&0&N_2\\\vdots&\vdots&&\vdots&\vdots\\0&0&\cdots&N_{q-1}&N_{q-1}\\N_1&N_2&\cdots&N_{q-1}&N\end{pmatrix},\tag{5}$$

$$C=\sum_{\alpha=1}^{N}\boldsymbol{x}_\alpha z_\alpha'=\left(\sum_\alpha \boldsymbol{y}_\alpha^{(1)}\quad\sum_\alpha \boldsymbol{y}_\alpha^{(2)}\cdots\sum_\alpha \boldsymbol{y}_\alpha^{(q-1)}\quad\sum_{i,\alpha}\boldsymbol{y}_\alpha^{(i)}\right).\tag{6}$$

Here $A_{22}=N$ and $C_2=\sum_{i,\alpha}\boldsymbol{y}_\alpha^{(i)}$. Thus $\hat{\boldsymbol{\beta}}_{2\omega}=\sum_{i,\alpha}\boldsymbol{y}_\alpha^{(i)}\cdot(1/N)=\bar{\boldsymbol{y}}$, say, and

$$\begin{aligned}N\hat{\boldsymbol{\Sigma}}_\omega&=\sum_\alpha \boldsymbol{x}_\alpha\boldsymbol{x}_\alpha'-\bar{\boldsymbol{y}}N\bar{\boldsymbol{y}}'\\&=\sum_{i,\alpha}\boldsymbol{y}_\alpha^{(i)}\boldsymbol{y}_\alpha^{(i)\prime}-N\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}'\\&=\sum_{i,\alpha}\left(\boldsymbol{y}_\alpha^{(i)}-\bar{\boldsymbol{y}}\right)\left(\boldsymbol{y}_\alpha^{(i)}-\bar{\boldsymbol{y}}\right)'.\end{aligned}\tag{7}$$

For $\hat{\boldsymbol{\Sigma}}_\Omega$, we use the formula $N\hat{\boldsymbol{\Sigma}}_\Omega=\sum\boldsymbol{x}_\alpha\boldsymbol{x}_\alpha'-\hat{\boldsymbol{\beta}}_\Omega A\hat{\boldsymbol{\beta}}_\Omega'=\sum\boldsymbol{x}_\alpha\boldsymbol{x}_\alpha'-CA^{-1}C'$ Let

$$D=\begin{pmatrix}1&0&\cdots&0&0\\0&1&\cdots&0&0\\\vdots&\vdots&&\vdots&\vdots\\0&0&\cdots&1&0\\-1&-1&\cdots&-1&1\end{pmatrix};\tag{8}$$

then

$$D^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}. \tag{9}$$

Thus

$$\begin{aligned} CA^{-1}C' &= CD'D^{-1\prime}A^{-1}D^{-1}DC' \\ &= CD'(DAD')^{-1}DC' \\ &= \left(\sum_\alpha y_\alpha^{(1)} \cdots \sum_\alpha y_\alpha^{(q)}\right) \begin{pmatrix} N_1 & 0 & \cdots & 0 \\ 0 & N_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & N_q \end{pmatrix}^{-1} \begin{pmatrix} \sum_\alpha y_\alpha^{(1)\prime} \\ \vdots \\ \sum_\alpha y_\alpha^{(q)\prime} \end{pmatrix} \\ &= \sum_i \left(\sum_\alpha y_\alpha^{(i)} \frac{1}{N_i} \sum_\gamma y_\gamma^{(i)\prime} \right) \\ &= \sum_i N_i \bar{y}^{(i)} \bar{y}^{(i)\prime}, \end{aligned} \tag{10}$$

where $\bar{y}^{(i)} = (1/N_i)\Sigma_\alpha y_\alpha^{(i)}$. Thus

$$\begin{aligned} N\hat{\Sigma}_\Omega &= \sum_{i,\alpha} y_\alpha^{(i)} y_\alpha^{(i)\prime} - \sum_i N_i \bar{y}^{(i)} \bar{y}^{(i)\prime} \\ &= \sum_{i,\alpha} \left(y_\alpha^{(i)} - \bar{y}^{(i)}\right)\left(y_\alpha^{(i)} - \bar{y}^{(i)}\right)'. \end{aligned} \tag{11}$$

It will be seen that $\hat{\Sigma}_\omega$ is the estimator of Σ when $\mu^{(1)} = \cdots = \mu^{(q)}$ and $\hat{\Sigma}_\Omega$ is the weighted average of the estimators of Σ based on the separate samples.

When the null hypothesis is true, $|N\hat{\Sigma}_\Omega| / |N\hat{\Sigma}_\omega|$ is distributed as $U_{p,q-1,n}$, where $n = N - q$. Therefore, the rejection region at the α significance level is

$$\lambda = \frac{|N\hat{\Sigma}_\Omega|}{|N\hat{\Sigma}_\omega|} < u_{p,q-1,n}(\alpha). \tag{12}$$

The left-hand side of (12) is (11) of Section 8.3, and

$$
(13)\quad N\hat{\Sigma}_\omega - N\hat{\Sigma}_\Omega = \sum_{i,\alpha} y_\alpha^{(i)} y_\alpha^{(i)\prime} - N\bar{y}\bar{y}' - \left(\sum_{i,\alpha} y_\alpha^{(i)} y_\alpha^{(i)\prime} - \sum_i N_i \bar{y}^{(i)} \bar{y}^{(i)\prime} \right)
$$

$$
= \sum N_i (\bar{y}^{(i)} - \bar{y})(\bar{y}^{(i)} - \bar{y})' = \boldsymbol{H},
$$

as implied by (4) and (5) of Section 8.4. Here $\boldsymbol{H}$ has the distribution $W(\boldsymbol{\Sigma}, q-1)$. It will be seen that when $p = 1$, this test reduces to the usual F-test

$$
(14)\qquad \frac{\sum N_i (\bar{y}^{(i)} - \bar{y})^2}{(y_\alpha^{(i)} - \bar{y}^{(i)})^2} \cdot \frac{n}{q-1} > F_{q-1,n}(\alpha).
$$

We give an example of the analysis. The data are taken from Barnard's study of Egyptian skulls (1935). The 4 $(=q)$ populations are Late Predynastic $(i = 1)$, Sixth to Twelfth $(i = 2)$, Twelfth to Thirteenth $(i = 3)$, and Ptolemaic Dynasties $(i = 4)$. The 4 $(=p)$ measurements (i.e., components of $y_\alpha^{(i)}$) are maximum breadth, basialveolar length, nasal height, and basibregmatic height. The numbers of observations are $N_1 = 91$, $N_2 = 162$, $N_3 = 70$, $N_4 = 75$. The data are sumn arized as

$$
(15)\quad \left(\bar{y}^{(1)}\ \ \bar{y}^{(2)}\ \ \bar{y}^{(3)}\ \ \bar{y}^{(4)}\right)
= \begin{pmatrix}
133.582418 & 134.265432 & 134.371429 & 135.306667 \\
98.307692 & 96.462963 & 95.857143 & 95.040000 \\
50.835165 & 51.148148 & 50.100000 & 52.093333 \\
133.000000 & 134.882716 & 133.642857 & 131.466667
\end{pmatrix}.
$$

$$
(16)\quad N\hat{\Sigma}_\Omega
= \begin{pmatrix}
9661.997470 & 445.573301 & 1130.623900 & 2148.584210 \\
445.573301 & 9073.115027 & 1239.211990 & 2255.812722 \\
1130.623900 & 1239.211990 & 3938.320351 & 1271.054662 \\
2148.584210 & 2255.812722 & 1271.054662 & 8741.508829
\end{pmatrix}.
$$

From these data we find

$$
(17)\quad N\hat{\Sigma}_\omega
= \begin{pmatrix}
9785.178098 & 214.197666 & 1217.929248 & 2019.820216 \\
214.197666 & 9559.460890 & 1131.716372 & 2381.126040 \\
1217.929248 & 1131.716372 & 4088.731856 & 1133.473898 \\
2019.820216 & 2381.126040 & 1133.473898 & 9382.242720
\end{pmatrix}.
$$

We shall use the likelihood ratio test. The ratio of determinants is

$$(18) \qquad U = \frac{|N\hat{\Sigma}_\Omega|}{|N\hat{\Sigma}_\omega|} = \frac{2.4269054 \times 10^5}{2.9544475 \times 10^5} = 0.8214344.$$

Here $N = 398$, $n = 394$, $p = 4$, and $q = 4$. Thus $k = 393$. Since n is very large, we may assume $-k \log U_{4,3,394}$ is distributed as χ^2 with 12 degrees of freedom (when the null hypothesis is true). Here $-k \log U = 77.30$. Since the 1% point of the χ^2_{12}-distribution is 26.2, the hypothesis of $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} = \boldsymbol{\mu}^{(3)} = \boldsymbol{\mu}^{(4)}$ is rejected.†

8.9. MULTIVARIATE ANALYSIS OF VARIANCE

The univariate analysis of variance has a direct generalization for vector variables leading to an analysis of vector sums of squares (i.e., sums such as $\sum x_\alpha x'_\alpha$). In fact, in the preceding section this generalization was considered for an analysis of variance problem involving a single classification.

As another example consider a two-way layout. Suppose that we are interested in the question whether the column effects are zero. We shall review the analysis for a scalar variable and then show the analysis for a vector variable. Let Y_{ij}, $i = 1, \dots, r$, $j = 1, \dots, c$, be a set of rc random variables. We assume that

$$(1) \qquad \mathscr{E} Y_{ij} = \mu + \lambda_i + \nu_j, \qquad i = 1, \dots, r, \quad j = 1, \dots, c,$$

with the restrictions

$$(2) \qquad \sum_{i=1}^{r} \lambda_i = \sum_{j=1}^{c} \nu_j = 0,$$

that the variance of Y_{ij} is σ^2, and that the Y_{ij} are independently normally distributed. To test that column effects are zero is to test that

$$(3) \qquad \nu_j = 0, \qquad j = 1, \dots, c.$$

This problem can be treated as a problem of regression by the introduction

†The above computations were given by Bartlett (1947).

of dummy fixed variates. Let

$$
\begin{aligned}
(4)\qquad z_{00,ij} &= 1, && \\
z_{k0,ij} &= 1, && k=i, \\
&= 0, && k \neq i, \\
z_{0k,ij} &= 1, && k=j, \\
&= 0, && k \neq j.
\end{aligned}
$$

Then (1) can be written

$$
(5)\qquad \mathscr{E}Y_{ij} = \mu z_{00,ij} + \sum_{k=1}^{r} \lambda_k z_{k0,ij} + \sum_{k=1}^{c} \nu_k z_{0k,ij}.
$$

The hypothesis is that the coefficients of $z_{0k,ij}$ are zero. Since the matrix of fixed variates here,

$$
(6)\qquad \begin{pmatrix} z_{00,11} & \cdots & z_{00,rc} \\ z_{10,11} & \cdots & z_{10,rc} \\ z_{20,11} & \cdots & z_{20,rc} \\ \vdots & & \vdots \\ z_{0c,11} & \cdots & z_{0c,rc} \end{pmatrix},
$$

is singular (for example, row 00 is the sum of rows $10, 20, \dots, r0$), one must elaborate the regression theory. When one does, one finds that the test criterion indicated by the regression theory is the usual F-test of analysis of variance.

Let

$$
\begin{aligned}
Y_{..} &= \frac{1}{rc} \sum_{i,j} Y_{ij}, \\
(7)\qquad Y_{i.} &= \frac{1}{c} \sum_{j} Y_{ij}, \\
Y_{.j} &= \frac{1}{r} \sum_{i} Y_{ij},
\end{aligned}
$$

and let

$$
\begin{aligned}
a &= \sum_{i,j} (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2 \\
&= \sum_{i,j} Y_{ij}^2 - c\sum_i Y_{i.}^2 - r\sum_j Y_{.j}^2 + rcY_{..}^2, \\
b &= r\sum_j (Y_{.j} - Y_{..})^2 \\
&= r\sum_j Y_{.j}^2 - rcY_{..}^2 .
\end{aligned} \tag{8}
$$

Then the F-statistic is given by

$$
F = \frac{b}{a} \cdot \frac{(c-1)(r-1)}{c-1} . \tag{9}
$$

Under the null hypothesis, this has the F-distribution with $c-1$ and $(r-1)\cdot(c-1)$ degrees of freedom. The likelihood ratio criterion for the hypothesis is the $rc/2$ power of

$$
\frac{a}{a+b} = \frac{1}{1 + \{(c-1)/[(r-1)(c-1)]\}F} . \tag{10}
$$

Now let us turn to the multivariate analysis of variance. We have a set of p-dimensional random vectors Y_{ij}, $i = 1, \dots, r$, $j = 1, \dots, c$, with expected values (1), where $\boldsymbol{\mu}$, the $\boldsymbol{\lambda}$'s, and the $\boldsymbol{\nu}$'s are vectors, and with covariance matrix $\boldsymbol{\Sigma}$, and they are independently normally distributed. Then the same algebra may be used to reduce this problem to the regression problem. We define $Y_{..}, Y_{i.}, Y_{.j}$ by (7) and

$$
\begin{aligned}
A &= \sum_{i,j} (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})(Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})' \\
&= \sum_{i,j} Y_{ij}Y_{ij}' - c\sum_i Y_{i.}Y_{i.}' - r\sum_j Y_{.j}Y_{.j}' + rcY_{..}Y_{..}', \\
B &= r\sum_j (Y_{.j} - Y_{..})(Y_{.j} - Y_{..})' \\
&= r\sum_j Y_{.j}Y_{.j}' - rcY_{..}Y_{..}' .
\end{aligned} \tag{11}
$$

Table 8.1

	Varieties					
Location	M	S	V	T	P	Sums
UF	81	105	120	110	98	514
	81	82	80	87	84	414
W	147	142	151	192	146	778
	100	116	112	148	108	584
M	82	77	78	131	90	458
	103	105	117	140	130	595
C	120	121	124	141	125	631
	99	62	96	126	76	459
GR	99	89	69	89	104	450
	66	50	97	62	80	355
D	87	77	79	102	96	441
	68	67	67	92	94	338
Sums	616	611	621	765	659	3272
	517	482	569	655	572	2795

A statistic analogous to (10) is

$$\frac{|A|}{|A+B|}. \tag{12}$$

Under the null hypothesis, this has the distribution of U for $p, n = (r-1)\cdot(c-1)$ and $q_1 = c - 1$ given in Section 8.4. In order for A to be nonsingular (with probability 1), we must require $p \le (r-1)(c-1)$.

As an example we use data first published by Immer, Hayes, and Powers (1934), and later used by Fisher (1947a), by Yates and Cochran (1938), and by Tukey (1949). The first component of the observation vector is the barley yield in a given year; the second component is the same measurement made the following year. Column indices run over the varieties of barley, and row indices over the locations. The data are given in Table 8.1 [e.g., $\begin{smallmatrix}81\\81\end{smallmatrix}$ in the upper left-hand corner indicates a yield of 81 in each year of variety M in location UF]. The numbers along the borders are sums.

We consider the square of (147, 100) to be

$$\begin{pmatrix}147\\100\end{pmatrix}(147 \quad 100) = \begin{pmatrix}21{,}609 & 14{,}700\\ 14{,}700 & 10{,}000\end{pmatrix}.$$

Then

$$\sum_{i,j} Y_{ij} Y_{ij}' = \begin{pmatrix} 380{,}944 & 315{,}381 \\ 315{,}381 & 277{,}625 \end{pmatrix}, \tag{13}$$

$$\sum_{j} (6Y_{\cdot j})(6Y_{\cdot j})' = \begin{pmatrix} 2{,}157{,}924 & 1{,}844{,}346 \\ 1{,}844{,}346 & 1{,}579{,}583 \end{pmatrix}, \tag{14}$$

$$\sum_{i} (5Y_{i\cdot})(5Y_{i\cdot})' = \begin{pmatrix} 1{,}874{,}386 & 1{,}560{,}145 \\ 1{,}560{,}145 & 1{,}353{,}727 \end{pmatrix}, \tag{15}$$

$$(30Y_{\cdot\cdot})(30Y_{\cdot\cdot})' = \begin{pmatrix} 10{,}750{,}984 & 9{,}145{,}240 \\ 9{,}145{,}240 & 7.812{,}025 \end{pmatrix}. \tag{16}$$

Then the *error sum of squares* is

$$A = \begin{pmatrix} 3279 & 802 \\ 802 & 4017 \end{pmatrix}, \tag{17}$$

the *row sum of squares* is

$$5\sum_{j} (Y_{i\cdot} - Y_{\cdot\cdot})(Y_{i\cdot} - Y_{\cdot\cdot})' = \begin{pmatrix} 18{,}011 & 7.188 \\ 7{,}188 & 10{,}345 \end{pmatrix}, \tag{18}$$

and the *column sum of squares* is

$$B = \begin{pmatrix} 2788 & 2550 \\ 2550 & 2863 \end{pmatrix}. \tag{19}$$

The test criterion is

$$\frac{|A|}{|A+B|} = \frac{\begin{vmatrix} 3279 & 802 \\ 802 & 4017 \end{vmatrix}}{\begin{vmatrix} 6067 & 3352 \\ 3352 & 6880 \end{vmatrix}} = 0.4107. \tag{20}$$

This result is to be compared with the significant point for $U_{2,4,20}$. Using the result of Section 8.4, we see that

$$\frac{1 - \sqrt{0.4107}}{\sqrt{0.4107}} \cdot \frac{19}{4} = 2.66$$

is to be compared with the significance point of $F_{3,38}$. This is significant at the 5% level. Our data show that there are differerces between varieties.

Now let us see that each F-test in the univariate analysis of variance has analogous tests in the multivariate analysis of variance. In the linear hypothesis model for the univariate analysis of variance, one assumes that the random variables $Y_1, \dots, Y_N$ have expected values that are linear combinations of unknown parameters

$$\mathscr{E} Y_\alpha = \sum_g \beta_g z_{g\alpha}, \tag{21}$$

where the β's are the parameters and the z's are the known coefficients. The variables $\{Y_\alpha\}$ are assumed to be normally and independently distributed with common variance σ^2. In this model there are a set of linear combinations, say $\sum_{\alpha=1}^N \gamma_{i\alpha} Y_\alpha$, where the γ's are known, such that

$$a = \sum_{i=1}^n \left(\sum_{\alpha=1}^N \gamma_{i\alpha} Y_\alpha \right)^2 = \sum_{\alpha,\beta=1}^N d_{\alpha\beta} Y_\alpha Y_\beta \tag{22}$$

is distributed as $\sigma^2\chi^2$ with n degrees of freedom. There is another set of linear combinations, say $\sum_\alpha \phi_{g\alpha} Y_\alpha$, where the ϕ's are known, such that

$$b = \sum_{g=1}^m \left(\sum_{\alpha=1}^N \phi_{g\alpha} Y_\alpha \right)^2 = \sum_{\alpha,\beta=1}^N c_{\alpha\beta} Y_\alpha Y_\beta \tag{23}$$

is distributed as $\sigma^2\chi^2$ with m degrees of freedom when the null hypothesis is true and as σ^2 times a noncentral χ^2 when the null hypothesis is not true; and in either case b is distributed independently of a. Then

$$\frac{b}{a} \cdot \frac{n}{m} = \frac{\sum c_{\alpha\beta} Y_\alpha Y_\beta}{\sum d_{\alpha\beta} Y_\alpha Y_\beta} \cdot \frac{n}{m} \tag{24}$$

has the F-distribution with m and n degrees of freedom, respectively, when the null hypothesis is true. The null hypothesis is that certain β's are zero.

In the multivariate analysis of variance, $Y_1, \dots, Y_N$ are vector variables with p components. The expected value of Y_α is given by (21) where $\boldsymbol{\beta}_g$ is a vector of p parameters. We assume that the $\{Y_\alpha\}$ are normally and independently distributed with common covariance matrix $\boldsymbol{\Sigma}$. The linear combinations $\sum \gamma_{i\alpha} Y_\alpha$ can be formed for the vectors. Then

$$A = \sum_{i=1}^n \left(\sum_{\alpha=1}^N \gamma_{i\alpha} Y_\alpha \right) \left(\sum_{\alpha=1}^N \gamma_{i\alpha} Y_\alpha \right)' = \sum_{\alpha,\beta=1}^N d_{\alpha\beta} Y_\alpha Y_\beta' \tag{25}$$

has the distribution $W(\Sigma, n)$. When the null hypothesis is true,

$$B = \sum_{g=1}^{m} \left(\sum_{\alpha=1}^{N} \phi_{g\alpha} Y_\alpha \right) \left(\sum_{\alpha=1}^{N} \phi_{g\alpha} Y_\alpha \right)' = \sum_{\alpha, \beta=1}^{N} c_{\alpha\beta} Y_\alpha Y_\beta' \tag{26}$$

has the distribution $W(\Sigma, m)$, and B is indeper dent of A. Then

$$\frac{|A|}{|A+B|} = \frac{\left| \sum d_{\alpha\beta} Y_\alpha Y_\beta' \right|}{\left| \sum d_{\alpha\beta} Y_\alpha Y_\beta' + \sum c_{\alpha\beta} Y_\alpha Y_\beta' \right|} \tag{27}$$

has the $U_{p,m,n}$-distribution.

The argument for the distribution of a and b involves showing that $\mathscr{E}\sum_\alpha \gamma_{i\alpha} Y_\alpha = 0$ and $\mathscr{E}\sum_\alpha \phi_{g\alpha} Y_\alpha = 0$ when certain β's are equal to zero as specified by the null hypothesis (as identities in the unspecified β's). Clearly this argument holds for the vector case as well. Secondly, one argues, in the univariate case, that there is an orthogonal matrix $\Psi = (\psi_{\alpha\beta})$ such that when the transformation $Y_\beta = \sum_\alpha \psi_{\beta\alpha} Z_\alpha$ is made

$$\begin{aligned} a &= \sum_{\alpha, \beta, \gamma, \delta} d_{\alpha\beta} \psi_{\alpha\gamma} \psi_{\beta\delta} Z_\gamma Z_\delta = \sum_{\alpha=1}^{n} Z_\alpha^2, \\ b &= \sum_{\alpha, \beta, \gamma, \delta} c_{\alpha\beta} \psi_{\alpha\gamma} \psi_{\beta\delta} Z_\gamma Z_\delta = \sum_{\alpha=n+1}^{n+m} Z_\alpha^2. \end{aligned} \tag{28}$$

Because the transformation is orthogonal, the $\{Z_\alpha\}$ are independently and normally distributed with common variance σ^2. Since the Z_α, $\alpha = 1, \dots, n$, must be linear combinations of $\sum_\alpha \gamma_{i\alpha} Y_\alpha$ and since Z_α, $\alpha = n+1, \dots, n+m$, must be linear combinations of $\sum_\alpha \phi_{g\alpha} Y_\alpha$, they must have means zero (under the null hypothesis). Thus a/σ^2 and b/σ^2 have the stated independent χ^2-distributions.

In the multivariate case the transformation $Y_\beta = \sum_\alpha \psi_{\beta\alpha} Z_\alpha$ is used, where Y_β and Z_α are vectors. Then

$$\begin{aligned} A &= \sum_{\alpha, \beta, \gamma, \delta} d_{\alpha\beta} \psi_{\alpha\gamma} \psi_{\beta\delta} Z_\gamma Z_\delta' = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha', \\ B &= \sum_{\alpha, \beta, \gamma, \delta} c_{\alpha\beta} \psi_{\alpha\gamma} \psi_{\beta\delta} Z_\gamma Z_\delta' = \sum_{\alpha=n+1}^{n+m} Z_\alpha Z_\alpha' \end{aligned} \tag{29}$$

because it follows from (28) that $\sum_{\alpha,\beta} d_{\alpha\beta} \psi_{\alpha\gamma} \psi_{\beta\delta} = 1$, $\gamma = \delta \le n$, and $= 0$ otherwise, and $\sum_{\alpha,\beta} c_{\alpha\beta} \psi_{\alpha\gamma} \psi_{\beta\delta} = 1$, $n+1 \le \gamma = \delta \le n+m$, and $= 0$ otherwise. Since Ψ is orthogonal, the $\{Z_\alpha\}$ are independently normally distributed with covariance matrix Σ. The same argument shows $\mathscr{E}Z_\alpha = \mathbf{0}$, $\alpha = 1, \ldots, n+m$, under the null hypothesis. Thus A and B are independently distributed according to $W(\boldsymbol{\Sigma}, n)$ and $W(\boldsymbol{\Sigma}, m)$, respectively.

8.10. SOME OPTIMAL PROPERTIES OF TESTS

8.10.1. Admissibility of Invariant Tests

In this chapter we have considered several tests of a linear hypothesis which are invariant with respect to transformations that leave the null hypothesis invariant. We raise the question of which invariant tests are good tests. In particular we ask for admissible procedures, that is, procedures that cannot be improved on in the sense of smaller probabilities of Type I and/or Type II error. The competing tests are not necessarily invariant. Clearly, if an invariant test is admissible in the class of all tests, it is admissible in the class of invariant tests.

Testing the general linear hypothesis as treated here is a generalization of testing the hypothesis concerning one mean vector as treated in Chapter 5. The invariant procedures in Chapter 8 are generalizations of the T^2-test. One way of showing a procedure is admissible is to display a prior distribution on the parameters such that the Bayes procedure is a given test procedure. This approach requires some ingenuity in constructing the prior, but the verification of the property given the prior is straightforward. Problems 8.26 and 8.27 show that the Bartlett–Nanda–Pillai trace criterion V and Wilks's likelihood ratio criterion U yield admissible tests. The disadvantage of this approach to admissibility is that one must invent a prior distribution for each procedure; a general theorem does not cover many cases.

The other approach to admissibility is to apply Stein's theorem (Theorem 5.6.5), which yields general results. The invariant tests can be stated in terms of the roots of the determinantal equation

$$|H - \lambda(H + G)| = 0, \tag{1}$$

where $H = \hat{\boldsymbol{\beta}}_1 A_{11\cdot2} \hat{\boldsymbol{\beta}}_1' = W_1 W_1'$ and $G = N\hat{\boldsymbol{\Sigma}}_\Omega = W_3 W_3'$. There is also a matrix $\hat{\boldsymbol{\beta}}_2$ (or W_2) associated with the nuisance parameters $\boldsymbol{\beta}_2$. For convenience, we define the canonical form in the following notation. Let $W_1 = X$ $(p \times m)$, $W_2 = Y$ $(p \times r)$, $W_3 = Z$ $(p \times n)$, $\mathscr{E}X = \boldsymbol{\Xi}$, $\mathscr{E}Y = \boldsymbol{H}$, and $\mathscr{E}Z = \mathbf{0}$; the columns are independently normally distributed with covariance matrix $\boldsymbol{\Sigma}$. The null hypothesis is $\boldsymbol{\Xi} = \mathbf{0}$, and the alternative hypothesis is $\boldsymbol{\Xi} \neq \mathbf{0}$.

The usual tests are given in terms of the (nonzero) roots of

$$|XX' - \lambda(ZZ' + XX')| = |XX' - \lambda(U - YY')| = 0. \tag{2}$$

where $U = XX' + YY' + ZZ'$. Expect for roots that are identically zero, the roots of (2) coincide with the nonzero characteristic roots of $X'(U - YY')^{-1}X$. Let $V = (X, Y, U)$ and

$$M(V) = X'(U - YY')^{-1}X. \tag{3}$$

The vector of ordered characteristic roots of $M(V)$ is denoted by

$$(\lambda_1, \ldots, \lambda_m)' = \boldsymbol{\lambda}(M(V)), \tag{4}$$

where $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$. Since the inclusion of zero roots (when $m > p$) causes no trouble in the sequel, we assume that the tests depend on $\boldsymbol{\lambda}(M(V))$.

The admissibility of these tests can be stated in terms of the geometric characteristics of the acceptance regions. Let

$$\begin{aligned} R^m_{\leq} &= \{\boldsymbol{\lambda} \in R^m \mid \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0\}, \\ R^m_{+} &= \{\boldsymbol{\lambda} \in R^m \mid \lambda_1 \geq 0, \ldots, \lambda_m \geq 0\}. \end{aligned} \tag{5}$$

It seems reasonable that if a set of sample roots leads to acceptance of the null hypothesis, then a set of smaller roots would as well (Figure 8.2).

Definition 8.10.1. *A region $A \subset R^m_{\leq}$ is* monotone *if $\boldsymbol{\lambda} \in A$, $\boldsymbol{\nu} \in R^m_{\leq}$, and $\nu_i \leq \lambda_i$, $i = 1, \ldots, m$, imply $\boldsymbol{\nu} \in A$.*

Definition 8.10.2. *For $A \subset R^m_{\leq}$ the* extended region *A^* is*

$$A^* = \bigcup_{\pi} \{(x_{\pi(1)}, \ldots, x_{\pi(m)})' \mid x \in A\}, \tag{6}$$

where π ranges over all permutations of $(1, \ldots, m)$.

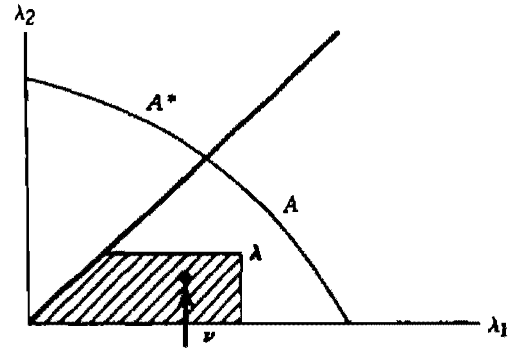

Figure 8.2. A monotone acceptance region.

The main result, first proved by Schwartz (1967), is the following theorem:

Theorem 8.10.1. *If the region $A \subset R^m_<$ is monotone and if the extended region A^* is closed and convex, then A is the acceptance region of an admissible test.*

Another characterization of admissible tests is given in terms of *majorization.*

Definition 8.10.3. *A vector* $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)'$ weakly majorizes *a vector* $\boldsymbol{\nu} = (\nu_1, \dots, \nu_m)'$ *if*

$$(7)\quad \lambda_{[1]} \geq \nu_{[1]},\ \lambda_{[1]} + \lambda_{[2]} \geq \nu_{[1]} + \nu_{[2]}, \dots,\ \lambda_{[1]} + \cdots + \lambda_{[m]} \geq \nu_{[1]} + \cdots + \nu_{[m]},$$

where $\lambda_{[i]}$ and $\nu_{[i]}$, $i = 1, \dots, m$, are the coordinates rearranged in nonascending order.

We use the notation $\boldsymbol{\lambda} \succ_w \boldsymbol{\nu}$ or $\boldsymbol{\nu} \prec_w \boldsymbol{\lambda}$ if $\boldsymbol{\lambda}$ weakly majorizes $\boldsymbol{\nu}$. If $\boldsymbol{\lambda}, \boldsymbol{\nu} \in \mathrm{R}^m_<$, then $\boldsymbol{\lambda} \succ_w \boldsymbol{\nu}$ is simply

$$(8)\quad \lambda_1 \geq \nu_1, \quad \lambda_1 + \lambda_2 \geq \nu_1 + \nu_2, \dots, \quad \lambda_1 + \cdots + \lambda_m \geq \nu_1 + \cdots + \nu_m.$$

If the last inequality in (7) is replaced by an equality, we say simply that $\boldsymbol{\lambda}$ *majorizes* $\boldsymbol{\nu}$ and denote this by $\boldsymbol{\lambda} \succ \boldsymbol{\nu}$ or $\boldsymbol{\nu} \prec \boldsymbol{\lambda}$. The theory of majorization and the related inequalities are developed in detail in Marshall and Olkin (1979).

Definition 8.10.4. *A region $A \subset R^m_<$ is* monotone in majorization *if $\boldsymbol{\lambda} \in A$, $\boldsymbol{\nu} \in R^m_<$, $\boldsymbol{\nu} \prec_w \boldsymbol{\lambda}$ imply $\boldsymbol{\nu} \in A$.* (See Figure 8.3.)

Theorem 8.10.2. *If a region $A \subset R^m_<$ is closed, convex, and monotone in majorization, then A is the acceptance region of an admissible test.*

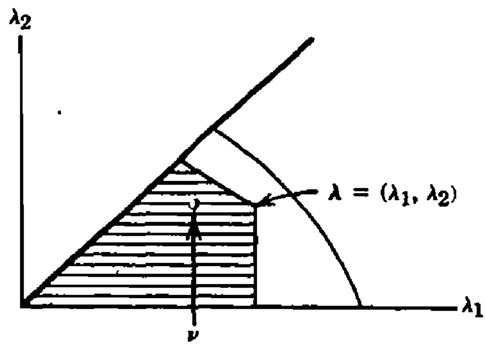

Figure 8.3. A region monotone in majorization.

Theorems 8.10.1 and 8.10.2 are equivalent; it will be convenient to prove Theorem 8.10.2 first. Then an argument about the extreme points of a certain convex set (Lemma 8.10.11) establishes the equivalence of the two theorems.

Theorem 5.6.5 (Stein's theorem) will be used because we can write the distribution of (X, Y, Z) in exponential form. Let $U = XX' + YY' + ZZ' = (u_{ij})$ and $\Sigma^{-1} = (\sigma^{ij})$. For a general matrix $C = (c_1, \dots, c_k)$, let $\mathrm{vec}(C) = (c_1', \dots, c_k')'$. The density of (X, Y, Z) can be written as

$$(9)\quad f(X,Y,Z) = K(\Xi, \mathrm{H}, \Sigma) \exp\{\mathrm{tr}\,\Xi'\Sigma^{-1}X + \mathrm{tr}\,\mathrm{H}'\Sigma^{-1}Y - \tfrac{1}{2}\mathrm{tr}\,\Sigma^{-1}U\}$$
$$= K(\Xi, \mathrm{H}, \Sigma)\exp\{\omega_{(1)}'y_{(1)} + \omega_{(2)}'y_{(2)} + \omega_{(3)}'y_{(3)}\},$$

where $K(X, \mathrm{H}, \Sigma)$ is a constant,

$$(10)\quad \begin{aligned} \omega_{(1)} &= \mathrm{vec}(\Sigma^{-1}\Xi), \qquad \omega_{(2)} = \mathrm{vec}(\Sigma^{-1}\mathrm{H}), \\ \omega_{(3)} &= -\tfrac{1}{2}(\sigma^{11}, 2\sigma^{12}, \dots, 2\sigma^{1p}, \sigma^{22}, \dots, \sigma^{pp})', \\ y_{(1)} &= \mathrm{vec}(X), \qquad y_{(2)} = \mathrm{vec}(Y), \\ y_{(3)} &= (u_{11}, u_{12}, \dots, u_{1p}, u_{22}, \dots, u_{pp})'. \end{aligned}$$

If we denote the mapping $(X, Y, Z) \to y = (y_{(1)}', y_{(2)}', y_{(3)}')'$ by g, $y = g(X, Y, Z)$, then the measure of a set A in the space of y is $m(A) = \mu(g^{-1}(A))$, where μ is the ordinary Lebesgue measure on $R^{p(m+r+n)}$. We note that (X, Y, U) is a sufficient statistic and so is $y = (y_{(1)}', y_{(2)}', y_{(3)}')'$. Because a test that is admissible with respect to the class of tests based on a sufficient statistic is admissible in the whole class of tests, we consider only tests based on a sufficient statistic. Then the acceptance regions of these tests are subsets in the space of y. The density of y given by the right-hand side of (9) is of the form of the exponential family, and therefore we can apply Stein's theorem. Furthermore, since the transformation $(X, Y, U) \to y$ is linear, we prove the convexity of an acceptance region of (X, Y, U). The acceptance region of an invariant test is given in terms of $\lambda(M(V)) = (\lambda_1, \dots, \lambda_m)'$. Therefore, in order to prove the admissibility of these tests we have to check that the inverse image of A, namely, $\tilde{A} = \{V \mid \lambda(M(V)) \in A\}$, satisfies the conditions of Stein's theorem, namely, is convex.

Suppose $V_i = (X_i, Y_i, U_i) \in \tilde{A}$, $i = 1, 2$, that is, $\lambda[M(V_i)] \in A$. By the convexity of A, $p\lambda[M(V_1)] + q\lambda[M(V_2)] \in A$ for $0 \le p = 1 - q \le 1$. To show $pV_1 + qV_2 \in \tilde{A}$, that is, $\lambda[M(pV_1 + qV_2)] \in A$, we use the property of monotonicity of majorization of A and the following theorem.

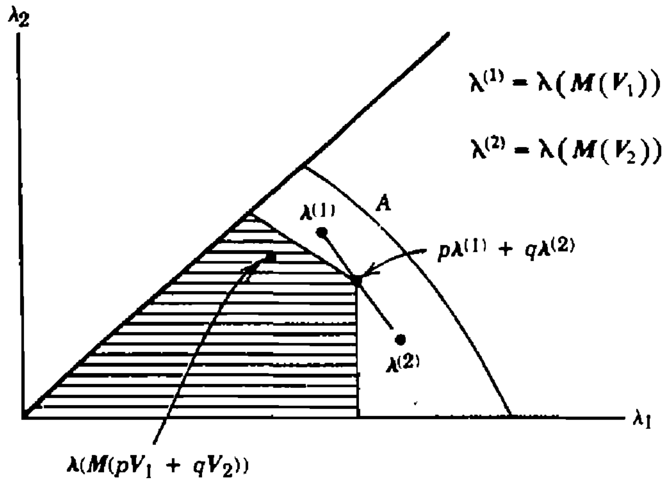

Figure 8.4. Theorem 8.10.3.

Theorem 8.10.3.

$$\lambda[M(pV_1 + gV_2)] \succ_w p\lambda[M(V_1)] + q\lambda[M(V_2)]. \tag{11}$$

The proof of Theorem 8.10.3 (Figure 8.4) follows from the pair of majorizations

$$\begin{aligned} \lambda[M(pV_1 + qV_2)] &\succ_w \lambda[pM(V_1) + qM(V_2)] \\ &\succ_w p\lambda[M(V_1)] + q\lambda[M(V_2)]. \end{aligned} \tag{12}$$

The second majorization in (12) is a special case of the following lemma.

Lemma 8.10.1. *For A and B symmetric,*

$$\lambda(A+B) \succ_w \lambda(A) + \lambda(B). \tag{13}$$

Proof. By Corollary A.4.2 of the Appendix,

$$\begin{aligned} \sum_{i=1}^{k} \lambda_i(A+B) &= \max_{R'R=I_k} \operatorname{tr} R'(A+B)R \\ &\le \max_{R'R=I_k} \operatorname{tr} R'AR + \max_{R'R=I_k} \operatorname{tr} R'BR \\ &= \sum_{i=1}^{k} \lambda_i(A) + \sum_{i=1}^{k} \lambda_i(B) \\ &= \sum_{i=1}^{k} \{\lambda_i(A) + \lambda_i(B)\}, \qquad k = 1, \dots, p. \end{aligned} \tag{14}$$

■

Let $A > B$ mean $A - B$ is positive definite and $A \geq B$ mean $A - B$ is positive semidefinite.

The first majorization in (12) follows from several lemmas.

Lemma 8.10.2

$$pU_1 + qU_2 - (pY_1 + qY_2)(pY_1 + qY_2)' \geq p(U_1 - Y_1Y_1') + q(U_2 - Y_2Y_2'). \tag{15}$$

Proof. The left-hand side minus the right-hand side is

$$\begin{aligned} &pY_1Y_1' + qY_2Y_2' - p^2Y_1Y_1' - q^2Y_2Y_2' - pq(Y_1Y_2' + Y_2Y_1') \\ &= p(1-p)Y_1Y_1' + q(1-q)Y_2Y_2' - pq(Y_1Y_2' + Y_2Y_1') \\ &= pq(Y_1 - Y_2)(Y_1 - Y_2)' \geq \mathbf{0}. \qquad \blacksquare \end{aligned} \tag{16}$$

Lemma 8.10.3. *If $A \geq B > \mathbf{0}$, then $A^{-1} \leq B^{-1}$.*

Proof. See Problem 8.31. ■

Lemma 8.10.4. *If $A > \mathbf{0}$, then $f(x, A) = x'A^{-1}x$ is convex in (x, A).*

Proof. See Problem 5.17. ■

Lemma 8.10.5. *If $A_1 > \mathbf{0}$, $A_2 > \mathbf{0}$, then*

$$(pB_1 + qB_2)'(pA_1 + qA_2)^{-1}(pB_1 + qB_2) \leq pB_1'A_1^{-1}B_1 + qB_2'A_2^{-1}B_2. \tag{17}$$

Proof. From Lemma 8.10.4 we have for all y

$$\begin{aligned} &py'B_1'A_1^{-1}B_1y + qy'B_2'A_2^{-1}B_2y \\ &\qquad - y'(pB_1 + qB_2)'(pA_1 + qA_2)^{-1}(pB_1 + qB_2)y \\ &= p(B_1y)'A_1^{-1}(B_1y) + q(B_2y)'A_2^{-1}(B_2y) \\ &\qquad - (pB_1y + qB_2y)'(pA_1 + qA_2)^{-1}(pB_1y + qB_2y) \\ &\geq 0. \qquad \blacksquare \end{aligned} \tag{18}$$

Thus the matrix of the quadratic form in y is positive semidefinite. ■

The relation as in (17) is sometimes called *matrix convexity*. [See Marshall and Olkin (1979).]

Lemma 8.10.6.

$$M(pV_1 + qV_2) \le pM(V_1) + qM(V_2), \tag{19}$$

where $V_1 = (X_1, Y_1, U_1)$, $V_2 = (X_2, Y_2, U_2)$, $U_1 - Y_1 Y_1' > 0$, $U_2 - Y_2 Y_2' > 0$, $0 \le p = 1 - q \le 1$.

Proof. Lemmas 8.10.2 and 8.10.3 show that

$$\begin{aligned} &\left[pU_1 + qU_2 - (pY_1 + qY_2)(pY_1 + qY_2)' \right]^{-1} \\ &\qquad \le \left[p(U_1 - Y_1 Y_1') + q(U_2 - Y_2 Y_2') \right]^{-1}. \end{aligned} \tag{20}$$

This implies

$$\begin{aligned} &M(pV_1 + qV_2) \\ &\qquad \le (pX_1 + qX_2)' \left[p(U_1 - Y_1 Y_1') + q(U_2 - Y_2 Y_2') \right]^{-1} (pX_1 + qX_2). \end{aligned} \tag{21}$$

Then Lemma 8.10.5 implies that the right-hand side of (21) is less than or equal to

$$pX_1'(U_1 - Y_1 Y_1')^{-1} X_1 + qX_2'(U_2 - Y_2 Y_2')^{-1} X_2 = pM(V_1) + qM(V_2). \tag{22}$$

■

Lemma 8.10.7. *If $A \le B$, then $\lambda(A) \prec_w \lambda(B)$.*

Proof. From Corollary A.4.2 of the Appendix,

$$\sum_{i=1}^{k} \lambda_i(A) = \max_{R'R = I_k} \operatorname{tr} R'AR \le \max_{R'R = I_k} \operatorname{tr} R'BR = \sum_{i=1}^{k} \lambda_i(B), \qquad k = 1, \dots, p. \tag{23}$$

■

From Lemma 8.10.7 we obtain the first majorization in (12) and hence Theorem 8.10.3, which in turn implies the convexity of $\tilde{A}$. Thus the acceptance region satisfies condition (i) of Stein's theorem.

Lemma 8.10.8. *For the acceptance region A of Theorem* 8.10.1 *or Theorem* 8.10.2, *condition* (ii) *of Stein's theorem is satisfied.*

Proof. Let $\boldsymbol{\omega}$ correspond to $(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Theta})$; then

$$\begin{aligned} \boldsymbol{\omega}' y &= \boldsymbol{\omega}_{(1)}' y_{(1)} + \boldsymbol{\omega}_{(2)}' y_{(2)} + \boldsymbol{\omega}_{(3)}' y_{(3)} \\ &= \operatorname{tr} \boldsymbol{\Phi}' X + \operatorname{tr} \boldsymbol{\Psi}' Y - \tfrac{1}{2} \operatorname{tr} \boldsymbol{\Theta} U, \end{aligned} \tag{24}$$

where Θ is symmetric. Suppose that $\{y \mid \omega' y > c\}$ is disjoint from $\tilde{A} = \{V \mid \lambda(M(V)) \in A\}$. We want to show that in this case Θ is positive semidefinite. If this were not true, then

$$\Theta = D \begin{bmatrix} I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & 0 \end{bmatrix} D', \tag{25}$$

where D is nonsingular and $-I$ is not vacuous. Let $X = (1/\gamma)X_0$, $Y = (1/\gamma)Y_0$,

$$U = (D')^{-1} \begin{pmatrix} I & 0 & 0 \\ 0 & \gamma I & 0 \\ 0 & 0 & I \end{pmatrix} D^{-1}, \tag{26}$$

and $V = (X, Y, U)$, where X_0, Y_0 are fixed matrices and γ is a positive number. Then

$$\omega' y = \frac{1}{\gamma} \operatorname{tr} \Phi' X_0 + \frac{1}{\gamma} \operatorname{tr} \Psi' Y_0 + \frac{1}{2} \operatorname{tr} \begin{pmatrix} -I & 0 & 0 \\ 0 & \gamma I & 0 \\ 0 & 0 & 0 \end{pmatrix} > c \tag{27}$$

for sufficiently large γ. On the other hand,

$$\begin{aligned} \lambda(M(V)) &= \lambda\{X'(U - YY')^{-1} X\} \\ &= \frac{1}{\gamma^2} \lambda \left\{ X_0' \left[(D')^{-1} \begin{pmatrix} I & 0 & 0 \\ 0 & \gamma I & 0 \\ 0 & 0 & I \end{pmatrix} D^{-1} - \frac{1}{\gamma^2} Y_0 Y_0' \right]^{-1} X_0 \right\} \\ &\to 0 \end{aligned} \tag{28}$$

as $\gamma \to \infty$. Therefore, $V \in \tilde{A}$ for sufficiently large γ. This is a contradiction. Hence Θ is positive semidefinite.

Now let ω_1 correspond to $(\Phi_1, 0, I)$, where $\Phi_1 \neq 0$. Then $I + \lambda\Theta$ is positive definite and $\Phi_1 + \lambda\Phi \neq 0$ for sufficiently large λ. Hence $\omega_1 + \lambda\omega \in \Omega - \Omega_0$ for sufficiently large λ. ∎

The preceding proof was suggested by Charles Stein.

By Theorem 5.6.5, Theorem 8.10.3 and Lemma 8.10.8 now imply Theorem 8.10.2.

To obtain Theorem 8.10.1 from Theorem 8.10.2, we use the following lemmas.

Lemma 8.10.9. *$A \subset R^m_{\leq}$ is convex and monotone in majorization if and only if A is monotone and A^* is convex.*

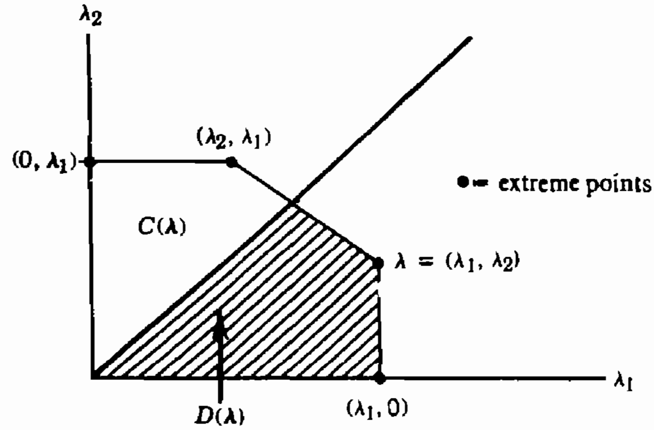

Figure 8.5

Proof. Necessity. If A is monotone in majorization, then it is obviously monotone. A^* is convex (see Problem 8.35).

Sufficiency. For $\lambda \in R^m_<$ let

$$C(\lambda) = \{x | x \in R^m_+, x \succ_w \lambda\},$$
$$D(\lambda) = \{x | x \in R^m_<, x \succ_w \lambda\}. \tag{29}$$

It will be proved in Lemma 8.10.10, Lemma 8.10.11, and its corollary that monotonicity of A and convexity of A^* implies $C(\lambda) \subset A^*$. Then $D(\lambda) = C(\lambda) \cap R^m_< \subset A^* \cap R^m_< = A$. Now suppose $\nu \in R^m_<$ and $\nu \prec_w \lambda$. Then $\nu \in D(\lambda) \subset A$. This shows that A is monotone in majorization. Furthermore, if A^* is convex, then $A = R^m_< \cap A^*$ is convex. (See Figure 8.5.) ■

Lemma 8.10.10. *Let C be compact and convex, and let D be convex. If the extreme points of C are contained in D, then $C \subset D$.*

Proof. Obvious. ■

Lemma 8.10.11. *Every extreme point of $C(\lambda)$ is of the form*

$$(\delta_{\pi(1)} \lambda_{\pi(1)}, \ldots, \delta_{\pi(m)} \lambda_{\pi(m)}), \tag{30}$$

where π is a permutation of $(1, \ldots, m)$ and $\delta_1 = \cdots = \delta_k = 1$, $\delta_{k+1} = \cdots = \delta_m = 0$ for some k.

Proof. $C(\lambda)$ is convex. (See Problem 8.34.) Now note that $C(\lambda)$ is permutation-symmetric, that is, if $(x_1, \ldots, x_m)' \in C(\lambda)$, then $(x_{\pi(1)}, \ldots, x_{\pi(m)})' \in C(\lambda)$ for any permutation π. Therefore, for any permutation π, $\pi(C(\lambda)) =$

$\{(x_{\pi(1)},\ldots,x_{\pi(m)})' | x \in C(\boldsymbol{\lambda})\}$ coincides with $C(\boldsymbol{\lambda})$. This implies that if $(x_1,\ldots,x_m)'$ is an extreme point of $C(\boldsymbol{\lambda})$, then $(x_{\pi(1)},\ldots,x_{\pi(m)})'$ is also an extreme point. In particular, $(x_{[1]},\ldots,x_{[m]}) \in R^m_<$ is an extreme point. Conversely, if $(x_1,\ldots,x_m) \in R^m_<$ is an extreme point of $C(\boldsymbol{\lambda})$, then $(x_{\pi(1)},\ldots,x_{\pi(m)})'$ is an extreme point.

We see that once we enumerate the extreme points of $C(\boldsymbol{\lambda})$ in $R^m_<$, the rest of the extreme points can be obtained by permutation.

Suppose $x \in R^m_<$. An extreme point, being the intersection of m hyperplanes, has to satisfy m or more of the following $2m$ equations:

$$
\begin{array}{lll}
 & E_1 : x_1 = 0, & F_1 : x_1 = \lambda_1, \\
 & E_2 : x_2 = 0, & F_2 : x_1 + x_2 = \lambda_1 + \lambda_2, \\
(31) & \vdots & \vdots \\
 & E_m : x_m = 0, & F_m : x_1 + \cdots + x_m = \lambda_1 + \cdots + \lambda_m.
\end{array}
$$

Suppose that k is the first index such that E_k holds. Then $x \in R^m_<$ implies $0 = x_k \geq x_{k+1} \geq \cdots \geq x_m \geq 0$. Therefore, $E_k,\ldots,E_m$ hold. The remaining $k-1 = m-(m-k+1)$ or more equations are among the F's. We order them as $F_{i_1},\ldots,F_{i_l}$, where $i_1 < \cdots < i_l$, $l \geq k-1$. Now $i_1 < \cdots < i_l$ implies $i_l \geq l$ with equality if and only if $i_1 = 1,\ldots,i_l = l$. In this case $F_1,\ldots,F_{k-1}$ hold ($l \geq k-1$). Now suppose $i_l > l$. Since $x_k = \cdots = x_m = 0$,

$$
(32) \qquad F_{i_l} : x_1 + \cdots + x_{k-1} = \lambda_1 + \cdots + \lambda_{k-1} + \cdots + \lambda_{i_l}.
$$

But $x_1 + \cdots + x_{k-1} \leq \lambda_1 + \cdots + \lambda_{k-1}$, and we have $\lambda_k + \cdots + \lambda_{i_l} = 0$. Therefore, $0 = \lambda_k + \cdots + \lambda_{i_l} \geq \lambda_k \geq \cdots \geq \lambda_m \geq 0$. In this case $F_{k-1},\ldots,F_m$ reduce to the same equation $x_1 + \cdots + x_{k-1} = \lambda_1 + \cdots + \lambda_{k-1}$. It follows that x satisfies $k-2$ more equations, which have to be $F_1,\ldots,F_{k-2}$. We have shown that in either case $E_k,\ldots,E_m,F_1,\ldots,F_{k-1}$ hold and this gives the point $\beta = (\lambda_1,\ldots,\lambda_{k-1},0,\ldots,0)$, which is in $R^m_< \cap C(\boldsymbol{\lambda})$. Therefore, β is an extreme point. ■

Corollary 8.10.1. $C(\boldsymbol{\lambda}) \subset A^*$.

Proof. If A is monotone, then A^* is monotone in the sense that if $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m)' \in A^*$, $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_m)'$, $\nu_i \leq \lambda_i$, $i = 1,\ldots,m$, then $\boldsymbol{\nu} \in A^*$. (See Problem 8.35.) Now the extreme points of $C(\boldsymbol{\lambda})$ given by (30) are in A^* because of permutation symmetry and monotonicity of A^*. Hence, by Lemma 8.10.10, $C(\boldsymbol{\lambda}) \subset A^*$. ■

Proof of Theorem 8.10.1. Immediate from Theorem 8.10.2 and Lemma 8.10.9. ■

Application of the theory of Schur-convex functions yields several corollaries to Theorem 8.10.2

Corollary 8.10.2. *Let g be continuous, nondecreasing, and convex in* [0, 1). *Let*

$$f(\boldsymbol{\lambda}) = f(\lambda_1, \dots, \lambda_m) = \sum_{i=1}^{m} g(\lambda_i). \tag{33}$$

Then a test with the acceptance region $A = \{\boldsymbol{\lambda} | f(\boldsymbol{\lambda}) \le c\}$ *is admissible.*

Proof. Being a sum of convex functions f is convex, and hence A is convex. A is closed because f is continuous. We want to show that if $f(x) \le c$ and $y \prec_w x$ $(x, y \in R_{\le}^m)$, then $f(y) \le c$. Let $\tilde{x}_k = \sum_{i=1}^{k} x_i$, $\tilde{y}_k = \sum_{i=1}^{k} y_i$. Then $y \prec_w x$ if and only if $\tilde{x}_k \ge \tilde{y}_k$, $k = 1, \dots, m$. Let $f(x) = h(\tilde{x}_1, \dots, \tilde{x}_m) = g(\tilde{x}_1) + \sum_{i=2}^{m} g(\tilde{x}_i - \tilde{x}_{i-1})$. It suffices to show that $h(\tilde{x}_1, \dots, \tilde{x}_m)$ is increasing in each $\tilde{x}_i$. For $i \le m - 1$ the convexity of g implies that

$$\begin{aligned} h(\tilde{x}_1, \dots, \tilde{x}_i + \varepsilon, \dots, \tilde{x}_m) &- h(\tilde{x}_1, \dots, \tilde{x}_i, \dots, \tilde{x}_m) \\ &= g(x_1 + \varepsilon) - g(x_i) - \{g(x_{i+1}) - g(x_{i+1} - \varepsilon)\} \ge 0. \end{aligned} \tag{34}$$

For $i = m$ the monotonicity of g implies

$$h(\tilde{x}_1, \dots, \tilde{x}_m + \varepsilon) - n(\tilde{x}_1, \dots, \tilde{x}_m) = g(x_m + \varepsilon) - g(x_m) \ge 0. \tag{35}$$ ■

Setting $g(\lambda) = -\log(1 - \lambda)$, $g(\lambda) = \lambda/(1 - \lambda)$, $g(\lambda) = \lambda$, respectively, shows that Wilks' likelihood ratio test, the Lawley–Hotelling trace test, and the Bartlett–Nanda–Pillai test are admissible. Admissibility of Roy's maximum root test $A: \lambda_1 \le c$ follows directly from Theorem 8.10.1 or Theorem 8.10.2. On the contrary, the minimum root test, $\lambda_t \le c$, where $t = \min(m, p)$, does not satisfy the convexity condition. The following theorem shows that this test is actually inadmissible.

Theorem 8.10.4. *A necessary condition for an invariant test to be admissible is that the extended region in the space of* $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_t}$ *is convex and monotone.*

We shall only sketch the proof of this theorem [following Schwartz (1967)]. Let $\sqrt{\lambda_i} = d_i$, $i = 1, \dots, t$, and let the density of $d_1, \dots, d_t$ be $f(\boldsymbol{d} | \boldsymbol{\nu})$, where $\boldsymbol{\nu} = (\nu_1, \dots, \nu_t)'$ is defined in Section 8.6.5 and $f(\boldsymbol{d} | \boldsymbol{\nu})$ is given in Chapter 13.

The ratio $f(d|\boldsymbol{\nu})/f(d|0)$ can be extended symmetrically to the unit cube $(0 \le d_i \le 1,\ i = 1, \dots, t)$. The extended ratio is then a convex function and is strictly increasing in each d_i. A proper Bayes procedure has an acceptance region

$$\int \frac{f(d|\boldsymbol{\nu})}{(d|0)}\, d\Pi(\boldsymbol{\nu}) \le c, \tag{36}$$

where $\Pi(\boldsymbol{\nu})$ is a finite measure on the space of $\boldsymbol{\nu}$'s. Then the symmetric extension of the set of $\boldsymbol{d}$ satisfying (36) is convex and monotone [as shown by Birnbaum (1955)]. The closure (in the weak* topology) of the set of Bayes procedures forms an essentially complete class [Wald (1950)]. In this case the limit of the convex monotone acceptance regions is convex and monotone. The exposition of admissibility here was developed by Anderson and Takemura (1982).

8.10.2. Unbiasedness of Tests and Monotonicity of Power Functions

A test T is called *unbiased* if the power achieves its minimum at the null hypothesis. When there is a natural parametrization and a notion of distance in the parameter space, the power function is *monotone* if the power increases as the distance between the alternative hypothesis and the null hypothesis increases. Note that monotonicity implies unbiasedness. In this section we shall show that the power functions of many of the invariant tests of the general linear hypothesis are monotone in the invariants of the parameters, namely, the roots; these can be considered as measures of distance.

To introduce the approach, we consider the acceptance interval $(-a, a)$ for testing the null hypothesis $\mu = 0$ against the alternative $\mu \ne 0$ on the basis of an observation from $N(\mu, \sigma^2)$. In Figure 8.6 the probabilities of acceptance are represented by the shaded regions for three values of μ. It is clear that the probability of acceptance decreases monotonically (or equivalently the power increases monotonically) as μ moves away from zero. In fact, this property depends only on the density function being unimodal and symmetric.

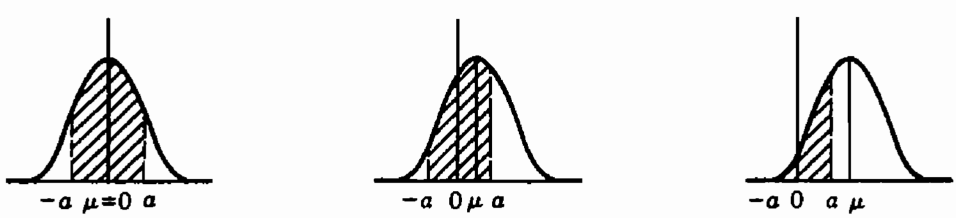

Figure 8.6. Three probabilities of acceptance.

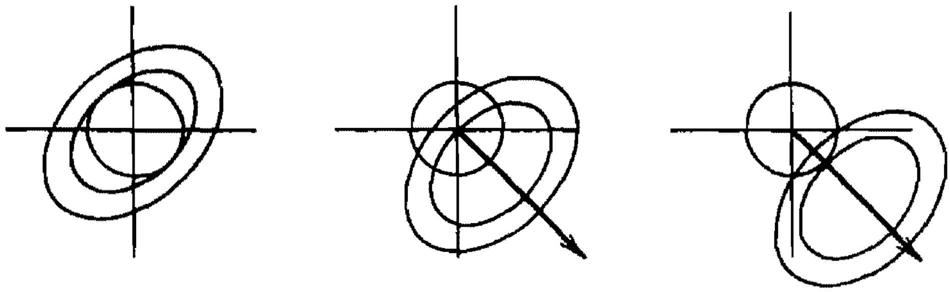

Figure 8.7. Acceptance regions.

In higher dimensions we generalize the interval by a symmetric convex set, and we ask that the density function be symmetric and unimodal in the sense that every contour of constant density surrounds a convex set. In Figure 8.7 we illustrate that in this case the probability of acceptance decreases monotonically. The following theorem is due to Anderson (1955b).

Theorem 8.10.5. *Let E be a convex set in n-space, symmetric about the origin. Let $f(x) \geq 0$ be a function such that* (i) $f(x) - f(-x)$, (ii) $\{x|f(x) \geq u\} = K_u$ *is convex for every u* $(0 < u < \infty)$, *and* (iii) $\int_E f(x)\,dx < \infty$. *Then*

$$\int_E f(x+ky)\,dx \geq \int_E f(x+y)\,dx \tag{37}$$

for $0 \leq k \leq 1$.

The proof of Theorem 8.10.5 is based on the following lemma.

Lemma 8.10.12. *Let E, F be convex and symmetric about the origin. Then*

$$V\{(E+ky) \cap F\} \geq V\{(E+y) \cap F\}, \tag{38}$$

where $0 \leq k \leq 1$ and V denotes the n-dimensional volume.

Proof. Consider the set $\alpha(E+y) + (1-\alpha)(E-y) = \alpha E + (1-\alpha)E + (2\alpha - 1)y$ which consists of points $\alpha(x+y) + (1-\alpha)(z-y)$ with $x, z \in E$ Let $\alpha_0 = (k+1)/2$, so that $2\alpha_0 - 1 = k$. Then by convexity of E we have

$$\alpha_0(E+y) + (1-\alpha_0)(E-y) \subset E + ky. \tag{39}$$

Hence by convexity of F

$$\alpha_0[(E+y) \cap F] + (1-\alpha_0)[(E-y) \cap F] \subset (E+ky) \cap F$$

and

$$V\{\alpha_0[(E+y)\cap F]+(1-\alpha_0)[(E-y)\cap F]\}\le V\{(E+ky)\cap F\}. \tag{40}$$

Now by the Brunn–Minkowski inequality [e.g., Bonnesen and Fenchel (1948), Section 48], we have

$$\begin{aligned}(41)\quad & V^{1/n}\{\alpha_0[(E+y)\cap F]+(1-\alpha_0)[(E-y)\cap F]\}\\ &\ge \alpha_0 V^{1/n}\{(E+y)\cap F\}+(1-\alpha_0)V^{1/n}\{(E-y)\cap F\}\\ &= \alpha_0 V^{1/n}\{(E+y)\cap F\}+(1-\alpha_0)V^{1/n}\{(-E+y)\cap(-F)\}\\ &= V^{1/n}\{(E+y)\cap F\}.\end{aligned}$$

The last equality follows from the symmetry of E and F. ■

Proof of Theorem 8.10.5. Let

$$H(u)=V\{(E+ky)\cap K_u\}, \tag{42}$$

$$H^*(u)=V\{(E+y)\cap K_u\}. \tag{43}$$

Then

$$\begin{aligned}(44)\quad \int_E f(x+y)\,dx &= \int_{E+y} f(x)\,dx\\ &= \int_{E+y}\int_0^\infty I_{\{0\le u\le f(x)\}}(u)\,du\,dx\\ &= \int_0^\infty\int_{E+y} I_{\{0\le u\le f(x)\}}(u)\,dx\,du\\ &= \int_0^\infty H^*(u)\,du.\end{aligned}$$

Similarly,

$$\int_E f(x+ky)\,dx=\int_0^\infty H(u)\,du. \tag{45}$$

By Lemma 8.10.12, $H(u)\ge H^*(u)$. Hence Theorem 8.10.5 follows from (44) and (45). ■

We start with the canonical form given in Section 8.10.1. We further simplify the problem as follows. Let $t = \min(m, p)$, and let $\nu_1, \dots, \nu_t$ ($\nu_1 \geq \nu_2 \geq \cdots \geq \nu_t$) be the nonzero characteristic roots of $\Xi' \Sigma^{-1} \Xi$, where $\Xi = \mathscr{E} X$.

Lemma 8.10.13. *There exist matrices B ($p \times p$) and F ($m \times m$) such that*

$$
\begin{aligned}
& B \Sigma B' = I_p, \qquad FF' = I_m, \\
(46) \qquad & B \Xi F' = \left(D_\nu^{\frac{1}{2}}, 0\right), && p \leq m, \\
& \qquad = \begin{pmatrix} D_\nu^{\frac{1}{2}} \\ 0 \end{pmatrix}, && p > m,
\end{aligned}
$$

where $D_\nu = \operatorname{diag}(\nu_1, \dots, \nu\ )$.

Proof. We prove this for the case $p \leq m$ and $\nu_p > 0$. Other cases can be proved similarly. By Theorem A.2.2 of the Appendix there is a matrix B such that

$$(47) \qquad B \Sigma B' = I, \qquad B \Xi \Xi' B' = D_\nu.$$

Let

$$(48) \qquad F_1 = D_\nu^{-\frac{1}{2}} B \Xi \quad (p \times m).$$

Then

$$(49) \qquad F_1 F_1' = I_p.$$

Let $F' = (F_1', F_2')$ be a full $m \times m$ orthogonal matrix. Then

$$(50) \qquad B \Xi F_2' = D_\nu^{\frac{1}{2}} F_1 F_2' = 0$$

and

$$(51) \qquad B \Xi F' = B \Xi (F_1', F_2') = B \Xi \left(\Xi' B' D_\nu^{-\frac{1}{2}}, F_2'\right) = \left(D_\nu^{\frac{1}{2}}, 0\right). \qquad \blacksquare$$

Now let

$$(52) \qquad U = BXF', \qquad V = BZ.$$

Then the columns of U, V are independently normally distributed with covariance matrix I and means when $p \leq m$

$$
\begin{aligned}
(53) \qquad & \mathscr{E} U = \left(D_\nu^{\frac{1}{2}}, 0\right), \\
& \mathscr{E} V = 0.
\end{aligned}
$$

Invariant tests are given in terms of characteristic roots $l_1, \dots, l_t$ $(l_1 \ge \cdots \ge l_t)$ of $U'(VV')^{-1}U$. Note that for the admissibility we used the characteristic roots of λ_i of $U'(UU' + VV')^{-1}U$ rather than $l_i = \lambda_i/(1-\lambda_i)$. Here it is more natural to use l_i, which corresponds to the parameter value ν_i. The following theorem is given by Das Gupta, Anderson, and Mudholkar (1964).

Theorem 8.10.6. *If the acceptance region of an invariant test is convex in the space of each column vector of U for each set of fixed values of V and of the other column vectors of U, then the power of the test increases monotonically in each ν_i.*

Proof. Since UU' is unchanged when any column vector of U is multiplied by -1, the acceptance region is symmetric about the origin in each of the column vectors of U. Now the density of $U = (u_{ij})$, $V = (v_{ij})$ is

$$(54)\quad f(U,V)$$
$$= (2\pi)^{-\frac{1}{2}(n+m)p} \exp\left[-\frac{1}{2}\left\{ \operatorname{tr} VV' + \sum_{i=1}^{t} \left(u_{ii} - \sqrt{\nu_i}\right)^2 + \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j \ne i}}^{m} u_{ij}^2 \right\} \right].$$

Applying Theorem 8.10.5 to (54), we see that the power increases monotonically in each $\sqrt{\nu_i}$. ■

Since the section of a convex set is convex, we have the following corollary.

Corollary 8.10.3. *If the acceptance region A of an invariant test is convex in U for each fixed V, then the power of the test increases monotonically in each ν_i.*

From this we see that Roy's maximum root test $A: l_1 \le K$ and the Lawley–Hotelling trace test $A: \operatorname{tr} U'(VV')^{-1}U \le K$ have power functions that are monotonically increasing in each ν_i.

To see that the acceptance region of the likelihood ratio test

$$(55)\qquad A: \prod_{i=1}^{t} (1 + l_i) \le K$$

satisfies the condition of Theorem 8.10.6 let

$$(56)\qquad \begin{aligned} (VV')^{-1} &= T'T, \qquad T: p \times p \\ U^* &= (u_1^*, \dots, u_m^*) = TU. \end{aligned}$$

Then

$$\prod_{i=1}^{t}(1+l_i)=\left|U'(VV')^{-1}U+I\right|=\left|U^{*\prime}U^*+I\right| \tag{57}$$

$$=\left|U^*U^{*\prime}+I\right|=\left|u_1^*u_1^{*\prime}+B\right|$$

$$=\left(u_1^{*\prime}B^{-1}u_1^*+1\right)|B|$$

$$=\left(u_1'T'B^{-1}Tu_1+1\right)|B|,$$

where $B=u_2^*u_2^{*\prime}+\cdots+u_m^*u_m^{*\prime}+I$. Since $T'B^{-1}T$ is positive definite, (55) is convex in u_1. Therefore, the likelihood ratio test has a power function which is monotone increasing in each ν_i.

The Bartlett–Nanda–Pillai trace test

$$A: \operatorname{tr} U'(UU'+VV')^{-1}U=\sum_{i=1}^{t}\frac{l_i}{1+l_i}\le K \tag{58}$$

has an acceptance region that is an ellipsoid if $K<1$ and is convex in each column u_i of U *provided* $K\le 1$. (See Problem 8.36.) For $K>1$ (58) may not be convex in each column of U. The reader can work out an example for $p=2$.

Eaton and Perlman (1974) have shown that if an invariant test is convex in U and $W=VV'$, then the power at $(\nu_1^0,\dots,\nu_t^0)$ is greater than at $(\nu_1,\dots,\nu_t)$ if $(\sqrt{\nu_1},\dots,\sqrt{\nu_t})\prec_w(\sqrt{\nu_1^0},\dots,\sqrt{\nu_t^0})$. We shall not prove this result. Roy's maximum root test and the Lawley–Hotelling trace test satisfy the condition, but the likelihood ratio and the Bartlett–Nanda–Pillai trace test do not.

Takemura has shown that if the acceptance region is convex in U and W, the set of $\sqrt{\nu_1},\dots,\sqrt{\nu_t}$ for which the power is not greater than a constant is monotone and convex.

It is enlightening to consider the contours of the power function, $\Pi(\sqrt{\nu_1},\dots,\sqrt{\nu_t})$. Theorem 8.10.6 does not exclude case (a) of Figure 8.8.

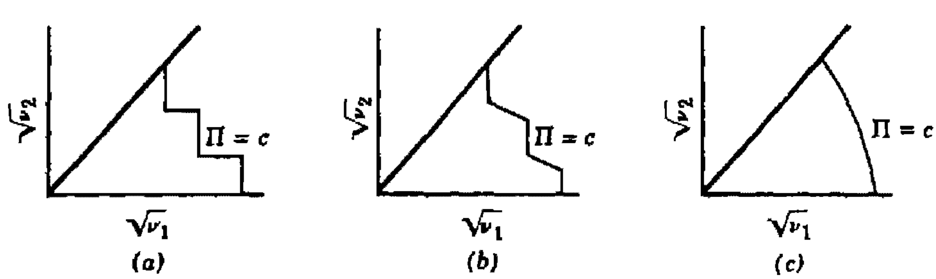

Figure 8.8. Contours of power functions.

and similarly the Eaton–Perlman result does not exclude (b). The last result guarantees that the contour looks like (c) for Roy's maximum root test and the Lawley–Hotelling trace test. These results relate to the fact that these two tests are more likely to detect alternative hypotheses where few $\boldsymbol{\nu}_i$'s are far from zero. In contrast with this, the likelihood ratio test and the Bartlett–Nanda–Pillai trace test are sensitive to the overall departure from the null hypothesis. It might be noted that the convexity in $\sqrt{\nu}$-space cannot be translated into the convexity in ν-space.

By using the noncentral density of l_i's which depends on the parameter values $\nu_1, \dots, \nu_t$, Perlman and Olkin (1980) showed that any invariant test with monotone acceptance region (in the space of roots) is unbiased. Note that this result covers all the standard tests considered earlier.

8.11. ELLIPTICALLY CONTOURED DISTRIBUTIONS

8.11.1. Observations Elliptically Contoured

The regression model of Section 8.2 can be written

$$x_\alpha = \boldsymbol{\beta} z_\alpha + e_\alpha, \qquad \alpha = 1, \dots, N, \tag{1}$$

where e_α is an unobserved disturbance with $\mathscr{E} e_\alpha = \mathbf{0}$ and $\mathscr{E} e_\alpha e_\alpha' = \boldsymbol{\Sigma}$. We assume that e_α has a density $|\boldsymbol{\Lambda}|^{-\frac{1}{2}} g(e' \boldsymbol{\Lambda}^{-1} e)$; then $\boldsymbol{\Sigma} = (\mathscr{E} R^2/p)\boldsymbol{\Lambda}$, where $R^2 = e_\alpha' \boldsymbol{\Lambda}^{-1} e_\alpha$. In general the exact distribution of $\boldsymbol{B} = \sum_{\alpha=1}^N x_\alpha z_\alpha' A^{-1}$ and $N\hat{\boldsymbol{\Sigma}} = \sum_{\alpha=1}^N (x_\alpha - Bz_\alpha)(x_\alpha - Bz_\alpha)'$ is difficult to obtain and cannot be expressed concisely. However, the expected value of B is $\boldsymbol{\beta}$, and the covariance matrix of $\operatorname{vec} \boldsymbol{B}$ is $\boldsymbol{\Sigma} \otimes A^{-1}$ with $A = \sum_{\alpha=1}^N z_\alpha z_\alpha'$. We can develop a large-sample distribution for $\boldsymbol{B}$ and $N\hat{\boldsymbol{\Sigma}}$.

Theorem 8.11.1. *Suppose $(1/N)A \to A_0$, $z_\alpha' z_\alpha <$ constant, $\alpha = 1, 2, \dots$, and either the e_α's are independent identically distributed or the e_α's are independent with $\mathscr{E}|e_\alpha' e_\alpha|^{2+\varepsilon} <$ constant for some $\varepsilon > 0$. Then $\boldsymbol{B} \xrightarrow{p} \boldsymbol{\beta}$ and $\sqrt{N} \operatorname{vec}(\boldsymbol{B} - \boldsymbol{\beta})$ has a limiting normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} \otimes A_0^{-1}$.*

Theorem 8.11.1 appears in Anderson (1971) as Theorem 5.5.13. There are many alternatives to its assumptions in the literature. Under its assumptions $\hat{\boldsymbol{\Sigma}}_\Omega \xrightarrow{p} \boldsymbol{\Sigma}$. This result permits a large-sample theory for the criteria for testing null hypotheses about $\boldsymbol{\beta}$.

Consider testing the null hypothesis

$$H: \boldsymbol{\beta} = \boldsymbol{\beta}^*, \tag{2}$$

where $\boldsymbol{\beta}^*$ is completely specified. In Section 8.3 a more general hypothesis was considered for $\boldsymbol{\beta}$ partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. However, as shown in that section by the transformation (4), the hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ can be reduced to a hypothesis of the form (1) above.

Let

$$G = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \boldsymbol{B}\boldsymbol{z}_\alpha)(\boldsymbol{x}_\alpha - \boldsymbol{B}\boldsymbol{z}_\alpha)' = N\hat{\boldsymbol{\Sigma}}_\Omega, \tag{3}$$

$$H = (\boldsymbol{B} - \boldsymbol{\beta})A(\boldsymbol{B} - \boldsymbol{\beta})'. \tag{4}$$

Lemma 8.11.1. *Under the conditions of Theorem* 8.11.1 *the limiting distribution of* H *is* $W(\boldsymbol{\Sigma}, q)$.

Proof. Write H as

$$H = \sqrt{N}(\boldsymbol{B} - \boldsymbol{\beta})\frac{1}{N}A\sqrt{N}(\boldsymbol{B} - \boldsymbol{\beta})'. \tag{5}$$

Then the lemma follows from Theorem 8.11.1 and (4) of Section 8.4. ∎

We can express the likelihood ratio criterion in the form

$$\begin{aligned} -2\log\lambda &= -N\log U = N\log|I + G^{-1}H| \\ &= N\log\left|I + \frac{1}{N}\left(\frac{1}{N}G\right)^{-1}H\right|. \end{aligned} \tag{6}$$

Theorem 8.11.2. *Under the conditions of Theorem* 8.11.1, *when the null hypothesis is true,*

$$-2\log\lambda \xrightarrow{d} \chi^2_{pq}. \tag{7}$$

Proof. We use the fact that $N\log|I + N^{-1}C| = \operatorname{tr} C + O_p(N^{-1})$ when $N \to \infty$, since $|I + xC| = 1 + x\operatorname{tr} C + O(x^2)$ (Theorem A.4.8).

We have

$$\begin{aligned} \operatorname{tr}\left(\frac{1}{N}G\right)^{-1}H &= N\sum_{i,j=1}^{p}\sum_{g,h=1}^{q} g^{ij}(b_{ig} - \beta_{ig})a_{gh}(b_{jh} - \beta_{jh}) \\ &= [\operatorname{vec}(B' - \boldsymbol{\beta}')]'\left(\frac{1}{N}G^{-1} \otimes A\right)\operatorname{vec}(B' - \boldsymbol{\beta}') \xrightarrow{d} \chi^2_{pq} \end{aligned} \tag{8}$$

because $(1/N)G \xrightarrow{p} \boldsymbol{\Sigma}$, $(1/N)A \to A_0$, and the limiting distribution of $\sqrt{N}\operatorname{vec}(B' - \boldsymbol{\beta}')$ is $N(\boldsymbol{\Sigma} \otimes A_0^{-1})$. ∎

Theorem 8.11.2 agrees with the first term of the asymptotic expansion of $-2\log\lambda$ given by Theorem 8.5.2 for sampling from a normal distribution. The test and confidence procedures discussed in Sections 8.3 and 8.4 can be applied using this χ^2-distribution.

The criterion $U=\lambda^{2/N}$ can be written as $U=\prod_{i=1}^{p}V_i$, where V_i is defined in (8) of Section 8.4. The term V_i has the form of U; that is, it is the ratio of the sum of squares of residuals of $x_{i\alpha}$ regressed on $x_{1\alpha},\dots,x_{i-1,\alpha},z_\alpha$ to the sum regressed on $x_{1\alpha},\dots,x_{i-1,\alpha}$. It follows that under the null hypothesis $V_1,\dots,V_p$ are asymptotically independent and $-N\log V_i\xrightarrow{d}\chi_q^2$. Thus $-N\log U=-N\sum_{i=1}^{p}\log V_i\xrightarrow{d}\chi_{pq}^2$. This argument justifies the step-down procedure asymptotically.

Section 8.6 gave several other criteria for the general linear hypothesis: the Lawley–Hotelling trace $\operatorname{tr}\boldsymbol{HG}^{-1}$, the Bartlett–Nanda–Pillai trace $\operatorname{tr}\boldsymbol{H}(\boldsymbol{G}+\boldsymbol{H})^{-1}$, and the Roy maximum root of $\boldsymbol{HG}^{-1}$ or $\boldsymbol{H}(\boldsymbol{G}+\boldsymbol{H})^{-1}$. The limiting distributions of $N\operatorname{tr}\boldsymbol{HG}^{-1}$ and $N\operatorname{tr}\boldsymbol{H}(\boldsymbol{G}+\boldsymbol{H})^{-1}$ are again χ_{pq}^2. The limiting distribution of the maximum characteristic root of $N\boldsymbol{HG}^{-1}$ or $N\boldsymbol{H}(\boldsymbol{G}+\boldsymbol{H})^{-1}$ is the distribution of the maximum characteristic root of $\boldsymbol{H}$ having the distributions $W(\boldsymbol{I},q)$ (Lemma 8.11.1). Significance points for these test criteria are available in Appendix B.

8.11.2. Elliptically Contoured Matrix Distributions

In Section 8.3.2 the $p\times N$ matrix of observations on the dependent variable was defined as $\boldsymbol{X}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_N)$, and the $q\times N$ matrix of observations on the independent variables as $\boldsymbol{Z}=(\boldsymbol{z}_1,\dots,\boldsymbol{z}_N)$; the two matrices are related by $\mathscr{E}\boldsymbol{X}=\boldsymbol{\beta}\boldsymbol{Z}$. Note that in this chapter the matrices of observations have N columns instead of N rows.

Let $\boldsymbol{E}=(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$ be a $p\times N$ random matrix with density $|\boldsymbol{\Lambda}|^{-N/2}g[\boldsymbol{F}^{-1}\boldsymbol{E}\boldsymbol{E}'(\boldsymbol{F}')^{-1}]$, where $\boldsymbol{\Lambda}=\boldsymbol{F}\boldsymbol{F}'$. Define $\boldsymbol{X}$ by

$$\boldsymbol{X}=\boldsymbol{\beta}\boldsymbol{Z}+\boldsymbol{E}. \tag{9}$$

In these terms the least squares estimator of $\boldsymbol{\beta}$ is

$$\boldsymbol{B}=\boldsymbol{X}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}=\boldsymbol{C}\boldsymbol{A}^{-1}, \tag{10}$$

where $\boldsymbol{C}=\boldsymbol{X}\boldsymbol{Z}'=\sum_{\alpha=1}^{N}\boldsymbol{x}_\alpha\boldsymbol{z}'_\alpha$ and $\boldsymbol{A}=\boldsymbol{Z}\boldsymbol{Z}'=\sum_{\alpha=1}^{N}\boldsymbol{z}_\alpha\boldsymbol{z}'_\alpha$. Note that the density of $\boldsymbol{E}$ is invariant with respect to multiplication on the right by $N\times N$ orthogonal matrices; that is, $\boldsymbol{E}'$ is left spherical. Then $\boldsymbol{E}'$ has the stochastic representation

$$\boldsymbol{E}'\overset{d}{=}\boldsymbol{U}\boldsymbol{T}\boldsymbol{F}', \tag{11}$$

where U has the uniform distribution on $U'U = I_p$, T is the lower triangular matrix with nonnegative diagonal elements satisfying $EE' = TT'$, and F is a lower triangular matrix with nonnegative diagonal elements satisfying $FF' = \Sigma$. We can write

$$(12) \quad B - \beta = EZ'A^{-1} \overset{d}{=} FT'U'Z'A^{-1},$$

$$(13) \quad H = (B - \beta)A(B - \beta)' = EZ'A^{-1}ZE' \overset{d}{=} FT'U'(Z'A^{-1}Z)UTF'.$$

$$(14) \quad \begin{aligned} G &= (X - \beta Z)(X - \beta Z)' - H = EE' - H \\ &= E(I_N - Z'A^{-1}Z)E' = FT'U'(I_N - Z'A^{-1}Z)UTF'. \end{aligned}$$

It was shown in Section 8.6 that the likelihood ratio criterion for $H: \beta = 0$, the Lawley–Hotelling trace criterion, the Bartlett–Nanda–Pillai trace criterion, and the Roy maximum root test are invariant with respect to linear transformations $x \to Kx$. Then Corollary 4.5.5 implies the following theorem.

Theorem 8.11.3. *Under the null hypothesis $\beta = 0$, the distribution of each invariant criterion when the distribution of E' is left spherical is the same as the distribution under normality.*

Thus the tests and confidence regions described in Section 8.7 are valid for left-spherical distributions E'.

The matrices $Z'A^{-1}Z$ and $I_N - Z'A^{-1}Z$ are idempotent of ranks q and $N - q$. There is an orthogonal matrix O_N such that

$$(15) \quad OZ'A^{-1}ZO' = \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix}, \qquad O(I_N - Z'A^{-1}Z)O' = \begin{bmatrix} 0 & 0 \\ 0 & I_{N-q} \end{bmatrix}.$$

The transformation $V = O'U$ is uniformly distributed on $V'V = I_p$, and

$$(16) \quad H = KV'\begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix}VK', \qquad G = KV'\begin{bmatrix} 0 & 0 \\ 0 & I_{N-q} \end{bmatrix}VK',$$

where $K = FT'$.

The trace criterion $\operatorname{tr} HG^{-1}$, for example, is

$$(17) \quad \operatorname{tr} HG^{-1} = V'\begin{bmatrix} I_1 & 0 \\ 0 & 0 \end{bmatrix}V\left(V'\begin{bmatrix} 0 & 0 \\ 0 & I_{N-q} \end{bmatrix}V\right)^{-1}.$$

The distribution of any invariant criterion depends only on U (or V), not on T.

Since $G+H=FT'TF'$, it is independent of U. A selection of a linear transformation of X can be made on the basis of $G+H$. Let D be a $p\times r$ matrix of rank r that may depend on $G+H$. Define $x_\alpha^*=D'x_\alpha$. Then $\mathscr{E}x_\alpha^*=(D'\boldsymbol{\beta})z_\alpha$, and the hypothesis $\boldsymbol{\beta}=0$ implies $D'\boldsymbol{\beta}=0$. Let $X^*=(x_1^*,\dots,x_N^*)=D'X$, $\boldsymbol{\beta}_D=D'\boldsymbol{\beta}$, $E_D=D'E$, $H_D=D'HD$, $G_D=D'GD$. Then $E_D'=E'D\overset{d}{=}UTF'D'$. The invariant test criteria for $\boldsymbol{\beta}_D=0$ are those for $\boldsymbol{\beta}=0$ and have the same distributions under the null hypothesis as for the normal distribution with p replaced by r.

PROBLEMS

8.1. (Sec. 8.2.2) Consider the following sample (for $N=8$):

Weight of grain	40	17	9	15	6	12	5	9
Weight of straw	53	19	10	29	13	27	19	30
Amount of fertilizer	24	11	5	12	7	14	11	18

Let $z_{2\alpha}=1$, and let $z_{1\alpha}$ be the amount of fertilizer on the αth plot. Estimate $\boldsymbol{\beta}$ for this sample. Test the hypothesis $\boldsymbol{\beta}_1=0$ at the 0.01 significance level.

8.2. (Sec. 8.2) Show that Theorem 3.2.1 is a special case of Theorem 8.2.1. [*Hint:* Let $q=1$, $z_\alpha=1$, $\boldsymbol{\beta}=\mu$.]

8.3. (Sec. 8.2) Prove Theorem 8.2.3.

8.4. (Sec. 8.2) Show that $\hat{\boldsymbol{\beta}}$ minimizes the generalized variance

$$\left|\sum_{\alpha=1}^{N}(x_\alpha-\boldsymbol{\beta}z_\alpha)(x_\alpha-\boldsymbol{\beta}z_\alpha)'\right|.$$

8.5. (Sec. 8.3) In the following data [Woltz, Reid, and Colwell (1948), used by R. L. Anderson and Bancroft (1952)] the variables are x_1, rate of cigarette burn; x_2, the percentage of nicotine; z_1, the percentage of nitrogen; z_2, of chlorine; z_3, of potassium; z_4, of phosphorus; z_5, of calcium; and z_6, of magnesium; and $z_7=1$; and $N=25$:

$$\sum_{\alpha=1}^{N}x_\alpha=\begin{pmatrix}42.20\\54.03\end{pmatrix},\qquad \sum_{\alpha=1}^{N}z_\alpha=\begin{pmatrix}53.92\\62.02\\56.00\\12.25\\89.79\\24.10\\25\end{pmatrix},$$

$$\sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})'=\begin{pmatrix}0.6690&0.4527\\0.4527&6.5921\end{pmatrix},$$

$$\sum_{\alpha=1}^{N} (z_\alpha - \bar{z})(z_\alpha - \bar{z})'$$

$$= \begin{pmatrix} 1.8311 & -0.3589 & -0.0125 & -0.0244 & 1.6379 & 0.5057 & 0 \\ -0.3589 & 8.8102 & -0.3469 & 0.0352 & 0.7920 & 0.2173 & 0 \\ -0.0125 & -0.3469 & 1.5818 & -0.0415 & -1.4278 & -0.4753 & 0 \\ -0.0244 & 0.0352 & -0.0415 & 0.0258 & 0.0043 & 0.0154 & 0 \\ 1.6379 & 0.7920 & -1.4278 & 0.0043 & 3.7248 & 0.9120 & 0 \\ 0.5057 & 0.2173 & -0.4753 & 0.0154 & 0.9120 & 0.3828 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\sum_{\alpha=1}^{N} (z_\alpha - \bar{z})(x_\alpha - \bar{x})' = \begin{pmatrix} 0.2501 & 2.6691 \\ -1.5136 & -2.0617 \\ 0.5007 & -0.9503 \\ -0.0421 & -0.0187 \\ -0.1914 & 3.4020 \\ -0.1586 & 1.1663 \\ 0 & 0 \end{pmatrix}.$$

(a) Estimate the regression of x_1 and x_2 on z_1, z_5, z_6, and z_7.

(b) Estimate the regression on all seven variables.

(c) Test the hypothesis that the regression on z_2, z_3, and z_4 is 0.

8.6. (Sec. 8.3) Let $q = 2$, $z_{1\alpha} = w_\alpha$ (scalar), $z_{2\alpha} = 1$. Show that the U-statistic for testing the hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$ is a monotonic function of a T^2-statistic, and give the T^2-statistic in a simple form. (See Problem 5.1.)

8.7. (Sec. 8.3) Let $z_{q\alpha} = 1$, let $q_2 = 1$, and let

$$A^* = \left[\sum_\alpha (z_{i\alpha} - \bar{z}_i)(z_{j\alpha} - \bar{z}_j)\right], \qquad i, j = 1, \dots, q_1 = q - 1.$$

Prove that

$$(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1)(A_{11} - A_{12}A_{22}^{-1}A_{21})(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1)' = (\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1)A^*(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1)'.$$

8.8. (Sec. 8.3) Let $q_1 = q_2$. How do you test the hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$?

8.9. (Sec. 8.3) Prove

$$\hat{\boldsymbol{\beta}}_{1\Omega} = \sum_\alpha x_\alpha \left(z_\alpha^{(1)} - A_{12}A_{22}^{-1}z_\alpha^{(2)}\right)' \left[\sum_\alpha \left(z_\alpha^{(1)} - A_{12}A_{22}^{-1}z_\alpha^{(2)}\right)\left(z_\alpha^{(1)} - A_{12}A_{22}^{-1}z_\alpha^{(2)}\right)'\right]^{-1}$$

$$= \left(C_1 - C_2A_{22}^{-1}A_{21}\right)\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}.$$

8.10. (Sec. 8.4) By comparing Theorem 8.2.2 and Problem 8.9, prove Lemma 8.4.1.

8.11. (Sec. 8.4) Prove Lemma 8.4.1 by showing that the density of $\hat{\boldsymbol\beta}_{1\Omega}$ and $\hat{\boldsymbol\beta}_{2\omega}$ is

$$K_1 \exp\left[-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\left(\hat{\boldsymbol\beta}_{1\Omega}-\boldsymbol\beta_1^*\right)A_{11\cdot 2}\left(\hat{\boldsymbol\beta}_{1\Omega}-\boldsymbol\beta_1^*\right)'\right]$$
$$\cdot K_2 \exp\left[-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\left(\hat{\boldsymbol\beta}_{2\omega}-\boldsymbol\beta_2\right)A_{22}\left(\hat{\boldsymbol\beta}_{2\omega}-\boldsymbol\beta_2\right)'\right].$$

8.12. (Sec. 8.4) Show that the cdf of $U_{3,3,n}$ is

$$I_u\left(\tfrac{1}{2}n-1,\tfrac{3}{2}\right)+\frac{\Gamma(n+2)\Gamma\left[\frac{1}{2}(n+1)\right]}{\Gamma(n-1)\Gamma\left(\frac{1}{2}n-1\right)\sqrt{\pi}}$$
$$\cdot\left\{\frac{2u^{\frac{1}{2}n-1}\sqrt{1-u}}{n(n-1)}+\frac{u^{\frac{1}{2}(n-1)}}{n-1}\left[\arcsin(2u-1)-\tfrac{1}{2}\pi\right]\right.$$
$$\left.+\frac{2u^{\frac{1}{2}n}}{n}\log\left(\frac{1+\sqrt{1-u}}{\sqrt{u}}\right)+\frac{2u^{\frac{1}{2}n-1}(1-u)^{\frac{3}{2}}}{3(n+1)}\right\}.$$

[*Hint:* Use Theorem 8.4.4. The region $\{0\le z_1\le 1, 0\le z_2\le 1, z_1^2z_2\le u\}$ is the union of $\{0\le z_1\le 1, 0\le z_2\le u\}$ and $\{0\le z_1\le u/z_2,\ u\le z_2\le 1\}$.]

8.13. (Sec. 8.4) Find $\Pr\{U_{4,3,n}\ge u\}$.

8.14. (Sec. 8.4) Find $\Pr\{U_{4,4,n}\ge u\}$.

8.15. (Sec. 8.4) For $p\le m$ find $\infty\mathscr{E}U^h$ from the density of G and H. [*Hint:* Use the fact that the density of $K+\sum_{i=1}^t V_iV_i'$ is $W(\Sigma, s+t)$ if the density of K is $W(\Sigma,s)$ and $V_1,\dots,V_t$ are independently distributed as $N(\mathbf{0},\Sigma)$.]

8.16. (Sec. 8.4)

(a) Show that when p is even, the characteristic function of $Y=\log U_{p,m,n}$, say $\phi(t)=\infty\mathscr{E}\,e^{itY}$, is the reciprocal of a polynomial.

(b) Sketch a method of inverting the characteristic function of Y by the method of residues.

(c) Show that the resulting density of U is a polynomial in $\sqrt{u}$ and $\log u$ with possibly a factor of $u^{-\frac{1}{2}}$.

8.17. (Sec. 8.5) Use the asymptotic expansion of the distribution to compute $\Pr\{-k\log U_{3,3,n}\le M^*\}$ for

(a) $n=8,\quad M^*=14.7$,

(b) $n=8,\quad M^*=21.7$,

(c) $n=16,\quad M^*=14.7$,

(d) $n=16,\quad M^*=21.7$.

(Either compute to the third decimal place or use the expansion to the k^{-4} term.)

8.18. (Sec. 8.5) In case $p = 3$, $q_1 = 4$, and $n = N - q = 20$, find the 50% significance point for $k \log U$ (a) using $-2\log\lambda$ as χ^2 and (b) using $-k\log U$ as χ^2. Using more terms of this expansion, evaluate the exact significance levels for your answers to (a) and (b).

8.19. (Sec. 8.6.5) Prove for $l_i \geq 0$, $i = 1, \dots, p$,

$$\sum_{i=1}^{p} \frac{l_i}{1 + l_i} \leq \log \prod_{i=1}^{p} (1 + l_i) \leq \sum_{i=1}^{p} l_i.$$

Comment: The inequalities imply an ordering of the values of the Bartlett–Nanda–Pillai trace, the negative logarithm of the likelihood ratio criterion, and the Lawley–Hotelling trace.

8.20. (Sec. 8.6) *The multivariate beta density.* Let H and G be independently distributed according to $W(\Sigma, m)$ and $W(\Sigma, n)$, respectively. Let C be a matrix such that $CC' = H + G$, and let

$$L = C^{-1} H C'^{-1}.$$

Show that the density of L is

$$\frac{\Gamma_p[\frac{1}{2}(m+n)]}{\Gamma_p(\frac{1}{2}m)\Gamma_p(\frac{1}{2}n)} |L|^{\frac{1}{2}(m-p-1)} |I - L|^{\frac{1}{2}(n-p-1)}$$

for L and $I - L$ positive definite, and 0 otherwise.

8.21. (Sec. 8.9) Let Y_{ij} (a p-component vector) be distributed according to $N(\boldsymbol{\mu}_{ij}, \Sigma)$, where $\mathscr{E} Y_{ij} = \boldsymbol{\mu}_{ij} = \boldsymbol{\mu} + \boldsymbol{\lambda}_i + \boldsymbol{\nu}_j + \boldsymbol{\gamma}_{ij}$, $\sum_i \boldsymbol{\lambda}_i = \mathbf{0} = \sum_j \boldsymbol{\nu}_j = \sum_i \boldsymbol{\gamma}_{ij} = \sum_j \boldsymbol{\gamma}_{ij}$; the $\boldsymbol{\gamma}_{ij}$ are the interactions. If m observations are made on each Y_{ij} (say $y_{ij1}, \dots, y_{ijm}$), how do you test the hypothesis $\boldsymbol{\lambda}_i = \mathbf{0}$, $i = 1, \dots, r$? How do you test the hypothesis $\boldsymbol{\gamma}_{ij} = \mathbf{0}$, $i = 1, \dots, r$, $j = 1, \dots, c$?

8.22. (Sec. 8.9) *The Latin square.* Let Y_{ij}, $i, j = 1, \dots, r$, be distributed according to $N(\boldsymbol{\mu}_{ij}, \Sigma)$, where $\mathscr{E} Y_{ij} = \boldsymbol{\mu}_{ij} = \boldsymbol{\gamma} + \boldsymbol{\lambda}_i + \boldsymbol{\nu}_j + \boldsymbol{\mu}_k$ and $k = j - i + 1 \pmod{r}$ with $\sum \boldsymbol{\lambda}_i = \sum \boldsymbol{\nu}_j = \sum \boldsymbol{\mu}_k = \mathbf{0}$.

(a) Give the univariate analysis of variance table for main effects and error (including sums of squares, numbers of degrees of freedom, and mean squares).

(b) Give the table for the vector case.

(c) Indicate in the vector case how to test the hypothesis $\boldsymbol{\lambda}_i = \mathbf{0}$, $i = 1, \dots, r$.

8.23. (Sec. 8.9) Let x_1 be the yield of a process and x_2 a quality measure. Let $z_1 = 1$, $z_2 = \pm 10°$ (temperature relative to average) $z_3 = \pm 0.75$ (relative measure of flow of one agent), and $z_4 = \pm 1.50$ (relative measure of flow of another agent). [See Anderson (1955a) for details.] Three observations were made on x_1

and x_2 for each possible triplet of values of z_2, z_3, and z_4. The estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 58.529 & -0.3829 & -5.050 & 2.308 \\ 98.675 & 0.1558 & 4.144 & -0.700 \end{pmatrix};$$

$s_1 = 3.090$, $s_2 = 1.619$, and $r = -0.6632$ can be used to compute S or $\hat{\boldsymbol{\Sigma}}$.

(a) Formulate an analysis of variance model for this situation.

(b) Find a confidence region for the effects of temperature (i.e., β_{12}, β_{22}).

(c) Test the hypothesis that the two agents have no effect on the yield and quantity.

8.24. (Sec. 8.6) Interpret the transformations referred to in Theorem 8.6.1 in the original terms; that is, $H: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ and $z_\alpha^{(1)}$.

8.25. (Sec. 8.6) Find the cdf of $\operatorname{tr} HG^{-1}$ for $p = 2$. [*Hint:* Use the distribution of the roots given in Chapter 13.]

8.26. (Sec. 8.10.1) *Bartlett–Nanda–Pillai V-test as a Bayes procedure.* Let $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_{m+n}$ be independently normally distributed with covariance matrix $\boldsymbol{\Sigma}$ and means $\mathscr{E}\boldsymbol{w}_i = \boldsymbol{\gamma}_i$, $i = 1, \dots, m$, $\mathscr{E}\boldsymbol{w}_i = \mathbf{0}$, $i = m+1, \dots, m+n$. Let Π_0 be defined by $[\boldsymbol{\Gamma}_1, \boldsymbol{\Sigma}] = [\mathbf{0}, (\boldsymbol{I} + \boldsymbol{CC}')^{-1}]$, where the $p \times m$ matrix $\boldsymbol{C}$ has a density proportional to $|\boldsymbol{I} + \boldsymbol{CC}'|^{-\frac{1}{2}(n+m)}$, and $\boldsymbol{\Gamma}_1 = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_m)$; let Π_1 be defined by $[\boldsymbol{\Gamma}_1, \boldsymbol{\Sigma}] = [(\boldsymbol{I} + \boldsymbol{CC}')^{-1}\boldsymbol{C}, (\boldsymbol{I} + \boldsymbol{CC}')^{-1}]$ where $\boldsymbol{C}$ has a density proportional to $|\boldsymbol{I} + \boldsymbol{CC}'|^{-\frac{1}{2}(n+m)} e^{\frac{1}{2}\operatorname{tr} \boldsymbol{C}'(\boldsymbol{I}+\boldsymbol{CC}')^{-1}\boldsymbol{C}}$.

(a) Show that the measures are finite for $n \ge p$ by showing $\operatorname{tr} \boldsymbol{C}'(\boldsymbol{I} + \boldsymbol{CC}')^{-1}\boldsymbol{C} < m$ and verifying that the integral of $|\boldsymbol{I} + \boldsymbol{CC}'|^{-\frac{1}{2}(n+m)}$ is finite. [*Hint:* Let $\boldsymbol{C} = (\boldsymbol{c}_1, \dots, \boldsymbol{c}_m)$, $\boldsymbol{D}_j = \boldsymbol{I} + \sum_{i=1}^{j} \boldsymbol{c}_i\boldsymbol{c}_i' = \boldsymbol{E}_j\boldsymbol{E}_j'$, $\boldsymbol{c}_j = \boldsymbol{E}_{j-1}\boldsymbol{d}_j$, $j = 1, \dots, m$ $(\boldsymbol{E}_0 = \boldsymbol{I})$. Show $|\boldsymbol{D}_j| = |\boldsymbol{D}_{j-1}|(1 + \boldsymbol{d}_j'\boldsymbol{d}_j)$ and hence $|\boldsymbol{D}_m| = \prod_{j=1}^m (1 + \boldsymbol{c}_j'\boldsymbol{d}_j)$. Then refer to Problem 5.15.]

(b) Show that the inequality (26) of Section 5.6 is equivalent to

$$\operatorname{tr}\left(\sum_{i=1}^{m+n} \boldsymbol{w}_i\boldsymbol{w}_i'\right)^{-1} \sum_{i=1}^{m} \boldsymbol{w}_i\boldsymbol{w}_i' \ge k.$$

Hence the Bartlett–Nanda–Pillai V-test is Bayes and thus admissible.

8.27. (Sec. 8.10.1) *Likelihood ratio test as a Bayes procedure.* Let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_{m+n}$ be independently normally distributed with covariance matrix $\boldsymbol{\Sigma}$ and means $\mathscr{E}\boldsymbol{w}_i = \boldsymbol{\gamma}_i$, $i = 1, \dots, m$, $\mathscr{E}\boldsymbol{w}_i = \mathbf{0}$, $i = m+1, \dots, m+n$, with $n \ge m + p$. Let Π_0 be defined by $[\boldsymbol{\Gamma}_1, \boldsymbol{\Sigma}] = [\mathbf{0}, (\boldsymbol{I} + \boldsymbol{CC}')^{-1}]$, where the $p \times m$ matrix $\boldsymbol{C}$ has a density proportional to $|\boldsymbol{I} + \boldsymbol{CC}'|^{-\frac{1}{2}(n+m)}$ and $\boldsymbol{\Gamma}_1 = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_m)$; let Π_1 be defined by

$$[\boldsymbol{\Gamma}_1, \boldsymbol{\Sigma}] = \left[(\boldsymbol{I} + \boldsymbol{CC}')^{-1}\boldsymbol{CD}, (\boldsymbol{I} + \boldsymbol{CC}')^{-1}\right],$$

where the m columns of D are conditionally independently normally distributed with means $\mathbf{0}$ and covariance matrix $[I - C'(I + CC')^{-1}C]^{-1}$, and C has (marginal) density proportional to

$$|I + CC'|^{-\frac{1}{2}(n+m)}|I - C'(I + CC')^{-1}C|^{-\frac{1}{2}m}.$$

(a) Show the measures are finite. [*Hint:* See Problem 8.26.]

(b) Show that the inequality (26) of Section 5.6 is equivalent to

$$\frac{\left|\sum_{i=1}^{m+n} w_i w_i'\right|}{\left|\sum_{i=m+1}^{m+n} w_i w_i'\right|} \geq k.$$

Hence the likelihood ratio test is Bayes and thus admissible.

8.28. (Sec. 8.10.1) *Admissibility of the likelihood ratio test.* Show that the acceptance region $|ZZ'|/|ZZ' + XX'| \geq c$ satisfies the conditions of Theorem 8.10.1. [*Hint:* The acceptance region can be written $\prod_{i=1}^{t} m_i > c$, where $m_i = 1 - \lambda_i$, $i = 1, \dots, t$.]

8.29. (Sec. 8.10.1) *Admissibility of the Lawley–Hotelling test.* Show that the acceptance region $\operatorname{tr} XX'(ZZ')^{-1} \leq c$ satisfies the conditions of Theorem 8.10.1.

8.30. (Sec. 8.10.1) *Admissibility of the Bartlett–Nanda–Pillai trace test.* Show that the acceptance region $\operatorname{tr} X'(ZZ' + XX')^{-1}X \leq c$ satisfies the conditions of Theorem 8.10.1.

8.31. (Sec. 8.10.1) Show that if A and B are positive definite and $A - B$ is positive semidefinite, then $B^{-1} - A^{-1}$ is positive semidefinite.

8.32. (Sec. 8.10.1) Show that the boundary of $\tilde{A}$ has m-measure 0. [*Hint:* Show that (closure of $\tilde{A}) \subset \tilde{A} \cup C$, where $C = \{V | U - YY'$ is singular$\}$.]

8.33. (Sec. 8.10.1) Show that if $A \subset R_<^m$ is convex and monotone in majorization, then A^* is convex. [*Hint:* Show

$$(px + qy)_\downarrow \succ_w px_\downarrow + qy_\downarrow,$$

where

$$z_\downarrow = (z_{[1]}, \dots, z_{[m]})' \in R_<^m.]$$

8.34. (Sec. 8.10.1) Show that $C(\boldsymbol{\lambda})$ is convex. [*Hint:* Follow the solution of Problem 8.33 to show $(px + qy) \prec_w \boldsymbol{\lambda}$ if $x \prec_w \boldsymbol{\lambda}$ and $y \prec_w \boldsymbol{\lambda}$.]

8.35. (Sec. 8.10.1) Show that if A is monotone, then A^* is monotone. [*Hint:* Use the fact that

$$x_{[k]} = \max_{i_1, \dots, i_k} \left\{\min(x_{i_1}, \dots, x_{i_k})\right\}.]$$

8.36. (Sec. 8.10.2) *Monotonicity of the power function of the Bartlett–Nanda–Pillai trace test*. Show that

$$\operatorname{tr}(uu' + B)(uu' + B + W)^{-1} \le K$$

is convex in u for fixed positive semidefinite B and positive definite $B + W$ if $0 \le K \le 1$. [*Hint:* Verify

$$(uu' + B + W)^{-1}$$
$$= (B + W)^{-1} - \frac{1}{1 + u'(B + W)^{-1}u}(B + W)^{-1}uu'(B + W)^{-1}.$$

The resulting quadratic form in u involves the matrix $(\operatorname{tr} A)I - A$ for $A = (B + W)^{-\frac{1}{2}}B(B + W)^{-\frac{1}{2}}$; show that this matrix is positive semidefinite by diagonalizing A.]

8.37. (Sec. 8.8) Let $x_\alpha^{(\nu)}$, $\alpha = 1, \dots, N_\nu$, be observations from $N(\mu^{(\nu)}, \Sigma)$, $\nu = 1, \dots, q$. What criterion may be used to test the hypothesis that

$$\mu^{(\nu)} = \sum_{h=1}^{m} \gamma_h c_{h\nu} + \mu,$$

where $c_{h\nu}$ are given numbers and γ_ν, μ are unknown vectors? [*Note:* This hypothesis (that the means lie on an m-dimensional hyperplane with ratios of distances known) can be put in the form of the general linear hypothesis.]

8.38. (Sec. 8.2) Let x_α be an observation from $N(\beta z_\alpha, \Sigma)$, $\alpha = 1, \dots, N$. Suppose there is a known fixed vector γ such that $\beta\gamma = 0$. How do you estimate β?

8.39. (Sec. 8.8) What is the largest group of transformations on $y_\alpha^{(i)}$, $\alpha = 1, \dots, N_i$, $i = 1, \dots, q$, that leaves (1) invariant? Prove the test (12) is invariant under this group.

CHAPTER 9

Testing Independence of Sets of Variates

9.1. INTRODUCTION

In this section we divide a set of p variates with a joint normal distribution into q subsets and ask whether the q subsets are mutually independent; this is equivalent to testing the hypothesis that each variable in one subset is uncorrelated with each variable in the others. We find the likelihood ratio criterion for this hypothesis, the moments of the criterion under the null hypothesis, some particular distributions, and an asymptotic expansion of the distribution.

The likelihood ratio criterion is invariant under linear transformations within sets; another such criterion is developed. Alternative test procedures are step-down procedures, which are not invariant, but are flexible. In the case of two sets, independence of the two sets is equivalent to the regression of one on the other being 0; the criteria for Chapter 8 are available. Some optimal properties of the likelihood ratio test are treated.

9.2. THE LIKELIHOOD RATIO CRITERION FOR TESTING INDEPENDENCE OF SETS OF VARIATES

Let the p-component vector X be distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We partition X into q subvectors with $p_1, p_2, \ldots, p_q$ components, respectively:

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

that is,

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(q)} \end{pmatrix}. \tag{1}$$

The vector of means $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are partitioned similarly,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \\ \vdots \\ \boldsymbol{\mu}^{(q)} \end{pmatrix}, \tag{2}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1q} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2q} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\Sigma}_{q1} & \boldsymbol{\Sigma}_{q2} & \cdots & \boldsymbol{\Sigma}_{qq} \end{pmatrix}. \tag{3}$$

The null hypothesis we wish to test is that the subvectors $X^{(1)}, \dots, X^{(q)}$ are mutually independently distributed, that is, that the density of X factors into the densities of $X^{(1)}, \dots, X^{(q)}$. It is

$$H: n(x|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{q} n\left(x^{(i)}|\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_{ii}\right). \tag{4}$$

If $X^{(1)}, \dots, X^{(q)}$ are independent subvectors,

$$\mathscr{E}\left(X^{(i)} - \boldsymbol{\mu}^{(i)}\right)\left(X^{(j)} - \boldsymbol{\mu}^{(j)}\right)' = \boldsymbol{\Sigma}_{ij} = \mathbf{0}, \qquad i \neq j. \tag{5}$$

(See Section 2.4.) Conversely, if (5) holds, then (4) is true. Thus the null hypothesis is equivalently $H: \boldsymbol{\Sigma}_{ij} = \mathbf{0}$, $i \neq j$. This can be stated alternatively as the hypothesis that $\boldsymbol{\Sigma}$ is of the form

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{qq} \end{pmatrix}. \tag{6}$$

Given a sample $x_1, \dots, x_N$ of N observations on X, the likelihood ratio criterion is

$$\lambda = \frac{\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}_0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)}{\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}, \tag{7}$$

where

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{\alpha=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}p} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (x_\alpha - \boldsymbol{\mu})} \tag{8}$$

and $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ is $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ij} = \mathbf{0}$, $i \neq j$, and where the maximum is taken with respect to all vectors $\boldsymbol{\mu}$ and positive definite $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$ (i.e., $\boldsymbol{\Sigma}_{ii}$). As derived in Section 5.2, Equation (6),

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}pN} |\hat{\boldsymbol{\Sigma}}_\Omega|^{\frac{1}{2}N}} e^{-\frac{1}{2}pN}, \tag{9}$$

where

$$\hat{\boldsymbol{\Sigma}}_\Omega = \frac{1}{N} A = \frac{1}{N} \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'. \tag{10}$$

Under the null hypothesis,

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0) = \prod_{i=1}^{q} L_i\left(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_{ii}\right), \tag{11}$$

where

$$L_i\left(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_{ii}\right) = \prod_{\alpha=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}p_i} |\boldsymbol{\Sigma}_{ii}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_\alpha^{(i)} - \boldsymbol{\mu}^{(i)})' \boldsymbol{\Sigma}_{ii}^{-1} (x_\alpha^{(i)} - \boldsymbol{\mu}^{(i)})}. \tag{12}$$

Clearly

$$\begin{aligned} \max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}_0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0) &= \prod_{i=1}^{q} \max_{\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_{ii}} L_i\left(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma}_{ii}\right) \\ &= \prod_{i=1}^{q} \frac{1}{(2\pi)^{\frac{1}{2}p_i N} |\hat{\boldsymbol{\Sigma}}_{ii\omega}|^{\frac{1}{2}N}} e^{-\frac{1}{2}p_i N} \\ &= \frac{1}{(2\pi)^{\frac{1}{2}pN} \prod_{i=1}^{q} |\hat{\boldsymbol{\Sigma}}_{ii\omega}|^{\frac{1}{2}N}} e^{-\frac{1}{2}pN}, \end{aligned} \tag{13}$$

where

$$\hat{\boldsymbol{\Sigma}}_{ii\omega} = \frac{1}{N} \sum_{\alpha=1}^{N} \left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)\left(x_\alpha^{(i)} - \bar{x}^{(i)}\right)'. \tag{14}$$

If we partition A and $\hat{\Sigma}_\Omega$ as we have Σ,

$$(15)\quad A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1q} \\ A_{21} & A_{22} & \cdots & A_{2q} \\ \vdots & \vdots & & \vdots \\ A_{q1} & A_{q2} & \cdots & A_{qq} \end{pmatrix}, \qquad \hat{\Sigma}_\Omega = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} & \cdots & \hat{\Sigma}_{1q} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} & \cdots & \hat{\Sigma}_{2q} \\ \vdots & \vdots & & \vdots \\ \hat{\Sigma}_{q1} & \hat{\Sigma}_{q2} & \cdots & \hat{\Sigma}_{qq} \end{pmatrix},$$

we see that $\hat{\Sigma}_{ii\omega} = \hat{\Sigma}_{ii} = (1/N)A_{ii}$.

The likelihood ratio criterion is

$$(16)\qquad \lambda = \frac{\max_{\mu,\Sigma_0} L(\mu, \Sigma_0)}{\max_{\mu,\Sigma} L(\mu, \Sigma)} = \frac{|\hat{\Sigma}_\Omega|^{\frac{1}{2}N}}{\prod_{i=1}^q |\hat{\Sigma}_{ii}|^{\frac{1}{2}N}} = \frac{|A|^{\frac{1}{2}N}}{\prod_{i=1}^q |A_{ii}|^{\frac{1}{2}N}}.$$

The critical region of the likelihood ratio test is

$$(17)\qquad \lambda \le \lambda(\varepsilon),$$

where $\lambda(\varepsilon)$ is a number such that the probability of (17) is ε with $\Sigma = \Sigma_0$. (It remains to show that such a number can be found.) Let

$$(18)\qquad V = \frac{|A|}{\prod_{i=1}^q |A_{ii}|}.$$

Then $\lambda = V^{\frac{1}{2}N}$ is a monotonic increasing function of V. The critical region (17) can be equivalently written as

$$(19)\qquad V \le V(\varepsilon).$$

Theorem 9.2.1. *Let $x_1, \dots, x_N$ be a sample of N observations drawn from $N(\mu, \Sigma)$, where x_α, μ, and Σ are partitioned into $p_1, \dots, p_q$ rows (and columns in the case of Σ) as indicated in* (1), (2), *and* (3). *The likelihood ratio criterion that the q sets of components are mutually independent is given by* (16), *where A is defined by* (10) *and partitioned according to* (15). *The likelihood ratio test is given by* (17) *and equivalently by* (19), *where V is defined by* (18) *and $\lambda(\varepsilon)$ or $V(\varepsilon)$ is chosen to obtain the significance level ε.*

Since $r_{ij} = a_{ij}/\sqrt{a_{ii}a_{jj}}$, we have

$$(20)\qquad |A| = |R| \prod_{i=1}^p a_{ii},$$

where

$$
(21) \qquad R = (r_{ij}) = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1q} \\ R_{21} & R_{22} & \cdots & R_{2q} \\ \vdots & \vdots & & \vdots \\ R_{q1} & R_{q2} & \cdots & R_{qq} \end{pmatrix}
$$

and

$$
(22) \qquad |A_{ii}| = |R_{ii}| \prod_{j=p_1+\cdots+p_{i-1}+1}^{p_1+\cdots+p_i} a_{jj}.
$$

Thus

$$
(23) \qquad V = \frac{|A|}{\prod |A_{ii}|} = \frac{|R|}{\prod |R_{ii}|}.
$$

That is, V can be expressed entirely in terms of sample correlation coefficients.

We can interpret the criterion V in terms of generalized variance. Each set $(x_{i1},\dots,x_{iN})$ can be considered as a vector in N-space; the let $(x_{i1} - \bar{x}_i,\dots,x_{iN}-\bar{x}_i) = z_i$, say, is the projection on the plane orthogonal to the equiangular line. The determinant $|A|$ is the p-dimensional volume squared of the parallelotope with $z_1,\dots,z_p$ as principal edges. The determinant $|A_{ii}|$ is the p_i-dimensional volume squared of the parallelotope having as principal edges the ith set of vectors. If each set of vectors is orthogonal to each other set (i.e., $R_{ij} = 0$, $i \neq j$), then the volume squared $|A|$ is the product of the volumes squared $|A_{ii}|$. For example, if $p = 2$, $p_1 = p_2 = 1$, this statement is that the area of a parallelogram is the product of the lengths of the sides if the sides are at right angles. If the sets are almost orthogonal, then $|A|$ is almost $\prod |A_{ii}|$, and V is almost 1.

The criterion has an invariance property. Let C_i be an arbitrary nonsingular matrix of order p_i and let

$$
(24) \qquad C = \begin{pmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & C_q \end{pmatrix}.
$$

Let $Cx_\alpha + d = x_\alpha^*$. Then the criterion for independence in terms of x_α^* is identical to the criterion in terms of x_α. Let $A^* = \sum_\alpha (x_\alpha^* - \bar{x}^*)(x_\alpha^* - \bar{x}^*)'$ be partitioned into submatrices A^*_{ij}. Then

$$(25)\qquad \begin{aligned} A^*_{ij} &= \sum_\alpha \left(x^{*(i)}_\alpha - \bar{x}^{*(i)}\right)\left(x^{*(j)}_\alpha - \bar{x}^{*(j)}\right)' \\ &= C_i \sum_\alpha \left(x^{(i)}_\alpha - \bar{x}^{(i)}\right)\left(x^{(j)}_\alpha - \bar{x}^{(j)}\right)' C_j' \\ &= C_i A_{ij} C_j' \end{aligned}$$

and $A^* = CAC'$. Thus

$$(26)\qquad \begin{aligned} V^* &= \frac{|A^*|}{\prod |A^*_{ii}|} = \frac{|CAC'|}{\prod |C_i A_{ii} C_i'|} \\ &= \frac{|C| \cdot |A| \cdot |C'|}{\prod |C_i| \cdot |A_{ii}| \cdot |C_i'|} = \frac{|A|}{\prod |A_{ii}|} = V \end{aligned}$$

for $|C| = \prod |C_i|$. Thus the test is invariant with respect to linear transformations within each set.

Narain (1950) showed that the test based on V is strictly unbiased; that is, the probability of rejecting the null hypothesis is greater than the significance level if the hypothesis is not true. [See also Daly (1940).]

9.3. THE DISTRIBUTION OF THE LIKELIHOOD RATIO CRITERION WHEN THE NULL HYPOTHESIS IS TRUE

9.3.1. Characterization of the Distribution

We shall show that under the null hypothesis the distribution of the criterion V is the distribution of a product of independent variables, each of which has the distribution of a criterion U for the linear hypothesis (Section 8.4).

Let

$$(1)\qquad V_i = \frac{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} & A_{1i} \\ \vdots & & \vdots & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} & A_{i-1,i} \\ A_{i1} & \cdots & A_{i,i-1} & A_{ii} \end{vmatrix}}{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{vmatrix} \cdot |A_{ii}|}, \qquad i = 2, \dots, q.$$

Then $V = V_2 V_3 \cdots V_q$. Note that V_i is the $N/2$th root of the likelihood ratio criterion for testing the null hypothesis

$$H_i : \Sigma_{i1} = \mathbf{0}, \ldots, \Sigma_{i,i-1} = \mathbf{0}, \tag{2}$$

that is, that $X^{(i)}$ is independent of $(X^{(1)\prime}, \ldots, X^{(i-1)\prime})'$. The null hypothesis H is the intersection of these hypotheses.

Theorem 9.3.1. *When H_i is true, V_i has the distribution of $U_{p_i, \bar{p}_i, n-\bar{p}_i}$, where $n = N - 1$ and $\bar{p}_i = p_1 + \cdots + p_{i-1}$, $i = 2, \ldots, q$.*

Proof. The matrix A has the distribution of $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, where $Z_1, \ldots, Z_n$ are independently distributed according to $N(\mathbf{0}, \Sigma)$ and Z_α is partitioned as $(Z_\alpha^{(1)\prime}, \ldots, Z_\alpha^{(q)\prime})'$. Then conditional on $Z_\alpha^{(1)} = z_\alpha^{(1)}, \ldots, Z_\alpha^{(i-1)} = z_\alpha^{(i-1)}$, $\alpha = 1, \ldots, n$, the subvectors $Z_1^{(i)}, \ldots, Z_n^{(i)}$ are independently distributed, $Z_\alpha^{(i)}$ having a normal distribution with mean

$$\boldsymbol{\beta}_i \begin{pmatrix} z_\alpha^{(1)} \\ \vdots \\ z_\alpha^{(i-1)} \end{pmatrix} \tag{3}$$

and covariance matrix

$$\Sigma_{ii} - \boldsymbol{\beta}_i \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1,i-1} \\ \vdots & & \vdots \\ \Sigma_{i-1,1} & \cdots & \Sigma_{i-1,i-1} \end{pmatrix} \boldsymbol{\beta}_i', \tag{4}$$

where

$$\boldsymbol{\beta}_i = (\Sigma_{i1} \quad \cdots \quad \Sigma_{i,i-1}) \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1,i-1} \\ \vdots & & \vdots \\ \Sigma_{i-1,1} & \cdots & \Sigma_{i-1,i-1} \end{pmatrix}^{-1}. \tag{5}$$

When the null hypothesis is not assumed, the estimator of $\boldsymbol{\beta}_i$ is (5) with Σ_{jk} replaced by A_{jk}, and the estimator of (4) is (4) with Σ_{jk} replaced by $(1/n)A_{jk}$ and $\boldsymbol{\beta}_i$ replaced by its estimator. Under $H_i : \boldsymbol{\beta}_i = \mathbf{0}$ and the covariance matrix (4) is Σ_{ii}, which is estimated by $(1/n)A_{ii}$. The $N/2$th root of the likelihood ratio criterion for H_i is

$$(6)\qquad \frac{\left| A_{ii} - (A_{i1}, \dots, A_{i,i-1}) \begin{pmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{pmatrix}^{-1} \begin{pmatrix} A_{1i} \\ \vdots \\ A_{i-1,i} \end{pmatrix} \right|}{|A_{ii}|}$$

$$= \frac{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} & A_{1i} \\ \vdots & & \vdots & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} & A_{i-1,i} \\ A_{i1} & \cdots & A_{i,i-1} & A_{ii} \end{vmatrix}}{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{vmatrix} \cdot |A_{ii}|},$$

which is V_i. This is the U-statistic for p_i dimensions, $\bar{p}_i$ components of the conditioning vector, and $n-\bar{p}_i$ degrees of freedom in the estimator of the covariance matrix. ∎

Theorem 9.3.2. *The distribution of V under the null hypothesis is the distribution of $V_2 V_3 \cdots V_q$, where $V_2, \dots, V_q$ are independently distributed with V_i having the distribution of $U_{p_i, \bar{p}_i, n-\bar{p}_i}$, where $\bar{p}_i = p_1 + \cdots + p_{i-1}$.*

Proof. From the proof of Theorem 9.3.1, we see that the distribution of V_i is that of $U_{p_i, \bar{p}_i, n-\bar{p}_i}$ not depending on the conditioning $z_\alpha^{(k)}$, $k = 1, \dots, i-1$, $\alpha = 1, \dots, n$. Hence the distribution of V_i does not depend on $V_2, \dots, V_{i-1}$. ∎

Theorem 9.3.3. *Under the null hypothesis V is distributed as $\prod_{i=2}^q \prod_{j=1}^{p_i} X_{ij}$, where the X_{ij}'s are independent and X_{ij} has the density $\beta[x \mid \frac{1}{2}(n-\bar{p}_i+1-j), \frac{1}{2}\bar{p}_i]$.*

Proof. This theorem follows from Theorems 9.3.2 and 8.4.1. ∎

9.3.2. Moments

Theorem 9.3.4. *When the null hypothesis is true, the* hth *moment of the criterion is*

$$(7)\qquad \mathscr{E} V^h = \prod_{i=2}^q \left\{ \prod_{j=1}^{p_i} \frac{\Gamma\left[\frac{1}{2}(n-\bar{p}_i+1-j)+h\right] \Gamma\left[\frac{1}{2}(n+1-j)\right]}{\Gamma\left[\frac{1}{2}(n-\bar{p}_i+1-j)\right] \Gamma\left[\frac{1}{2}(n+1-j)+h\right]} \right\}.$$

Proof. Because $V_2, \ldots, V_q$ are independent,

$$\mathscr{E}V^h = \mathscr{E}V_2^h \mathscr{E}V_3^h \cdots \mathscr{E}V_q^h. \tag{8}$$

Theorem 9.3.2 implies $\mathscr{E}V_i^h = \mathscr{E}U^h_{p_i, \bar{p}_i, n-\bar{p}_i}$. Then the theorem follows by substituting from Theorem 8.4.3. ∎

If the p_i are even, say $p_i = 2r_i$, $i > 1$, then by using the duplication formula $\Gamma(\alpha + \frac{1}{2})\Gamma(\alpha + 1) = \sqrt{\pi}\,\Gamma(2\alpha + 1)2^{-2\alpha}$ for the gamma function we can reduce the hth moment of V to

$$\begin{aligned}
\mathscr{E}V^h &= \prod_{i=2}^{q} \left\{ \prod_{k=1}^{r_i} \frac{\Gamma(n + 1 - \bar{p}_i - 2k + 2h)\Gamma(n + 1 - 2k)}{\Gamma(n + 1 - \bar{p}_i - 2k)\Gamma(n + 1 - 2k + 2h)} \right\} \\
&= \prod_{i=2}^{q} \left\{ \prod_{k=1}^{r_i} B^{-1}(n + 1 - \bar{p}_i - 2k, \bar{p}_i) \right. \\
&\qquad \left. \cdot \int_0^1 x^{n+1-\bar{p}_i-2k+2h-1}(1 - x)^{\bar{p}_i - 1}\, dx \right\}.
\end{aligned} \tag{9}$$

Thus V is distributed as $\prod_{i=2}^{q}\{\prod_{k=1}^{r_i} Y_{ik}^2\}$, where the Y_{ik} are independent, and Y_{ik} has density $\beta(y; n + 1 - \bar{p}_i - 2k, \bar{p}_i)$.

In general, the duplication formula for the gamma function can be used to reduce the moments as indicated in Section 8.4.

9.3.3. Some Special Distributions

If $q = 2$, then V is distributed as $U_{p_2, p_1, n-p_1}$. Special cases have been treated in Section 8.4, and references to the literature given. The distribution for $p_1 = p_2 = p_3 = 1$ is given in Problem 9.2, and for $p_1 = p_2 = p_3 = 2$ in Problem 9.3. Wilks (1935) gave the distributions for $p_1 = p_2 = 1$, for $p_3 = p - 2$,† for $p_1 = 1$, $p_2 = p_3 = 2$, for $p_1 = 1$, $p_2 = 2$, $p_3 = 3$, for $p_1 = 1$, $p_2 = 2$, $p_3 = 4$, and for $p_1 = p_2 = 2$, $p_3 = 3$. Consul (1967a) treated the case $p_1 = 2$, $p_2 = 3$, p_3 even.

Wald and Brookner (1941) gave a method for deriving the distribution if not more than one p_i is odd. It can be seen that the same result can be obtained by integration of products of beta functions after using the duplication formula to reduce the moments.

Mathai and Saxena (1973) gave the exact distribution for the general case. Mathai and Katiyar (1979) gave exact significance points for $p = 3(1)10$ and $n = 3(1)20$ for significance levels of 5% and 1% (of $-k \log V$ of Section 9.4).

†In Wilks's formula $\Gamma[\frac{1}{2}(N - 2 - i)]$ should be $\Gamma[\frac{1}{2}(n - 2 - i)]$.

9.4. AN ASYMPTOTIC EXPANSION OF THE DISTRIBUTION OF THE LIKELIHOOD RATIO CRITERION

The hth moment of $\lambda = V^{\frac{1}{2}N}$ is

$$(1) \qquad \mathscr{E}\lambda^h = K \frac{\prod_{i=1}^{p} \Gamma\{\frac{1}{2}[N(1+h)-i]\}}{\prod_{i=1}^{q}\{\prod_{j=1}^{p_i} \Gamma\{\frac{1}{2}[N(1+h)-j]\}\}},$$

where K is chosen so that $\mathscr{E}\lambda^0 = 1$. This is of the form of (1) of Section 8.5 with

$$(2) \qquad \begin{aligned} &a = p, \qquad b = p, \qquad x_k = \frac{N}{2}, \quad \xi_k = \frac{-k}{2}, \quad k = 1, \dots, p, \\ &y_j = \frac{N}{2}, \qquad \eta_j = \frac{-j + p_1 + \cdots + p_{i-1}}{2}, \\ &j = p_1 + \cdots + p_{i-1} + 1, \dots, p_1 + \cdots + p_i, \quad i = 1, \dots, q. \end{aligned}$$

Then $f = \frac{1}{2}[p(p+1) - \Sigma p_i(p_i + 1)] = \frac{1}{2}(p^2 - \Sigma p_i^2)$, $\beta_k = \varepsilon_j = \frac{1}{2}(1-\rho)N$.

In order to make the second term in the expansion vanish we take ρ as

$$(3) \qquad \rho = 1 - \frac{2(p^3 - \Sigma p_i^3) + 9(p^2 - \Sigma p_i^2)}{6N(p^2 - \Sigma p_i^2)}.$$

Let

$$(4) \qquad k = \rho N = N - \frac{3}{2} - \frac{p^3 - \Sigma p_i^3}{3(p^2 - \Sigma p_i^2)}.$$

Then $\omega_2 = \gamma_2 / k^2$, where [as shown by Box (1949)]

$$(5) \qquad \gamma_2 = \frac{p^4 - \Sigma p_i^4}{48} - \frac{5(p^2 - \Sigma p_i^2)}{96} - \frac{(p^3 - \Sigma p_i^3)^2}{72(p^2 - \Sigma p_i^2)}.$$

We obtain from Section 8.5 the following expansion:

$$(6) \qquad \Pr\{-k \log V \le v\} = \Pr\{\chi_f^2 \le v\} + \frac{\gamma_2}{k^2}\left[\Pr\{\chi_{f+4}^2 \le v\} - \Pr\{\chi_f^2 \le v\}\right] + O(k^{-3}).$$

Table 9.1

p	f	v	γ_2	N	k	γ_2/k^2	Second Term
4	6	12.592	$\frac{11}{24}$	15	$\frac{71}{6}$	0.0033	-0.0007
5	10	18.307	$\frac{15}{8}$	15	$\frac{69}{6}$	0.0142	-0.0021
6	15	24.996	$\frac{235}{48}$	15	$\frac{67}{6}$	00393	-0.0043
				16	$\frac{73}{6}$	0.0331	-0.0036

If $q = 2$, we obtain further terms in the expansion by using the results of Section 8.5.

If $p_i = 1$, we have

$$f = \tfrac{1}{2}p(p-1),$$

$$k = N - \frac{2p+11}{6},$$

(7)
$$\gamma_2 = \frac{p(p-1)}{288}(2p^2 - 2p - 13),$$

$$\gamma_3 = \frac{p(p-1)}{3240}(p-2)(2p-1)(p+1);$$

other terms are given by Box (1949). If $p_i = 2$ $(p = 2q)$

$$f = 2q(q-1),$$

(8)
$$k = N - \frac{4q+13}{6},$$

$$\gamma_2 = \frac{q(q-1)}{72}(8q^2 - 8q - 7).$$

Table 9.1 gives an indication of the order of approximation of (6) for $p_i = 1$. In each case v is chosen so that the first term is 0.95.

If $q = 2$, the approximate distributions given in Sections 8.5.3 and 8.5.4 are available. [See also Nagao (1973c).]

9.5. OTHER CRITERIA

In case $q = 2$, the criteria considered in Section 8.6 can be used with $\boldsymbol{G} + \boldsymbol{H}$ replaced by $\boldsymbol{A}_{11}$ and $\boldsymbol{H}$ replaced by $\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$, or $\boldsymbol{G} + \boldsymbol{H}$ replaced by $\boldsymbol{A}_{22}$ and $\boldsymbol{H}$ replaced by $\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$.

The null hypothesis of independence is that $\Sigma - \Sigma_0 = 0$, where Σ_0 is defined in (6) of Section 9.2. An appropriate test procedure will reject the null hypothesis if the elements of $A - A_0$ are large compared to the elements of the diagonal blocks of A_0 (where A_0 is composed of diagonal blocks A_{ii} and off-diagonal blocks of 0). Let the nonsingular matrix B_{ii} be such that $B_{ii}A_{ii}B'_{ii} = I$, that is, $A_{ii}^{-1} = B'_{ii}B_{ii}$, and let B_0 be the matrix with B_{ii} as the ith diagonal block and 0's as off-diagonal blocks. Then $B_0 A_0 B'_0 = I$ and

$$(1)\quad B_0(A - A_0)B'_0 = \begin{pmatrix} 0 & B_{11}A_{12}B'_{22} & \cdots & B_{11}A_{1q}B'_{qq} \\ B_{22}A_{21}B'_{11} & 0 & \cdots & B_{22}A_{2q}B'_{qq} \\ \vdots & \vdots & & \vdots \\ B_{qq}A_{q1}B'_{11} & B_{qq}A_{q2}B'_{22} & \cdots & 0 \end{pmatrix}.$$

This matrix is invariant with respect to transformations (24) of Section 9.2 operating on A. A different choice of B_{ii} amounts to multiplying (1) on the left by Q_0 and on the right by Q'_0, where Q_0 is a matrix with orthogonal diagonal blocks and off-diagonal blocks of 0's. A test procedure should reject the null hypothesis if some measure of the numerical values of the elements of (1) is too large. The likelihood ratio criterion is the $N/2$ power of $|B_0(A - A_0)B'_0 + I| = |B_0 A B'_0|$.

Another measure, suggested by Nagao (1973a), is

$$(2)\quad \tfrac{1}{2}\operatorname{tr}[B_0(A - A_0)B'_0]^2 = \tfrac{1}{2}\operatorname{tr}[(A - A_0)A_0^{-1}]^2 = \tfrac{1}{2}\operatorname{tr}(AA_0^{-1} - I)^2$$
$$= \tfrac{1}{2}\sum_{\substack{i,j=1 \\ i\neq j}}^{q} \operatorname{tr} A_{ij}A_{jj}^{-1}A_{ji}A_{ii}^{-1}.$$

For $q = 2$ this measure is the average of the Bartlett–Nanda–Pillai trace criterion with $G + H$ replaced by A_{11} and H replaced by $A_{12}A_{22}^{-1}A_{21}$ and the same criterion with $G + H$ replaced by A_{22} and H replaced by $A_{21}A_{11}^{-1}A_{12}$.

This criterion multiplied by n or N has a limiting χ^2-distribution with number of degrees of freedom $f = \frac{1}{2}(p^2 - \sum_{i=1}^{q} p_i^2)$, which is the same number as for $-N \log V$. Nagao obtained an asymptotic expansion of the distribution:

$$(3)\quad \Pr\left\{\tfrac{1}{2}n \operatorname{tr}(AA_0^{-1} - I)^2 \le x\right\}$$
$$= \Pr\{\chi_f^2 \le x\}$$
$$+ \frac{1}{n}\left[\frac{1}{12}\left(p^3 - 3p\sum_{i=1}^{q} p_i^2 + 2\sum_{i=1}^{q} p_i^3\right)\Pr\{\chi_{f+6}^2 \le x\}\right.$$

$$+\frac{1}{8}\left(-2p^3+4p\sum_{i=1}^{q}p_i^2-2\sum_{i=1}^{q}p_i^3-p^2+\sum_{i=1}^{q}p_i^2\right)\Pr\{\chi_{f+4}^2\le x\}$$

$$+\frac{1}{4}\left(p^3-p\sum_{i=1}^{q}p_i^2+p^2-\sum_{i=1}^{q}p_i^3\right)\Pr\{\chi_{f+2}^2\le x\}$$

$$-\frac{1}{24}\left(2p^3-2\sum_{i=1}^{q}p_i^3+3p^2-3\sum_{i=1}^{q}p_i^2\right)\Pr\{\chi_f^2\le x\}\Bigg]+O(n^{-2}).$$

9.6. STEP-DOWN PROCEDURES

9.6.1. Step-down by Blocks

It was shown in Section 9.3 that the $N/2$th root of the likelihood ratio criterion, namely V, is the product of $q-1$ of these criteria, that is, $V_2,\dots,V_q$. The ith subcriterion V_i provides a likelihood ratio test of the hypothesis H_i [(2) of Section 9.3] that the ith subvector is independent of the preceding $i-1$ subvectors. Under the null hypothesis H $[=\bigcap_{i=2}^{q}H_i]$, these $q-1$ criteria are independent (Theorem 9.3.2). A step-down testing procedure is to accept the null hypothesis if

$$V_i\ge v_i(\varepsilon_i),\qquad i=2,\dots,q. \tag{1}$$

and reject the null hypothesis if $V_i<v_i(\varepsilon_i)$ for any i. Here $v_i(\varepsilon_i)$ is the number such that the probability of (1) when H_i is true is $1-\varepsilon_i$. The significance level of the procedure is ε satisfying

$$1-\varepsilon=\prod_{i=2}^{q}(1-\varepsilon_i). \tag{2}$$

The subtests can be done sequentially, say, in the order $2,\dots,q$. As soon as a subtest calls for rejection, the procedure is terminated; if no subtest leads to rejection, H is accepted. The ordering of the subvectors is at the discretion of the investigator as well as the ordering of the tests.

Suppose, for example, that measurements on an individual are grouped into physiological measurements, measurements of intelligence, and measurements of emotional characteristics. One could test that intelligence is independent of physiology and then that emotions are independent of physiology and intelligence, or the order of these could be reversed. Alternatively, one could test that intelligence is independent of emotions and then that physiology is independent of these two aspects, or the order reversed. There is a third pair of procedures.

Other criteria for the linear hypothesis discussed in Section 8.6 can be used to test the component hypotheses $H_2, \dots, H_q$ in a similar fashion. When H_i is true, the criterion is distributed independently of $X_\alpha^{(1)}, \dots, X_\alpha^{(i-1)}$, $\alpha = 1, \dots, N$, and hence independently of the criteria for $H_2, \dots, H_{i-1}$.

9.6.2. Step-down by Components

In Section 8.4.5 we discussed a componentwise step-down procedure for testing that a submatrix of regression coefficients was a specified matrix. We adapt this procedure to test the null hypothesis H_i cast in the form

$$(3)\qquad H_i : (\Sigma_{i1}\ \Sigma_{i2}\ \cdots\ \Sigma_{i,i-1}) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1,i-1} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2,i-1} \\ \vdots & \vdots & & \vdots \\ \Sigma_{i-1,1} & \Sigma_{i-1,2} & \cdots & \Sigma_{i-1,i-1} \end{pmatrix}^{-1} = \mathbf{0},$$

where $\mathbf{0}$ is of order $p_i \times \bar{p}_i$. The matrix in (3) consists of the coefficients of the regression of $X^{(i)}$ on $(X^{(1)\prime}, \dots, X^{(i-1)\prime})'$.

For $i = 2$, we test in sequence whether the regression of X_{p_1+1} on $X^{(1)} = (X_1, \dots, X_{p_1})'$ is $\mathbf{0}$, whether the regression of X_{p_1+2} on $X^{(1)}$ is $\mathbf{0}$ in the regression of X_{p_1+2} on $X^{(1)}$ and $X_{p_1+1}, \dots,$ and whether the regression of $X_{p_1+p_2}$ on $X^{(1)}$ is $\mathbf{0}$ in the regression of $X_{p_1+p_2}$ on $X^{(1)}, X_{p_1+1}, \dots, X_{p_1+p_2-1}$. These hypotheses are equivalently that the first, second, ..., and p_2th rows of the matrix in (3) for $i = 2$ are $\mathbf{0}$-vectors.

Let $A_{ii}^{(k)}$ be the $k \times k$ matrix in the upper left-hand corner of A_{ii}, let $A_{ij}^{(k)}$ consist of the upper k rows of A_{ij}, and let $A_{ji}^{(k)}$ consist of the first k columns of A_{ji}, $k = 1, \dots, p_i$. Then the criterion for testing that the first row of (3) is $\mathbf{0}$ is

$$(4)$$

$$X_{i1} = \left| A_{ii}^{(1)} - \left(A_{i1}^{(1)} \cdots A_{i,i-1}^{(1)}\right) \begin{pmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{pmatrix}^{-1} \begin{pmatrix} A_{1i}^{(1)} \\ \vdots \\ A_{i-1,i}^{(1)} \end{pmatrix} \right| \div A_{ii}^{(1)}$$

$$= \frac{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} & A_{1i}^{(1)} \\ \vdots & & \vdots & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} & A_{i-1,i}^{(1)} \\ A_{i1}^{(1)} & \cdots & A_{i,i-1}^{(1)} & A_{ii}^{(1)} \end{vmatrix}}{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{vmatrix} A_{ii}^{(1)}}.$$

For $k > 1$, the criterion for testing that the kth row of the matrix in (3) is $\mathbf{0}$ is [see (8) in Section 8.4]

(5)

$$X_{ik} = \frac{\left| A_{ii}^{(k)} - \left(A_{i1}^{(k)} \cdots A_{i,i-1}^{(k)}\right) \begin{pmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{pmatrix}^{-1} \begin{pmatrix} A_{1i}^{(k)} \\ \vdots \\ A_{1,i-1}^{(k)} \end{pmatrix} \right|}{\left| A_{ii}^{(k-1)} - \left(A_{i1}^{(k-1)} \cdots A_{i,i-1}^{(k-1)}\right) \begin{pmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{pmatrix}^{-1} \begin{pmatrix} A_{1i}^{(k-1)} \\ \vdots \\ A_{1,i-1}^{(k-1)} \end{pmatrix} \right|}$$

$$\div \frac{|A_{ii}^{(k)}|}{|A_{ii}^{(k-1)}|}$$

$$= \frac{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} & A_{1i}^{(k)} \\ \vdots & & \vdots & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} & A_{i-1,i}^{(k)} \\ A_{i1}^{(k)} & \cdots & A_{i,i-1}^{(k)} & A_{ii}^{(k)} \end{vmatrix}}{\begin{vmatrix} A_{11} & \cdots & A_{1,i-1} & A_{1i}^{(k-1)} \\ \vdots & & \vdots & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} & A_{i-1,i}^{(k-1)} \\ A_{i1}^{(k-1)} & \cdots & A_{i,i-1}^{(k-1)} & A_{ii}^{(k-1)} \end{vmatrix}} \cdot \frac{|A_{ii}^{(k-1)}|}{|A_{ii}^{(k)}|},$$

$$k = 2, \ldots, p_i, \quad i = 2, \ldots, q.$$

Under the null hypothesis the criterion has the beta density $\beta[x; \frac{1}{2}(n - \bar{p}_i + 1 - j), \frac{1}{2}p_i]$. For given i, the criteria $X_{i1}, \ldots, X_{ip_i}$ are independent (Theorem 8.4.1). The sets for different i are independent by the argument in Section 9.6.1.

A step-down procedure consists of a sequence of tests based on $X_{21}, \ldots, X_{2p_2}, X_{31}, \ldots, X_{qp_q}$. A particular component test leads to rejection if

$$\frac{1 - X_{ij}}{X_{ij}} \frac{n - \bar{p}_i + 1 - j}{p_i} > F_{p_i, n-\bar{p}_i+1-j}(\varepsilon_{ij}). \tag{6}$$

The significance level is ε, where

$$1 - \varepsilon = \prod_{i=2}^{q} \prod_{j=1}^{p_i} (1 - \varepsilon_{ij}). \tag{7}$$

The sequence of subvectors and the sequence of components within each subvector is at the discretion of the investigator.

The criterion V_i for testing H_i is $V_i = \prod_{k=1}^{p_i} X_{ik}$, and criterion for the null hypothesis H is

$$(8) \qquad V = \prod_{i=2}^{q} V_i = \prod_{i=2}^{q} \prod_{k=1}^{p_i} X_{ik}.$$

These are the random variables described in Theorem 9.3.3.

9.7. AN EXAMPLE

We take the following example from an industrial time study [Abruzzi (1950)]. The purpose of the study was to investigate the length of time taken by various operators in a garment factory to do several elements of a pressing operation. The entire pressing operation was divided into the following six elements:

1. Pick up and position garment.
2. Press and repress short dart.
3. Reposition garment on ironing board.
4. Press three-quarters of length of long dart.
5. Press balance of long dart.
6. Hang garment on rack.

In this case x_α is the vector of measurements on individual α. The component $x_{i\alpha}$ is the time taken to do the ith element of the operation. N is 76. The data (in seconds) are summarized in the sample mean vector and covariance matrix:

$$(1) \qquad \bar{x} = \begin{pmatrix} 9.47 \\ 25.56 \\ 13.25 \\ 31.44 \\ 27.29 \\ 8.80 \end{pmatrix},$$

$$(2) \qquad S = \begin{pmatrix} 2.57 & 0.85 & 1.56 & 1.79 & 1.33 & 0.42 \\ 0.85 & 37.00 & 3.34 & 13.47 & 7.59 & 0.52 \\ 1.56 & 3.34 & 8.44 & 5.77 & 2.00 & 0.50 \\ 1.79 & 13.47 & 5.77 & 34.01 & 10.50 & 1.77 \\ 1.33 & 7.59 & 2.00 & 10.50 & 23.01 & 3.43 \\ 0.42 & 0.52 & 0.50 & 1.77 & 3.43 & 4.59 \end{pmatrix}.$$

The sample standard deviations are (1.604, 6.041, 2.903, 5.832, 4.798, 2.141). The sample correlation matrix is

$$R = \begin{pmatrix} 1.000 & 0.088 & 0.334 & 0.191 & 0.173 & 0.123 \\ 0.088 & 1.000 & 0.186 & 0.384 & 0.262 & 0.040 \\ 0.334 & 0.186 & 1.000 & 0.343 & 0.144 & 0.080 \\ 0.191 & 0.384 & 0.343 & 1.000 & 0.375 & 0.142 \\ 0.173 & 0.262 & 0.144 & 0.375 & 1.000 & 0.334 \\ 0.123 & 0.040 & 0.080 & 0.142 & 0.334 & 1.000 \end{pmatrix}. \tag{3}$$

The investigators are interested in testing the hypothesis that the six variates are mutually independent. It often happens in time studies that a new operation is proposed in which the elements are combined in a different way; the new operation may use some of the elements several times and some elements may be omitted. If the times for the different elements in the operation for which data are available are independent, it may reasonably be assumed that they will be independent in a new operation. Then the distribution of time for the new operation can be estimated by using the means and variances of the individual items.

In this problem the criterion V is $V = |R| = 0.472$. Since the sample size is large we can use asymptotic theory: $k = \frac{433}{6}$, $f = 15$, and $-k \log V = 54.1$. Since the significance point for the χ^2-distribution with 15 degrees of freedom is 30.6 at the 0.01 significance level, we find the result significant. We reject the hypothesis of independence; we cannot consider the times of the elements independent.

9.8. THE CASE OF TWO SETS OF VARIATES

In the case of two sets of variates ($q = 2$), the random vector X, the observation vector x_α, the mean vector $\boldsymbol{\mu}$, and the covariance matrix $\boldsymbol{\Sigma}$ are partitioned as follows:

$$\begin{aligned} X &= \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, & x_\alpha &= \begin{pmatrix} x_\alpha^{(1)} \\ x_\alpha^{(2)} \end{pmatrix}, \\ \boldsymbol{\mu} &= \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}, & \boldsymbol{\Sigma} &= \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \end{aligned} \tag{1}$$

The null hypothesis of independence specifies that $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, that is, that $\boldsymbol{\Sigma}$ is of the form

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \tag{2}$$

The test criterion is

$$V = \frac{|A|}{|A_{11}| \cdot |A_{22}|}. \tag{3}$$

It was shown in Section 9.3 that when the null hypothesis is true, this criterion is distributed as $U_{p_1, p_2, N-1-p_2}$, the criterion for testing a hypothesis about regression coefficients (Chapter 8). We now wish to study further the relationship between testing the hypothesis of independence of two sets and testing the hypothesis that regression of one set on the other is zero.

The conditional distribution of $X_\alpha^{(1)}$ given $X_\alpha^{(2)} = x_\alpha^{(2)}$ is $N[\mu^{(1)} + \beta(x_\alpha^{(2)} - \mu^{(2)}), \Sigma_{11\cdot 2}] = N[\beta(x_\alpha^{(2)} - \bar{x}^{(2)}) + \nu, \Sigma_{11\cdot 2}]$, where $\beta = \Sigma_{12}\Sigma_{22}^{-1}$, $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, and $\nu = \mu^{(1)} + \beta(\bar{x}^{(2)} - \mu^{(2)})$. Let $X_\alpha^* = X_\alpha^{(1)}$, $z_\alpha^{*\prime} = [(x_\alpha^{(2)} - \bar{x}^{(2)})' \; 1]$, $\beta^* = (\beta \;\; \nu)$, and $\Sigma^* = \Sigma_{11\cdot 2}$. Then the conditional distribution of X_α^* is $N(\beta^* z_\alpha^*, \Sigma^*)$. This is exactly the distribution studied in Chapter 8.

The null hypothesis that $\Sigma_{12} = 0$ is equivalent to the null hypothesis $\beta = 0$. Considering $x_\alpha^{(2)}$ fixed, we know from Chapter 8 that the criterion (based on the likelihood ratio criterion) for testing this hypothesis is

$$U = \frac{\left|\sum\left(x_\alpha^* - \hat{\beta}_\Omega^* z_\alpha^*\right)\left(x_\alpha^* - \hat{\beta}_\Omega^* z_\alpha^*\right)'\right|}{\left|\sum\left(x_\alpha^* - \hat{\beta}_{2\omega}^* z_\alpha^{*(2)}\right)\left(x_\alpha^* - \hat{\beta}_{2\omega}^* z_\alpha^{*(2)}\right)'\right|}, \tag{4}$$

where

$$\begin{aligned}
& z_\alpha^{*(2)} = 1, \\
(5) \qquad & \hat{\beta}_{2\omega}^* = \hat{\nu} = \bar{x}^* = \bar{x}^{(1)}, \\
& \hat{\beta}_\Omega^* = \left(\hat{\beta}_{1\Omega}^* \;\; \hat{\beta}_{2\Omega}^*\right) \\
& \quad = \left(\sum x_\alpha^* z_\alpha^{*(1)\prime} \quad \sum x_\alpha^* z_\alpha^{*(2)\prime}\right) \begin{pmatrix} \sum z_\alpha^{*(1)} z_\alpha^{*(1)\prime} & \sum z_\alpha^{*(1)} z_\alpha^{*(2)\prime} \\ \sum z_\alpha^{*(2)} z_\alpha^{*(1)\prime} & \sum z_\alpha^{*(2)} z_\alpha^{*(2)\prime} \end{pmatrix}^{-1} \\
& \quad = \left(A_{12} \quad N\bar{x}^{(1)}\right) \begin{pmatrix} A_{22} & 0 \\ 0 & N \end{pmatrix}^{-1} \\
& \quad = \left(A_{12}A_{22}^{-1} \quad \bar{x}^{(1)}\right).
\end{aligned}$$

The matrix in the denominator of U is

$$\sum_{\alpha=1}^{N} \left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)\left(\bar{x}_\alpha^{(1)} - \bar{x}^{(1)}\right)' = A_{11}. \tag{6}$$

The matrix in the numerator is

$$\sum_{\alpha=1}^{N}\left[x_\alpha^{(1)}-\bar{x}^{(1)}-A_{12}A_{22}^{-1}\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]\left[x_\alpha^{(1)}-\bar{x}^{(1)}-A_{12}A_{22}^{-1}\left(x_\alpha^{(2)}-\bar{x}^{(2)}\right)\right]' \tag{7}$$

$$=A_{11}-A_{12}A_{22}^{-1}A_{21}.$$

Therefore,

$$U=\frac{|A_{11}-A_{12}A_{22}^{-1}A_{21}|}{|A_{11}|}=\frac{|A|}{|A_{11}|\cdot|A_{22}|}, \tag{8}$$

which is exactly V.

Now let us see why it is that when the null hypothesis is true the distribution of $U=V$ does not depend on whether the $X_\alpha^{(2)}$ are held fixed. It was shown in Chapter 8 that when the null hypothesis is true the distribution of U depends only on p, q_1, and $N-q_2$, not on z_α. Thus the conditional distribution of V given $X_\alpha^{(2)}=x_\alpha^{(2)}$ does not depend on $x_\alpha^{(2)}$; the joint distribution of V and $X_\alpha^{(2)}$ is the product of the distribution of V and the distribution of $X_\alpha^{(2)}$, and the marginal distribution of V is this conditional distribution. This shows that the distribution of V (under the null hypothesis) does not depend on whether the $X_\alpha^{(2)}$ are fixed or have any distribution (normal or not).

We can extend this result to show that if $q>2$, the distribution of V under the null hypothesis of independence does not depend on the distribution of one set of variates, say $X_\alpha^{(1)}$. We have $V=V_2\cdots V_q$, where V_i is defined in (1) of Section 9.3. When the null hypothesis is true, V_q is distributed independently of $X_\alpha^{(1)},\dots,X_\alpha^{(q-1)}$ by the previous result. In turn we argue that V_j is distributed independently of $X_\alpha^{(1)},\dots,X_\alpha^{(j-1)}$. Thus $V_2\cdots V_q$ is distributed independently of $X_\alpha^{(1)}$.

Theorem 9.8.1. *Under the null hypothesis of independence, the distribution of V is that given earlier in this chapter if $q-1$ sets are jointly normally distributed, even though one set is not normally distributed.*

In the case of two sets of variates, we may be interested in a measure of association between the two sets which is a generalization of the correlation coefficient. The square of the correlation between two scalars X_1 and X_2 can be considered as the ratio of the variance of the regression of X_1 on X_2 to the variance of X_1; this is $\mathscr{V}(\beta X_2)/\mathscr{V}(X_1)=\beta^2\sigma_{22}/\sigma_{11}=(\sigma_{12}^2/\sigma_{22})/\sigma_{11}=\rho_{12}^2$. A corresponding measure for vectors $X^{(1)}$ and $X^{(2)}$ is the ratio of the generalized variance of the regression of $X^{(1)}$ on $X^{(2)}$ to the generalized variance of $X^{(1)}$, namely,

$$\text{(9)} \qquad \frac{|\mathscr{E}\boldsymbol{\beta}X^{(2)}(\boldsymbol{\beta}X^{(2)})'|}{|\boldsymbol{\Sigma}_{11}|} = \frac{|\boldsymbol{\beta}\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}'|}{|\boldsymbol{\Sigma}_{11}|} = \frac{|\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|}{|\boldsymbol{\Sigma}_{11}|}$$

$$= (-1)^{p_1}\frac{\begin{vmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{vmatrix}}{|\boldsymbol{\Sigma}_{11}|}.$$

If $p_1 = p_2$, the measure is

$$\text{(10)} \qquad \frac{|\boldsymbol{\Sigma}_{12}|^2}{|\boldsymbol{\Sigma}_{11}|\cdot|\boldsymbol{\Sigma}_{22}|}.$$

In a sense this measure shows how well $X^{(1)}$ can be predicted from $X^{(2)}$.

In the case of two scalar variables X_1 and X_2 the coefficient of alienation is $\sigma_{1\cdot 2}^2/\sigma_1^2$, where $\sigma_{1\cdot 2}^2 = \mathscr{E}(X_1 - \beta X_2)^2$ is the variance of X_1 about its regression on X_2 when $\mathscr{E}X_1 = \mathscr{E}X_2 = 0$ and $\mathscr{E}(X_1|X_2) = \beta X_2$. In the case of two vectors $X^{(1)}$ and $X^{(2)}$, the regression matrix is $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, and the generalized variance of $X^{(1)}$ about its regression on $X^{(2)}$ is

$$\text{(11)} \qquad |\mathscr{E}\{(X^{(1)} - \boldsymbol{\beta}X^{(2)})(X^{(1)} - \boldsymbol{\beta}X^{(2)})'\}| = |\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}| = \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_{22}|}.$$

Since the generalized variance of $X^{(1)}$ is $|\mathscr{E}X^{(1)}X^{(1)\prime}| = |\boldsymbol{\Sigma}_{11}|$, the *vector coefficient of alienation* is

$$\text{(12)} \qquad \frac{|\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|}{|\boldsymbol{\Sigma}_{11}|} = \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_{11}|\cdot|\boldsymbol{\Sigma}_{22}|}.$$

The sample equivalent of (12) is simply V.

A measure of association is 1 minus the coefficient of alienation. Either of these two measures of association can be modified to take account of the number of components. In the first case, one can take the p_1th root of (9); in the second case, one can subtract the p_1th root of the coefficient of alienation from 1. Another measure of association is

$$\text{(13)} \qquad \frac{\operatorname{tr}\mathscr{E}[\boldsymbol{\beta}X^{(2)}(\boldsymbol{\beta}X^{(2)})'](\mathscr{E}X^{(1)}X^{(1)\prime})^{-1}}{p} = \frac{\operatorname{tr}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}}{p}.$$

This measure of association ranges between 0 and 1. If $X^{(1)}$ can be predicted exactly from $X^{(2)}$ for $p_1 \le p_2$ (i.e., $\boldsymbol{\Sigma}_{11\cdot 2} = \mathbf{0}$), then this measure is 1. If no linear combination of $X^{(1)}$ can be predicted exactly, this measure is 0.

9.9. ADMISSIBILITY OF THE LIKELIHOOD RATIO TEST

The admissibility of the likelihood ratio test in the case of the 0–1 loss function can be proved by showing that it is the Bayes procedure with respect to an appropriate a priori distribution of the parameters. (See Section 5.6.)

Theorem 9.9.1. *The likelihood ratio test of the hypothesis that $\mathbf{\Sigma}$ is of the form* (6) *of Section* 9.2 *is Bayes and admissible if $N > p + 1$.*

Proof. We shall show that the likelihood ratio test is equivalent to rejection of the hypothesis when

$$\frac{\int f(\boldsymbol{x}|\boldsymbol{\theta})\Pi_1(d\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})\Pi_0(d\boldsymbol{\theta})} \geq c, \tag{1}$$

where $\boldsymbol{x}$ represents the sample, $\boldsymbol{\theta}$ represents the parameters ($\boldsymbol{\mu}$ and $\mathbf{\Sigma}$), $f(\boldsymbol{x}|\boldsymbol{\theta})$ is the density, and Π_1 and Π_0 are proportional to probability measures of $\boldsymbol{\theta}$ under the alternative and null hypotheses, respectively. Specifically, the left-hand side is to be proportional to the square root of $\prod_{i=1}^{q}|\boldsymbol{A}_{ii}|/|\boldsymbol{A}|$.

To define Π_1, let

$$\boldsymbol{\mu} = (\boldsymbol{I} + \boldsymbol{V}\boldsymbol{V}')^{-1}\boldsymbol{V}Y, \qquad \mathbf{\Sigma} = (\boldsymbol{I} + \boldsymbol{V}\boldsymbol{V}')^{-1}, \tag{2}$$

where the p-component random vector $\boldsymbol{V}$ has the density proportional to $(1 + \boldsymbol{v}'\boldsymbol{v})^{-\frac{1}{2}n}$, $n = N - 1$, and the conditional distribution of Y given $\boldsymbol{V} = \boldsymbol{v}$ is $N[0, (1 + \boldsymbol{v}'\boldsymbol{v})/N]$. Note that the integral of $(1 + \boldsymbol{v}'\boldsymbol{v})^{-\frac{1}{2}n}$ is finite if $n > p$ (Problem 5.15). The numerator of (1) is then

$$\begin{aligned}
&\text{const} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\boldsymbol{I} + \boldsymbol{v}\boldsymbol{v}'|^{\frac{1}{2}N} \\
&\cdot \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{N}\left[\boldsymbol{x}_\alpha - (\boldsymbol{I} + \boldsymbol{v}\boldsymbol{v}')^{-1}\boldsymbol{v}y\right]'(\boldsymbol{I} + \boldsymbol{v}\boldsymbol{v}')\left[\boldsymbol{x}_\alpha - (\boldsymbol{I} + \boldsymbol{v}\boldsymbol{v}')^{-1}\boldsymbol{v}y\right]\right\} \\
&\cdot (1 + \boldsymbol{v}'\boldsymbol{v})^{-\frac{1}{2}n} \cdot (1 + \boldsymbol{v}'\boldsymbol{v})^{-\frac{1}{2}} \exp\left\{\frac{-\frac{1}{2}Ny^2}{1 + \boldsymbol{v}'\boldsymbol{v}}\right\} d\boldsymbol{v}\, dy.
\end{aligned} \tag{3}$$

The exponent in the integrand of (3) is -2 times

$$(4)\quad \sum_{\alpha=1}^{N} x_\alpha'(I+vv')x_\alpha - 2yv'\sum_{\alpha=1}^{N} x_\alpha + Ny^2v'(I+vv')^{-1}v + \frac{Ny^2}{1+v'v}$$

$$= \sum_{\alpha=1}^{N} x_\alpha' x_\alpha + v'\sum_{\alpha=1}^{N} x_\alpha x_\alpha' v - 2yv'N\bar{x} + Ny^2$$

$$= \operatorname{tr} A + v'Av + N\bar{x}'\bar{x} + N(y-\bar{x}'v)^2,$$

where $A = \sum_{\alpha=1}^{N} x_\alpha x_\alpha' - N\bar{x}\bar{x}'$. We have used $v'(I+vv')^{-1}v + (1+v'v)^{-1} = 1$ [from $(I+vv')^{-1} = I - (1+v'v)^{-1}vv'$]. Using $|I+vv'| = 1+v'v$ (Corollary A.3.1), we write (3) as

$$(5)\quad \text{const}\, e^{-\frac{1}{2}\operatorname{tr} A - \frac{1}{2}N\bar{x}'\bar{x}} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-\frac{1}{2}v'Av}\,dv = \text{const}|A|^{-\frac{1}{2}} e^{-\frac{1}{2}\operatorname{tr} A - \frac{1}{2}N\bar{x}'\bar{x}}.$$

To define Π_0 let $\mathbf{\Sigma}$ have the form of (6) of Section 9.2. Let

$$(6)\quad \left[\boldsymbol{\mu}^{(i)}, \Sigma_{ii}\right] = \left[(I+V^{(i)}V^{(i)\prime})^{-1}V^{(i)}Y_i, (I+V^{(i)}V^{(i)\prime})^{-1}\right], \qquad i=1,\dots,q,$$

where the p_i-component random vector $V^{(i)}$ has density proportional to $(1+v^{(i)\prime}v^{(i)})^{-\frac{1}{2}n}$, and the conditional distribution of Y_i given $V^{(i)} = v^{(i)}$ is $N[0,(1+v^{(i)\prime}v^{(i)})/N]$, and let $(V_1,Y_1),\dots,(V_q,Y_q)$ be mutually independent. Then the denominator of (1) is

$$(7)\quad \prod_{i=1}^{q} \text{const}|A_{ii}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\left(\operatorname{tr} A_{ii} + N\bar{x}^{(i)\prime}\bar{x}^{(i)}\right)\right]$$

$$= \text{const}\left(\prod_{i=1}^{q}|A_{ii}|^{-\frac{1}{2}}\right)\exp\left[-\tfrac{1}{2}(\operatorname{tr} A + N\bar{x}'\bar{x})\right].$$

The left-hand side of (1) is then proportional to the square root of $\prod_{i=1}^{q}|A_{ii}|/|A|$. ■

This proof has been adapted from that of Kiefer and Schwartz (1965).

9.10. MONOTONICITY OF POWER FUNCTIONS OF TESTS OF INDEPENDENCE OF SETS

Let $Z_\alpha = [Z_\alpha^{(1)\prime}, Z_\alpha^{(2)\prime}]'$, $\alpha = 1,\dots,n$, be distributed according to

$$(1)\qquad N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{Z}_{21} & \mathbf{Z}_{22}\end{pmatrix}\right].$$

We want to test $H: \Sigma_{12} = 0$. We suppose $p_1 \le p_2$ without loss of generality. Let $\rho_1, \dots, \rho_{p_1}$ ($\rho_1 \ge \cdots \ge \rho_{p_1}$) be the (population) canonical correlation coefficients. (The ρ_i^2's are the characteristic roots of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, Chapter 12.) Let $R = \operatorname{diag}(\rho_1, \dots, \rho_{p_1})$ and $\Delta = [R, 0]$ $(p_1 \times p_2)$.

Lemma 9.10.1. *There exist matrices* B_1 $(p_1 \times p_1)$, B_2 $(p_2 \times p_2)$ *such that*

$$B_1 \Sigma_{11} B_1' = I_{p_1}, \qquad B_2 \Sigma_{22} B_2' = I_{p_2}, \qquad B_1 \Sigma_{12} B_2' = \Delta. \tag{2}$$

Proof. Let $m = p_2$, $B = B_1$, $F' = \Sigma_{22}^{\frac{1}{2}} B_2'$, $\Xi = \Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ in Lemma 8.10.13. Then $F'F = B_2 \Sigma_{22} B_2' = I_{p_2}$, $B_1 \Sigma_{12} B_2' = B_1 \Xi F = \Delta$. ■

(This lemma is also contained in Section 12.2.)

Let $x_\alpha = B_1 Z_\alpha^{(1)}$, $y_\alpha = B_2 Z_\alpha^{(2)}$, $\alpha = 1, \dots, n$, and $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_n)$. Then $(x_\alpha', y_\alpha')'$, $\alpha = 1, \dots, n$, are independently distributed according to

$$N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_p & \Delta \\ \Delta' & I_q \end{pmatrix}\right]. \tag{3}$$

The hypothesis $H: \Sigma_{12} = 0$ is equivalent to $H: \Delta = 0$ (i.e., all the canonical correlation coefficients $\rho_1, \dots, \rho_{p_1}$ are zero). Now given Y, the vectors x_α, $\alpha = 1, \dots, n$, are conditionally independently distributed according to $N(\Delta y_\alpha, I - \Delta\Delta') = N(\Delta y_\alpha, I - R^2)$. Then $x_\alpha^* = (I_{p_1} - R^2)^{-\frac{1}{2}} x_\alpha$ is distributed according to $N(M y_\alpha, I_{p_1})$ where

$$\begin{aligned} M &= (D, 0), \\ D &= \operatorname{diag}(\delta_1, \dots, \delta_{p_1}), \\ \delta_i &= \rho_i / (1 - \rho_i^2)^{\frac{1}{2}}, \qquad i = 1, \dots, p_1. \end{aligned} \tag{4}$$

Note that δ_i^2 is a characteristic root of $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\cdot 2}^{-1}$, where $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Invariant tests depend only on the (sample) canonical correlation coefficients $r_i = \sqrt{c_i}$, where

$$c_i = \lambda_i\left[(X^*X^{*\prime})^{-1}(X^*Y')(YY')^{-1}(YX^{*\prime})\right]. \tag{5}$$

Let

$$\begin{aligned} S_h &= X^* Y'(YY')^{-1} Y X^{*\prime}, \\ S_e &= X^* X^{*\prime} - S_h = X^*\left[I - Y'(YY')^{-1} Y\right] X^{*\prime}. \end{aligned} \tag{6}$$

Then

$$\lambda_i(S_h S_e^{-1}) = \frac{c}{1-c_i}. \tag{7}$$

Now given Y, the problem reduces to the MANOVA problem and we can apply Theorem 8.10.6 as follows. There is an orthogonal transformation (Section 8.3.3) that carries X^* to (U, V) such that $S_h = UU'$, $S_e = VV'$, $U = (u_1, \ldots, u_{p_2})$, V is $p_1 \times (n - p_2)$, u_i has the distribution $N(\delta_i \varepsilon_i, I)$, $i = 1, \ldots, p_1$ (ε_i being the ith column of I), and $N(0, I)$, $i = p_1 + 1, \ldots, p_2$, and the columns of V are independently distributed according to $N(0, I)$. Then $c_1, \ldots, c_{p_1}$ are the characteristic roots of $UU'(VV')^{-1}$, and their distribution depends on the characteristic roots of $MYY'M'$, say, $\tau_1^2, \ldots, \tau_{p_1}^2$. Now from Theorem 8.10.6, we obtain the following lemma.

Lemma 9.10.2. *If the acceptance region of an invariant test is convex in each column of U, given V and the other columns of U, then the conditional power given Y increases in each characteristic root τ_i^2 of $MYY'M'$.*

Lemma 9.10.3. *If $A \geq B$, then $\lambda_i(A) \geq \lambda_i(B)$.*

Proof. By the minimax property of the characteristic roots [see, e.g., Courant and Hilbert (1953)],

$$\lambda_i(A) = \max_{S_i} \min_{x \in S_i} \frac{x'Ax}{x'x} \geq \max_{S_i} \min_{x \in S_i} \frac{x'Bx}{x'x} = \lambda_i(B), \tag{8}$$

where S_i ranges over i-dimensional subspaces. ∎

Now Lemma 9.10.3 applied to $MYY'M'$ shows that for every j, τ_j^2 is an increasing function of $\delta_i = \rho_i/(1-\rho_i^2)^{\frac{1}{2}}$ and hence of ρ_i. Since the marginal distribution of Y does not depend on the ρ_i's, by taking the unconditional power we obtain the following theorem.

Theorem 9.10.1. *An invariant test for which the acceptance region is convex in each column of U for each set of fixed V and other columns of U has a power function that is monotonically increasing in each ρ_i.*

9.11. ELLIPTICALLY CONTOURED DISTRIBUTIONS

9.11.1. Observations Elliptically Contoured

Let $x_1, \ldots, x_N$ be N observations on a random vector X with density

$$|\Lambda|^{-\frac{1}{2}} g\left[(x-\nu)'\Lambda^{-1}(x-\nu)\right], \tag{1}$$

where $\mathscr{E}R^4 < \infty$ and $R^2 = (x - \nu)'\Lambda^{-1}(x - \nu)$. Then $\mathscr{E}X = \nu$ and $\mathscr{E}(X - \nu)(X - \nu)' = \Sigma = (\mathscr{E}R^2/p)\Lambda$. Let

$$\bar{x} = \frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha, \qquad S = \frac{1}{N-1}\sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'. \tag{2}$$

Then

$$\sqrt{N}\operatorname{vec}(S - \Sigma) \xrightarrow{d} N\left[0, (\kappa+1)(I_{p^2} + K_{pp})(\Sigma \otimes \Sigma) + \kappa \operatorname{vec}\Sigma(\operatorname{vec}\Sigma)'\right] \tag{3}$$

where $1 + \kappa = p\,\mathscr{E}R^4/[(p+2)(\mathscr{E}R^2)^2]$.

The likelihood ratio criterion for testing the null hypothesis $\Sigma_{ij} = 0$, $i \neq j$, is the $N/2$th power of $U = \prod_{i=2}^{q} V_i$, where V_i is the U-criterion for testing the null hypothesis $\Sigma_{1i} = 0, \dots, \Sigma_{i-1,i} = 0$ and is given by (1) and (6) of Section 9.3. The form of V_i is that of the likelihood ratio criterion U of Chapter 8 with X replaced by $X^{(i)}$, $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_i$ given by (5) of Section 9.3, Z by

$$\bar{X}^{(i-1)} = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(i-1)} \end{bmatrix}, \tag{4}$$

and Σ by Σ_{ii} under the null hypothesis $\boldsymbol{\beta}_i = 0$. The subvector $\bar{X}^{(i-1)}$ is uncorrelated with $X^{(i)}$, but not independent of $X^{(i)}$ unless $(\bar{X}^{(i-1)\prime}, X^{(i)\prime})'$ is normal. Let

$$\bar{A}^{(i-1)} = \begin{bmatrix} A_{11} & \cdots & A_{1,i-1} \\ \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} \end{bmatrix}, \tag{5}$$

$$\bar{A}^{(i,i-1)} = (A_{i1}, \dots, A_{i,i-1}) = \bar{A}^{(i-1,i)\prime} \tag{6}$$

with similar definitions of $\bar{\Sigma}^{(i-1)}$, $\bar{\Sigma}^{(i,i-1)}$, $\bar{S}^{(i-1)}$, and $\bar{S}^{(i,i-1)}$. We write $V_i = |G_i| / |G_i + H_i|$, where

$$\begin{aligned} H_i &= \bar{A}^{(i,i-1)}\left(\bar{A}^{(i-1)}\right)^{-1}\bar{A}^{(i-1,i)} \\ &= (N-1)\bar{S}^{(i,i-1)}\left(\bar{S}^{(i-1)}\right)^{-1}\bar{S}^{(i-1,i)}, \end{aligned} \tag{7}$$

$$G_i = A_{ii} - H_i = (N-1)S_{ii} - H_i. \tag{8}$$

Theorem 9.11.1. *When X has the density* (1) *and the null hypothesis is true, the limiting distribution of H_i is $W[(1+\kappa)\Sigma_{ii}, \bar{p}_i]$, where $\bar{p}_i = p_1 + \cdots + p_{i-1}$*

and p_j is the number of components of $X^{(j)}$.

Proof. Since $\bar{\Sigma}^{(i,i-1)} = 0$, we have $\mathscr{E}\bar{S}^{(i,i-1)} = 0$ and

$$\mathscr{E}s_{jk}s_{lm} = \left(\frac{\kappa}{N} + \frac{1}{N-1}\right)\sigma_{jl}\sigma_{km} \tag{9}$$

if $j, l \le \bar{p}_i$ and $k, m > \bar{p}_i$ or if $j, l > \bar{p}_i$, and $k, m \le \bar{p}_i$, *and* $\mathscr{E}s_{jk}s_{lm} = 0$ otherwise (Theorem 3.6.1). We can write

$$\mathscr{E}\operatorname{vec}\bar{S}^{(i,i-1)}(\operatorname{vec}\bar{S}^{(i,i-1)})' = \left(\frac{\kappa}{N} + \frac{1}{N-1}\right)\left(\Sigma_{ii} \otimes \bar{\Sigma}^{(i-1)}\right). \tag{10}$$

Since $S^{(i-1)} \to_p \Sigma^{(i-1)}$ and $\sqrt{N}\operatorname{vec}S^{(i,i-1)}$ has a limiting normal distribution, Theorem 9.10.1 follows by (2) of Section 8.4. ■

Theorem 9.11.2. *Under the conditions of Theorem* 9.11.1 *when the null hypothesis is true*

$$-N\log V_i \xrightarrow{d} (1+\kappa)\chi^2_{\bar{p}_i p_i}. \tag{11}$$

Proof. We can write $V_i = |I + N^{-1}(\frac{1}{N}G_i)^{-1}H_i|$ and use $N\log|I + N^{-1}C| = \operatorname{tr}C + O_p(N^{-1})$ and

$$\begin{aligned}\operatorname{tr}\left(\frac{1}{N}G_i\right)^{-1}H_i &= N\sum_{i,j=\bar{p}_i+1}^{\bar{p}_{i-1}}\sum_{g,h=1}^{\bar{p}_i} g^{ij}s_{ig}\bar{s}^{gh}s_{jh} \\ &= N(\operatorname{vec}S^{(i,i-1)})'\left[\left(\frac{1}{N}G_i\right)^{-1} \otimes \bar{S}_{ii}^{-1}\right]\operatorname{vec}S^{(i-1,i)}.\end{aligned} \tag{12}$$

■

Because $X^{(i)}$ is uncorrelated with $\bar{X}^{(i-1)}$ when the null hypothesis $H_i: \bar{\Sigma}^{(i,i-1)} = 0$, V_i is asymptotically independent of $V_2, \dots, V_{i-1}$. When the null hypotheses $H_2, \dots, H_i$ are true, V_i is asymptotically independent of $V_2, \dots, V_{i-1}$. It follows from Theorem 9.10.2 that

$$-N\log V = -N\sum_{i=2}^{q}\log V_i \xrightarrow{d} \chi^2_f, \tag{13}$$

where $f = \sum_{i=2}^{p}\bar{p}_i p_i = \frac{1}{2}[p(p+1) - \sum_{i=1}^{q}p_i(p_i+1)]$. The likelihood ratio test of Section 9.2 can be carried out on an asymptotic basis.

Let $A_0 = \operatorname{diag}(A_{11}, \dots, A_{qq})$. Then

$$\frac{1}{2}\operatorname{tr}\left(AA_0^{-1} - I\right)^2 = \frac{1}{2}\sum_{\substack{i,j=1 \\ i \neq j}}^{g} \operatorname{tr} A_{ij}A_{jj}^{-1}A_{ji}A_{ii}^{-1} \tag{14}$$

has the χ_f^2-distribution when $\Sigma = \operatorname{diag}(\Sigma_{11}, \dots, \Sigma_{qq})$. The step-down procedure of Section 9.6.1 is also justified on an asymptotic basis.

9.11.2. Elliptically Contoured Matrix Distributions

Let Y $(p \times N)$ have the density $g(\operatorname{tr} YY')$. The matrix Y is vector-spherical; that is, $\operatorname{vec} Y$ is spherical and has the stochastic representation $\operatorname{vec} Y = R \operatorname{vec} U_{p \times N}$, where $R^2 = (\operatorname{vec} Y)' \operatorname{vec} Y = \operatorname{tr} YY'$ and $\operatorname{vec} U_{p\times N}$ has the uniform distribution on the unit sphere $(\operatorname{vec} U_{p\times N})' \operatorname{vec} U_{p\times N} = 1$. (We use the notation $U_{p\times N}$ to distinguish from U uniform on the space $UU' = I_p$).

Let

$$X = \nu \varepsilon_N' + CY, \tag{15}$$

where $\Lambda = CC'$ and C is lower triangular. Then X has the density

$$\begin{aligned} &|\Lambda|^{-N/2} g\left[\operatorname{tr} C^{-1}(X - \nu\varepsilon_N')(X' - \varepsilon_N \nu')(C')\right]^{-1} \\ &\quad = |\Lambda|^{-N/2} g\left[\operatorname{tr}(X' - \varepsilon_n \nu')\Lambda^{-1}(X - \nu\varepsilon_N')\right]. \end{aligned} \tag{16}$$

Consider the null hypothesis $\Sigma_{ij} = \mathbf{0}$, $i \neq j$, or alternatively $\Lambda_{ij} = \mathbf{0}$, $i \neq j$, or alternatively, $R_{ij} = \mathbf{0}$, $i \neq j$. Then $C = \operatorname{diag}(C_{11}, \dots, C_{qq})$.

Let $M = I_N - (1/N)\varepsilon_N \varepsilon_N'$; since $M^2 = M$, M is an idempotent matrix with $N-1$ characteristic roots 1 and one root 0. Then $A = XMX'$ and $A_{ii} = X^{(i)}MX^{(i)\prime}$. The likelihood function is

$$|\Lambda|^{-n/2} g\{\operatorname{tr} \Lambda^{-1}[A + N(\bar{x} - \nu)(\bar{x} - \nu)']\}. \tag{17}$$

The matrix A and the vector $\bar{x}$ are sufficient statistics, and the likelihood ratio criterion for the hypothesis H is $(|A| / \prod_{i=1}^{g} |A_{ii}|)^{N/2}$, the same as for normality. See Anderson and Fang (1990b).

Theorem 9.11.3. *Let $f(X)$ be a vector-valued function of X $(p \times N)$ such that*

$$f(X + \nu\varepsilon_N') = f(X) \tag{18}$$

for all ν *and*

$$f(KX) = f(X) \tag{19}$$

for all $K = \operatorname{diag}(K_{11}, \ldots, K_{qq})$. *Then the distribution of* $f(X)$, *where* X *has the arbitrary density* (16), *is the same as the distribution of* $f(X)$, *where* X *has the normal density* (16).

Proof. The proof is similar to the proof of Theorem 4.5.4. ■

It follows from Theorem 9.11.3 that V has the same distribution under the null hypothesis H when X has the density (16) and for X normally distributed since V is invariant under the transformation $X \to KX$. Similarly, V_t and the criterion (14) are invariant, and hence have the distribution under normality.

PROBLEMS

9.1. (Sec. 9.3) Prove

$$\mathscr{E}V^h = \frac{\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)+h\right]\prod_{i=1}^{q}\left\{\prod_{j=1}^{p_i}\Gamma\left[\frac{1}{2}(n+1-j)\right]\right\}}{\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]\prod_{i=1}^{q}\left\{\prod_{j=1}^{p_i}\Gamma\left[\frac{1}{2}(n+1-j)+h\right]\right\}}$$

by integration of $V^h w(A|\Sigma_0, n)$. *Hint:* Show

$$\mathscr{E}V^h = \frac{K(\Sigma_0, n)}{K(\Sigma_0, n+2h)} \int \cdots \int \prod_{i=1}^{q} |A_{ii}|^{-h} w(A, \Sigma_0, n+2h)\, dA,$$

where $K(\Sigma, n)$ is defined by $w(A|\Sigma, n) = K(\Sigma, n)|A|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\operatorname{tr}\Sigma^{-1}A}$. Use Theorem 7.3.5 to show

$$\mathscr{E}V^h = \frac{K(\Sigma_0, n)}{K(\Sigma_0, n+2h)} \prod_{i=1}^{q}\left[\frac{K(\Sigma_{ii}, n+2h)}{K(\Sigma_{ii}, n)} \int \cdots \int w(A_{ii}|\Sigma_{ii}, n)\, dA_{ii}\right].$$

9.2. (Sec. 9.3) Prove that if $p_1 = p_2 = p_3 = 1$ [Wilks (1935)]

$$\Pr\{V \le v\} = I_v\left[\tfrac{1}{2}(n-1), \tfrac{1}{2}\right] + 2B^{-1}\left[\tfrac{1}{2}(n-1), \tfrac{1}{2}\right]\sin^{-1}\sqrt{1-v}.$$

[*Hint:* Use Theorem 9.3.3 and $\Pr\{V \le v\} = 1 - \Pr\{v \le V\}$.]

9.3. (Sec. 9.3) Prove that if $p_1 = p_2 = p_3 = 2$ [Wilks (1935)]

$$\begin{aligned}\Pr\{V \le v\} &= I_{\sqrt{v}}(n-5,4) \\ &\quad + B^{-1}(n-5,4)v^{\frac{1}{2}(n-5)}\Big\{n/6 - \tfrac{3}{2}(n-1)\sqrt{v} - \tfrac{3}{2}(n-4)v \\ &\qquad + \left(\tfrac{17}{6}n - \tfrac{15}{2}\right)v^{3/2} \\ &\qquad - \tfrac{3}{2}(n-2)v\log v - \tfrac{1}{2}(n-3)v^{3/2}\log v\Big\}.\end{aligned}$$

[*Hint:* Use (9).]

9.4. (Sec. 9.3) Derive some of the distributions obtained by Wilks (1935) and referred to at the end of Section 9.3.3. [*Hint:* In addition to the results for Problems 9.2 and 9.3, use those of Section 9.3.2.]

9.5. (Sec. 9.4) For the case $p_i = 2$, express k and γ_2. Compute the second term of (6) when v is chosen so that the first term is 0.95 for $p = 4$ and 6 and $N = 15$.

9.6. (Sec. 9.5) Prove that if $\boldsymbol{BAB}' = \boldsymbol{CAC}' = \boldsymbol{I}$ for $\boldsymbol{A}$ positive definite and $\boldsymbol{B}$ and $\boldsymbol{C}$ nonsingular then $\boldsymbol{B} = \boldsymbol{QC}$ where $\boldsymbol{Q}$ is orthogonal.

9.7. (Sec. 9.5) Prove N times (2) has a limiting χ^2-distribution with f degrees of freedom under the null hypothesis.

9.8. (Sec. 9.8) Give the sample vector coefficient of alienation and the vector correlation coefficient.

9.9. (Sec. 9.8) If y is the sample vector coefficient of alienation and z the square of the vector correlation coefficient, find $\mathscr{E}y^g z^h$ when $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

9.10. (Sec. 9.9) Prove

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{\left(1 + \sum_{i=1}^{p} v_i^2\right)^{\frac{1}{2}r}} \, dv_1 \cdots dv_p < \infty$$

if $p < n$. [*Hint:* Let $y_j = w_j\sqrt{1 + \sum_{i=j+1}^{p} y_i^2}$, $j = 1, \ldots, p-1$, in turn.]

9.11. Let x_1 = arithmetic speed, x_2 = arithmetic power, x_3 = intellectual interest, x_4 = social interest, x_5 = activity interest. Kelley (1928) observed the following correlations between batteries of tests identified as above, based on 109 pupils:

$$\begin{pmatrix} 1.0000 & 0.4249 & -0.0552 & -0.0031 & 0.1927 \\ 0.4249 & 1.0000 & -0.0416 & 0.0495 & 0.0687 \\ -0.0552 & -0.0416 & 1.0000 & 0.7474 & 0.1691 \\ -0.0031 & 0.0495 & 0.7474 & 1.0000 & 0.2653 \\ 0.1927 & 0.0687 & 0.1691 & 0.2653 & 1.0000 \end{pmatrix}.$$

Let $x^{(1)\prime} = (x_1, x_2)$ and $x^{(2)\prime} = (x_3, x_4, x_5)$. Test the hypothesis that $x^{(1)}$ is independent of $x^{(2)}$ at the 1% significance level.

9.12. Carry out the same exercise on the data in Problem 3.42.

9.13. Another set of time-study data [Abruzzi (1950)] is summarized by the correlation matrix based on 188 observations:

$$\begin{pmatrix} 1.00 & -0.27 & 0.06 & 0.07 & 0.02 \\ -0.27 & 1.00 & -0.01 & -0.02 & -0.02 \\ 0.06 & -0.01 & 1.00 & -0.07 & -0.04 \\ 0.07 & -0.02 & -0.07 & 1.00 & -0.10 \\ 0.02 & -0.02 & -0.04 & -0.10 & 1.00 \end{pmatrix}.$$

Test the hypothesis that $\sigma_{ij} = 0$, $i \neq j$, at the 5% significance level.

CHAPTER 10

Testing Hypotheses of Equality of Covariance Matrices and Equality of Mean Vectors and Covariance Matrices

10.1. INTRODUCTION

In this chapter we study the problems of testing hypotheses of equality of covariance matrices and equality of both covariance matrices and mean vectors. In each case (except one) the problem and tests considered are multivariate generalizations of a univariate problem and test. Many of the tests are likelihood ratio tests or modifications of likelihood ratio tests. Invariance considerations lead to other test procedures.

First, we consider equality of covariance matrices and equality of covariance matrices and mean vectors of several populations without specifying the common covariance matrix or the common covariance matrix and mean vector. The multivariate analysis of variance with random factors is considered in this context. Later we treat the equality of a covariance matrix to a given matrix and also simultaneous equality of a covariance matrix to a given matrix and equality of a mean vector to a given vector. One other hypothesis considered, the equality of a covariance matrix to a given matrix except for a proportionality factor, has only a trivial corresponding univariate hypothesis.

In each case the class of tests for a class of hypotheses leads to a confidence region. Families of simultaneous confidence intervals for covariances and for ratios of covariances are given.

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

The application of the tests for elliptically contoured distributions is treated in Section 10.11.

10.2. CRITERIA FOR TESTING EQUALITY OF SEVERAL COVARIANCE MATRICES

In this section we study several normal distributions and consider using a set of samples, one from each population, to test the hypothesis that the covariance matrices of these populations are equal. Let $x_\alpha^{(g)}$, $\alpha = 1, \dots, N_g$, $g = 1, \dots, q$, be an observation from the gth population $N(\mu^{(g)}, \Sigma_g)$. We wish to test the hypothesis

$$H_1 : \Sigma_1 = \cdots = \Sigma_q. \tag{1}$$

Let $\sum_{g=1}^q N_g = N$,

$$A_g = \sum_{\alpha=1}^{N_g} \left(x_\alpha^{(g)} - \bar{x}^{(g)}\right)\left(x_\alpha^{(g)} - \bar{x}^{(g)}\right)', \qquad g = 1, \dots, q, \tag{2}$$

$$A = \sum_{g=1}^{q} A_g.$$

First we shall obtain the likelihood ratio criterion. The likelihood function is

$$L = \prod_{g=1}^{q} \frac{1}{(2\pi)^{\frac{1}{2}pN_g} |\Sigma_g|^{\frac{1}{2}N_g}} \exp\left[-\frac{1}{2} \sum_{\alpha=1}^{N_g} \left(x_\alpha^{(g)} - \mu^{(g)}\right)' \Sigma_g^{-1} \left(x_\alpha^{(g)} - \mu^{(g)}\right)\right]. \tag{3}$$

The space Ω is the parameter space in which each Σ_g is positive definite and $\mu^{(g)}$ any vector. The space ω is the parameter space in which $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_q$ (positive definite) and $\mu^{(g)}$ is any vector. The maximum likelihood estimators of $\mu^{(g)}$ and Σ_g in Ω are given by

$$\hat{\mu}_\Omega^{(g)} = \bar{x}^{(g)}, \qquad \hat{\Sigma}_{g\Omega} = \frac{1}{N_g} A_g, \qquad g = 1, \dots, q. \tag{4}$$

The maximum likelihood estimators of $\mu^{(g)}$ in ω are given by (4), $\hat{\mu}_\omega^{(g)} = \bar{x}^{(g)}$, since the maximizing values of $\mu^{(g)}$ are the same regardless of Σ_g. The function to be maximized with respect to $\Sigma_1 = \cdots = \Sigma_q = \Sigma$, say, is

$$\frac{1}{(2\pi)^{\frac{1}{2}pN} |\Sigma|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2} \sum_{g=1}^{q} \sum_{\alpha=1}^{N_g} \left(x_\alpha^{(g)} - \bar{x}^{(g)}\right)' \Sigma^{-1} \left(x_\alpha^{(g)} - \bar{x}^{(g)}\right)\right]. \tag{5}$$

By Lemma 3.2.2, the maximizing value of Σ is

$$\hat{\Sigma}_\omega = \frac{1}{N} A, \tag{6}$$

and the maximum of the likelihood function is

$$\frac{1}{(2\pi)^{\frac{1}{2}pN} |\hat{\Sigma}_\omega|^{\frac{1}{2}N}} e^{-\frac{1}{2}pN}. \tag{7}$$

The likelihood ratio criterion for testing (1) is

$$\lambda_1 = \frac{\prod_{g=1}^q |\hat{\Sigma}_{g\Omega}|^{\frac{1}{2}N_g}}{|\hat{\Sigma}_\omega|^{\frac{1}{2}N}} = \frac{\prod_{g=1}^q |A_g|^{\frac{1}{2}N_g}}{|A|^{\frac{1}{2}N}} \cdot \frac{N^{\frac{1}{2}pN}}{\prod_{g=1}^q N_g^{\frac{1}{2}pN_g}}. \tag{8}$$

The critical region is

$$\lambda_1 \le \lambda_1(\varepsilon), \tag{9}$$

where $\lambda_1(\varepsilon)$ is defined so that (9) holds with probability ε when (1) is true.

Bartlett (1937a) has suggested modifying λ_1 in the univariate case by replacing sample numbers by the numbers of degrees of freedom of the A_g. Except for a numerical constant, the statistic he proposes is

$$V_1 = \frac{\prod_{g=1}^q |A_g|^{\frac{1}{2}n_g}}{|A|^{\frac{1}{2}n}}, \tag{10}$$

where $n_g = N_g - 1$ and $n = \sum_{g=1}^q n_g = N - q$. The numerator is proportional to a power of a weighted geometric mean of the sample generalized variances, and the denominator is proportional to a power of the determinant of a weighted arithmetic mean of the sample covariance matrices.

In the scalar case ($p = 1$) of two samples the criterion (10) is

$$\frac{(n_1)^{\frac{1}{2}n_1}(n_2)^{\frac{1}{2}n_2}(s_1^2)^{\frac{1}{2}n_1}(s_2^2)^{\frac{1}{2}n_2}}{(n_1 s_1^2 + n_2 s_2^2)^{\frac{1}{2}(n_1+n_2)}} = \frac{(n_1)^{\frac{1}{2}n_1}(n_2)^{\frac{1}{2}n_2} F^{\frac{1}{2}n_1}}{(n_1 F + n_2)^{\frac{1}{2}(n_1+n_2)}}, \tag{11}$$

where s_1^2 and s_2^2 are the usual unbiased estimators of σ_1^2 and σ_2^2 (the two population variances) and

$$F = \frac{s_1^2}{s_2^2}. \tag{12}$$

Thus the critical region

$$V_1 \le V_1(\varepsilon) \tag{13}$$

is based on the F-statistic with n_1 and n_2 degrees of freedom, and the inequality (13) implies a particular method of choosing $F_1(\varepsilon)$ and $F_2(\varepsilon)$ for the critical region

$$F \le F_1(\varepsilon), \qquad F \ge F_2(\varepsilon). \tag{14}$$

Brown (1939) and Scheffé (1942) have shown that (14) yields an unbiased test.

Bartlett gave a more intuitive argument for the use of V_1 in place of λ_1. He argues that if N_1, say, is small, A_1 is given too much weight in λ_1, and other effects may be missed. Perlman (1980) has shown that the test based on V_1 is unbiased.

If one assumes

$$\mathscr{E} X_\alpha^{(g)} = \boldsymbol{\beta}_g z_\alpha^{(g)}, \tag{15}$$

where $z_\alpha^{(g)}$ consists of k_g components, and if one estimates the matrix $\boldsymbol{\beta}_g$, defining

$$A_g = \sum_{\alpha=1}^{N_g} \left(x_\alpha^{(g)} - \hat{\boldsymbol{\beta}}_g z_\alpha^{(g)}\right)\left(x_\alpha - \hat{\boldsymbol{\beta}}_g z_\alpha^{(g)}\right)', \tag{16}$$

one uses (10) with $n_g = N_g - k_g$.

The statistical problem (parameter space Ω and null hypothesis ω) is invariant with respect to changes of location within populations and a common linear transformation

$$X^{*(g)} = CX^{(g)} + \boldsymbol{\nu}^{(g)}, \qquad g = 1, \dots, q, \tag{17}$$

where C is nonsingular. Each matrix A_g is invariant under change of location, and the modified criterion (10) is invariant:

$$V_1^* = \frac{\prod_{g=1}^q |A_g^*|^{\frac{1}{2}n_g}}{|A^*|^{\frac{1}{2}n}} = \frac{\prod_{g=1}^q |CA_gC'|^{\frac{1}{2}n_g}}{|CAC'|^{\frac{1}{2}n}} = \frac{\prod_{g=1}^q |A_g|^{\frac{1}{2}n_g}}{|A|^{\frac{1}{2}n}} = V_1. \tag{18}$$

Similarly, the likelihood ratio criterion (8) is invariant.

An alternative invariant test procedure [Nagao (1973a)] is based on the criterion

$$(19)\qquad \tfrac{1}{2}\sum_{g=1}^{q} n_g \operatorname{tr}\left(S_g S^{-1} - I\right)^2 = \tfrac{1}{2}\sum_{g=1}^{q} n_g \operatorname{tr}(S_g - S)S^{-1}(S_g - S)S^{-1},$$

where $S_g = (1/n_g)A_g$ and $S = (1/n)A$. (See Section 7.8.)

10.3. CRITERIA FOR TESTING THAT SEVERAL NORMAL DISTRIBUTIONS ARE IDENTICAL

In Section 8.8 we considered testing the equality of mean vectors when we assumed the covariance matrices were the same; that is, we tested

$$(1)\qquad H_2 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} = \cdots = \boldsymbol{\mu}^{(q)} \qquad \text{given} \quad \Sigma_1 = \Sigma_2 \cdots = \Sigma_q.$$

The test of the assumption in H_2 was considered in Section 10.2. Now let us consider the hypothesis that both means and covariances are the same; this is a combination of H_1 and H_2. We test

$$(2)\qquad H : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} = \cdots = \boldsymbol{\mu}^{(q)}, \ \Sigma_1 = \Sigma_2 = \cdots = \Sigma_q.$$

As in Section 10.2, let $x_\alpha^{(g)}$, $\alpha = 1, \dots, N_g$, be an observation from $N(\boldsymbol{\mu}^{(g)}, \Sigma_g)$, $g = 1, \dots, q$. Then Ω is the unrestricted parameter space of $\{\boldsymbol{\mu}^{(g)}, \Sigma_g\}$, $g = 1, \dots, q$, where Σ_g is positive definite, and ω^* consists of the space restricted by (2).

The likelihood function is given by (3) of Section 10.2. The hypothesis H_1 of Section 10.2 is that the parameter point falls in ω; the hypothesis H_2 of Section 8.8 is that the parameter point falls in ω^* given it falls in $\omega \supset \omega^*$; and the hypothesis H here is that the parameter point falls in ω^* given that it is in Ω.

We use the following lemma:

Lemma 10.3.1. *Let y be an observation vector on a random vector with density $f(z, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector in a space Ω. Let H_a be the hypothesis $\boldsymbol{\theta} \in \Omega_a \subset \Omega$, let H_b be the hypothesis $\boldsymbol{\theta} \in \Omega_b$, $\subset \Omega_a$, given $\boldsymbol{\theta} \in \Omega_a$, and let H_{ab} be the hypothesis $\boldsymbol{\theta} \in \Omega_b$, given $\boldsymbol{\theta} \in \Omega$. If λ_a, the likelihood ratio criterion for testing H_a, λ_b for H_b, and λ_{ab} for H_{ab} are uniquely defined for the observation vector y, then*

$$(3)\qquad \lambda_{ab} = \lambda_a \lambda_b.$$

Proof. The lemma follows from the definitions:

$$\lambda_a = \frac{\max_{\theta \in \Omega_a} f(y, \theta)}{\max_{\theta \in \Omega} f(y, \theta)}, \tag{4}$$

$$\lambda_b = \frac{\max_{\theta \in \Omega_b} f(y, \theta)}{\max_{\theta \in \Omega_a} f(y, \theta)}, \tag{5}$$

$$\lambda_{ab} = \frac{\max_{\theta \in \Omega_b} f(y, \theta)}{\max_{\theta \in \Omega} f(y, \theta)}. \quad \blacksquare \tag{6}$$

Thus the likelihood ratio criterion for the hypothesis H is the product of the likelihood ratio criteria for H_1 and H_2,

$$\lambda = \lambda_1 \lambda_2 = \left(\prod_{g=1}^{q} \frac{|A_g|^{\frac{1}{2}N_g}}{N_g^{\frac{1}{2}pN_g}} \right) \frac{N^{\frac{1}{2}pN}}{|B|^{\frac{1}{2}N}}, \tag{7}$$

where

$$\begin{aligned} B &= \sum_{g=1}^{q} \sum_{\alpha=1}^{N_g} \left(x_\alpha^{(g)} - \bar{x} \right)\left(x_\alpha^{(g)} - \bar{x} \right)' \\ &= A + \sum_{g=1}^{q} N_g (\bar{x}^{(g)} - \bar{x})(\bar{x}^{(g)} - \bar{x})'. \end{aligned} \tag{8}$$

The critical region is defined by

$$\lambda \leq \lambda(\varepsilon), \tag{9}$$

where $\lambda(\varepsilon)$ is chosen so that the probability of (9) under H is ε.

Let

$$V_2 = \frac{|A|^{\frac{1}{2}n}}{|B|^{\frac{1}{2}n}} = \lambda_2^{n/N}; \tag{10}$$

this is equivalent to λ_2 for testing H_2, which is λ of (12) of Section 8.8. We might consider

$$V = V_1 V_2 = \frac{\prod_{g=1}^{q} |A_g|^{\frac{1}{2}n_g}}{|B|^{\frac{1}{2}n}}. \tag{11}$$

However, Perlman (1980) has shown that the likelihood ratio test is unbiased.

10.4. DISTRIBUTIONS OF THE CRITERIA

10.4.1. Characterization of the Distributions

First let us consider V_1 given by (10) of Section 10.2. If

$$V_{1g} = \frac{|A_1 + \cdots + A_{g-1}|^{\frac{1}{2}(n_1 + \cdots + n_{g-1})} |A_g|^{\frac{1}{2}n_g}}{|A_1 + \cdots + A_g|^{\frac{1}{2}(n_1 + \cdots + n_g)}}, \qquad g = 2, \ldots, q. \tag{1}$$

then

$$V_1 = \prod_{g=2}^{q} V_{1g}. \tag{2}$$

Theorem 10.4.1. *$V_{12}, V_{13}, \ldots, V_{1q}$ defined by (1) are independent when $\Sigma_1 = \cdots = \Sigma_q$ and $n_g \geq p$, $g = 1, \ldots, q$.*

The theorem is a consequence of the following lemma:

Lemma 10.4.1. *If A and B are independently distributed according to $W(\Sigma, m)$ and $W(\Sigma, n)$, respectively, $n \geq p$, $m \geq p$, and C is such that $C(A + B)C' = I$, then $A + B$ and CAC' are independently distributed; $A + B$ has the Wishart distribution with $m + n$ degrees of freedom, and CAC' has the multivariate beta distribution with n and m degrees of freedom.*

Proof of Lemma. The density of $D = A + B$ and $E = CAC'$ is found by replacing A and B in their joint density by $C^{-1}EC'^{-1}$ and $D - C^{-1}EC'^{-1} = C^{-1}(I - E)C'^{-1}$, respectively, and multiplying by the Jacobian, which is $\operatorname{mod}|C|^{-(p+1)} = |D|^{\frac{1}{2}(p+1)}$, to obtain

$$\begin{aligned} & K(\Sigma, m) K(\Sigma, n) |C^{-1}EC'^{-1}|^{\frac{1}{2}(m-p-1)} \\ & \qquad \cdot |C^{-1}(I - E)C'^{-1}|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\operatorname{tr} \Sigma^{-1} D} |D|^{\frac{1}{2}(p+1)} \\ & = K(\Sigma, m + n) |D|^{\frac{1}{2}(m+n-p-1)} e^{-\frac{1}{2}\operatorname{tr} \Sigma^{-1} D} \\ & \qquad \cdot \frac{\Gamma_p\left[\frac{1}{2}(m + n)\right]}{\Gamma_p\left(\frac{1}{2}m\right)\Gamma_p\left(\frac{1}{2}n\right)} |E|^{\frac{1}{2}(m-p-1)} |I - E|^{\frac{1}{2}(n-p-1)} \end{aligned} \tag{3}$$

for D, E, and $I - E$ positive definite. ■

Proof of Theorem. If we let $A_1 + \cdots + A_g = D_g$ and $C_g(A_1 + \cdots + A_{g-1})C_g' = E_g$, where $C_g D_g C_g' = I$, $g = 2, \ldots, q$, then

$$(4) \qquad V_{1g} = \frac{|C_g^{-1} E_g C_g'^{-1}|^{\frac{1}{2}(n_1 + \cdots + n_{g-1})} |C_g^{-1}(I - E_g) C_g'^{-1}|^{\frac{1}{2}n_g}}{|C_g^{-1} C_g'^{-1}|^{\frac{1}{2}(n_1 + \cdots + n_g)}}$$

$$= |E_g|^{\frac{1}{2}(n_1 + \cdots + n_{g-1})} |I - E_g|^{\frac{1}{2}n_g}, \qquad g = 2, \ldots, q,$$

and $E_2, \ldots, E_q$ are independent by Lemma 10.4.1. ■

We shall now find a characterization of the distribution of V_{1g}. A statistic V_{1g} is of the form

$$(5) \qquad \frac{|B|^b |C|^c}{|B + C|^{b+c}}.$$

Let B_i and C_i be the upper left-hand square submatrices of B and C, respectively, of order i. Define $b_{(i)}$ and $c_{(i)}$ by

$$(6) \qquad B_i = \begin{pmatrix} B_{i-1} & b_{(i)} \\ b_{(i)}' & b_{ii} \end{pmatrix}, \qquad C_i = \begin{pmatrix} C_{i-1} & c_{(i)} \\ c_{(i)}' & c_{ii} \end{pmatrix}, \qquad i = 2, \ldots, p.$$

Then (5) is ($B_0 = C_0 = I$, $b_{(1)} = c_{(1)} = 0$)

(7)

$$\frac{|B|^b |C|^c}{|B + C|^{b+c}} = \prod_{i=1}^{p} \frac{|B_i|^b |C_i|^c}{|B_{i-1}|^b |C_{i-1}|^c} \cdot \frac{|B_{i-1} + C_{i-1}|^{b+c}}{|B_i + C_i|^{b+c}}$$

$$= \prod_{i=1}^{p} \frac{\left(b_{ii} - b_{(i)}' B_{i-1}^{-1} b_{(i)}\right)^b \left(c_{ii} - c_{(i)}' C_{i-1}^{-1} c_{(i)}\right)^c}{\left[b_{ii} + c_{ii} - \left(b_{(i)} + c_{(i)}\right)' (B_{i-1} + C_{i-1})^{-1} \left(b_{(i)} + c_{(i)}\right)\right]^{b+c}}$$

$$= \prod_{i=1}^{p} \left\{ \frac{b_{ii\cdot i-1}^b c_{ii\cdot i-1}^c}{(b_{ii\cdot i-1} + c_{ii\cdot i-1})^{b+c}} \right.$$

$$\left. \cdot \frac{(b_{ii\cdot i-1} + c_{ii\cdot i-1})^{b+c}}{\left[b_{ii\cdot i-1} + c_{ii\cdot i-1} + b_{(i)}' B_{i-1}^{-1} b_{(i)} + c_{(i)}' C_{i-1}^{-1} c_{(i)} - \left(b_{(i)} + c_{(i)}\right)' (B_{i-1} + C_{i-1})^{-1} \left(b_{(i)} + c_{(i)}\right)\right]^{b+c}} \right\},$$

where $b_{ii\cdot i-1} = b_{ii} - b'_{(i)}B_{i-1}^{-1}b_{(i)}$ and $c_{ii\cdot i-1} = c_{ii} - c'_{(i)}C_{i-1}^{-1}c_{(i)}$. The second term for $i = 1$ is defined as 1.

Now we want to argue that the ratios on the right-hand side of (7) are statistically independent when B and C are independently distributed according to $W(\Sigma, m)$ and $W(\Sigma, n)$, respectively. It follows from Theorem 4.3.3 that for B_{i-1} fixed $b_{(i)}$ and $b_{ii\cdot i-1}$ are independently distributed according to $N(\beta_{(i)}, \sigma_{ii\cdot i-1}B_{i-1}^{-1})$ and $\sigma_{ii\cdot i-1}\chi^2$ with $m-(i-1)$ degrees of freedom, respectively. Lemma 10.4.1 implies that the first term (which is a function of $b_{ii\cdot i-1}/c_{ii\cdot i-1}$) is independent of $b_{ii\cdot i-1} + c_{ii\cdot i-1}$.

We apply the following lemma:

Lemma 10.4.2. *For B_{i-1} and C_{i-1} positive definite*

$$(8)\quad b'_{(i)}B_{i-1}^{-1}b_{(i)} + c'_{(i)}C_{i-1}^{-1}c_{(i)} - \left(b_{(i)} + c_{(i)}\right)'\left(B_{i-1} + C_{i-1}\right)^{-1}\left(b_{(i)} + c_{(i)}\right)$$
$$= \left(B_{i-1}^{-1}b_{(i)} - C_{i-1}^{-1}c_{(i)}\right)'\left(B_{i-1}^{-1} + C_{i-1}^{-1}\right)^{-1}\left(B_{i-1}^{-1}b_{(i)} - C_{i-1}^{-1}c_{(i)}\right).$$

Proof. Use of $(B^{-1} + C^{-1})^{-1} = [C^{-1}(B + C)B^{-1}]^{-1} = B(B + C)^{-1}C$ shows the left-hand side of (8) is (omitting i and $i-1$)

$$(9)$$
$$b'B^{-1}\left(B^{-1} + C^{-1}\right)^{-1}\left(B^{-1} + C^{-1}\right)b + c'\left(B^{-1} + C^{-1}\right)\left(B^{-1} + C^{-1}\right)^{-1}C^{-1}c$$
$$-(b + c)'B^{-1}\left(B^{-1} + C^{-1}\right)^{-1}C^{-1}(b + c)$$
$$= b'B^{-1}\left(B^{-1} + C^{-1}\right)^{-1}B^{-1}b + c'C^{-1}\left(B^{-1} + C^{-1}\right)^{-1}C^{-1}c$$
$$-b'B^{-1}\left(B^{-1} + C^{-1}\right)^{-1}C^{-1}c - c'C^{-1}\left(B^{-1} + C^{-1}\right)B^{-1}b,$$

which is the right-hand side of (8). ■

The denominator of the ith second term in (7) is the numerator plus (8). The conditional distribution of $B_{i-1}^{-1}b_{(i)} - C_{i-1}^{-1}c_{(i)}$ is normal with mean $B_{i-1}^{-1}\beta_{(i)} - C_{i-1}^{-1}\gamma_{(i)}$ and covariance matrix $\sigma_{ii\cdot i-1}(B_{i-1}^{-1} + C_{i-1}^{-1})$. The covariance matrix is $\sigma_{ii\cdot i-1}$ times the inverse of the second matrix on the right-hand side of (8). Thus (8) is distributed as $\sigma_{ii\cdot i-1}\chi^2$ with $i-1$ degrees of freedom, independent of B_{i-1}, C_{i-1}, $b_{ii\cdot i-1}$, and $c_{ii\cdot i-1}$.

Then

$$(10)\qquad \frac{b_{ii\cdot i-1}^{b}c_{ii\cdot i-1}^{2}}{\left(b_{ii\cdot i-1} + c_{ii\cdot i-1}\right)^{b+c}} = \left(\frac{b_{ii\cdot i-1}}{b_{ii\cdot i-1} + c_{ii\cdot i-1}}\right)^{b}\left(\frac{c_{ii\cdot i-1}}{b_{ii\cdot i-1} + c_{ii\cdot i-1}}\right)^{c}$$

is distributed as $X_i^b(1 - X_i)^c$, where X_i has the $\beta[\frac{1}{2}(m - i + 1), \frac{1}{2}(n - i + 1)]$

distribution, $i = 1, \dots, p$. Also

$$(11) \qquad \left[\frac{b_{ii\cdot i-1} + c_{ii\cdot i-1}}{b_{ii\cdot i-1} + c_{ii\cdot i-1} + (8)}\right]^{b+c}, \qquad i = 2, \dots, p,$$

is distributed as Y_i^{b+c}, where Y_i has the $\beta[\frac{1}{2}(m+n) - i + 1, \frac{1}{2}(i-1)]$ distribution. Then (5) is distributed as $\prod_{i=1}^p X_i^b(1-X_i)^c \prod_{i=2}^p Y_i^{b+c}$, and the factors are mutually independent.

Theorem 10.4.2.

$$(12) \qquad V_1 = \prod_{g=2}^{q} \left\{ \prod_{i=1}^{p} X_{ig}^{\frac{1}{2}(n_1 + \cdots + n_{g-1})} (1 - X_{ig})^{\frac{1}{2}n_g} \cdot \prod_{i=2}^{p} Y_{ig}^{\frac{1}{2}(n_1 + \cdots + n_g)} \right\},$$

where the X's and Y's are independent, X_{ig} has the $\beta[\frac{1}{2}(n_1 + \cdots + n_{g-1} - i + 1), \frac{1}{2}(n_g - i + 1)]$ distribution, and Y_{ig} has the $\beta[\frac{1}{2}(n_1 + \cdots + n_g) - i + 1, \frac{1}{2}(i-1)]$ distribution.

Proof. The factors $V_{12}, \dots, V_{1q}$ are independent by Theorem 10.4.1. Each term V_{1g} is decomposed according to (7), and the factors are independent. ∎

The factors of V_1 can be interpreted as test criteria for subhypotheses. The term depending on X_{i2} is the criterion for testing the hypothesis that $\sigma_{ii\cdot i-1}^{(1)} = \sigma_{ii\cdot i-1}^{(2)}$, and the term depending on Y_{i2} is the criterion for testing $\boldsymbol{\sigma}_{(i)}^{(1)} = \boldsymbol{\sigma}_{(i)}^{(2)}$ given $\sigma_{ii\cdot i-1}^{(1)} = \sigma_{ii\cdot i-1}^{(2)}$, and $\boldsymbol{\Sigma}_{i-1,1} = \boldsymbol{\Sigma}_{i-1,2}$. The terms depending on X_{ig} and Y_{ig} similarly furnish criteria for testing $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_g$ given $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_{g-1}$.

Now consider the likelihood ratio criterion λ given by (7) of Section 10.3 for testing the hypothesis $\boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(q)}$ and $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q$. It is equivalent to the criterion

$$(13) \qquad W = \frac{\prod_{g=1}^q |\boldsymbol{A}_g|^{\frac{1}{2}N_g}}{|\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_q|^{\frac{1}{2}(N_1 + \cdots + N_g)}} \cdot \frac{|\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_q|^{\frac{1}{2}N}}{|\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_q + \sum_{g=1}^q N_g(\bar{\boldsymbol{x}}^{(g)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(g)} - \bar{\boldsymbol{x}})'|^{\frac{1}{2}N}}.$$

The two factors of (13) are independent because the first factor is independent of $\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_q$ (by Lemma 10.4.1 and the proof of Theorem 10.4.1) and of $\bar{\boldsymbol{x}}^{(1)}, \dots, \bar{\boldsymbol{x}}^{(q)}$.

Theorem 10.4.3

$$W = \prod_{g=2}^{q} \left\{ \prod_{i=1}^{p} X_{ig}^{\frac{1}{2}(N_1 + \cdots + N_{g-1})} (1 - X_{ig})^{\frac{1}{2}N_g} \prod_{i=2}^{p} Y_{ig}^{\frac{1}{2}(N_1 + \cdots + N_g)} \right\} \prod_{i=1}^{p} Z_i^{\frac{1}{2}N}, \tag{14}$$

where the X's, Y's, and Z's are independent, X_{ig} has the $\beta[\frac{1}{2}(n_1 + \cdots + n_{g-1} - i + 1), \frac{1}{2}(n_g - i + 1)]$ distribution, Y_{ig} has the $\beta[\frac{1}{2}(n_1 + \cdots + n_g) - i + 1, \frac{1}{2}(i - 1)]$ distribution, and Z_i has the $\beta[\frac{1}{2}(n + 1 - i), \frac{1}{2}(q - 1)]$ distribution.

Proof. The characterization of the first factor in (13) corresponds to that of V_1 with the exponents of X_{ig} and $1 - X_{ig}$ modified by replacing n_g by N_g. The second term in $U_{p,q-1,n}^{\frac{1}{2}N}$, and its characterization follows from Theorem 8.4.1. ∎

10.4.2. Moments of the Distributions

We now find the moments of V_1 and of W. Since $0 \le V_1 \le 1$ and $0 \le W \le 1$, the moments determine the distributions uniquely. The hth moment of V_1 we find from the characterization of the distribution in Theorem 10.4.2:

(15)

$$\begin{aligned}
\mathscr{E} V_1^h &= \prod_{g=2}^{q} \left\{ \prod_{i=1}^{p} \mathscr{E} X_{ig}^{\frac{1}{2}(n_1 + \cdots + n_{g-1})h} (1 - X_{ig})^{\frac{1}{2}n_g h} \prod_{i=2}^{p} \mathscr{E} Y_{ig}^{\frac{1}{2}(n_1 + \cdots + n_g)h} \right\} \\
&= \prod_{g=2}^{q} \left\{ \prod_{i=1}^{p} \frac{\Gamma\left[\frac{1}{2}(n_1 + \cdots + n_{g-1})(1+h) - \frac{1}{2}(i-1)\right]}{\Gamma\left[\frac{1}{2}(n_1 + \cdots + n_{g-1} - i + 1)\right]} \right. \\
&\qquad \cdot \frac{\Gamma\left[\frac{1}{2}n_g(1+h) - \frac{1}{2}(i-1)\right] \Gamma\left[\frac{1}{2}(n_1 + \cdots + n_g) - i + 1\right]}{\Gamma\left[\frac{1}{2}(n_g - i + 1)\right] \Gamma\left[\frac{1}{2}(n_1 + \cdots + n_g)(1+h) - i + 1\right]} \\
&\qquad \left. \cdot \prod_{i=2}^{p} \frac{\Gamma\left[\frac{1}{2}(n_1 + \cdots + n_g)(1+h) - i + 1\right] \Gamma\left[\frac{1}{2}(n_1 + \cdots + n_g - i + 1)\right]}{\Gamma\left[\frac{1}{2}(n_1 + \cdots + n_g) - i + 1\right] \Gamma\left[\frac{1}{2}(n_1 + \cdots + n_g)(1+h) - \frac{1}{2}(i-1)\right]} \right\} \\
&= \prod_{i=1}^{p} \left\{ \frac{\Gamma\left[\frac{1}{2}(n + 1 - i)\right]}{\Gamma\left[\frac{1}{2}(n + hn + 1 - i)\right]} \prod_{g=1}^{q} \frac{\Gamma\left[\frac{1}{2}(n_g + hn_g + 1 - i)\right]}{\Gamma\left[\frac{1}{2}(n_g + 1 - i)\right]} \right\} \\
&= \frac{\Gamma_p(\frac{1}{2}n)}{\Gamma_p\left[\frac{1}{2}(n + hn)\right]} \prod_{g=1}^{q} \frac{\Gamma_p\left[\frac{1}{2}(n_g + hn_g)\right]}{\Gamma_p(\frac{1}{2}n_g)}.
\end{aligned}$$

The hth moment of W can be found from its representation in Theorem 10.4.3. We have

(16)

$$\mathscr{E}W^h = \prod_{g=2}^{q}\prod_{i=1}^{p}\mathscr{E}X_{ig}^{\frac{1}{2}(N_1+\cdots+N_{g-1})h}(1-X_{ig})^{\frac{1}{2}N_g h}\prod_{i=2}^{p}\mathscr{E}Y_{ig}^{\frac{1}{2}(N_1+\cdots+N_g)h}\,\mathscr{E}U_{p,q-1,n}^{\frac{1}{2}Nh}$$

$$= \prod_{g=2}^{q}\left\{\prod_{i=1}^{p}\frac{\Gamma\left[\frac{1}{2}(n_1+\cdots+n_{g-1}+1-i)+\frac{1}{2}h(N_1+\cdots+N_{g-1})\right]}{\Gamma\left[\frac{1}{2}(n_1+\cdots+n_{g-1}+1-i)\right]\Gamma\left[\frac{1}{2}(n_g+1-i)\right]}\right.$$

$$\cdot\frac{\Gamma\left[\frac{1}{2}(n_g+1-i+N_g h)\right]\Gamma\left[\frac{1}{2}(n_1+\cdots+n_g)-i+1\right]}{\Gamma\left[\frac{1}{2}(n_1+\cdots+n_g)+\frac{1}{2}h(N_1+\cdots+N_g)+1-i\right]}$$

$$\cdot\prod_{i=2}^{p}\frac{\Gamma\left[\frac{1}{2}(n_1+\cdots+n_g)+\frac{1}{2}h(N_1+\cdots+N_g)+1-i\right]}{\Gamma\left[\frac{1}{2}(n_1+\cdots+n_g)+1-i\right]}$$

$$\left.\cdot\frac{\Gamma\left[\frac{1}{2}(n_1+\cdots+n_g+1-i)\right]}{\Gamma\left[\frac{1}{2}(n_1+\cdots+n_g+1-i)+\frac{1}{2}h(N_1+\cdots+N_g)\right]}\right\}$$

$$\cdot\prod_{i=1}^{p}\frac{\Gamma\left[\frac{1}{2}(n+1-i+hN)\right]\Gamma\left[\frac{1}{2}(N-i)\right]}{\Gamma\left[\frac{1}{2}(n+1-i)\right]\Gamma\left[\frac{1}{2}(N+hN-i)\right]}$$

$$= \prod_{i=1}^{p}\left\{\prod_{g=1}^{q}\frac{\Gamma\left[\frac{1}{2}(N_g+hN_g-i)\right]}{\Gamma\left[\frac{1}{2}(N_g-i)\right]}\right\}\frac{\Gamma\left[\frac{1}{2}(N-i)\right]}{\Gamma\left[\frac{1}{2}(N+hN-i)\right]}$$

$$= \frac{\Gamma_p(\frac{1}{2}n)}{\Gamma_p(\frac{1}{2}n+\frac{1}{2}hN)}\prod_{g=1}^{q}\frac{\Gamma_p\left[\frac{1}{2}(n_g+hN_g)\right]}{\Gamma_g(\frac{1}{2}n_g)}.$$

We summarize in the following theorem:

Theorem 10.4.4. *Let V_1 be the criterion defined by* (10) *of Section* 10.2 *for testing the hypothesis that $H_1 : \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_q$, where A_g is n_g times the sample covariance matrix and $n_g + 1$ is the size of the sample from the gth population; let W be the criterion defined by* (13) *for testing the hypothesis $H : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q$ and H_1, where $B = A + \Sigma_g N_g(\bar{x}^{(g)} - \bar{x})(\bar{x}^{(g)} - \bar{x})'$. The* hth *moment of V_1 when H_1 is true is given by* (15). *The* hth *moment of W, the criterion for testing H, is given by* (16).

This theorem was first proved by Wilks (1932). See Problem 10.5 for an alternative approach.

If p is even, say $p = 2r$, we can use the duplication formula for the gamma function $[\Gamma(\alpha + \frac{1}{2})\Gamma(\alpha + 1) = \sqrt{\pi}\,\Gamma(2\alpha + 1)2^{-2\alpha}]$. Then

$$(17)\qquad \mathscr{E}V_1^h = \prod_{j=1}^{r} \left\{ \left[\prod_{g=1}^{q} \frac{\Gamma(n_g + hn_g + 1 - 2j)}{\Gamma(n_g + 1 - 2j)} \right] \frac{\Gamma(n + 1 - 2j)}{\Gamma(n + hn + 1 - 2j)} \right\}$$

and

$$(18)\qquad \mathscr{E}W^h = \prod_{j=1}^{r} \left\{ \left[\prod_{g=1}^{q} \frac{\Gamma(n_g + hN_g + 1 - 2j)}{\Gamma(n_g + 1 - 2j)} \right] \frac{\Gamma(N - 2j)}{\Gamma(N + hN - 2j)} \right\}.$$

In principle the distributions of the factors can be integrated to obtain the distributions of V_1 and W. In Section 10.6 we consider V_1 when $p = 2$, $q = 2$ (the case of $p = 1, q = 2$ being a function of an F-statistic). In other cases, the integrals become unmanageable. To find probabilities we use the asymptotic expansion given in the next section. Box (1949) has given some other approximate distributions.

10.4.3. Step-down Tests

The characterizations of the distributions of the criteria in terms of independent factors suggests testing the hypotheses H_1 and H by testing component hypotheses sequentially. First, we consider testing $H_1 : \Sigma_1 = \Sigma_2$ for $q = 2$. Let

$$(19)\qquad X_{(i)}^{(g)} = \begin{pmatrix} X_{(i-1)}^{(g)} \\ X_i^{(g)} \end{pmatrix}, \qquad \boldsymbol{\mu}_{(i)}^{(g)} = \begin{pmatrix} \boldsymbol{\mu}_{(i-1)}^{(g)} \\ \mu_i^{(g)} \end{pmatrix}, \qquad \Sigma_i^{(g)} = \begin{bmatrix} \Sigma_{(i-1)}^{(g)} & \boldsymbol{\sigma}_{(i)}^{(g)} \\ \boldsymbol{\sigma}_{(i)}^{(g)\prime} & \sigma_{ii}^{(g)} \end{bmatrix},$$

$$i = 2, \dots, p, \quad g = 1, 2.$$

The conditional distribution of $X_i^{(g)}$ given $X_{(i-1)}^{(g)} = \boldsymbol{x}_{(i-1)}^{(g)}$ is

$$(20)\qquad N\left[\mu_i^{(g)} + \boldsymbol{\sigma}_{(i)}^{(g)\prime} \Sigma_{i-1}^{-1} \left(\boldsymbol{x}_{(i-1)}^{(g)} - \boldsymbol{\mu}_{(i-1)}^{(g)} \right), \sigma_{ii\cdot i-1}^{(g)} \right],$$

where $\sigma_{ii\cdot i-1}^{(g)} = \sigma_{ii}^{(g)} - \boldsymbol{\sigma}_{(i)}^{(g)\prime} \Sigma_{i-1}^{(i)}\boldsymbol{\sigma}_{(i)}^{(g)}$. It is assumed that the components of X have been numbered in descending order of importance. At the ith step the component hypothesis $\sigma_{ii\cdot i-1}^{(1)} = \sigma_{ii\cdot i-1}^{(2)}$ is tested at significance level ε_i by means of an F-test based on $s_{ii\cdot i-1}^{(1)}/s_{ii\cdot i-1}^{(2)}$; S_1 and S_2 are partitioned like $\Sigma^{(1)}$ and $\Sigma^{(2)}$. If that hypothesis is accepted, then the hypothesis $\boldsymbol{\sigma}_{(i)}^{(1)} = \boldsymbol{\sigma}_{(i)}^{(2)}$ (or $\Sigma_{i-1}^{(1)-1}\boldsymbol{\sigma}_{(i)}^{(1)} = \Sigma_{i-1}^{(2)-1}\boldsymbol{\sigma}_{(i)}^{(2)}$) is tested at significance level δ_i on the assumption that $\Sigma_{i-1}^{(1)} = \Sigma_{i-1}^{(2)}$ (a hypothesis previously accepted). The criterion is

$$(21)\qquad \frac{\left(S_{i-1}^{(1)-1}s_{(i)}^{(1)} - S_{i-1}^{(2)-1}s_{(i)}^{(2)}\right)'\left(S_{i-1}^{(1)-1} + S_{i-1}^{(2)-1}\right)^{-1}\left(S_{i-1}^{(1)-1}s_{(i)}^{(1)} - S_{i-1}^{(2)-1}s_{(i)}^{(2)}\right)}{(i-1)s_{ii\cdot i-1}},$$

where $(n_1+n_2-2i+2)s_{ii\cdot i-1} = (n_1-i+1)s_{ii\cdot i-1}^{(1)} + (n_2-i+1)s_{ii\cdot i-1}^{(2)}$. Under the null hypothesis (21) has the F-distribution with $i-1$ and n_1+n_2-2i+2 degrees of freedom. If this hypothesis is accepted, the $(i+1)$st step is taken. The overall hypothesis $\Sigma_1 = \Sigma_2$ is accepted if the $2p-1$ component hypotheses are accepted. (At the first step, $\sigma_{(1)}^{(g)}$ is vacuous.) The overall significance level is

$$(22) \qquad 1-\prod_{i=1}^{p}(1-\varepsilon_i)\prod_{i=2}^{p}(1-\delta_i).$$

If any component null hypothesis is rejected, the overall hypothesis is rejected.

If $q>2$, the null hypotheses $H_1:\Sigma_1=\cdots=\Sigma_q$ is broken down into a sequence of hypotheses $[1/(g-1)](\Sigma_1+\cdots+\Sigma_{g-1})=\Sigma_g$ and tested sequentially. Each such matrix hypothesis is tested as $\Sigma_1=\Sigma_2$ with S_2 replaced by S_g and S_1 replaced by $[1/(n_1+\cdots+n_{g-1})](A_1+\cdots+A_{g-1})$.

In the case of the hypothesis H, consider first $q=2$, $\Sigma_1=\Sigma_2$, and $\boldsymbol{\mu}^{(1)}=\boldsymbol{\mu}^{(2)}$. One can test $\Sigma_1=\Sigma_2$. The steps for testing $\boldsymbol{\mu}^{(1)}=\boldsymbol{\mu}^{(2)}$ consist of t-tests for $\mu_i^{(1)}=\mu_i^{(2)}$ based on the conditional distribution of $X_i^{(1)}$ and $X_i^{(2)}$ given $\boldsymbol{x}_{(i-1)}^{(1)}$ and $\boldsymbol{x}_{(i-1)}^{(2)}$. Alternatively one can test in sequence the equality of the conditional distributions of $X_i^{(1)}$ and $X_i^{(2)}$ given $\boldsymbol{x}_{(i-1)}^{(1)}$ and $\boldsymbol{x}_{(i-1)}^{(2)}$.

For $q>2$, the hypothesis $\Sigma_1=\cdots=\Sigma_q$ can be tested, and then $\boldsymbol{\mu}_1=\cdots=\boldsymbol{\mu}_q$. Alternatively, one can test $[1/(g-1)](\Sigma_1+\cdots+\Sigma_{g-1})=\Sigma_g$ and $[1/(g-1)](\boldsymbol{\mu}^{(1)}+\cdots+\boldsymbol{\mu}^{(g-1)})=\boldsymbol{\mu}^{(g)}$.

10.5. ASYMPTOTIC EXPANSIONS OF THE DISTRIBUTIONS OF THE CRITERIA

Again we make use of Theorem 8.5.1 to obtain asymptotic expansions of the distributions of V_1 and of λ. We assume that $n_g=k_g n$, where $\sum_{g=1}^q k_g=1$. The asymptotic expansion is in terms of n increasing with $k_1,\dots,k_q$ fixed. (We could assume only $\lim n_g/n=k_g>0$.)

The hth moment of

$$(1) \qquad \lambda_1^* = V_1\cdot\frac{n^{\frac12 pn}}{\prod_{g=1}^q n_g^{\frac12 pn_g}} = V_1\cdot\prod_{g=1}^q\left(\frac{n}{n_g}\right)^{\frac12 pn_g} = \left[\prod_{g=1}^q\left(\frac{1}{k_g}\right)^{k_g}\right]^{\frac12 pn} V_1$$

is

$$(2) \qquad \mathscr{E}\lambda_1^{*h} = K\left(\frac{\prod_{j=1}^p\left(\frac12 n\right)^{\frac12 n}}{\prod_{g=1}^q\prod_{i=1}^p\left(\frac12 n_g\right)^{\frac12 n_g}}\right)^h \frac{\prod_{g=1}^q\prod_{i=1}^p\Gamma\left[\frac12 n_g(1+h)+\frac12(1-i)\right]}{\prod_{j=1}^p\Gamma\left[\frac12 n(1+h)+\frac12(1-j)\right]}$$

This is of the form of (1) of Section 8.6 with

$$
\begin{aligned}
&b = p, \qquad y_j = \tfrac{1}{2}n, \qquad \eta_j = \tfrac{1}{2}(1-j), && j = 1, \dots, p, \\
(3)\qquad &a = pq, \qquad x_k = \tfrac{1}{2}n_g, \qquad k = (g-1)p+1, \dots, gp, && g = 1, \dots, q, \\
&\xi_k = \tfrac{1}{2}(1-i), \qquad k = i, p+i, \dots, (q-1)p+i, && i = 1, \dots, p
\end{aligned}
$$

Then

$$
\begin{aligned}
(4)\qquad f &= -2\left[\sum \xi_k - \sum \eta_j - \tfrac{1}{2}(a-b)\right] \\
&= -\left[q\sum_{i=1}^{p}(1-i) - \sum_{j=1}^{p}(1-j) - (qp-p)\right] \\
&= -\left[-q\tfrac{1}{2}p(p-1) + \tfrac{1}{2}p(p-1) - (q-1)p\right] \\
&= \tfrac{1}{2}(q-1)p(p+1),
\end{aligned}
$$

$\varepsilon_j = \frac{1}{2}(1-\rho)n$, $j = 1, \dots, p$, and $\beta_k = \frac{1}{2}(1-\rho)n_g = \frac{1}{2}(1-\rho)k_g n$, $k = (g-1)p + 1, \dots, gp$.

In order to make the second term in the expansion vanish, we take ρ as

$$
(5)\qquad \rho = 1 - \left(\sum_{g=1}^{q}\frac{1}{n_g} - \frac{1}{n}\right)\frac{2p^2+3p-1}{6(p+1)(q-1)}.
$$

Then

$$
(6)\qquad \omega_2 = \frac{p(p+1)\left[(p-1)(p+2)\left(\sum_{g=1}^{q}\frac{1}{n_g^2} - \frac{1}{n^2}\right) - 6(q-1)(1-\rho)^2\right]}{48\rho^2}.
$$

Thus

$$
\begin{aligned}
(7)\qquad &\Pr\{-2\rho\log\lambda_1^* \le z\} \\
&\quad = \Pr\{\chi_f^2 \le z\} + \omega_2\left[\Pr\{\chi_{f+4}^2 \le z\} - \Pr\{\chi_f^2 \le z\}\right] + O(n^{-3}).
\end{aligned}
$$

Let $\lambda = WN^{\frac{1}{2}pN}\prod_{g=1}^{q}N_g^{-\frac{1}{2}pN_g}$. The hth moment is

$$
(8)\qquad \mathscr{E}\lambda^h = K\left[\frac{\prod_{j=1}^{p}\left(\frac{1}{2}N\right)^{\frac{1}{2}N}}{\prod_{g=1}^{q}\prod_{i=1}^{p}\left(\frac{1}{2}N_g\right)^{\frac{1}{2}N_g}}\right]^h \frac{\prod_{g=1}^{q}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}N_g(1+h) - \frac{1}{2}i\right]}{\prod_{j=1}^{p}\Gamma\left[\frac{1}{2}N(1+h) - \frac{1}{2}j\right]}.
$$

This is the form (1) of Section 8.5 with

$$b = p, \qquad y_j = \tfrac{1}{2}N = \tfrac{1}{2}\sum_{g=1}^{q} N_g, \qquad \eta_j = -\tfrac{1}{2}j, \qquad j = 1,\ldots,p,$$

$$(9) \qquad a = pq, \qquad x_k = \tfrac{1}{2}N_g, \qquad k = (g-1)p+1,\ldots,gp, \qquad g = 1,\ldots,q,$$

$$\xi_k = -\tfrac{1}{2}i, \qquad k = i, p+i,\ldots,(q-1)p+i, \quad i = 1,\ldots,p.$$

The basic number of degrees of freedom is $f = \frac{1}{2}p(p+3)(q-1)$. We use (11) of Section 8.5 with $\beta_k = (1-\rho)x_k$ and $\varepsilon_j = (1-\rho)y_j$. To make $\omega_1 = 0$, we take

$$(10) \qquad \rho = 1 - \left(\sum_{g=1}^{q}\frac{1}{N_g} - \frac{1}{N}\right)\frac{2p^2+9p+11}{6(q-1)(p+3)}.$$

Then

$$(11) \quad \omega_2 = \frac{p(p+3)}{48\rho^2}\left[\sum_{g=1}^{q}\left(\frac{1}{N_g^2} - \frac{1}{N^2}\right)(p+1)(p+2) - 6(1-\rho)^2(q-1)\right]$$

The asymptotic expansion of the distribution of $-2\rho\log\lambda$ is

$$(12) \quad \Pr\{-2\rho\log\lambda \le z\}$$
$$= \Pr\{\chi_f^2 \le z\} + \omega_2\left[\Pr\{\chi_{f+4}^2 \le z\} - \Pr\{\chi_f^2 \le z\}\right] + O(n^{-3}).$$

Box (1949) considered the case of λ_1^* in considerable detail. In addition to this expansion he considered the use of (13) of Section 8.6. He also gave an F-approximation.

As an example, we use one given by E. S. Pearson and Wilks (1933). The measurements are made on tensile strength (X_1) and hardness (X_2) of aluminum die castings. There are 12 observations in each of five samples. The observed sums of squares and cross-products in the five samples are

$$A_1 = \begin{pmatrix} 78.948 & 214.18 \\ 214.18 & 1247.18 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 223.695 & 657.62 \\ 657.62 & 2519.31 \end{pmatrix},$$

$$(13) \qquad A_3 = \begin{pmatrix} 57.448 & 190.63 \\ 190.63 & 1241.78 \end{pmatrix},$$

$$A_4 = \begin{pmatrix} 187.618 & 375.91 \\ 375.91 & 1473.44 \end{pmatrix},$$

$$A_5 = \begin{pmatrix} 88.456 & 259.18 \\ 259.18 & 1171.73 \end{pmatrix},$$

and the sum of these is

$$(14) \qquad \sum A_i = \begin{pmatrix} 636.165 & 1697.52 \\ 1697.52 & 7653.44 \end{pmatrix}.$$

The $-\log \lambda_1^*$ is 5.399. To use the asymptotic expansion we find $\rho = 152/165 = 0.9212$ and $\omega_2 = 0.0022$. Since ω_2 is small, we can consider $-2\rho \log \lambda_1^*$ as χ^2 with 12 degrees of freedom. Our observed criterion, therefore, is clearly not significant.

Table B.5 [due to Korin (1969)] gives 5% significance points for $-2\log \lambda_1^*$ for $N_1 = \cdots = N_q$ for various q, small values of N_g, and $p = 2(1)6$.

The limiting distribution of the criterion (19) of Section 10.1 is also χ_f^2. An asymptotic expansion of the distribution was given by Nagao (1973b) to terms of order $1/n$ involving χ^2-distributions with f, $f+2$, $f+4$, and $f+6$ degrees of freedom.

10.6. THE CASE OF TWO POPULATIONS

10.6.1. Invariant Tests

When $q = 2$, the null hypothesis H_1 is $\Sigma_1 = \Sigma_2$. It is invariant with respect to transformations

$$(1) \qquad x^{*(1)} = Cx^{(1)} + \nu^{(1)}, \qquad x^{*(2)} = Cx^{(2)} + \nu^{(2)},$$

where C is nonsingular. The maximal invariant of the parameters under the transformation of locations ($C = I$) is the pair of covariance matrices Σ_1, Σ_2, and the maximal invariant of the sufficient statistics $\bar{x}^{(1)}, S_1, \bar{x}^{(2)}, S_2$ is the pair of matrices S_1, S_2 (or equivalently A_1, A_2). The transformation (1) induces the transformations $\Sigma_1^* = C\Sigma_1 C'$, $\Sigma_2^* = C\Sigma_2 C'$, $S_1^* = CS_1C'$, and $S_2^* = CS_2C'$. The roots $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ of

$$(2) \qquad |\Sigma_1 - \lambda\Sigma_2| = 0$$

are invariant under these transformations since

$$(3) \qquad |\Sigma_1^* - \lambda\Sigma_2^*| = |C\Sigma_1C' - \lambda C\Sigma_2C'| = |CC'| \cdot |\Sigma_1 - \lambda\Sigma_2|.$$

Moreover, the roots are the only invariants because there exists a nonsingular matrix C such that

$$(4) \qquad C\Sigma_1C' = \Lambda, \qquad C\Sigma_2C' = I,$$

where Λ is the diagonal matrix with λ_i as the ith diagonal element, $i = 1, \dots, p$. (See Theorem A.2.2 of the Appendix.) Similarly, the maximal invariants of S_1 and S_2 are the roots $l_1 \geq l_2 \geq \cdots \geq l_p$ of

$$|S_1 - lS_2| = 0. \tag{5}$$

Theorem 10.6.1. *The maximal invariant of the parameters of* $N(\mu^{(1)}, \Sigma_1)$ *and* $N(\mu^{(2)}, \Sigma_2)$ *under the transformation* (1) *is the set of roots* $\lambda_1 \geq \cdots \geq \lambda_p$ *of* (2). *The maximal invariant of the sufficient statistics* $\bar{x}^{(1)}, S_1, \bar{x}^{(2)}, S_2$ *is the set of roots* $l_1 \geq \cdots \geq l_p$ *of* (5).

Any invariant test criterion can be expressed in terms of the roots $l_1, \ldots, l_p$. The criterion V_1 is $n_1^{\frac{1}{2}pn_1} n_2^{\frac{1}{2}pn_2}$ times

$$\frac{|S_1|^{\frac{1}{2}n_1}|S_2|^{\frac{1}{2}n_2}}{|n_1 S_1 + n_2 S_2|^{\frac{1}{2}n}} = \frac{|L|^{\frac{1}{2}n_1}|I|^{\frac{1}{2}n_2}}{|n_1 L + n_2 I|^{\frac{1}{2}n}} = \prod_{i=1}^{p} \frac{l_i^{\frac{1}{2}n_1}}{(n_1 l_i + n_2)^{\frac{1}{2}n}}, \tag{6}$$

where L is the diagonal matrix with l_i as the ith diagonal element. The null hypothesis is rejected if the smaller roots are too small or if the larger roots are too large, or both.

The null hypothesis is that $\lambda_1 = \cdots = \lambda_p = 1$. Any useful invariant test of the null hypothesis has a rejection region in the space of $l_1, \ldots, l_p$ that includes the points that in some sense are far from $l_1 = \cdots = l_p = 1$. The power of an invariant test depends on the parameters through the roots $\lambda_1, \ldots, \lambda_p$.

The criterion (19) of Section 10.2 is (with $nS = n_1 S_1 + n_2 S_2$)

$$\begin{aligned}
(7)\qquad & \tfrac{1}{2}n_1 \operatorname{tr}\left[(S_1 - S)S^{-1}\right]^2 + \tfrac{1}{2}n_2 \operatorname{tr}\left[(S_2 - S)S^{-1}\right]^2 \\
&= \tfrac{1}{2}n_1 \operatorname{tr}\left[C(S_1 - S)C'(CSC')^{-1}\right]^2 \\
&\quad + \tfrac{1}{2}n_2 \operatorname{tr}\left[C(S_2 - S)C'(CSC')^{-1}\right]^2 \\
&= \tfrac{1}{2}n_1 \operatorname{tr}\left[\left\{L - \left(\frac{n_1}{n}L + \frac{n_2}{n}I\right)\right\}\left(\frac{n_1}{n}L + \frac{n_2}{n}I\right)^{-1}\right]^2 \\
&\quad + \tfrac{1}{2}n_2 \operatorname{tr}\left[\left\{I - \left(\frac{n_1}{n}L + \frac{n_2}{n}I\right)\right\}\left(\frac{n_1}{n}L + \frac{n_2}{n}I\right)^{-1}\right]^2 \\
&= \tfrac{1}{2}n_1 n_2 n \sum_{i=1}^{p} \frac{(l_i - 1)^2}{(n_1 l_i + n_2)^2}.
\end{aligned}$$

This criterion is a measure of how close $l_1, \ldots, l_p$ are to 1; the hypothesis is rejected if the measure is too large. Under the null hypothesis, (7) has the χ^2-distribution with $f = \frac{1}{2}p(p+1)$ degrees of freedom as $n_1 \to \infty$, $n_2 \to \infty$, and n_1/n_2 approaches a positive constant. Nagao (1973b) gives an asymptotic expansion of this distribution to terms of order $1/n$.

Roy (1953) suggested a test based on the largest and smallest roots, l_1 and l_p. The procedure is to reject the null hypothesis if $l_1 > k_1$ or if $l_p < k_p$, where k_1 and k_p are chosen so that the probability of rejection when $\Lambda = I$ is the desired significance level. Roy (1957) proposed determining k_1 and k_p so that the test is locally unbiased, that is, that the power functions have a relative minimum at $\Lambda = I$. Since it is hard to determine k_1 and k_p on this basis, other proposals have been made. The limit k_1 can be determined so that $\Pr\{l_1 > k_1 | H_1\}$ is one-half the significance level, or $\Pr\{l_p < k_p | H_1\}$ is one-half of the significance level, or $k_1 + k_p = 2$, or $k_1 k_p = 1$. In principle k_1 and k_p can be determined from the distribution of the roots, given in Section 13.2. Schuurmann, Waikar, and Krishnaiah (1975) and Chu and Pillai (1979) give some exact values of k_1 and k_p for small values of p. Chu and Pillai (1979) also make some power comparisons of several test procedures.

In the case of $p = 1$ the only invariant of the sufficient statistics is S_1/S_2, which is the usual F-statistic with n_1 and n_2 degrees of freedom. The criterion V_1 is $(A_1/A_2)^{\frac{1}{2}n_1}[1 + A_1/A_2)]^{-\frac{1}{2}n}$; the critical region V_1 less than a constant is equivalent to a two-tailed critical region for the F-statistic. The quantity $n(B-A)/A$ has an independent F-distribution with 1 and n degrees of freedom. (See Section 10.3.)

In the case of $p = 2$, the hth moment of V_1 is, from (15) of Section 10.4,

$$\begin{aligned} (8) \qquad \mathscr{E} V_1^h &= \frac{\Gamma(n_1 + hn_1 - 1)\Gamma(n_2 + hn_2 - 1)\Gamma(n-1)}{\Gamma(n_1 - 1)\Gamma(n_2 - 1)\Gamma(n + hn - 1)} \\ &= \mathscr{E}\left[X_1^{n_1}(1 - X_1)^{n_2} X_2^{n_1+n_2}\right]^h, \end{aligned}$$

where X_1 and X_2 are independently distributed according to $\beta(x|n_1 - 1, n_2 - 1)$ and $\beta(x|n_1 + n_2 - 2, 1)$, respectively. Then $\Pr\{V_1 \le v\}$ can be found by integration. (See Problems 10.8 and 10.9.)

Anderson (1965a) has shown that a confidence interval for $a'\Sigma_1 a/a'\Sigma_2 a$ for all a with confidence coefficient ε is given by $(l_p/U, l_1/L)$, where $\Pr\{(n_2 - p + 1)L \le n_2 F_{n_1, n_2 - p + 1}\}\Pr\{(n_1 - p + 1)F_{n_1 - p + 1, n_2} \le n_1 U\} = 1 - \varepsilon$.

10.6.2. Components of Variance

In Section 8.8 we considered what is equivalent to the one-way analysis of variance with fixed effects. We can write the model in the balanced case $(N_1 = N_2 = \cdots = N_q)$ as

$$\begin{aligned} (9) \qquad X_\alpha^{(g)} &= \boldsymbol{\mu}^{(g)} + U_\alpha^{(g)} \\ &= \boldsymbol{\mu} + \boldsymbol{\nu}_g + U_\alpha^{(g)}, \qquad \alpha = 1, \dots, M, \quad g = 1, \dots, q. \end{aligned}$$

where $\mathscr{E}U^{(g)} = \mathbf{0}$ and $\mathscr{E}U^{(g)}U^{(g)\prime} = \mathbf{\Sigma}$, $\boldsymbol{\nu}_g = \boldsymbol{\mu}^{(g)} - \boldsymbol{\mu}$, and $\boldsymbol{\mu} = (1/q)\Sigma_{g=1}^q \boldsymbol{\mu}^{(g)}$ $(\Sigma_{g=1}^q \boldsymbol{\nu}_g = \mathbf{0})$. The null hypothesis of no effect is $\boldsymbol{\nu}_1 = \cdots = \boldsymbol{\nu}_q = \mathbf{0}$. Let $\bar{\boldsymbol{x}}^{(g)} = (1/M)\Sigma_{\alpha=1}^M \boldsymbol{x}_\alpha^{(g)}$ and $\bar{\boldsymbol{x}} = (1/q)\Sigma_{g=1}^q \bar{\boldsymbol{x}}^{(g)}$. The analysis of variance table is

Source	Sum of Squares	Degrees of Freedom
Effect	$H = M \sum_{g=1}^{q} (\bar{\boldsymbol{x}}^{(g)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(g)} - \bar{\boldsymbol{x}})'$	$q-1$
Error	$G = \sum_{g=1}^{q} \sum_{\alpha=1}^{M} (\boldsymbol{x}_\alpha^{(g)} - \bar{\boldsymbol{x}}^{(g)})(\boldsymbol{x}_\alpha^{(g)} - \bar{\boldsymbol{x}}^{(g)})'$	$q(M-1)$
Total	$\sum_{g=1}^{q} \sum_{\alpha=1}^{M} (\boldsymbol{x}_\alpha^{(g)} - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha^{(g)} - \bar{\boldsymbol{x}})'$	$qM-1$

Invariant tests of the null hypothesis of no effect are based on the roots of $|H - mG| = 0$ or of $|S_h - lS_e| = 0$, where $S_h = [1/(q-1)]H$ and $S_e = [1/q(M-1)]G$. The null hypothesis is rejected if one or more of the roots is too large. The error matrix G has the distribution $W(\mathbf{\Sigma}, q(M-1))$. The effects matrix H has the distribution $W(\mathbf{\Sigma}, q-1)$ when the null hypothesis is true and has the noncentral Wishart distribution when the null hypothesis is not true; its expected value is

$$(10) \qquad \mathscr{E}H = (q-1)\mathbf{\Sigma} + M\sum_{g=1}^{q} (\boldsymbol{\mu}^{(g)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(g)} - \boldsymbol{\mu})'$$

$$= (q-1)\mathbf{\Sigma} + M\sum_{g=1}^{q} \boldsymbol{\nu}_g \boldsymbol{\nu}_g'.$$

The MANOVA model with random effects is

$$(11) \qquad X_\alpha^{(g)} = \boldsymbol{\mu} + V_g + U_\alpha^{(g)}, \qquad \alpha = 1, \dots, M, \quad g = 1, \dots, q,$$

where V_g has the distribution $N(\mathbf{0}, \mathbf{\Theta})$. Then $X_\alpha^{(g)}$ has the distribution $N(\boldsymbol{\mu}, \mathbf{\Sigma} + \mathbf{\Theta})$. The null hypothesis of no effect is

$$(12) \qquad \mathbf{\Theta} = \mathbf{0}.$$

In this model G again has the distribution $W(\mathbf{\Sigma}, q(M-1))$. Since $\bar{X}^{(g)} = \boldsymbol{\mu} + V_g + \bar{U}^{(g)}$ has the distribution $N(\boldsymbol{\mu}, (1/M)\mathbf{\Sigma} + \mathbf{\Theta})$, H has the distribution $W(\mathbf{\Sigma} + M\mathbf{\Theta}, q-1)$. The null hypothesis (12) is equivalent to the equality of the covariance matrices in these two Wishart distributions; that is, $\Sigma_* = \Sigma + M\Theta$. The matrices G and H correspond to A_1 and A_2 in Section 10.6.1. However, here the alternative to the null hypothesis is that $(\Sigma + M\Theta) - \Sigma$ is positive semidefinite, rather than $\Sigma_1 \neq \Sigma_2$. The null hypothesis is to be rejected if H is too large relative to G. Any of the criteria presented in Section 10.2 can be used to test the null hypothesis here, and its distribution under the null hypothesis is the same as given there.

The likelihood ratio criterion for testing $\Theta = \mathbf{0}$ must take into account the fact that Θ is positive semidefinite; that is, the maximum likelihood estimators of Σ and $\Sigma + M\Theta$ under Ω must be such that the estimator of Θ is positive semidefinite. Let $l_1 > l_2 > \cdots > l_p$ be the roots of

$$\left| H - l \frac{1}{M-1} G \right| = 0. \tag{13}$$

(Note $\{1/[q(M-1)]\}G$ and $(1/q)H$ maximize the likelihood without regard to Θ being positive definite.) Let $l_i^* = l_i$ if $l_i > 1$, and let $l_i^* = 1$ if $l_i \leq 1$. Then the likelihood ratio criterion for testing the hypothesis $\Theta = \mathbf{0}$ against the alternative Θ positive semidefinite and $\Theta \neq \mathbf{0}$ is

$$M^{\frac{1}{2}qMp} \prod_{i=1}^{p} \frac{l_i^{*\frac{1}{2}q}}{(l_i^* + M - 1)^{\frac{1}{2}qM}} = M^{\frac{1}{2}qMk} \prod_{i=1}^{k} \frac{l_i^{\frac{1}{2}q}}{(l_i + M - 1)^{\frac{1}{2}qM}}, \tag{14}$$

where k is the number of roots of (13) greater than 1. [See Anderson (1946b), (1984a), (1989a), Morris and Olkin (1964), and Klotz and Putter (1969).]

10.7. TESTING THE HYPOTHESIS THAT A COVARIANCE MATRIX IS PROPORTIONAL TO A GIVEN MATRIX; THE SPHERICITY TEST

10.7.1. The Hypothesis

In many statistical analyses that are considered univariate, the assumption is made that a set of random variables are independent and have a common variance. In this section we consider a test of these assumptions based on repeated sets of observations.

More precisely, we use a sample of p-component vectors $x_1, \ldots, x_N$ from $N(\mu, \Sigma)$ to test the hypothesis $H: \Sigma = \sigma^2 I$, where σ^2 is not specified. The hypothesis can be given an algebraic interpretation in terms of the characteristic roots of Σ, that is, the roots of

$$|\Sigma - \phi I| = 0. \tag{1}$$

The hypothesis is true if and only if all the roots of (1) are equal.† Another way of putting it is that the arithmetic mean of roots $\phi_1, \ldots, \phi_p$ is equal to the geometric mean, that is,

$$\frac{\prod_{i=1}^{p} \phi_i^{1/p}}{\sum_{i=1}^{p} \phi_i/p} = \frac{|\Sigma|^{1/p}}{\operatorname{tr} \Sigma/p} = 1. \tag{2}$$

The lengths squared of the principal axes of the ellipsoids of constant density are proportional to the roots ϕ_i (see Chapter 11); the hypothesis specifies that these are equal, that is, that the ellipsoids are spheres.

The hypothesis H is equivalent to the more general form $\boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Psi}_0$, with $\boldsymbol{\Psi}_0$ specified, having observation vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ from $N(\boldsymbol{\nu}, \boldsymbol{\Psi})$. Let C be a matrix such that

$$C\boldsymbol{\Psi}_0 C' = \boldsymbol{I}, \tag{3}$$

and let $\boldsymbol{\mu}^* = C\boldsymbol{\nu}$, $\boldsymbol{\Sigma}^* = C\boldsymbol{\Psi}C'$, $\boldsymbol{x}_\alpha^* = C\boldsymbol{y}_\alpha$. Then $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_N^*$ are observations from $N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, and the hypothesis is transformed into $H: \boldsymbol{\Sigma}^* = \sigma^2 \boldsymbol{I}$.

10.7.2. The Criterion

In the canonical form the hypothesis H is a combination of the hypothesis $H_1 : \boldsymbol{\Sigma}$ is diagonal or the components of X are independent and H_2 : the diagonal elements of $\boldsymbol{\Sigma}$ are equal given that $\boldsymbol{\Sigma}$ is diagonal or the variances of the components of X are equal given that the components are independent. Thus by Lemma 10.3.1 the likelihood ratio criterion λ for H is the product of the criterion λ_1 for H_1 and λ_2 for H_2. From Section 9.2 we see that the criterion for H_1 is

$$\lambda_1 = \frac{|A|^{\frac{1}{2}N}}{\prod a_{ii}^{\frac{1}{2}N}} = |r_{ij}|^{\frac{1}{2}N}, \tag{4}$$

where

$$A = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})' = (a_{ij}) \tag{5}$$

and $r_{ij} = a_{ij}/\sqrt{a_{ii} a_{jj}}$. We use the results of Section 10.2 to obtain λ_2 by considering the ith component of $\boldsymbol{x}_\alpha$ as the αth observation from the ith population. (p here is q in Section 10.2; N here is N_g there; pN here is N

†This follows from the fact that $\boldsymbol{\Sigma} = \boldsymbol{O}'\boldsymbol{\Phi}\boldsymbol{O}$, where $\boldsymbol{\Phi}$ is a diagonal matrix with roots as diagonal elements and $\boldsymbol{O}$ is an orthogonal matrix.

there.) Thus

$$(6) \qquad \lambda_2 = \frac{\prod_i \left[\sum_\alpha (x_{i\alpha} - \bar{x}_i)^2\right]^{\frac{1}{2}N}}{\left[\sum_{i,\alpha}(x_{i\alpha} - \bar{x}_i)^2/p\right]^{\frac{1}{2}pN}}$$

$$= \frac{\prod a_{ii}^{\frac{1}{2}N}}{(\operatorname{tr} A/p)^{\frac{1}{2}pN}}.$$

Thus the criterion for H is

$$(7) \qquad \lambda = \lambda_1 \lambda_2 = \frac{|A|^{\frac{1}{2}N}}{(\operatorname{tr} A/p)^{\frac{1}{2}pN}}.$$

It will be observed that λ resembles (2). If $l_1, \dots, l_p$ are the roots of

$$(8) \qquad |S - lI| = 0,$$

where $S = (1/n)A$, the criterion is a power of the ratio of the geometric mean to the arithmetic mean,

$$(9) \qquad \lambda = \left(\frac{\prod l_i^{1/p}}{\sum l_i/p}\right)^{\frac{1}{2}pN}.$$

Now let us go back to the hypothesis $\boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Psi}_0$, given observation vectors $y_1, \dots, y_N$ from $N(\boldsymbol{\nu}, \boldsymbol{\Psi})$. In the transformed variables $\{\boldsymbol{x}_\alpha^*\}$ the criterion is $|A^*|^{\frac{1}{2}N}(\operatorname{tr} A^*/p)^{-\frac{1}{2}pN}$, where

$$(10) \qquad A^* = \sum_{\alpha=1}^N (\boldsymbol{x}_\alpha^* - \bar{\boldsymbol{x}}^*)(\boldsymbol{x}_\alpha^* - \bar{\boldsymbol{x}}^*)'$$

$$= C \sum_{\alpha=1}^N (y_\alpha - \bar{y})(y_\alpha - \bar{y})'C'$$

$$= CBC',$$

where

$$(11) \qquad B = \sum_{\alpha=1}^N (y_\alpha - \bar{y})(y_\alpha - \bar{y})'.$$

From (3) we have $\Psi_0 = C^{-1}(C')^{-1} = (C'C)^{-1}$. Thus

$$|A^*| = \frac{|B|}{|\Psi_0|} = |B\Psi_0^{-1}|,$$

$$\text{(12)} \qquad \operatorname{tr} A^* = \operatorname{tr} CBC' = \operatorname{tr} BC'C = \operatorname{tr} B\Psi_0^{-1}.$$

The results can be summarized.

Theorem 10.7.1. *Given a set of p-component observation vectors* $y_1, \dots, y_N$ *from* $N(\boldsymbol{\nu}, \Psi)$, *the likelihood ratio criterion for testing the hypothesis* $H: \Psi = \sigma^2 \Psi_0$, *where* Ψ_0 *is specified and* σ^2 *is not specified, is*

$$\text{(13)} \qquad \frac{|B\Psi_0^{-1}|^{\frac{1}{2}N}}{\left(\operatorname{tr} B\Psi_0^{-1}/p\right)^{\frac{1}{2}pN}}.$$

Mauchly (1940) gave this criterion and its moments under the null hypothesis.

The maximum likelihood estimator of σ^2 under the null hypothesis is $\operatorname{tr} B\Psi_0^{-1}/(pN)$, which is $\operatorname{tr} A/(pN)$ in canonical form; an unbiased estimator is $\operatorname{tr} B\Psi_0^{-1}/[p(N-1)]$ or $\operatorname{tr} A/[p(N-1)]$ in canonical form [Hotelling (1951)]. Then $\operatorname{tr} B\Psi_0^{-1}/\sigma^2$ has the χ^2-distribution with $p(N-1)$ degrees of freedom.

10.7.3. The Distribution and Moments of the Criterion

The distribution of the likelihood ratio criterion under the null hypothesis can be characterized by the facts that $\lambda = \lambda_1 \lambda_2$ and λ_1 and λ_2 are independent and by the characterizations of λ_1 and λ_2. As was observed in Section 7.6, when Σ is diagonal the correlation coefficients $\{r_{ij}\}$ are distributed independently of the variances $\{a_{ii}/(N-1)\}$. Since λ_1 depends only on $\{r_{ij}\}$ and λ_2 depends only on $\{a_{ii}\}$, they are independently distributed when the null hypothesis is true. Let $W = \lambda^{2/N}$, $W_1 = \lambda_1^{2/N}$, $W_2 = \lambda_2^{2/N}$. From Theorem 9.3.3, we see that W_1 is distributed as $\prod_{i=2}^p X_i$, where $X_2, \dots, X_p$ are independent and X_i has the density $\beta[x|\frac{1}{2}(n-i+1), \frac{1}{2}(i-1)]$, where $n = N-1$. From Theorem 10.4.2 with $W_2 = p^p V_1^{2/n}$, we find that W_2 is distributed as $p^p \prod_{j=2}^p Y_j^{j-1}(1-Y_j)$, where $Y_2, \dots, Y_p$ are independent and Y_j has the density $\beta(y|\frac{1}{2}n(j-1), \frac{1}{2}n)$. Then W is distributed as $W_1 W_2$, where W_1 and W_2 are independent.

The moments of W can be found from this characterization or from Theorems 9.3.4 and 10.4.4. We have

$$\mathscr{E}W_1^h = \frac{\Gamma^p(\frac{1}{2}n)}{\Gamma^p(\frac{1}{2}n+h)}\frac{\Gamma_p(\frac{1}{2}n+h)}{\Gamma_p(\frac{1}{2}n)}, \tag{14}$$

$$\mathscr{E}W_2^h = p^{hp}\frac{\Gamma^p(\frac{1}{2}n+h)\Gamma(\frac{1}{2}pn)}{\Gamma^p(\frac{1}{2}n)\Gamma(\frac{1}{2}pn+ph)}. \tag{15}$$

It follows that

$$\mathscr{E}W^h = p^{hp}\frac{\Gamma(\frac{1}{2}pn)}{\Gamma(\frac{1}{2}pn+ph)}\frac{\Gamma_p(\frac{1}{2}n+h)}{\Gamma_p(\frac{1}{2}n)}. \tag{16}$$

For $p=2$ we have

$$\begin{aligned}\mathscr{E}W^h &= 4^h\frac{\Gamma(n)}{\Gamma(n+2h)}\prod_{i=1}^{2}\frac{\Gamma[\frac{1}{2}(n+1-i)+h]}{\Gamma[\frac{1}{2}(n+1-i)]} \\ &= \frac{\Gamma(n)\Gamma(n-1+2h)}{\Gamma(n+2h)\Gamma(n-1)} = \frac{n-1}{n-1+2h} \\ &= (n-1)\int_0^1 z^{n-2+2h}\,dz,\end{aligned} \tag{17}$$

by use of the duplication formula for the gamma function. Thus W is distributed as Z^2, where Z has the density $(n-1)z^{n-2}$, and W has the density $\frac{1}{2}(n-1)w^{\frac{1}{2}(n-3)}$. The cdf is

$$\Pr\{W \le w\} = F(w) = w^{\frac{1}{2}(n-1)}. \tag{18}$$

This result can also be found from the joint distribution of l_1, l_2, the roots of (8). The density for $p=3$, 4, and 6 has been obtained by Consul (1967b). See also Pillai and Nagarsenkar (1971).

10.7.4. Asymptotic Expansion of the Distribution

From (16) we see that the rth moment of $W^{\frac{1}{2}n} = Z$, say, is

$$\mathscr{E}Z^r = Kp^{\frac{1}{2}npr}\frac{\prod_{i=1}^{p}\Gamma[\frac{1}{2}n(1+r)+\frac{1}{2}(1-i)]}{\Gamma[\frac{1}{2}pn(1+r)]}. \tag{19}$$

This is of the form of (1), Section 8.5, with

$$(20)\qquad \begin{aligned} &a = p, \quad && x_k = \tfrac{1}{2}n, \quad && \xi_k = \tfrac{1}{2}(1-k), \quad && k = 1, \ldots, p, \\ &b = 1, && y_1 = \tfrac{1}{2}np, && \eta_1 = 0. && \end{aligned}$$

Thus the expansion of Section 8.5 is valid with $f = \frac{1}{2}p(p+1) - 1$. To make the second term in the expansion zero we take ρ so

$$(21)\qquad 1 - \rho = \frac{2p^2 + p + 2}{6pn}.$$

Then

$$(22)\qquad \omega_2 = \frac{(p+2)(p-1)(p-2)(2p^3 + 6p^2 + 3p + 2)}{288p^2n^2\rho^2}.$$

Thus the cdf of W is found from

$$(23)\qquad \begin{aligned} &\Pr\{-2\rho \log Z \le z\} \\ &\quad = \Pr\{-n\rho \log W \le z\} \\ &\quad = \Pr\{\chi_f^2 \le z\} + \omega_2\left(\Pr\{\chi_{f+4}^2 \le z\} - \Pr\{\chi_f^2 \le z\}\right) + O(n^{-3}). \end{aligned}$$

Factors $c(n, p, \varepsilon)$ have been tabulated in Table B.6 such that

$$(24)\qquad \Pr\left\{-n\rho \log W \le c(n, p, \varepsilon)\,\chi^2_{\frac{1}{2}p(n+1)-1}(\varepsilon)\right\} = \varepsilon.$$

Nagarsenkar and Pillai (1973a) have tables for W.

10.7.5. Invariant Tests

The null hypothesis $H: \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ is invariant with respect to transformations $\boldsymbol{X}^* = c\boldsymbol{Q}\boldsymbol{X} + \boldsymbol{\nu}$, where c is a scalar and $\boldsymbol{Q}$ is an orthogonal matrix. The invariant of the sufficient statistic under shift of location is $\boldsymbol{A}$, the invariants of $\boldsymbol{A}$ under orthogonal transformations are the characteristic roots $l_1, \ldots, l_p$, and the invariants of the roots under scale transformations are functions that are homogeneous of degree 0, such as the ratios of roots, say $l_1/l_2, \ldots, l_{p-1}/l_p$. Invariant tests are based on such functions; the likelihood ratio criterion is such a function.

Nagao (1973a) proposed the criterion

$$
\begin{aligned}
(25)\qquad & \tfrac{1}{2} n \operatorname{tr}\left(S-\frac{\operatorname{tr} S}{p} I\right) \frac{p}{\operatorname{tr} S}\left(S-\frac{\operatorname{tr} S}{p} I\right) \frac{p}{\operatorname{tr} S} \\
& =\tfrac{1}{2} n \operatorname{tr}\left(\frac{p}{\operatorname{tr} S} S-I\right)^{2}=\tfrac{1}{2} n\left[\frac{p^{2}}{(\operatorname{tr} S)^{2}} \operatorname{tr} S^{2}-p\right] \\
& =\tfrac{1}{2} n\left[\frac{p}{\left(\sum_{i=1}^{p} l_{i}\right)^{2}} \sum_{i=1}^{p} l_{i}^{2}-p\right]=\tfrac{1}{2} n \frac{\sum_{i=1}^{p}\left(l_{i}-\bar{l}\right)^{2}}{\bar{l}^{2}},
\end{aligned}
$$

where $\bar{l}=\sum_{i=1}^{p} l_i/p$. The left-hand side of (25) is based on the loss function $L_q(\boldsymbol{\Sigma}, \boldsymbol{G})$ of Section 7.8; the right-hand side shows it is proportional to the square of the coefficient of variation of the characteristic roots of the sample covariance matrix $\boldsymbol{S}$. Another criterion is l_1/l_p. Percentage points have been given by Krishnaiah and Schuurmann (1974).

10.7.6. Confidence Regions

Given observations $y_1, \ldots, y_N$ from $N(\boldsymbol{\nu}, \boldsymbol{\Psi})$, we can test $\boldsymbol{\Psi}=\sigma^2 \boldsymbol{\Psi}_0$ for any specified $\boldsymbol{\Psi}_0$. From this family of tests we can set up a confidence region for $\boldsymbol{\Psi}$. If any matrix is in the confidence region, all multiples of it are. This kind of confidence region is of interest if all components of y_α are measured in the same unit, but the investigator wants a region independent of this common unit. The confidence region of confidence $1-\varepsilon$ consists of all matrices $\boldsymbol{\Psi}^*$ satisfying

$$
(26)\qquad \frac{\left|\boldsymbol{B} \boldsymbol{\Psi}^{*-1}\right|}{\left[\left(\operatorname{tr} \boldsymbol{B} \boldsymbol{\Psi}^{*-1}\right)/p\right]^{p}} \geq \lambda^{2/N}(\varepsilon),
$$

where $\lambda(\varepsilon)$ is the ε significance level for the criterion.

Consider the case of $p=2$. If the common unit of measurement is irrelevant, the investigator is interested in $\tau=\psi_{11}/\psi_{22}$ and $\rho=\psi_{12}/\sqrt{\psi_{11}\psi_{22}}$. In this case

$$
\begin{aligned}
(27)\qquad \boldsymbol{\Psi}^{-1} & =\frac{1}{\psi_{11} \psi_{22}\left(1-\rho^{2}\right)}\begin{pmatrix}
\psi_{22} & -\rho \sqrt{\psi_{11} \psi_{22}} \\
-\rho \sqrt{\psi_{11} \psi_{22}} & \psi_{11}
\end{pmatrix} \\
& =\frac{1}{\psi_{11}\left(1-\rho^{2}\right)}\begin{pmatrix}
1 & -\rho \sqrt{\tau} \\
-\rho \sqrt{\tau} & \tau
\end{pmatrix} .
\end{aligned}
$$

The region in terms of τ and ρ is

$$4\frac{(b_{11}b_{22}-b_{12}^2)(1-\rho^2)^{\tau}}{(b_{11}+\tau b_{22}-2\rho\sqrt{\tau}b_{12})^2} \geq \lambda^{2/N}(\varepsilon). \tag{28}$$

Hickman (1953) has given an example of such a confidence region.

10.8. TESTING THE HYPOTHESIS THAT A COVARIANCE MATRIX IS EQUAL TO A GIVEN MATRIX

10.8.1. The Criteria

If Y is distributed according to $N(\boldsymbol{\nu}, \boldsymbol{\Psi})$, we wish to test H_1 that $\boldsymbol{\Psi} = \boldsymbol{\Psi}_0$, where $\boldsymbol{\Psi}_0$ is a given positive definite matrix. By the argument of the preceding section we see that this is equivalent to testing the hypothesis $H_1: \boldsymbol{\Sigma} = \boldsymbol{I}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of a vector X distributed according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Given a sample $x_1, \dots, x_N$, the likelihood ratio criterion is

$$\lambda_1 = \frac{\max_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{I})}{\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}, \tag{1}$$

where the likelihood function is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{1}{2}pN}|\boldsymbol{\Sigma}|^{-\frac{1}{2}N}\exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_\alpha-\boldsymbol{\mu})\right]. \tag{2}$$

Results in Chapter 3 show that

$$\begin{aligned} \lambda_1 &= \frac{(2\pi)^{-\frac{1}{2}pN}\exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})'(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})\right]}{(2\pi)^{-\frac{1}{2}pN}|(1/N)\boldsymbol{A}|^{-\frac{1}{2}N}e^{-\frac{1}{2}pN}} \\ &= \left(\frac{e}{N}\right)^{\frac{1}{2}pN}|\boldsymbol{A}|^{\frac{1}{2}N}e^{-\frac{1}{2}\operatorname{tr}\boldsymbol{A}}, \end{aligned} \tag{3}$$

where

$$\boldsymbol{A} = \sum_{\alpha}(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})'. \tag{4}$$

Sugiura and Nagao (1968) have shown that the likelihood ratio test is biased, but the modified likelihood ratio test based on

$$\lambda_1^* = \left(\frac{e}{n}\right)^{\frac{1}{2}pn}|\boldsymbol{A}|^{\frac{1}{2}n}e^{-\frac{1}{2}\operatorname{tr}\boldsymbol{A}} = e^{\frac{1}{2}pn}\left(|\boldsymbol{S}|e^{-\operatorname{tr}\boldsymbol{S}}\right)^{\frac{1}{2}n}, \tag{5}$$

where $S = (1/n)A$, is unbiased. Note that

$$(6) \qquad -\frac{2}{n}\log \lambda_1^* = \operatorname{tr} S - \log|S| - p = L_l(\boldsymbol{I}, S),$$

where $L_l(\boldsymbol{I}, S)$ is the loss function for estimating $\boldsymbol{I}$ by S defined in (2) of Section 7.8. In terms of the characteristic roots of S the criterion (6) is a constant plus

$$(7) \qquad \sum_{i=1}^{p} l_i - \log \prod_{i=1}^{p} l_i - p = \sum_{i=1}^{p} (l_i - \log l_i - 1);$$

for each i the minimum of (7) is at $l_i = 1$.

Using the algebra of the preceding section, we see that given $y_1, \dots, y_N$ as observation vectors of p components from $N(\boldsymbol{\nu}, \boldsymbol{\Psi})$, the modified likelihood ratio criterion for testing the hypothesis $H_1 : \boldsymbol{\Psi} = \boldsymbol{\Psi}_0$, where $\boldsymbol{\Psi}_0$ is specified, is

$$(8) \qquad \lambda_1^* = \left(\frac{e}{n}\right)^{\frac{1}{2}pn} |B\boldsymbol{\Psi}_0^{-1}|^{\frac{1}{2}n} e^{-\frac{1}{2}\operatorname{tr} B\boldsymbol{\Psi}_0^{-1}},$$

where

$$(9) \qquad B = \sum_{\alpha=1}^{N} (y_\alpha - \bar{y})(y_\alpha - \bar{y})'.$$

10.8.2. The Distribution and Moments of the Modified Likelihood Ratio Criterion

The null hypothesis $H_1 : \boldsymbol{\Sigma} = \boldsymbol{I}$ is the intersection of the null hypothesis of Section 10.7, $H : \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, and the null hypothesis $\sigma^2 = 1$ given $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. The likelihood ratio criterion for H_1 given by (3) is the product of (7) of Section 10.7 and

$$(10) \qquad \left(\frac{\operatorname{tr} A}{pN}\right)^{\frac{1}{2}pN} e^{-\frac{1}{2}\operatorname{tr} A + \frac{1}{2}pN},$$

which is the likelihood ratio criterion for testing the hypothesis $\sigma^2 = 1$ given $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. The modified criterion λ_1^* is the product of $|A|^{\frac{1}{2}n}/(\operatorname{tr} A/p)^{\frac{1}{2}pn}$ and

$$(11) \qquad \left(\frac{\operatorname{tr} A}{pn}\right)^{\frac{1}{2}pn} e^{-\frac{1}{2}\operatorname{tr} A + \frac{1}{2}pn};$$

these two factors are independent (Lemma 10.4.1). The characterization of the distribution of the modified criterion can be obtained from Section

10.7.3. The quantity $\operatorname{tr} A$ has the χ^2-distribution with np degrees of freedom under the null hypothesis.

Instead of obtaining the moments and characteristic function of λ_1^* [defined by (5)] from the preceding characterization, we shall find them by use of the fact that A has the distribution $W(\Sigma, n)$. We shall calculate

$$\mathscr{E}\lambda_1^{*h} = \int \cdots \int \left(\frac{e^{\frac{1}{2}pn}}{n^{\frac{1}{2}pn}} |A|^{\frac{1}{2}n} e^{-\frac{1}{2}\operatorname{tr} A} \right)^h w(A|\Sigma, n)\, dA \tag{12}$$

$$= \frac{e^{\frac{1}{2}pnh}}{n^{\frac{1}{2}pnh}} \int \cdots \int |A|^{\frac{1}{2}nh} e^{-\frac{1}{2}h \operatorname{tr} A} w(A|\Sigma, n)\, dA.$$

Since

(13)

$$|A|^{\frac{1}{2}nh} e^{-\frac{1}{2}h\operatorname{tr} A} w(A|\Sigma, n) = \frac{|A|^{\frac{1}{2}(n+nh-p-1)} e^{-\frac{1}{2}(\operatorname{tr}\Sigma^{-1}A + \operatorname{tr} hA)}}{2^{\frac{1}{2}pn}|\Sigma|^{\frac{1}{2}n}\Gamma_p(\frac{1}{2}n)}$$

$$= \frac{2^{\frac{1}{2}pnh}\Gamma_p\left[\frac{1}{2}n(1+h)\right]}{|\Sigma^{-1} + hI|^{\frac{1}{2}(n+nh)}|\Sigma|^{\frac{1}{2}n}\Gamma_p(\frac{1}{2}n)}$$

$$\cdot \frac{|\Sigma^{-1} + hI|^{\frac{1}{2}(n+nh)}|A|^{\frac{1}{2}(n+nh-p-1)} e^{-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}+hI)A}}{2^{\frac{1}{2}p(n+nh)}\Gamma_p\left[\frac{1}{2}n(1+h)\right]}$$

$$= \frac{2^{\frac{1}{2}pnh}|\Sigma|^{\frac{1}{2}nh}\Gamma_p\left[\frac{1}{2}n(1+h)\right]}{|I + h\Sigma|^{\frac{1}{2}(n+nh)}\Gamma_p(\frac{1}{2}n)}$$

$$\cdot w\left(A|(\Sigma^{-1} + hI)^{-1}, n + nh\right),$$

the hth moment of λ_1^* is

$$\mathscr{E}\lambda_1^{*h} = \left(\frac{2e}{n}\right)^{\frac{1}{2}phn} \frac{|\Sigma|^{\frac{1}{2}nh}\prod_{j=1}^p \Gamma\left[\frac{1}{2}(n+nh+1-j)\right]}{|I + h\Sigma|^{\frac{1}{2}(n+nh)}\prod_{j=1}^p \Gamma\left[\frac{1}{2}(n+1-j)\right]}. \tag{14}$$

Then the characteristic function of $-2\log\lambda^*$ is

$$\mathscr{E} e^{-2it\log\lambda_1^*} = \mathscr{E}\lambda_1^{*-2it} \tag{15}$$

$$= \left(\frac{2e}{n}\right)^{-ipnt} \frac{|\Sigma|^{-int}}{|I - 2it\Sigma|^{\frac{1}{2}n-int}} \cdot \prod_{j=1}^p \frac{\Gamma\left[\frac{1}{2}(n+1-j) - int\right]}{\Gamma\left[\frac{1}{2}(n+1-j)\right]}.$$

When the null hypothesis is true, $\mathbf{\Sigma} = \mathbf{I}$, and

(16)
$$\mathscr{E} e^{-2it \log \lambda_1^*} = \left(\frac{2e}{n}\right)^{-ipnt} (1-2it)^{-\frac{1}{2}p(n-2int)} \prod_{j=1}^{p} \frac{\Gamma\left[\frac{1}{2}(n+1-j) - int\right]}{\Gamma\left[\frac{1}{2}(n+1-j)\right]}.$$

This characteristic function is the product of p terms such as

$$\phi_j(t) = \left(\frac{2e}{n}\right)^{-int} (1-2it)^{-\frac{1}{2}(n-2int)} \frac{\Gamma\left[\frac{1}{2}(n+1-j) - int\right]}{\Gamma\left[\frac{1}{2}(n+1-j)\right]}. \tag{17}$$

Thus $-2\log \lambda_1^*$ is distributed as the sum of p independent variates, the characteristic function of the jth being (17). Using Stirling's approximation for the gamma function, we have

$$\begin{aligned}
\phi_j(t) &\sim 2^{-int} e^{-int} n^{int} (1-2it)^{\frac{1}{2}(2int-n)} \\
&\quad \cdot \frac{e^{-[\frac{1}{2}(n+1-j)-int]}\left[\frac{1}{2}(n+1-j) - int\right]^{\frac{1}{2}(n-j)-int}}{e^{-[\frac{1}{2}(n+1-j)]}\left[\frac{1}{2}(n-j+1)\right]^{\frac{1}{2}(n-j)}} \\
&= (1-2it)^{-\frac{1}{2}j} \left(1 - \frac{it(j-1)}{\frac{1}{2}(n-j+1)(1-2it)}\right)^{\frac{1}{2}(n+1-j)-int} \\
&\quad \cdot \left(1 - \frac{2j-1}{n(1-2it)}\right)^{-int}.
\end{aligned} \tag{18}$$

As $n \to \infty$, $\phi_j(t) \to (1-2it)^{-\frac{1}{2}j}$, which is the characteristic function of χ_j^2 (χ^2 with j degrees of freedom). Thus $-2\log \lambda_1^*$ is asymptotically distributed as $\sum_{j=1}^{p} \chi_j^2$, which is χ^2 with $\sum_{j=1}^{p} j = \frac{1}{2}p(p+1)$ degrees of freedom. The distribution of λ_1^* can be further expanded [Korin (1968), Davis (1971)] as

$$\begin{aligned}
&\Pr\{-2\rho \log \lambda_1^* \le z\} \\
&\quad = \Pr\{\chi_f^2 \le z\} + \frac{\gamma_2}{\rho^2 N^2}\left(\Pr\{\chi_{f+4}^2 \le z\} - \Pr\{\chi_f^2 \le z\}\right) + O(N^{-3}),
\end{aligned} \tag{19}$$

where

$$\rho = 1 - \frac{2p^2 + 3p - 1}{6N(p+1)}, \tag{20}$$

$$\gamma_2 = \frac{p(2p^4 + 6p^3 + p^2 - 12p - 13)}{288(p+1)}. \tag{21}$$

Nagarsenker and Pillai (1973b) found exact distributions and tabulated 5% and 1% significant points, as did Davis and Field (1971), for $p = 2(1)10$ and $n = 6(1)30(5)50, 60, 120$. Table B.7 [due to Korin (1968)] gives some 5% and 1% significance points of $-2\log \lambda_1^*$ for small values of n and $p = 2(1)10$.

10.8.3. Invariant Tests

The null hypothesis $H: \Sigma = I$ is invariant with respect to transformations $X^* = QX + v$, where Q is an orthogonal matrix. The invariants of the sufficient statistics are the characteristic roots $l_1, \dots, l_p$ of S, and the invariants of the parameters are the characteristic roots of Σ. Invariant tests are based on the roots of S; the modified likelihood ratio criterion is one of them. Nagao (1973a) suggested the criterion

$$\frac{1}{2}n\,\mathrm{tr}(S - I)^2 = \frac{1}{2}n \sum_{i=1}^{p} (l_i - 1)^2. \tag{22}$$

Under the null hypothesis this criterion has a limiting χ^2-distribution with $\frac{1}{2}p(p+1)$ degrees of freedom.

Roy (1957), Section 6.4, proposed a test based on the largest and smallest characteristic roots l_1 and l_p: Reject the null hypothesis if

$$l_p < l \quad \text{or} \quad l_1 > u, \tag{23}$$

where

$$\Pr\{l < l_p, l_1 < u | \Sigma = I\} = 1 - \varepsilon \tag{24}$$

and ε is the significance level. Clemm, Krishnaiah, and Waikar (1973) give tables of $u = 1/l$. See also Schuurman and Waikar (1973).

10.8.4. Confidence Bounds for Quadratic Forms

The test procedure based on the smallest and largest characteristic roots can be inverted to give confidence bounds on qudaratic forms in Σ. Suppose nS has the distribution $W(\Sigma, n)$. Let C be a nonsingular matrix such that $\Sigma = C'C$. Then $nS^* = nC'^{-1}SC^{-1}$ has the distribution $W(I, n)$. Since $l_p^* \le a'S^*a/a'a < l_1^*$ for all a, where l_p^* and l_1^* are the smallest and largest characteristic roots of S^* (Sections 11.2 and A.2),

$$\Pr\left\{l \le \frac{a'S^*a}{a'a} \le u \quad \forall a \ne \mathbf{0}\right\} = 1 - \varepsilon, \tag{25}$$

where

$$\Pr\{l \le l_p^* \le l_1^* \le u\} = 1 - \varepsilon. \tag{26}$$

Let $a = Cb$. Then $a'a = b'C'Cb = b'\Sigma b$ and $a'S^*a = b'C'S^*Cb = b'Sb$. Thus (25) is

$$1 - \varepsilon = \Pr\left\{ l \le \frac{b'Sb}{b'\Sigma b} \le u \quad \forall b \ne 0 \right\} \tag{27}$$

$$= \Pr\left\{ \frac{b'Sb}{u} \le b'\Sigma b \le \frac{b'Sb}{l} \quad \forall b \right\}.$$

Given an observed S, one can assert

$$\frac{b'Sb}{u} \le b'\Sigma b \le \frac{b'Sb}{l} \qquad \forall b \tag{28}$$

with confidence $1 - \varepsilon$.

If b has 1 in the ith position and 0's elsewhere, (28) is $s_{ii}/u \le \sigma_{ii} \le s_{ii}/l$. If b has 1 in the ith position, -1 in the jth position, $i \ne j$, and 0's elsewhere, then (28) is

$$\frac{s_{ii} + s_{jj} - 2s_{ij}}{u} \le \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij} \le \frac{s_{ii} + s_{jj} - 2s_{ij}}{l}. \tag{29}$$

Manipulation of tnese inequalities yields

$$\frac{s_{ij}}{l} - \frac{s_{ii} + s_{jj}}{2}\left(\frac{1}{l} - \frac{1}{u}\right) \le \sigma_{ij} \le \frac{s_{ij}}{u} + \frac{s_{ii} + s_{jj}}{2}\left(\frac{1}{l} - \frac{1}{u}\right), \qquad i \ne j. \tag{30}$$

We can obtain simultaneously confidence intervals on all elements of Σ.

From (27) we can obtain

$$1 - \varepsilon = \Pr\left\{ \frac{1}{u}\frac{b'Sb}{b'b} \le \frac{b'\Sigma b}{b'b} \le \frac{1}{l}\frac{b'Sb}{b'b} \quad \forall b \right\} \tag{31}$$

$$\le \Pr\left\{ \frac{1}{u}\min_a \frac{a'Sa}{a'a} \le \frac{b'\Sigma b}{b'b} \le \frac{1}{l}\max_a \frac{a'Sa}{a'a} \quad \forall b \right\}$$

$$= \Pr\left\{ \frac{1}{u} l_p \le \lambda_p \le \lambda_1 \le \frac{1}{l} l_1 \right\},$$

where l_1 and l_p are the largest and smallest characteristic roots of S and λ_1 and λ_p are the largest and smallest characteristic roots of Σ. Then

$$\frac{1}{u} l_p \le \lambda(\Sigma) \le \frac{1}{l} l_1 \tag{32}$$

is a confidence interval for all characteristic roots of Σ with confidence at least $1 - \varepsilon$. In Section 11.6 we give tighter bounds on $\lambda(\Sigma)$ with exact confidence.

10.9. TESTING THE HYPOTHESIS THAT A MEAN VECTOR AND A COVARIANCE MATRIX ARE EQUAL TO A GIVEN VECTOR AND MATRIX

In Chapter 3 we pointed out that if $\boldsymbol{\Psi}$ is known, $(\bar{\boldsymbol{y}} - \boldsymbol{\nu}_0)'\boldsymbol{\Psi}_0^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\nu}_0)$ is suitable for testing

$$(1) \qquad H_2 : \boldsymbol{\nu} = \boldsymbol{\nu}_0, \qquad \text{given} \quad \boldsymbol{\Psi} \equiv \boldsymbol{\Psi}_0.$$

Now let us combine H_1 of Section 10.8 and H_2, and test

$$(2) \qquad H : \boldsymbol{\nu} = \boldsymbol{\nu}_0, \qquad \boldsymbol{\Psi} = \boldsymbol{\Psi}_0,$$

on the basis of a sample $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N$ from $N(\boldsymbol{\nu}, \boldsymbol{\Psi})$.

Let

$$(3) \qquad \boldsymbol{X} = \boldsymbol{C}(\boldsymbol{Y} - \boldsymbol{\nu}_0),$$

where

$$(4) \qquad \boldsymbol{C}\boldsymbol{\Psi}_0\boldsymbol{C}' = \boldsymbol{I}.$$

Then $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ constitutes a sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and the hypothesis is

$$(5) \qquad H : \boldsymbol{\mu} = \boldsymbol{0}, \qquad \boldsymbol{\Sigma} = \boldsymbol{I}.$$

The likelihood ratio criterion for $H_2 : \boldsymbol{\mu} = \boldsymbol{0}$, given $\boldsymbol{\Sigma} = \boldsymbol{I}$, is

$$(6) \qquad \lambda_2 = e^{-\frac{1}{2}N\bar{\boldsymbol{x}}'\bar{\boldsymbol{x}}}.$$

The likelihood ratio criterion for H is (by Lemma 10.3.1)

$$(7) \qquad \begin{aligned} \lambda = \lambda_1\lambda_2 &= \left(\frac{e}{N}\right)^{\frac{1}{2}pN} |\boldsymbol{A}|^{\frac{1}{2}N} e^{-\frac{1}{2}\operatorname{tr}\boldsymbol{A}} e^{-\frac{1}{2}N\bar{\boldsymbol{x}}'\bar{\boldsymbol{x}}} \\ &= \left(\frac{e}{N}\right)^{\frac{1}{2}pN} |\boldsymbol{A}|^{\frac{1}{2}N} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{A} + N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}')} \\ &= \left(\frac{e}{N}\right)^{\frac{1}{2}pN} |\boldsymbol{A}|^{\frac{1}{2}N} e^{-\frac{1}{2}\Sigma \boldsymbol{x}_\alpha'\boldsymbol{x}_\alpha}. \end{aligned}$$

The likelihood ratio test (rejecting H if λ is less than a suitable constant) is unbiased [Srivastava and Khatri (1979), Theorem 10.4.5]. The two factors λ_1 and λ_2 are independent because λ_1 is a function of $\boldsymbol{A}$ and λ_2 is a function of $\bar{\boldsymbol{x}}$, and $\boldsymbol{A}$ and $\bar{\boldsymbol{x}}$ are independent. Since

$$(8) \qquad \mathscr{E}\lambda_2^h = \mathscr{E}e^{-\frac{1}{2}hN\Sigma\bar{x}_i^2} = \mathscr{E}e^{-\frac{1}{2}h\chi_p^2} = (1+h)^{-\frac{1}{2}p},$$

the hth moment of λ is

$$\mathscr{E}\lambda^h = \mathscr{E}\lambda_1^h \mathscr{E}\lambda_2^h = \left(\frac{2e}{N}\right)^{\frac{1}{2}pNh} \frac{1}{(1+h)^{\frac{1}{2}pN(1+h)}} \frac{\Gamma_p\left[\frac{1}{2}(n+Nh)\right]}{\Gamma_p\left(\frac{1}{2}n\right)} \tag{9}$$

under the null hypothesis. Then

$$-2\log\lambda = -2\log\lambda_1 - 2\log\lambda_2 \tag{10}$$

has asymptotically the χ^2-distribution with $f = p(p+1)/2 + p$ degrees of freedom. In fact, an asymptotic expansion of the distribution [Davis (1971)] of $-2\rho\log\lambda$ is

$$\begin{aligned}(11)\quad &\Pr\{-2\rho\log\lambda \le z\}\\ &= \Pr\{\chi_f^2 \le z\} + \frac{\gamma_2}{\rho^2 N^2}\left(\Pr\{\chi_{f+4}^2 \le z\} - \Pr\{\chi_f^2 \le z\}\right) + O(N^{-3}),\end{aligned}$$

where

$$\rho = 1 - \frac{2p^2 + 9p - 11}{6N(p+3)}, \tag{12}$$

$$\gamma_2 = \frac{p(2p^4 + 18p^3 + 49p^2 + 36p - 13)}{288(p-3)}. \tag{13}$$

Nagarsenker and Pillai (1974) used the moments to derive exact distributions and tabulated the 5% and 1% significance points for $p = 2(1)6$ and $N = 4(1)20(2)40(5)100$.

Now let us return to the observations $y_1, \dots, y_N$. Then

$$\begin{aligned}(14)\quad \sum_\alpha x_\alpha' x_\alpha &= \sum_\alpha (y_\alpha - \nu_0)' C' C (y_\alpha - \nu_0)\\ &= \sum_\alpha (y_\alpha - \nu_0)' \Psi_0^{-1} (y_\alpha - \nu_0)\\ &= \operatorname{tr} A + N\bar{x}'\bar{x}\\ &= \operatorname{tr}(B\Psi_0^{-1}) + N(\bar{y} - \nu_0)'\Psi_0^{-1}(\bar{y} - \nu_0)\end{aligned}$$

and

$$|A| = |B\Psi_0^{-1}|. \tag{15}$$

Theorem 10.9.1. *Given the p-component observation vectors $y_1, \dots, y_N$ from $N(\nu, \Psi)$, the likelihood ratio criterion for testing the hypothesis $H: \nu = \nu_0$,*

$\Psi = \Psi_0$, *is*

$$\lambda = \left(\frac{e}{N}\right)^{\frac{1}{2}pN} |B\Psi_0^{-1}|^{\frac{1}{2}N} e^{-\frac{1}{2}[\operatorname{tr} B\Psi_0^{-1} + N(\bar{y}-\nu_0)'\Psi_0^{-1}(\bar{y}-\nu_0)]}. \tag{16}$$

When the null hypothesis is true, $-2\log\lambda$ is asymptotically distributed as χ^2 with $\frac{1}{2}p(p+1)+p$ degrees of freedom.

10.10. ADMISSIBILITY OF TESTS

We shall consider some Bayes solutions to the problem of testing the hypothesis

$$\Sigma_1 = \cdots = \Sigma_q \tag{1}$$

as in Section 10.2. Under the alternative hypothesis, let

$$[\mu^{(g)}, \Sigma_g] = \left[(I + C_g C_g')^{-1} C_g y^{(g)}, (I + C_g C_g)^{-1}\right], \qquad g = 1, \ldots, q, \tag{2}$$

where the $p \times r_g$ matrix C_g has density proportional to $|I + C_g C_g'|^{-\frac{1}{2}n_g}$, $n_g = N_g - 1$, the r_g-component random vector $y^{(g)}$ has the conditional normal distribution with mean $\mathbf{0}$ and covariance matrix $(1/N_g)[I_{r_g} - C_g'(I_p + C_g C_g')^{-1} C_g]^{-1}$ given C_g, and $(C_1, y^{(1)}), \ldots, (C_q, y^{(q)})$ are independently distributed. As we shall see, we need to choose suitable integers $r_1, \ldots, r_q$. Note that the integral of $|I + C_g C_g'|^{-\frac{1}{2}n_g}$ is finite if $n_g \geq p + r_g$. Then the numerator of the Bayes ratio is

$$\begin{aligned}
(3)\quad \text{const} \prod_{g=1}^{q} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} & |I + C_g C_g'|^{\frac{1}{2}N_g} \\
& \cdot \exp\left\{ -\frac{1}{2} \sum_{\alpha=1}^{N_g} \left[x_\alpha^{(g)} - (I + C_g C_g')^{-1} C_g y^{(g)} \right]' \right. \\
& \qquad \left. \cdot (I + C_g C_g') \left[x_\alpha^{(g)} - (I + C_g C_g')^{-1} C_g y^{(g)} \right] \right\} \\
& \cdot |I + C_g C_g'|^{-\frac{1}{2}n_g} \left| I - C_g'(I + C_g C_g')^{-1} C_g \right|^{\frac{1}{2}} \\
& \qquad \cdot \exp\left\{ -\tfrac{1}{2} N_g y^{(g)\prime} \left[I - C_g'(I + C_g C_g')^{-1} C_g^{(g)} \right] y^{(g)} \right\} dy^{(g)}\, dC_g
\end{aligned}$$

$$= \text{const} \prod_{g=1}^{q} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2} \left[\sum_{\alpha=1}^{N_g} x_\alpha^{(g)\prime} (I + C_g C_g') x_\alpha^{(g)} \right.\right.$$

$$\left.\left. - 2 y^{(g)\prime} C_g' \sum_{\alpha=1}^{N_g} x_\alpha^{(g)} + N_g y^{(g)\prime} y^{(g)} \right] \right\} dy^{(g)}\, dC_g$$

$$= \text{const} \prod_{g=1}^{q} \exp\left\{ -\frac{1}{2} \sum_{\alpha=1}^{N_g} x_\alpha^{(g)\prime} x_\alpha^{(g)} \right\} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}$$

$$\cdot \exp\left\{ -\tfrac{1}{2} N_g \left(y^{(g)} - C_g' \bar{x}^{(g)} \right)' \left(y^{(g)} - C_g' \bar{x}^{(g)} \right) - \tfrac{1}{2} \operatorname{tr} C_g' A_g C_g \right\} dy^{(g)}\, dC_g$$

$$= \text{const} \prod_{g=1}^{q} \exp\left\{ -\tfrac{1}{2} \left[\operatorname{tr} A_g + N_g \bar{x}^{(g)\prime} \bar{x}^{(g)} \right] \right\} |A_g|^{-\frac{1}{2} r_g}.$$

Under the null hypothesis let

$$(4) \qquad \left[\mu^{(g)}, \Sigma_g \right] = \left[(I + CC')^{-1} C y^{(g)}, (I + CC')^{-1} \right],$$

where the $p \times r$ matrix C has density proportional to $|I + CC'|^{-\frac{1}{2}n}$, $n = \sum_{g=1}^{q} n_g$, the r-component vector $y^{(g)}$ has the conditional normal distribution with mean 0 and covariance matrix $(1/N_g)[I_r - C'(I_p + CC')^{-1} C]^{-1}$ given C, and $y^{(1)}, \dots, y^{(g)}$ are conditionally independent. Note that the integral of $|I + CC'|^{-\frac{1}{2}n}$ is finite if $n \ge p + r$. The denominator of the Bayes ratio is

(5)

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{g=1}^{q} \left[|I + CC'|^{\frac{1}{2} N_g} \exp\left\{ -\frac{1}{2} \sum_{\alpha=1}^{N_g} \left[x_\alpha^{(g)} - (I + CC')^{-1} C y^{(g)} \right]' \right.\right.$$

$$\left. \cdot (I + CC') \left[x_\alpha^{(g)} - (I + CC')^{-1} C y^{(g)} \right] \right\}$$

$$\cdot |I + CC'|^{-\frac{1}{2} n_g} \left| I - C'(I + CC')^{-1} C \right|^{\frac{1}{2}}$$

$$\left. \cdot \exp\left\{ -\tfrac{1}{2} N_g y^{(g)\prime} \left[I - C'(I + CC')^{-1} C \right] y^{(g)} \right\} dy^{(g)} \right] dC$$

$$= \text{const} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{g=1}^{q} \exp\left\{ -\frac{1}{2} \left[\sum_{\alpha=1}^{N_g} x_\alpha^{(g)\prime} (I + CC') x_\alpha^{(g)} \right.\right.$$

$$\left.\left. - 2y^{(g)\prime} C' \sum_{\alpha=1}^{N_g} x_\alpha^{(g)} + N_g y^{(g)\prime} y^{(g)} \right] \right\} dy^{(g)}\, dC$$

$$= \text{const} \exp\left\{ -\frac{1}{2} \sum_{g=1}^{q} \sum_{\alpha=1}^{N_g} x_\alpha^{(g)\prime} x_\alpha^{(g)} \right\} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}$$

$$\cdot \exp\left\{ -\frac{1}{2} \sum_{g=1}^{q} N_g \left(y^{(g)} - C'\bar{x}^{(g)} \right)' \left(y^{(g)} - C'\bar{x}^{(g)} \right) - \tfrac{1}{2} \operatorname{tr} C'AC \right\} \prod_{g=1}^{q} dy^{(g)}\, dC$$

$$= \text{const} \exp\left\{ -\frac{1}{2} \left(\operatorname{tr} A + \sum_{g=1}^{q} N_g \bar{x}^{(g)\prime} \bar{x}^{(g)} \right) \right\} |A|^{-\frac{1}{2}r}.$$

The Bayes test procedure is to reject the hypothesis if

$$\frac{|A|^{\frac{1}{2}r}}{\prod_{g=1}^{q} |A_g|^{\frac{1}{2}r_g}} \geq c. \tag{6}$$

For invariance we want $\sum_{g=1}^{q} r_g = r$.

The binding constraint on the choice of $r_1, \dots, r_q$ is $r_g \leq n_g - p$, $g = 1, \dots, q$. It is possible in some special cases to choose $r_1, \dots, r_q$ so that $(r_1, \dots, r_q)$ is proportional to $(N_1, \dots, N_q)$ and hence yield the likelihood ratio test or proportional to $(n_1, \dots, n_q)$ and hence yield the modified likelihood ratio test, but since $r_1, \dots, r_q$ have to be integers, it may not be possible to choose them in either such way. Next we consider an extension of this approach that involves the choice of numbers $t_1, \dots, t_q$, and t as well as $r_1, \dots, r_q$, and r.

Suppose $2(p-1) < n_g$, $g = 1, \dots, q$, and take $r_g \geq p$. Let t_g be a real number such that $2p - 1 < r_g + t_g + p < n_g + 1$, and let t be a real number such that $2p - 1 < r + t + p < n + 1$. Under the alternative hypothesis let the marginal density of C_g be proportional to $|C_g C_g'|^{\frac{1}{2}t_g} |I + C_g C_g'|^{-\frac{1}{2}n_g}$, $g = 1, \dots, q$, and under the null hypothesis let the marginal density of C be proportional to $|CC'|^{\frac{1}{2}t} |I + CC'|^{-\frac{1}{2}n}$. (The conditions on $t_1, \dots, t_q$, and t ensure that the purported densities have finite integrals; see Problem 10.18.) Then the Bayes procedure is to reject the null hypothesis if

$$\frac{|A|^{\frac{1}{2}(r+t)}}{\prod_{g=1}^{q} |A_g|^{\frac{1}{2}(r_g+t_g)}} \geq c. \tag{7}$$

For invariance we want $t = \sum_{g=1}^{q} t_g$. If $t_1, \dots, t_q$ are taken so $r_g + t_g = kN_g$ and $p - 1 < kN_g < N_g - p$, $g = 1, \dots, q$, for some k, then (7) is the likelihood ratio test; if $r_g + t_g = kn_g$ and $p - 1 < kn_g < n_g + 1 - p$, $g = 1, \dots, q$, for some k, then (7) is the modified test [i.e., $(p-1)/\min_g N_g < k < 1 - p/\min_g N_g$].

Theorem 10.10.1. *If $2p < N_g + 1$, $g = 1, \dots, q$, then the likelihood ratio test and the modified likelihood ratio test of the null hypothesis* (1) *are admissible.*

Now consider the hypothesis

$$\mu^{(1)} = \cdots = \mu^{(q)}, \qquad \Sigma_1 = \cdots = \Sigma_q. \tag{8}$$

The alternative hypothesis has been treated before. For the null hypothesis let

$$\left[\mu^{(g)}, \Sigma_g\right] = \left[(I + CC')Cy, (I + CC')^{-1}\right], \tag{9}$$

where the $p \times r$ matrix C has the density proportional to $|I + CC'|^{-\frac{1}{2}(N-1)}$ and the r-component vector y has the conditional normal distribution with mean 0 and covariance matrix $(1/N)[I - C'(I + CC')^{-1}C]^{-1}$ given C. Then the Bayes procedure is to reject the null hypothesis (8) if

$$\frac{\left|\sum_{g=1}^{q} A_g + \sum_{g=1}^{q} N_g (\bar{y}^{(g)} - \bar{y})(\bar{y}^{(g)} - \bar{y})'\right|^{\frac{1}{2}r}}{\prod_{g=1}^{q} |A_g|^{\frac{1}{2}r_g}} \geq c. \tag{10}$$

If $2p < N_g + 1$, $g = 1, \dots, q$, the prior distribution can be modified as before to obtain the likelihood ratio test and modified likelihood ratio test.

Theorem 10.10.2. *If $2p < N_g + 1$, $g = 1, \dots, q$, the likelihood ratio test and modified likelihood ratio test of the null hypothesis* (8) *are admissible.*

For more details see Kiefer and Schwartz (1965).

10.11. ELLIPTICALLY CONTOURED DISTRIBUTIONS

10.11.1. Observations Elliptically Contoured

Let $x_\alpha^{(g)}$, $\alpha = 1, \dots, N_g$, be N_g observations on $X^{(g)}$ having the density

$$|\Lambda_g|^{-\frac{1}{2}} g\left[(x - \nu^{(g)})' \Lambda_g^{-1} (x - \nu^{(g)})\right], \tag{1}$$

where $\mathscr{E}[(X - \nu^{(g)})' \Lambda_g^{-1} (X - \nu^{(g)})]^2 = \mathscr{E} R_g^4 < \infty$, $g = 1, \dots, q$. Note that the same function $g(\cdot)$ is used for the density in all q populations. Define N, A_g,

$g = 1, \dots, q$, and A by (1) of Section 10.2. Let $S_g = (1/n_g)A_g$, where $n_g = N_g - 1$, and $S = (1/n)A$, where $n = \sum_{g=1}^q n_g$.

Since the likelihood ratio criterion λ_1 is invariant under the transformation $x^{(g)} = Cx^{(g)} + \boldsymbol{\nu}^{(g)}$, under the null hypothesis we can take $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q = \boldsymbol{I}$ and $\boldsymbol{\nu}^{(1)} = \cdots = \boldsymbol{\nu}^{(g)} = \boldsymbol{0}$. Then

$$
\begin{aligned}
(2)\quad -2\log\lambda_1 &= -\left[\sum_{g=1}^{q} N_g \log|\hat{\boldsymbol{\Sigma}}_{g\Omega}| - N\log|\hat{\boldsymbol{\Sigma}}_{\omega}|\right] \\
&= -\left[\sum_{g=1}^{g} N_g \log\left|\boldsymbol{I} + \left(\hat{\boldsymbol{\Sigma}}_{g\Omega} - \boldsymbol{I}\right)\right| - N\log\left|\boldsymbol{I} + \left(\hat{\boldsymbol{\Sigma}}_{\omega} - \boldsymbol{I}\right)\right|\right] \\
&= -\left\{\sum_{g=1}^{q} N_g\left[\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{g\Omega} - \boldsymbol{I}\right) - \tfrac{1}{2}\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{g\Omega} - \boldsymbol{I}\right)^2 + O_p\left(N_g^{-3}\right)\right]\right. \\
&\qquad \left. - N\left[\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{\omega} - \boldsymbol{I}\right) - \tfrac{1}{2}\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{\omega} - \boldsymbol{I}\right)^2 + O_p\left(N^{-3}\right)\right]\right\} \\
&= \tfrac{1}{2}\sum_{g=1}^{q} N_g \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{g\Omega} - \boldsymbol{I}\right)^2 - \tfrac{1}{2}N\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{\omega} - \boldsymbol{I}\right)^2 + O_p\left(N^{-3}\right) \\
&= \tfrac{1}{2}\sum_{g=1}^{q} N_g\left[\operatorname{vec}\left(\hat{\boldsymbol{\Sigma}}_{g\Omega} - \boldsymbol{I}\right)\right]'\operatorname{vec}\left(\hat{\boldsymbol{\Sigma}}_{g\Omega} - \boldsymbol{I}\right) \\
&\qquad - \tfrac{1}{2}N\left[\operatorname{vec}\left(\hat{\boldsymbol{\Sigma}}_{\omega} - \boldsymbol{I}\right)\right]'\operatorname{vec}\left(\hat{\boldsymbol{\Sigma}}_{\omega} - \boldsymbol{I}\right) + O_p\left(N^{-3}\right).
\end{aligned}
$$

By Theorem 3.6.2

$$
(3)\quad \sqrt{N_g}\operatorname{vec}\left(S_g - \boldsymbol{I}_p\right) \xrightarrow{d} N\left[\boldsymbol{0}, (\kappa + 1)\left(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{pp}\right) + \kappa \operatorname{vec}\boldsymbol{I}_p\left(\operatorname{vec}\boldsymbol{I}_p\right)'\right],
$$

and $n_g S_g = N_g\hat{\boldsymbol{\Sigma}}_{g\Omega}$, $g = 1, \dots, q$, are independent. Let $N_g = k_g N$, $g = 1, \dots, q$, $\sum_{g=1}^q k_g = 1$, and let $N \to \infty$. In terms of this asymptotic theory the limiting distribution of $\operatorname{vec}(S_1 - \boldsymbol{I}), \dots, \operatorname{vec}(S_q - \boldsymbol{I})$ is the same as the distribution of $\bar{\boldsymbol{y}}^{(1)}, \dots, \bar{\boldsymbol{y}}^{(q)}$ of Section 8.8, with $\boldsymbol{\Sigma}$ of Section 8.8 replaced by $(\kappa + 1)(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{pp}) + \kappa \operatorname{vec}\boldsymbol{I}_p(\operatorname{vec}\boldsymbol{I}_p)'$.

When $\boldsymbol{\Sigma} = \boldsymbol{I}$, the variance of the limiting distribution of $\sqrt{N_g}(s_{ii}^{(g)} - 1)$ is $3\kappa + 2$; the covariance of the limiting distribution of $\sqrt{N_g}(s_{ii}^{(g)} - 1)$ and $\sqrt{N_g}(s_{jj}^{(g)} - 1)$, $i \neq j$, is κ; the variance of $s_{ij}^{(g)}$, $i \neq j$, is $\kappa + 1$; the set $\sqrt{N}(s_{11}^{(g)} - 1), \dots, \sqrt{N}(s_{pp}^{(g)} - 1)$ is independent of the set $(s_{ij}^{(g)})$, $i \neq j$; and the $s_{ij}^{(g)}$, $i < j$, are mutually uncorrelated (as in Section 7.9.1).

Let $\bar{y}_g = \operatorname{vec}(\hat{\Sigma}_{g\Omega} - I)$ and $\bar{y} = \operatorname{vec}(\hat{\Sigma}_\omega - I)$. Then $\bar{y} = \sum_{g=1}^q (N_g/N)\bar{y}_g$ and

$$(4) \qquad -2\log\lambda = \tfrac{1}{2}\sum_{g=1}^q N_g(\bar{y}_g - \bar{y})'(\bar{y}_g - \bar{y})$$

$$= \operatorname{tr}\tfrac{1}{2}\sum_{g=1}^q N_g(\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'$$

$$= \tfrac{1}{2}\operatorname{tr}\left(\sum_{g=1}^q N_g\bar{y}_g\bar{y}_g - N\bar{y}\bar{y}'\right).$$

Let Q be a $q \times q$ orthogonal matrix with last column $(\sqrt{N_1/N}, \dots, \sqrt{N_q/N})'$. Define

$$(5) \qquad (w_1, \dots, w_q) = \left(\sqrt{N_1}\,\bar{y}_1, \dots, \sqrt{N_g}\,\bar{y}_g\right)Q.$$

Then $w_q = \sqrt{N}\bar{y}$ and

$$(6) \qquad \sum_{g=1}^q N_g\bar{y}_g\bar{y}_g' - N\bar{y}\bar{y} = \sum_{g=1}^{q-1} w_g w_g'.$$

In these terms

$$(7) \qquad -2\log\lambda_1 = \tfrac{1}{2}\sum_{g=1}^{q-1} w_g'w_g + O_p(N^{-3}),$$

and $w_1, \dots, w_{q-1}$ are asymptotically independent, w_g having the covariance matrix of $\sqrt{N}\bar{y}_g$; that is, $(\kappa+1)(I_{p^2} + K_{pp}) + \kappa \operatorname{vec} I_p (\operatorname{vec} I_p)$. Then $w_g'w_g = \sum_{i,j=1}^p (w_{ij}^{(g)})^2 = \sum_{i=1}^p (w_{ii}^{(g)})^2 + 2\sum_{i<j}(w_{ij}^{(g)})^2$. The covariance matrix of $w_{11}^{(g)}, \dots, w_{pp}^{(g)}$ is $2(\kappa+1)I_p + \kappa\varepsilon\varepsilon$, where $\varepsilon = (1, \dots, 1)'$. The characteristic roots of this matrix are $2(\kappa+1)$ of multiplicity $p-1$ and a single root of $2(\kappa+1) + p\kappa$. Thus $\sum_{i=1}^p (w_{ii}^{(g)})^2$ has the distribution of $2(\kappa+1)\chi_{p-1}^2 + [2(\kappa+1) + p\kappa]\chi_1^2$. The distribution of $2\sum_{i<j}(w_{ij}^{(g)})^2$ is $2(\kappa+1)\chi_{p(p-1)/2}^2$.

Theorem 10.11.1. *When sampling from* (1) *and the null hypothesis is true,*

$$(8) \qquad -2\log\lambda_1 \xrightarrow{d} (\kappa+1)\chi_{(q-1)(p-1)(p+2)/2}^2 + [(\kappa+1) + p\kappa/2]\,\chi_{q-1}^2.$$

When $\kappa = 0$, $-2\log\lambda_1 \xrightarrow{d} \chi_{(q-1)p(p+1)/2}^2$ is in agreement with (12) of Section 10.5. The validity of the distributions derived in Section 10.4 depend on the observations being normally distributed; Theorem 10.11.1 shows that even the asymptotic theory depends on nonnormality.

The likelihood criteria for testing the null hypothesis (2) of Section 10.3 is the product $\lambda_1\lambda_2$ or V_1V_2. Lemma 10.4.1 states that under normality V_1 and V_2 (or equivalently λ_1 and λ_2) are independent. In the elliptically contoured case we want to show that $\log V_1$ and $\log V_2$ are asymptotically independent.

Lemma 10.11.1. *Let $A_1 = n_1S_1$ and $A_2 = n_2S_2$ be defined by* (2) *of Section* 10.2 *with $\Sigma_1 = \Sigma_2 = I$. Then $A_1(A_1+A_2)^{-1}$ and A_1+A_2 are asymptotically independent.*

Proof. Let $(1\sqrt{n_g})(A_g - n_gI) = W_g$, $g = 1,2$. Then

(9)

$$\sqrt{n_1}\left[A_1(A_1+A_2)^{-1} - \frac{n_1}{n_1+n_2}I\right] = \frac{n_1n_2}{(n_1+n_2)^2}W_1 - \frac{n_1\sqrt{n_1n_2}}{(n_1+n_2)^2}W_2 + O_p(1),$$

(10)

$$\sqrt{n_1+n_2}\,[A_1+A_2-(n_1+n_2)I] = \sqrt{\frac{n_1}{n_1+n_2}}\,W_1 + \sqrt{\frac{n_2}{n_1+n_2}}\,W_2 + O_p(1).$$

Then

$$\mathscr{E}\operatorname{vec}\left(\frac{n_1n_2}{(n_1+n_2)^2}W_1 - \frac{n_1\sqrt{n_1n_2}}{(n_1+n_2)^2}W_2\right) \tag{11}$$
$$\left[\operatorname{vec}\left(\sqrt{\frac{n_1}{n_1+n_2}}\,W_1 + \sqrt{\frac{n_2}{n_1+n_2}}\,W_2\right)\right]' = 0. \qquad \blacksquare$$

By application of Lemma 10.11.1 in succession to A_1 and A_1+A_2, to A_1+A_2 and $A_1+A_2+A_3$, etc., we establish that $A_1A^{-1}, A_2A^{-1}, \ldots, A_qA^{-1}$ are independent of $A = A_1 + \cdots + A_q$. It follows that V_1 and V_2 are asymptotically independent.

Theorem 10.11.2. *When $\Sigma_1 = \cdots = \Sigma_g$ and $\mu^{(1)} = \cdots = \mu^{(g)}$,*

$$-2\log\lambda_1\lambda_2 = -2\log\lambda_1 - 2\log\lambda_2 \tag{12}$$
$$\xrightarrow{d} (\kappa+1)\chi^2_{(q-1)(p-1)(p+2)/2} + [(\kappa+1)+p\kappa/2]\,\chi^2_{q-1} + \chi^2_{p(q-1)}.$$

The hypothesis of sphericity is that $\Sigma = \sigma^2 I$ (or $\Lambda = \lambda I$). The criteron is $\lambda_1 \lambda_2$, where

$$\lambda_1 = \left(\frac{|A|}{\prod_{i=1}^{p} a_{ii}} \right)^{N/2}, \qquad \lambda_2 = \left[\frac{\prod_{i=1}^{p} a_{ii}}{\left(\frac{\operatorname{tr} A}{p} \right)^p} \right]^{N/2}. \tag{13}$$

The first factor is the criteron for independence of the components of X, and the second is that the variances of the components are equal. For the first we set $q = p$ and $p_i = 1$ in Theorem 9.10, and for the second we set $q = p$ and $p = 1$. Thus

$$-2 \log(\lambda_1 \lambda_2) \xrightarrow{d} (1 + \kappa) \chi^2_{p-1} + \tfrac{1}{2}(3\kappa + 2) \chi^2_{p-1}. \tag{14}$$

10.11.2. Elliptically Contoured Matrix Distributions

Consider the density

$$\begin{aligned} &\prod_{g=1}^{q} |\Lambda_g|^{-N_g/2} g\left[\operatorname{tr} \sum_{g=1}^{q} \Lambda_g^{-1} \left(X^{(g)} - \nu_g \varepsilon'_{N_g} \right) \left(X^{(g)} - \nu_g \varepsilon'_{N_g} \right)' \right] \\ &= \prod_{g=1}^{q} |\Lambda_g|^{-N_g/2} g\left[\operatorname{tr} \sum_{g=1}^{q} \Lambda_g^{-1} A_g + \sum_{g=1}^{q} N_g \left(\bar{x}^{(g)} - \nu^{(g)} \right)' \Lambda_g^{-1} \left(\bar{x}^{(g)} - \nu^{(g)} \right) \right]. \end{aligned} \tag{15}$$

In this density $(A_g, \bar{x}_g)$, $g = 1, \dots, q$, is a sufficient set of statistics, and the likelihood ratio criterion is (8) of Section 10.2, the same as for normality [Anderson and Fang (1990b)].

Theorem 10.11.3. *Let $f(X)$ be a vector-valued function of $X = (X^{(1)}, \dots, X^{(q)})$ $(p \times N)$ such that*

$$f\left(X^{(1)} + \nu^{(1)} \varepsilon'_{N_1}, \dots, X^{(q)} + \nu^{(q)} \varepsilon'_{N_q} \right) = f\left(X^{(1)}, \dots, X^{(q)} \right) \tag{16}$$

for every $(\nu^{(1)}, \dots, \nu^{(q)})$ and

$$f\left(CX^{(1)}, \dots, CX^{(q)} \right) = f\left(X^{(1)}, \dots, X^{(q)} \right) \tag{17}$$

for every nonsingular C. Then the distribution of $f(X)$ where X has the arbitrary density (15) with $\Lambda_1 = \cdots = \Lambda_q$ is the same as the distribution of $f(X)$ where X has the normal density (15).

The proof of Theorem 10.11.3 is similar to the proof of Theorem 4.5.4. The theorem implies that the distribution of the criterion V_1 of (10) of Section 10.2 when the density of X is (15) with $\Lambda_1 = \cdots = \Lambda_q$ is the same as for normality. Hence the distributions and their asymptotic expansions are those discussed in Sections 10.4 and 10.5.

Corollary 10.11.1. *Let $f(X)$ be a vector-valued function of X $(p \times N)$ such that*

$$f(X + \boldsymbol{\nu}\boldsymbol{\varepsilon}_N') = f(X) \tag{18}$$

for every $\boldsymbol{\nu}$ and (17) holds. Then the distribution of $f(X)$, where X has the arbitrary density (15) with $\Lambda_1 = \cdots = \Lambda_q$ and $\boldsymbol{\nu}^{(1)} = \cdots = \boldsymbol{\nu}^{(q)}$, is the same as the distribution of $f(X)$, where X has the normal density (15).

If follows that the distribution of the criterion λ of (7) or V of (11) of Section 10.3 is the same for the density (15) as for X being normally distributed.

Let X $(p \times N)$ have the density

$$|\Lambda|^{-N/2} g\left[\operatorname{tr} \Lambda^{-1}(X - \boldsymbol{\nu}\boldsymbol{\varepsilon}_N')(X - \boldsymbol{\nu}\boldsymbol{\varepsilon}_N')'\right]. \tag{19}$$

Then the likelihood ratio criterion for testing the null hypothesis $\Lambda = \lambda I$ for some $\lambda > 0$ is (7) of Section 10.7, and its distribution under the null hypothesis is the same as for X being normally distributed.

For more detail see Anderson and Fang (1990b) and Fang and Zhang (1990).

PROBLEMS

10.1. (Sec. 10.2) Sums of squares and cross-products of deviations from the means of four measurements are given below (from Table 3.4). The populations are *Iris versicolor* (1), *Iris setosa* (2), and *Iris virginica* (3); each sample consists of 50 observations:

$$A_1 = \begin{pmatrix} 13.0552 & 4.1740 & 8.9620 & 2.7332 \\ 4.1740 & 4.8250 & 4.0500 & 2.0190 \\ 8.9620 & 4.0500 & 10.8200 & 3.5820 \\ 2.7332 & 2.0190 & 3.5820 & 1.9162 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 6.0882 & 4.8616 & 0.8014 & 0.5062 \\ 4.8616 & 7.0408 & 0.5732 & 0.4556 \\ 0.8014 & 0.5732 & 1.4778 & 0.2974 \\ 0.5062 & 0.4556 & 0.2974 & 0.5442 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 19.8128 & 4.5944 & 14.8612 & 2.4056 \\ 4.5944 & 5.0962 & 3.4976 & 2.3338 \\ 14.8612 & 3.4976 & 14.9248 & 2.3924 \\ 2.4056 & 2.3338 & 2.3924 & 3.6962 \end{pmatrix}.$$

(a) Test the hypothesis $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ at the 5% significance level.

(b) Test the hypothesis $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ at the 5% significance level.

10.2. (Sec. 10.2)

(a) Let $Y^{(g)}$, $g = 1, \dots, q$, be a set of random vectors each with p components. Suppose

$$\mathscr{E} Y^{(g)} = \mathbf{0}, \qquad \mathscr{E} Y^{(g)} Y^{(h)\prime} = \delta_{gh} \mathbf{\Sigma}_g.$$

Let C be an orthogonal matrix of order q such that each element of the last row is

$$c_{qh} = 1/\sqrt{q}\,.$$

Define

$$Z^{(g)} = \sum_{h=1}^{q} c_{gh} Y^{(h)}, \qquad g = 1, \dots, q.$$

Show that

$$\mathscr{E} Z^{(q)} Z^{(g)\prime} = \mathbf{0}, \qquad g = 1, \dots, q-1,$$

if and only if

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_q.$$

(b) Let $X_\alpha^{(g)}$, $\alpha = 1, \dots, N$, be a random sample from $N(\boldsymbol{\mu}^{(g)}, \mathbf{\Sigma}_g)$, $g = 1, \dots, q$. Use the result from (a) to construct a test of the hypothesis

$$H: \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_q,$$

based on a test of independence of $Z^{(q)}$ and the set $Z^{(1)}, \dots, Z^{(q-1)}$. Find the exact distribution of the criterion for the case $p = 2$.

10.3. (Sec. 10.2) *Unbiasedness of the modified likelihood ratio test of* $\sigma_1^2 = \sigma_2^2$. Show that (14) is unbiased. [*Hint:* Let $G = n_1 F/n_2$, $r = \sigma_1^2/\sigma_2^2$, and $c_1 < c_2$ be the solutions to $G^{\frac{1}{2}n_1}(1+G)^{-\frac{1}{2}(n_1+n_2)} = k$, the critical value for the modified likelihood ratio criterion. Then

$$\Pr\{\text{Acceptance} \mid \sigma_1^2/\sigma_2^2 = r\} = \text{const} \int_{c_1}^{c_2} r^{\frac{1}{2}n_1} G^{\frac{1}{2}n_1 - 1} (1 + rG)^{-\frac{1}{2}(n_1+n_2)}\, dG$$

$$= \text{const} \int_{rc_1}^{rc_2} H^{\frac{1}{2}n_1 - 1} (1 + H)^{-\frac{1}{2}(n_1+n_2)}\, dH.$$

Show that the derivative of the above with respect to r is positive for $0 < r < 1$, 0 for $r = 1$, and negative for $r > 1$.]

10.4. (Sec. 10.2) Prove that the limiting distribution of (19) is χ_f^2, where $f = \frac{1}{2}p(p+1)(q-1)$. [*Hint:* Let $\Sigma = I$. Show that the limiting distribution of (19) is the limiting distribution of

$$\frac{1}{2}\sum_{i=1}^{p}\sum_{g=1}^{q} n_g\left(s_{ii}^{(g)} - s_{ii}\right)^2 + \sum_{i<j}\sum_{g=1}^{q} n_g\left(s_{ij}^{(g)} - s_{ij}\right)^2,$$

where $S^{(g)} = (s_{ij}^{(g)})$, $S = (s_{ij})$, and the $\sqrt{n_g}(s_{ij}^{(g)} - \delta_{ij})$, $i \le j$, are independent in the limiting distribution, the limiting distribution of $\sqrt{n_g}(s_{ii}^{(g)} - 1)$ is $N(0,2)$, and the limiting distribution of $\sqrt{n_g}\, s_{ij}^{(g)}$, $i<j$, is $N(0,1)$.]

10.5. (Sec. 10.4) Prove (15) by integration of Wishart densities. [*Hint:* $\mathscr{E}V_1^h = \mathscr{E}\prod_{g=1}^{q}|A_g|^{\frac{1}{2}hn_g}|A|^{-\frac{1}{2}hn}$ can be written as the integral of a constant times $|A|^{-\frac{1}{2}hn}\prod_{g=1}^{q} w(A_g|\Sigma, n_g + hn_g)$. Integration over $\sum_{g=1}^{q} A_g = A$ gives a constant times $w(A|\Sigma, n)$.]

10.6. (Sec. 10.4) Prove (16) by integration of Wishart and normal densities. [*Hint:* $\sum_{g=1}^{q} N_g(\bar{x}^{(g)} - \bar{x})(\bar{x}^{(g)} - \bar{x})'$ is distributed as $\sum_{f=1}^{q-1} y_f y_f'$. Use the hint of Problem 10.5.]

10.7. (Sec. 10.6) Let $x_1^{(\nu)}, \dots, x_N^{(\nu)}$ be observations from $N(\mu^{(\nu)}, \Sigma_\nu)$, $\nu = 1,2$, and let $A_\nu = \sum(x_\alpha^{(\nu)} - \bar{x}^{(\nu)})(x_\alpha^{(\nu)} - \bar{x}^{(\nu)})'$.

(a) Prove that the likelihood ratio test for $H: \Sigma_1 = \Sigma_2$ is equivalent to rejecting H if

$$T = \frac{|A_1| \cdot |A_2|}{|A_1 + A_2|^2} \le C.$$

(b) Let $d_1^2, d_2^2, \dots, d_p^2$ be the roots of $|\Sigma_1 - \lambda\Sigma_2| = 0$, and let

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix}.$$

Show that T is distributed as $|B_1| \cdot |B_2| / |B_1 + B_2|^2$, where B_1 is distributed according to $W(D^2, N-1)$ and B_2 is distributed according to $W(I, N-1)$. Show that T is distributed as $|DC_1D| \cdot |C_2| / |DC_1D + C_2|^2$, where C_i is distributed according to $W(I, N-1)$.

10.8. (Sec. 10.6) For $p = 2$ show

$$\begin{aligned}\Pr\{V_1 \le v\} &= I_a(n_1 - 1, n_2 - 1) \\ &\quad + B^{-1}(n_1 - 1, n_2 - 1)\, v^{(n_1+n_2-2)/n}\int_a^b x_1^{-2n_2/n}(1 - x_1)^{-n_1/n}\, dx_1 \\ &\quad + 1 - I_b(n_1 - 1, n_2 - 1),\end{aligned}$$

where $a<b$ are the two roots of $x_1^{n_1}(1-x_1)^{n_2} = v \le n_1^{n_1} n_2^{n_2}/n^n$. [*Hint:* This follows from integrating the density defined by (8).]

10.9. (Sec. 10.6) For $p=2$ and $n_1 = n_2 = m$, say, show

$$\Pr\{V_1 \le v\}$$

$$= 2I_a(m-1, m-1) + 2B^{-1}(m-1, m-1)v^{1-(1/m)} \log \frac{1+\sqrt{1-4v^{1/m}}}{1-\sqrt{1-4v^{1/m}}},$$

where $a = \frac{1}{2}[1-\sqrt{1-4v^{1/m}}]$.

10.10. (Sec. 10.7) Find the distribution of W for $p=2$ under the null hypothesis (a) directly from the distribution of A and (b) from the distribution of the characteristic roots (Chapter 13).

10.11. (Sec. 10.7) Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be a sample from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. What is the likelihood ratio criterion for testing the hypothesis $\boldsymbol{\mu} = k\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma} = k^2\boldsymbol{\Sigma}_0$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are specified and k is unspecified?

10.12. (Sec. 10.7) Let $\mathbf{x}_1^{(1)}, \dots, \mathbf{x}_{N_1}^{(1)}$ be a sample from $N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_1)$, and $\mathbf{x}_1^{(2)}, \dots, \mathbf{x}_{N_2}^{(2)}$ be a sample from $N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_2)$. What is the likelihood ratio criterion for testing the hypothesis that $\boldsymbol{\Sigma}_1 = k^2\boldsymbol{\Sigma}_2$, where k is unspecified? What is the likelihood ratio criterion for testing the hypothesis that $\boldsymbol{\mu}^{(1)} = k\boldsymbol{\mu}^{(2)}$ and $\boldsymbol{\Sigma}_1 = k^2\boldsymbol{\Sigma}_2$, where k is unspecified?

10.13. (Sec. 10.7) Let $\mathbf{x}_\alpha$ of p components, $\alpha = 1, \dots, N$, be observations from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We define the following hypotheses:

$$H: \boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\Sigma} = k^2\boldsymbol{\Sigma}_0,$$

$$H_1: \boldsymbol{\Sigma} = k^2\boldsymbol{\Sigma}_0,$$

$$H_2: \boldsymbol{\mu} = \mathbf{0}, \qquad \text{given that} \quad \boldsymbol{\Sigma} = k^2\boldsymbol{\Sigma}_0.$$

In each case k^2 is unspecified, but $\boldsymbol{\Sigma}_0$ is specified. Find the likelihood ratio criterion λ_2 for testing H_2. Give the asymptotic distribution of $-2\log\lambda_2$ under H_2. Obtain the exact distribution of a suitable monotonic function of λ_2 under H_2.

10.14. (Sec. 10.7) Find the likelihood ratio criterion λ for testing H of Problem 10.13 (given $\mathbf{x}_1, \dots, \mathbf{x}_N$). What is the asymptotic distribution of $-2\log\lambda$ under H?

10.15. (Sec. 10.7) Show that $\lambda = \lambda_1\lambda_2$, where λ is defined in Problem 10.14, λ_2 is defined in Problem 10.13, and λ_1 is the likelihood ratio criterion for H_1 in Problem 10.13. Are λ_1 and λ_2 independently distributed under H? Prove your answer.

10.16. (Sec. 10.7) Verify that $\operatorname{tr} B \Psi_0^{-1}$ has the χ^2-distribution with $p(N-1)$ degrees of freedom.

10.17. (Sec. 10.7.1) *Admissibility of sphericity test.* Prove that the likelihood ratio test of sphericity is admissible. [*Hint:* Under the null hypothesis let $\boldsymbol{\Sigma} = [1/(1+\eta^2)]\boldsymbol{I}$, and let η have the density $(1+\eta^2)^{-\frac{1}{2}np}(\eta^2)^{p-\frac{1}{2}}$.]

10.18. (Sec. 10.10.1) Show that for $r \ge p$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left| \sum_{i=1}^{r} \mathbf{x}_i \mathbf{x}_i' \right|^{\frac{1}{2}t} \left| I + \sum_{i=1}^{r} \mathbf{x}_i \mathbf{x}_i' \right|^{-\frac{1}{2}n} \prod_{i=1}^{r} d\mathbf{x}_i < \infty$$

if $2p - 1 < t + r + p < n + 1$. [*Hint:* $|A|/|I+A| \le 1$ if A is positive semidefinite. Also, $|\sum_{i=1}^{r} \mathbf{x}_i \mathbf{x}_i'|$ has the distribution of $\chi_r^2 \chi_{r-1}^2 \cdots \chi_{r-p+1}^2$ if $\mathbf{x}_1, \dots, \mathbf{x}_r$ are independently distributed according to $N(\mathbf{0}, \boldsymbol{I})$.]

10.19. (Sec. 10.10.1) Show

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |CC'|^{\frac{1}{2}t} e^{-\frac{1}{2}\operatorname{tr} C'AC} \, dC = \text{const} |A|^{-\frac{1}{2}(t+r)},$$

where C is $p \times r$. [*Hint:* CC' has the distribution $W(A^{-1}, r)$ if C has a density proportional to $e^{-\frac{1}{2}\operatorname{tr} C'AC}$.]

10.20. (Sec. 10.10.1) Using Problem 10.18, complete the proof of Theorem 10.10.1.

CHAPTER 11

Principal Components

11.1. INTRODUCTION

Principal components are linear combinations of random or statistical variables which have special properties in terms of variances. For example, the first principal component is the normalized linear combination (the sum of squares of the coefficients being one) with maximum variance. In effect, transforming the original vector variable to the vector of principal components amounts to a rotation of coordinate axes to a new coordinate system that has inherent statistical properties. This choosing of a coordinate system is to be contrasted with the many problems treated previously where the coordinate system is irrelevant.

The principal components turn out to be the characteristic vectors of the covariance matrix. Thus the study of principal components can be considered as putting into statistical terms the usual developments of characteristic roots and vectors (for positive semidefinite matrices).

From the point of view of statistical theory, the set of principal components yields a convenient set of coordinates, and the accompanying variances of the components characterize their statistical properties. In statistical practice, the method of principal components is used to find the linear combinations with large variance. In many exploratory studies the number of variables under consideration is too large to handle. Since it is the deviations in these studies that are of interest, a way of reducing the number of variables to be treated is to discard the linear combinations which have small variances and study only those with large variances. For example, a physical anthropologist may make dozens of measurements of lengths and breadths of

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

each of a number of individuals, such measurements as ear length, ear breadth, facial length, facial breadth, and so forth. He may be interested in describing and analyzing how individuals differ in these kinds of physiological characteristics. Eventually he will want to *explain* these differences, but first he wants to know what measurements or combinations of measurements show considerable variation; that is, which should have further study. The principal components give a new set of linearly combined measurements. It may be that most of the variation from individual to individual resides in three linear combinations; then the anthropologist can direct his study to these three quantities; the other linear combinations vary so little from one person to the next that study of them will tell little of individual variation.

Hotelling (1933), who developed many of these ideas, gave a rather thorough discussion.

In Section 11.2 we define principal components in the population to have the properties described above; they define an orthogonal transformation to a diagonal covariance matrix. The maximum likelihood estimators have similar properties in the sample (Section 11.3). A brief discussion of computation is given in Section 11.4, and a numerical example is carried out in Section 11.5. Asymptotic distributions of the coefficients of the sample principal components and the sample variances are derived and applied to obtain large-sample tests and confidence intervals for individual parameters (Section 11.6); exact confidence bounds are found for the characteristic roots of a covariance matrix. In Section 11.7 we consider other tests of hypotheses about these roots.

11.2. DEFINITION OF PRINCIPAL COMPONENTS IN THE POPULATION

Suppose the random vector X of p components has the covariance matrix Σ. Since we shall be interested only in variances and covariances in this chapter, we shall assume that the mean vector is 0. Moreover, in developing the ideas and algebra here, the actual distribution of X is irrelevant except for the covariance matrix; however, if X is normally distributed, more meaning can be given to the principal components.

In the following treatment we shall not use the usual theory of characteristic roots and vectors; as a matter of fact, that theory will be derived implicitly. The treatment will include the cases where Σ is singular (i.e., positive semidefinite) and where Σ has multiple roots.

Let β be a p-component column vector such that $\beta'\beta = 1$. The variance of $\beta' X$ is

$$\mathscr{E}(\beta'X)^2 = \mathscr{E}\beta'XX'\beta = \beta'\Sigma\beta. \tag{1}$$

To determine the normalized linear combination $\boldsymbol{\beta}'X$ with maximum variance, we must find a vector $\boldsymbol{\beta}$ satisfying $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$ which maximizes (1). Let

$$(2) \qquad \phi = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} - \lambda(\boldsymbol{\beta}'\boldsymbol{\beta} - 1) = \sum_{i,j} \beta_i \sigma_{ij} \beta_j - \lambda\Big(\sum_i \beta_i^2 - 1\Big),$$

where λ is a Lagrange multiplier. The vector of partial derivatives $(\partial\phi/\partial\beta_i)$ is

$$(3) \qquad \frac{\partial\phi}{\partial\boldsymbol{\beta}} = 2\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}$$

(by Theorem A.4.3 of the Appendix). Since $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}$ and $\boldsymbol{\beta}'\boldsymbol{\beta}$ have derivatives everywhere in a region containing $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$, a vector $\boldsymbol{\beta}$ maximizing $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}$ must satisfy the expression (3) set equal to $\mathbf{0}$; that is

$$(4) \qquad (\boldsymbol{\Sigma} - \lambda I)\boldsymbol{\beta} = \mathbf{0}.$$

In order to get a solution of (4) with $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$ we must have $\boldsymbol{\Sigma} - \lambda I$ singular; in other words, λ must satisfy

$$(5) \qquad |\boldsymbol{\Sigma} - \lambda I| = 0.$$

The function $|\boldsymbol{\Sigma} - \lambda I|$ is a polynomial in λ of degree p. Therefore (5) has p roots; let these be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. [$\boldsymbol{\beta}'$ complex conjugate in (6) proves λ real.] If we multiply (4) on the left by $\boldsymbol{\beta}'$, we obtain

$$(6) \qquad \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} = \lambda\boldsymbol{\beta}'\boldsymbol{\beta} = \lambda.$$

This shows that if $\boldsymbol{\beta}$ satisfies (4) (and $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$), then the variance of $\boldsymbol{\beta}'X$ [given by (1)] is λ. Thus for the maximum variance we should use in (4) the largest root λ_1. Let $\boldsymbol{\beta}^{(1)}$ be a normalized solution of $(\boldsymbol{\Sigma} - \lambda_1 I)\boldsymbol{\beta} = \mathbf{0}$. Then $U_1 = \boldsymbol{\beta}^{(1)\prime}X$ is a normalized linear combination with maximum variance. [If $\boldsymbol{\Sigma} - \lambda_1 I$ is of rank $p-1$, then there is only one solution to $(\boldsymbol{\Sigma} - \lambda_1 I)\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$.]

Now let us find a normalized combination $\boldsymbol{\beta}'X$ that has maximum variance of all linear combinations uncorrelated with U_1. Lack of correlation means

$$(7) \qquad 0 = \mathscr{E}\boldsymbol{\beta}'XU_1 = \mathscr{E}\boldsymbol{\beta}'XX'\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)} = \lambda_1\boldsymbol{\beta}'\boldsymbol{\beta}^{(1)}$$

since $\boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)} = \lambda_1\boldsymbol{\beta}^{(1)}$. Thus $\boldsymbol{\beta}\ X$ is orthogonal to U in both the statistical sense (of lack of correlation) and the geometric sense (of the inner product of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{(1)}$ being zero). (That is, $\lambda_1\boldsymbol{\beta}'\boldsymbol{\beta}^{(1)} = 0$ only if $\boldsymbol{\beta}'\boldsymbol{\beta}^{(1)} = 0$ when $\lambda_1 \neq 0$, and $\lambda_1 \neq 0$ if $\boldsymbol{\Sigma} \neq \mathbf{0}$; the case of $\boldsymbol{\Sigma} = \mathbf{0}$ is trivial and is not treated.)

We now want to maximize

$$\phi_2 = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} - \lambda(\boldsymbol{\beta}'\boldsymbol{\beta} - 1) - 2\nu_1\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)}, \tag{8}$$

where λ and ν_1 are Lagrange multipliers. The vector of partial derivatives is

$$\frac{\partial\phi_2}{\partial\boldsymbol{\beta}} = 2\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta} - 2\nu_1\boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)}, \tag{9}$$

and we set this equal to $\mathbf{0}$. From (9) we obtain by multiplying on the left by $\boldsymbol{\beta}^{(1)\prime}$

$$0 = 2\boldsymbol{\beta}^{(1)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}^{(1)\prime}\boldsymbol{\beta} - 2\nu_1\boldsymbol{\beta}^{(1)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta}^{(1)} = -2\nu_1\lambda_1, \tag{10}$$

by (7). Therefore, $\nu_1 = 0$ and $\boldsymbol{\beta}$ must satisfy (4), and therefore λ must satisfy (5). Let $\lambda_{(2)}$ be the maximum of $\lambda_1, \ldots, \lambda_p$ such that there is a vector $\boldsymbol{\beta}$ satisfying $(\boldsymbol{\Sigma} - \lambda_{(2)}I)\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$, and (7); call this vector $\boldsymbol{\beta}^{(2)}$ and the corresponding linear combination $U_2 = \boldsymbol{\beta}^{(2)\prime}X$. (It will be shown eventually that $\lambda_{(2)} = \lambda_2$. We define $\lambda_{(1)} = \lambda_1$.)

This procedure is continued; at the $(r+1)$st step, we want to find a vector $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}'X$ has maximum variance of all normalized linear combinations which are uncorrelated with $U_1, \ldots, U_r$, that is, such that

$$0 = \mathscr{E}\boldsymbol{\beta}'XU_i = \mathscr{E}\boldsymbol{\beta}'XX'\boldsymbol{\beta}^{(i)} = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^{(i)} = \lambda_{(i)}\boldsymbol{\beta}'\boldsymbol{\beta}^{(i)}, \qquad i = 1, \ldots, r. \tag{11}$$

We want to maximize

$$\phi_{r+1} = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} - \lambda(\boldsymbol{\beta}'\boldsymbol{\beta} - 1) - 2\sum_{i=1}^{r}\nu_i\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}^{(i)}, \tag{12}$$

where λ and $\nu_1, \ldots, \nu_r$ are Lagrange multipliers. The vector of partial derivatives is

$$\frac{\partial\phi_{r+1}}{\partial\boldsymbol{\beta}} = 2\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta} - 2\sum_{i=1}^{r}\nu_i\boldsymbol{\Sigma}\boldsymbol{\beta}^{(i)}, \tag{13}$$

and we set this equal to $\mathbf{0}$. Multiplying (13) on the left by $\boldsymbol{\beta}^{(j)\prime}$, we obtain

$$0 = 2\boldsymbol{\beta}^{(j)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta}^{(j)\prime}\boldsymbol{\beta} - 2\nu_j\boldsymbol{\beta}^{(j)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta}^{(j)}. \tag{14}$$

If $\lambda_{(j)} \neq 0$, this gives $-2\nu_j\lambda_{(j)} = 0$ and $\nu_j = 0$. If $\lambda_{(j)} = 0$, then $\boldsymbol{\Sigma}\boldsymbol{\beta}^{(j)} = \lambda_{(j)}\boldsymbol{\beta}^{(j)} = \mathbf{0}$ and the jth term in the sum in (13) vanishes Thus $\boldsymbol{\beta}$ must satisfy (4), and therefore λ must satisfy (5).

Let $\lambda_{(r+1)}$ be the maximum of $\lambda_1, \ldots, \lambda_p$ such that there is a vector $\boldsymbol{\beta}$ satisfying $(\boldsymbol{\Sigma} - \lambda_{(r+1)} \boldsymbol{I})\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\beta}'\boldsymbol{\beta}' = 1$, and (11); call this vector $\boldsymbol{\beta}^{(r+1)}$, and the corresponding linear combination $U_{r+1} = \boldsymbol{\beta}^{(r+1)\prime}\boldsymbol{X}$. If $\lambda_{(r+1)} = 0$ and $\lambda_{(j)} = 0$, $j \neq r+1$, then $\boldsymbol{\beta}^{(j)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta}^{(r+1)} = 0$ does not imply $\boldsymbol{\beta}^{(j)\prime}\boldsymbol{\beta}^{(r+1)} = 0$. However, $\boldsymbol{\beta}^{(r+1)}$ can be replaced by a linear combination of $\boldsymbol{\beta}^{(r+1)}$ and the $\boldsymbol{\beta}^{(j)}$'s with $\lambda_{(j)}$'s being 0, so that the new $\boldsymbol{\beta}^{(r+1)}$ is orthogonal to all $\boldsymbol{\beta}^{(j)}$, $j = 1, \ldots, r$. This procedure is carried on until at the $(m+1)$st stage one cannot find a vector $\boldsymbol{\beta}$ satisfying $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$, (4), and (11). Either $m = p$ or $m < p$ since $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}$ must be linearly independent.

We shall now show that the inequality $m < p$ leads to a contradiction. If $m < p$ there exist $p - m$ vectors, say $\boldsymbol{e}_{m+1}, \ldots, \boldsymbol{e}_p$, such that $\boldsymbol{\beta}^{(i)\prime}\boldsymbol{e}_j = 0$, $\boldsymbol{e}_i'\boldsymbol{e}_j = \delta_{ij}$. (This follows from Lemma A.4.2 in the Appendix.) Let $(\boldsymbol{e}_{m+1}, \ldots, \boldsymbol{e}_p) = \boldsymbol{E}$. Now we shall show that there exists a $(p-m)$-component vector $\boldsymbol{c}$ and a number θ such that $\boldsymbol{E}\boldsymbol{c} = \sum c_i \boldsymbol{e}_i$ is a solution to (4) with $\lambda = \theta$. Consider a root of $|\boldsymbol{E}'\boldsymbol{\Sigma}\boldsymbol{E} - \theta \boldsymbol{I}| = 0$ and a corresponding vector $\boldsymbol{c}$ satisfying $\boldsymbol{E}'\boldsymbol{\Sigma}\boldsymbol{E}\boldsymbol{c} = \theta\boldsymbol{c}$. The vector $\boldsymbol{\Sigma}\boldsymbol{E}\boldsymbol{c}$ is orthogonal to $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}$ (since $\boldsymbol{\beta}^{(i)\prime}\boldsymbol{\Sigma}\boldsymbol{E}\boldsymbol{c} = \lambda_{(i)}\boldsymbol{\beta}^{(i)\prime}\sum c_j \boldsymbol{e}_j = \lambda_{(i)}\sum c_j \boldsymbol{\beta}^{(i)\prime}\boldsymbol{e}_j = 0$) and therefore is a vector in the space spanned by $\boldsymbol{e}_{m+1}, \ldots, \boldsymbol{e}_p$ and can be written as $\boldsymbol{E}\boldsymbol{g}$ [where $\boldsymbol{g}$ is a $(p-m)$-component vector]. Multiplying $\boldsymbol{\Sigma}\boldsymbol{E}\boldsymbol{c} = \boldsymbol{E}\boldsymbol{g}$ on the left by $\boldsymbol{E}'$, we obtain $\boldsymbol{E}'\boldsymbol{\Sigma}\boldsymbol{E}\boldsymbol{c} = \boldsymbol{E}'\boldsymbol{E}\boldsymbol{g} = \boldsymbol{g}$. Thus $\boldsymbol{g} = \theta\boldsymbol{c}$, and we have $\boldsymbol{\Sigma}(\boldsymbol{E}\boldsymbol{c}) = \theta(\boldsymbol{E}\boldsymbol{c})$. Then $(\boldsymbol{E}\boldsymbol{c})'\boldsymbol{X}$ is uncorrelated with $\boldsymbol{\beta}^{(j)\prime}\boldsymbol{X}$, $j = 1, \ldots, m$, and thus leads to a new $\boldsymbol{\beta}^{(m+1)}$. Since this contradicts the assumption that $m < p$, we must have $m = p$.

Let $\mathbf{B} = (\boldsymbol{\beta}^{(1)} \ \cdots \ \boldsymbol{\beta}^{(p)})$ and

$$\Lambda = \begin{pmatrix} \lambda_{(1)} & 0 & \cdots & 0 \\ 0 & \lambda_{(2)} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_{(p)} \end{pmatrix}, \tag{15}$$

The equations $\boldsymbol{\Sigma}\boldsymbol{\beta}^{(r)} = \lambda_{(r)}\boldsymbol{\beta}^{(r)}$ can be written in matrix form as

$$\boldsymbol{\Sigma}\mathbf{B} = \mathbf{B}\Lambda, \tag{16}$$

and the equations $\boldsymbol{\beta}^{(r)\prime}\boldsymbol{\beta}^{(r)} = 1$ and $\boldsymbol{\beta}^{(r)\prime}\boldsymbol{\beta}^{(s)} = 0$, $r \neq s$, can be written as

$$\mathbf{B}'\mathbf{B} = \boldsymbol{I}. \tag{17}$$

From (16) and (17) we obtain

$$\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B} = \Lambda. \tag{18}$$

From the fact that

$$(19)\qquad \begin{aligned} |\Sigma - \lambda I| &= |\beta'|\cdot|\Sigma - \lambda I|\cdot|\beta| \\ &= |\beta'\Sigma\beta - \lambda\beta'\beta| = |\Lambda - \lambda I| \\ &= \Pi(\lambda_{(i)} - \lambda) \end{aligned}$$

we see that the roots of (19) are the diagonal elements of Λ; that is, $\lambda_{(1)} = \lambda_1, \lambda_{(2)} = \lambda_2, \ldots, \lambda_{(p)} = \lambda_p$.

We have proved the following theorem:

Theorem 11.2.1. *Let the p-component random vector X have $\mathscr{E}X = 0$ and $\mathscr{E}XX' = \Sigma$. Then there exists an orthogonal linear transformation*

$$(20)\qquad U = \beta' X$$

such that the covariance matrix of U is $\mathscr{E}UU' = \Lambda$ and

$$(21)\qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix},$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$ are the roots of (5). *The* rth *column of β, $\beta^{(r)}$, satisfies $(\Sigma - \lambda_r I)\beta^{(r)} = 0$. The* rth *component of U, $U_r = \beta^{(r)\prime}X$, has maximum variance of all normalized linear combinations uncorrelated with $U_1, \ldots, U_{r-1}$.*

The vector U is defined as the vector of principal components of X. It will be observed that we have proved Theorem A.2.1 of Appendix A for B positive semidefinite, and indeed, the proof holds for any symmetric B. It might be noted that once the transformation to $U_1, \ldots, U_p$ has been made, it is obvious that U_1 is the normalized linear combination with maximum variance, for if $U^* = \Sigma c_i U_i$, where $\Sigma c_i^2 = 1$ (U^* also being a normalized linear combination of the X's), then $\operatorname{Var}(U^*) = \Sigma c_i^2 \lambda_i = \lambda_1 + \Sigma_{i=2}^p c_i^2(\lambda_i - \lambda_1)$ (since $c_1^2 = 1 - \Sigma_2^p c_i^2$), which is clearly maximum for $c_i^2 = 0$, $i = 2, \ldots, p$. Similarly, U_2 is the normalized linear combination uncorrelated with U_1 which has maximum variance ($U^* = \Sigma c_i U_i$ being uncorrelated with U_1 implying $c_1 = 0$); in turn the maximal properties of $U_3, \ldots, U_p$ are verified.

Some other consequences can be derived.

Corollary 11.2.1. *Suppose $\lambda_{r+1} = \cdots = \lambda_{r+m} = \nu$ (i.e., ν is a root of multiplicity m); then $\Sigma - \nu I$ is of rank $p - m$. Furthermore $\beta^* = (\beta^{(r+1)} \cdots \beta^{(r+m)})$ is uniquely determined except for multiplication on the right by an orthogonal matrix.*

Proof. From the derivation of the theorem we have $(\boldsymbol{\Sigma} - \nu \boldsymbol{I})\boldsymbol{\beta}^{(i)} = \boldsymbol{0}$, $i = r+1, \dots, r+m$; that is, $\boldsymbol{\beta}^{(r+1)}, \dots, \boldsymbol{b}^{(r+m)}$ are m linearly independent solutions of $(\boldsymbol{\Sigma} - \nu \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{0}$. To show that there cannot be another linearly independent solution, take $\sum_{i=1}^{p} x_i \boldsymbol{\beta}^{(i)}$, where the x_i are scalars. If it is a solution, we have $\nu \sum x_i \boldsymbol{\beta}^{(i)} = \boldsymbol{\Sigma}(\sum x_i \boldsymbol{\beta}^{(i)}) = \sum x_i \boldsymbol{\Sigma} \boldsymbol{\beta}^{(i)} = \sum x_i \lambda_i \boldsymbol{\beta}^{(i)}$. Since $\nu x_i = \lambda_i x_i$, we must have $x_i = 0$ unless $i = r+1, \dots, r+m$. Thus the rank is $p - m$.

If $\boldsymbol{\beta}^*$ is one set of solutions to $(\boldsymbol{\Sigma} - \nu \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{0}$, then any other set of solutions are linear combinations of the others, that is, are $\boldsymbol{\beta}^* A$ for A nonsingular. However, the orthogonality conditions $\boldsymbol{\beta}^{*\prime}\boldsymbol{\beta}^* = I$ applied to the linear combinations give $I = (\boldsymbol{\beta}^* A)'(\boldsymbol{\beta}^* A) = A'\boldsymbol{\beta}^{*\prime}\boldsymbol{\beta}^* A = A'A$, and thus A must be orthogonal. ∎

Theorem 11.2.2. *An orthogonal transformation $\boldsymbol{V} = \boldsymbol{C}\boldsymbol{X}$ of a random vector $\boldsymbol{X}$ leaves invariant the generalized variance and the sum of the variances of the components.*

Proof. Let $\mathscr{E}\boldsymbol{X} = \boldsymbol{0}$ and $\mathscr{E}\boldsymbol{X}\boldsymbol{X}' = \boldsymbol{\Sigma}$. Then $\mathscr{E}\boldsymbol{V} = \boldsymbol{0}$ and $\mathscr{E}\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'$. The generalized variance of $\boldsymbol{V}$ is

$$|\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'| = |\boldsymbol{C}| \cdot |\boldsymbol{\Sigma}| \cdot |\boldsymbol{C}'| = |\boldsymbol{\Sigma}| \cdot |\boldsymbol{C}\boldsymbol{C}'| = |\boldsymbol{\Sigma}|, \tag{22}$$

which is the generalized variance of $\boldsymbol{X}$. The sum of the variances of the components of $\boldsymbol{V}$ is

$$\sum \mathscr{E} V_i^2 = \operatorname{tr}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}') = \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{C}'\boldsymbol{C}) = \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{I}) = \operatorname{tr}\boldsymbol{\Sigma} = \sum \mathscr{E} X_i^2. \tag{23}$$ ∎

Corollary 11.2.2. *The generalized variance of the vector of principal components is the generalized variance of the original vector, and the sum of the variances of the principal components is the sum of the variances of the original variates.*

Another approach to the above theory can be based on the surfaces of constant density of the normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}$ (nonsingular). The density is

$$\frac{1}{(2\pi)^{\frac{1}{2}p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}, \tag{24}$$

and surfaces of constant density are ellipsoids

$$\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = C. \tag{25}$$

A principal axis of this ellipsoid is defined as the line from $-\boldsymbol{y}$ to $\boldsymbol{y}$, where $\boldsymbol{y}$ is a point on the ellipsoid where the squared distance $\boldsymbol{x}'\boldsymbol{x}$ has a stationary point. Using the method of Lagrange multipliers, we determine the stationary points by considering

$$\psi = x'x - \lambda x'\Sigma^{-1}x, \tag{26}$$

where λ is a Lagrange multiplier. We differentiate ψ with respect to the components of x, and the derivatives set equal to 0 are

$$\frac{\partial\psi}{\partial x} = 2x - 2\lambda\Sigma^{-1}x = \mathbf{0}, \tag{27}$$

or

$$x = \lambda\Sigma^{-1}x. \tag{28}$$

Multiplication by Σ gives

$$\Sigma x = \lambda x. \tag{29}$$

This equation is the same as (4) and the same algebra can be developed. Thus the vectors $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(p)}$ give the principal axis of the ellipsoid. The transformation $u = \boldsymbol{\beta}'x$ is a rotation of the coordinate axes so that the new axes are in the direction of the principal axes of the ellipsoid. In the new coordinates the ellipsoid is

$$u'\Lambda^{-1}u = \sum \frac{u_i^2}{\lambda_i} = C. \tag{30}$$

Thus the length of the ith principal axis is $2\sqrt{\lambda_i C}$.

A third approach to the same results is in terms of *planes of closest fit* [Pearson (1901)]. Consider a plane through the origin, $\boldsymbol{\alpha}'x = 0$, where $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$. The distance of a point x from this plane is $\boldsymbol{\alpha}'x$. Let us find the coefficients of a plane such that the expected distance squared of a random point X from the plane is a minimum, where $\mathscr{E}X = \mathbf{0}$ and $\mathscr{E}XX' = \Sigma$. Thus we wish to minimize $\mathscr{E}(\boldsymbol{\alpha}'X)^2 = \mathscr{E}\boldsymbol{\alpha}'XX'\boldsymbol{\alpha} = \boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}$, subject to the restriction $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$. Comparison with the first approach immediately shows that the solution is $\boldsymbol{\alpha} = \boldsymbol{\beta}^{(p)}$.

Analysis into principal components is most suitable when all the components of X are measured in the same units. If they are not measured in the same units, the rationale of maximizing $\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}$ relative to $\boldsymbol{\beta}'\boldsymbol{\beta}$ is questionable; in fact, the analysis will depend on the various units of measurement. Suppose Δ is a diagonal matrix, and let $Y = \Delta X$. For example, one component of X may be measured in inches and the corresponding component of Y may be measured in feet; another component of X may be in pounds and the corresponding one of Y in ounces. The covariance matrix of Y is $\mathscr{E}YY' = \mathscr{E}\Delta XX'\Delta = \Delta\Sigma\Delta = \Psi$, say. Then analysis of Y into principal components involves maximizing $\mathscr{E}(\gamma'Y)^2 = \gamma'\Psi\gamma$ relative to $\gamma'\gamma$ and leads to the equation $0 = (\Psi - \nu I)\gamma = (\Delta\Sigma\Delta - \nu I)\gamma$, where ν must satisfy $|\Psi - \nu I| = 0$. Multiplication on the left by Δ^{-1} gives

$$0 = (\Sigma - \nu\Delta^{-2})(\Delta\gamma). \tag{31}$$

Let $\Delta\gamma = \alpha$; that is, $\gamma'Y = \gamma'\Delta X = \alpha'X$. Then (31) results from maximizing $\mathscr{E}(\alpha'X)^2 = \alpha'\Sigma\alpha$ relative to $\alpha'\Delta^{-2}\alpha$. This last quadratic form is a weighted sum of squares, the weights being the diagonal elements of Δ^{-2}.

It might be noted that if Δ^{-2} is taken to be the matrix

$$\Delta^{-2} = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{pmatrix}, \tag{32}$$

then Ψ is the matrix of correlations.

11.3. MAXIMUM LIKELIHOOD ESTIMATORS OF THE PRINCIPAL COMPONENTS AND THEIR VARIANCES

A primary problem of statistical inference in principal component analysis is to estimate the vectors $\beta^{(1)}, \dots, \beta^{(p)}$ and the scalars $\lambda_1, \dots, \lambda_p$. We apply the algebra of the preceding section to an estimate of the covariance matrix.

Theorem 11.3.1. *Let $x_1, \dots, x_N$ be N $(>p)$ observations from $N(\mu, \Sigma)$, where Σ is a matrix with p different characteristic roots. Then a set of maximum likelihood estimators of $\lambda_1, \dots, \lambda_p$ and $\beta^{(1)}, \dots, \beta^{(p)}$ defined in Theorem* 11.2.1 *consists of the roots $k_1 > \cdots > k_p$ of*

$$|\hat{\Sigma} - kI| = 0 \tag{1}$$

and a set of corresponding vectors $b^{(1)}, \dots, b^{(p)}$ satisfying

$$\left(\hat{\Sigma} - k_i I\right)b^{(i)} = 0, \tag{2}$$

$$b^{(i)\prime}b^{(i)} = 1, \tag{3}$$

where $\hat{\Sigma}$ is the maximum likelihood estimate of Σ.

Proof. When the roots of $|\Sigma - \lambda I| = 0$ are different, each vector $\beta^{(i)}$ is uniquely defined except that $\beta^{(i)}$ can be replaced by $-\beta^{(i)}$. If we require that the first nonzero component of $\beta^{(i)}$ be positive, then $\beta^{(i)}$ is uniquely defined, and μ, Λ, β is a single-valued function of μ, Σ. By Corollary 3.2.1, the set of maximum likelihood estimates of μ, Λ, β is the same function of $\hat{\mu}, \hat{\Sigma}$. This function is defined by (1), (2), and (3) with the corresponding restriction that the first nonzero component of $b^{(i)}$ must be positive. [It can be shown that if $|\Sigma| \neq 0$, the probability is 1 that the roots of (1) are different, because the conditions on $\hat{\Sigma}$ for the roots to have multiplicities higher than 1 determine a region in the space of $\hat{\Sigma}$ of dimensionality less than $\frac{1}{2}p(p+1)$; see Okamoto (1973).] From (18) of Section 11.2 we see that

$$\Sigma = \beta\Lambda\beta' = \sum \lambda_i \beta^{(i)}\beta^{(i)\prime}, \tag{4}$$

and by the same algebra

$$\hat{\Sigma} = \sum k_i b^{(i)} b^{(i)\prime}. \tag{5}$$

Replacing $b^{(i)}$ by $-b^{(i)}$ clearly does not change $\sum k_i b^{(i)} b^{(i)\prime}$. Since the likelihood function depends only on $\hat{\Sigma}$ (see Section 3.2), the maximum of the likelihood function is attained by taking any set of solutions of (2) and (3). ∎

It is possible to assume explicitly arbitrary multiplicities of roots of Σ. If these multiplicities are not all unity, the maximum likelihood estimates are not defined as in Theorem 11.3.1. [See Anderson (1963a).] As an example suppose that we assume that the equation $|\Sigma - \lambda I| = 0$ has one root of multiplicity p. Let this root be λ_1. Then by Corollary 11.2.1, $\Sigma - \lambda_1 I$ is of rank 0; that is, $\Sigma - \lambda_1 I = 0$ or $\Sigma = \lambda_1 I$. If X is distributed according to $N(\mu, \Sigma) = N(\mu, \lambda_1 I)$, the components of X are independently distributed with variance λ_1. Thus the maximum likelihood estimator of λ_1 is

$$\hat{\lambda}_1 = \frac{1}{pN} \sum_{i=1}^{p} \sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)^2, \tag{6}$$

and $\hat{\Sigma} = \hat{\lambda}_1 I$, and $\hat{\beta}$ can be any orthogonal matrix. It might be pointed out that in Section 10.7 we considered a test of the hypothesis that $\Sigma = \lambda_1 I$ (with λ_1 unspecified), that is, the hypothesis is that Σ has one characteristic root of multiplicity p.

In most applications of principal component analysis it can be assumed that the roots of Σ are different. It might also be pointed out that in some uses of this method the algebra is applied to the matrix of correlation coefficients rather than to the covariance matrix. In general this leads to different roots and vectors.

11.4. COMPUTATION OF THE MAXIMUM LIKELIHOOD ESTIMATES OF THE PRINCIPAL COMPONENTS

There are several ways of computing the characteristic roots and characteristic vectors (principal components) of a matrix Σ or $\hat{\Sigma}$. We shall indicate some of them.

One method for small p involves expanding the determinantal equation

$$0 = |\Sigma - \lambda I| \tag{1}$$

and solving the resulting pth-degree equation in λ (e.g., by Newton's method or the secant method) for the roots $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. Then $\Sigma - \lambda_i I$ is of rank $p-1$, and a solution of $(\Sigma - \lambda_i I)\boldsymbol{\beta}^{(i)} = \mathbf{0}$ can be obtained by taking $\beta_j^{(i)}$ as the cofactor of the element in the first (or any other fixed) column and jth row of $\Sigma - \lambda_i I$.

The second method iterates using the equation for a characteristic root and the corresponding characteristic vector

$$\Sigma \boldsymbol{x} = \lambda \boldsymbol{x}, \tag{2}$$

where we have written the equation for the population. Let $\boldsymbol{x}_{(0)}$ be any vector not orthogonal to the first characteristic vector, and define

$$\boldsymbol{x}_{(i)} = \Sigma \boldsymbol{y}_{(i-1)}, \qquad \boldsymbol{y}_{(i)} = \frac{1}{\sqrt{\boldsymbol{x}'_{(i)}\boldsymbol{x}_{(i)}}}\boldsymbol{x}_{(i)}, \qquad i = 0, 1, 2, \ldots. \tag{3}$$

It can be shown (Problem 11.12) that

$$\lim_{i\to\infty} \boldsymbol{y}_{(i)} = \pm \boldsymbol{\beta}^{(1)}, \qquad \lim_{i\to\infty} \boldsymbol{x}'_{(i)}\boldsymbol{x}_{(i)} = \lambda_1^2. \tag{4}$$

The rate of convergence depends on the ratio λ_2/λ_1; the closer this ratio is to 1, the slower the convergence.

To find the second root and vector define

$$\Sigma_2 = \Sigma - \lambda_1 \boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime}. \tag{5}$$

Then

$$\begin{aligned} \Sigma_2 \boldsymbol{\beta}^{(i)} &= \Sigma \boldsymbol{\beta}^{(i)} - \lambda_1 \boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime}\boldsymbol{\beta}^{(i)} \\ &= \Sigma \boldsymbol{\beta}^{(i)} = \lambda_i \boldsymbol{\beta}^{(i)} \end{aligned} \tag{6}$$

if $i \neq 1$, and

$$\Sigma_2 \beta^{(1)} = 0. \tag{7}$$

Thus λ_2 is the largest root of Σ_2 and $\beta^{(2)}$ is the corresponding vector. The iteration process is now applied to Σ_2 to find λ_2 and $\beta^{(2)}$. Defining $\Sigma_3 = \Sigma_2 - \lambda_2 \beta^{(2)} \beta^{(2)\prime}$, we can find λ_3 and $\beta^{(3)}$, and so forth.

There are several ways in which the labor of the iteration procedure may be reduced. One is to raise Σ to a power before proceeding with the iteration. Thus one can use Σ^2, defining

$$x_{(i)} = \Sigma^2 y_{(i-1)}, \qquad y_{(i)} = \frac{x_{(i)}}{\sqrt{x'_{(i)} x_{(i)}}}, \qquad i = 0, 1, 2, \dots. \tag{8}$$

This procedure will give twice as rapid convergence as the use of (3). Using $\Sigma^4 = \Sigma^2 \Sigma^2$ will lead to convergence four times as rapid, and so on. It should be noted that since Σ^2 is symmetric, there are only $p(p+1)/2$ elements to be found.

Efficient computation, however, uses other methods. One method is the QR or QL algorithm. Let $\Sigma_0 = \Sigma$. Define recursively the orthogonal Q_i and lower triangular L_i by $\Sigma_i = Q_i L_i$ and $\Sigma_{i+1} = L_i Q_i$ $(= Q_i' \Sigma_i Q_i)$, $i = 1, 2, \dots$. (The Gram–Schmidt orthogonalization is a way of finding Q_i and L_i; the QR method replaces a lower triangular matrix L by an upper triangular matrix R.) If the characteristic roots of Σ are distinct, $\lim_{i \to \infty} \Sigma_{i+1} = \Lambda^*$, where Λ^* is the diagonal matrix with the roots usually ordered in ascending order. The characteristic vectors are the columns of $\lim_{i \to \infty} Q_i' Q_{i-1}' \cdots Q_1'$ (which is computed recursively).

A more efficient algorithm (for the symmetric Σ) uses a sequence of Householder transformations to carry Σ to tridiagonal form. A *Householder matrix* is $H = I - 2\alpha\alpha'$ where $\alpha'\alpha = 1$. Such a matrix is orthogonal and symmetric. A Householder transformation of the symmetric matrix Σ is $H\Sigma H$. It is symmetric and has the same characteristic roots as Σ; its characteristic vectors are H times those of Σ.

A *tridiagonal matrix* is one with all entries 0 except on the main diagonal, the first superdiagonal, and the first subdiagonal. A sequence of $p-2$ Householder transformations carries the symmetric Σ to tridiagonal form. (The first one inserts 0's into the last $p-2$ entries of the first column and row of $H\Sigma H$, etc. See Problem 11.13.)

The QL method is applied to the tridiagonal form. At the ith step let the tridiagonal matrix be $T_0^{(i)}$; let $P_j^{(i)}$ be a block-diagonal matrix (Givens matrix)

$$P_j^{(i)} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & \cos\theta_j & -\sin\theta_j & 0 \\ 0 & \sin\theta_j & \cos\theta_j & 0 \\ 0 & 0 & 0 & I \end{bmatrix}, \tag{9}$$

where $\cos\theta_j$ is the jth and $j+1$st diagonal element; and let $T_j^{(i)} = P_{p-j}^{(i)} T_{j-1}^{(i)}$, $j = 1, \dots, p-1$. Here θ_j is chosen so that the element in position $j, j+1$ in T_j is 0. Then $P^{(i)} = P_1^{(i)} P_2^{(i)} \cdots P_{p-1}^{(i)}$ is orthogonal and $P^{(i)} T_0^{(i)} = R^{(i)}$ is lower triangular. Then $T_0^{(i+1)} = R^{(i)} P^{(i)\prime}$ $(= P^{(i)} T_0^{(i)} P^{(i)\prime})$ is symmetric and tridiagonal. It converges to Λ^* (if the roots are all different). For more details see Chapters II/2 and II/3 of Wilkinson and Reinsch (1971), Chapter 5 of Wilkinson (1965), and Chapters 5, 7, and 8 of Golub and Van Loan (1989). A sequence of one-sided Householder transformation ($H\ \Sigma$) can carry Σ to R (upper triangular), thus effecting the QR decomposition.

11.5. AN EXAMPLE

In Table 3.4 we presented three samples of observations on varieties of iris [Fisher (1936)]; as an example of principal component analysis we use one of those samples, namely *Iris versicolor*. There are 50 observations ($N = 50$, $n = N - 1 = 49$). Each observation consists of four measurements on a plant: x_1 is sepal length, x_2 is sepal width, x_3 is petal length, and x_4 is petal width. The observed sums of squares and cross products of deviations from means are

$$A = \sum_{\alpha=1}^{50} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' = \begin{pmatrix} 13.0552 & 4.1740 & 8.9620 & 2.7332 \\ 4.1740 & 4.8250 & 4.0500 & 2.0190 \\ 8.9620 & 4.0500 & 10.8200 & 3.5820 \\ 2.7332 & 2.0190 & 3.5820 & 1.9162 \end{pmatrix}, \tag{1}$$

and an estimate of Σ is

$$S = \tfrac{1}{49} A = \begin{pmatrix} 0.266\,433 & 0.085\,184 & 0.182\,899 & 0.055\,780 \\ 0.085\,184 & 0.098\,469 & 0.082\,653 & 0.041\,204 \\ 0.182\,899 & 0.082\,653 & 0.220\,816 & 0.073\,102 \\ 0.055\,780 & 0.041\,204 & 0.073\,102 & 0.039\,106 \end{pmatrix}. \tag{2}$$

We use the iterative procedure to find the first principal component, by computing in turn $z^{(j)} = Sz^{(j-1)}$. As an initial approximation, we use $z^{(0)\prime} = (1,0,1,0)$. It is not necessary to normalize the vector at each iteration; but to compare successive vectors, we compute $z_i^{(j)}/z_i^{(j-1)} = r_i^{(j)}$, each of which is an approximation to l_1, the largest root of S. After seven iterations, $r_i^{(7)}$ agree to within two units in the fifth decimal place (fifth significant figure). This vector is normalized, and S is applied to the normalized vector. The ratios, $r_i^{(8)}$, agree to within two units in the sixth place; the value of l_1 is (nearly accurate to the sixth place) $l_1 = 0.487\,875$. The normalized eighth iterated vector is our estimate of $\boldsymbol{\beta}^{(1)}$, namely,

$$b^{(1)} = \begin{pmatrix} 0.686\,7244 \\ 0.305\,3463 \\ 0.623\,6628 \\ 0.214\,9837 \end{pmatrix}. \tag{3}$$

This vector agrees with the normalized seventh iterate to about one unit in the sixth place. It should be pointed out that l_1 and $b^{(1)}$ have to be calculated more accurately than l_2 and $b^{(2)}$, and so forth. The trace of S is 0.624 824, which is the sum of the roots. Thus l_1 is more than three times the sum of the other roots.

We next compute

$$(4)\quad S_2 = S - l_1 b^{(1)} b^{(1)\prime}$$

$$= \begin{pmatrix} 0.036\,3559 & -0.017\,1179 & -0.026\,0502 & -0.016\,2472 \\ -0.017\,1179 & 0.052\,9813 & -0.010\,2546 & 0.009\,1777 \\ -0.026\,0502 & -0.010\,2546 & 0.031\,0544 & 0.007\,6890 \\ -0.016\,2472 & 0.009\,1777 & 0.007\,6890 & 0.016\,5574 \end{pmatrix},$$

and iterate $z^{(j)} = S_2 z^{(j-1)}$, using $z^{(0)\prime} = (0,1,0,0)$. (In the actual computation S_2 was multiplied by 10 and the first row and column were multiplied by -1.) In this case the iteration does not proceed as rapidly; as will be seen, the ratio of l_2 to l_3 is approximately 1.32. On the last iteration, the ratios agree to within four units in the fifth significant figure. We obtain $l_2 = 0.072\,3828$ and

$$b^{(2)} = \begin{pmatrix} -0.669\,033 \\ 0.567\,484 \\ 0.343\,309 \\ 0.335\,307 \end{pmatrix}. \tag{5}$$

The third principal component is found from $S_3 = S_2 - l_2 b^{(2)} b^{(2)\prime}$, and the fourth from $S_4 = S_3 - l_3 b^{(3)} b^{(3)\prime}$.

The results may be summarized as follows:

(6) $$(l_1, l_2, l_3, l_4) = (0.4879, 0.0724, 0.0548, 0.0098),$$

(7) $$B = \begin{pmatrix} 0.6867 & -0.6690 & -0.2651 & 0.1023 \\ 0.3053 & 0.5675 & -0.7296 & -0.2289 \\ 0.6237 & 0.3433 & 0.6272 & -0.3160 \\ 0.2150 & 0.3353 & 0.0637 & 0.9150 \end{pmatrix}.$$

The sum of the four roots is $\sum_{i=1}^{4} l_i = 0.6249$, compared with the trace of the sample covariance matrix, $\operatorname{tr} S = 0.624\,824$. The first accounts for 78% of the total variance in the four measurements; the last accounts for a little more than 1%. In fact, the variance of $0.7x_1 + 0.3x_2 + 0.6x_3 + 0.2x_4$ (an approximation to the first principal component) is 0.478, which is almost 77% of the total variance. If one is interested in studying the variations in conditions that lead to variations of (x_1, x_2, x_3, x_4), one can look for variations in conditions that lead to variations of $0.7x_1 + 0.3x_2 + 0.6x_3 + 0.2x_4$. It is not very important if the other variations in (x_1, x_2, x_3, x_4) are neglected in exploratory investigations.

11.6. STATISTICAL INFERENCE

11.6.1. Asymptotic Distributions

In Section 13.3 we shall derive the exact distribution of the sample characteristic roots and vectors when the population covariance matrix is I or proportional to I, that is, in the case of all population roots equal. The exact distribution of roots and vectors when the population roots are not all equal involves a multiply infinite series of zonal polynomials; that development is beyond the scope of this book. [See Muirhead (1982).] We derive the asymptotic distribution of the roots and vectors when the population roots are all different (Theorem 13.5.1) and also when one root is multiple (Theorem 13.5.2). Since it can usually be assumed that the population roots are different unless there is information to the contrary, we summarize here Theorem 13.5.1.

As earlier, let the characteristic roots of Σ be $\lambda_1 > \cdots > \lambda_p$ and the corresponding characteristic vectors be $\beta^{(1)}, \dots, \beta^{(p)}$, normalized so $\beta^{(i)\prime}\beta^{(i)} = 1$ and satisfying $\beta_{1i} \geq 0$, $i = 1, \dots, p$. Let the roots and vectors of S be $l_1 > \cdots > l_p$ and $b^{(1)}, \dots, b^{(p)}$ normalized so $b^{(i)\prime}b^{(i)} = 1$ and satisfying $b_{1i} \geq 0$, $i = 1, \dots, p$. Let $d_i = \sqrt{n}(l_i - \lambda_i)$ and $g^{(i)} = \sqrt{n}(b^{(i)} - \beta^{(i)})$, $i = 1, \dots, p$. Then in the limiting normal distribution the sets $d_1, \dots, d_p$ and $g^{(1)}, \dots, g^{(p)}$ are independent and $d_1, \dots, d_p$ are mutually independent. The element d_i has the limiting distribution $N(0, 2\lambda_i^2)$. The covariances of $g^{(1)}, \dots, g^{(p)}$ in the limiting distribution are

(1) $$\mathscr{C}(g^{(i)}) = \sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \boldsymbol{\beta}^{(k)} \boldsymbol{\beta}^{(k)\prime},$$

(2) $$\mathscr{C}(g^{(i)}, g^{(j)}) = -\frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \boldsymbol{\beta}^{(j)} \boldsymbol{\beta}^{(i)\prime}, \qquad i \neq j.$$

See Theorem 13.5.1.

In making inferences about a single ordered root, one treats l_i as approximately normal with mean λ_i and variance $2\lambda_i^2/n$. Since l_i is a consistent estimate of λ_i, the limiting distribution of

(3) $$\sqrt{n}\, \frac{l_i - \lambda_i}{\sqrt{2}\, l_i}$$

is $N(0, 1)$. A two-tailed test of the hypothesis $\lambda_i = \lambda_i^0$ has the (asymptotic) acceptance region

(4) $$-z(\varepsilon) \leq \sqrt{\frac{n}{2}}\, \frac{l_i - \lambda_i^0}{\lambda_i^0} \leq z(\varepsilon),$$

where the value of the $N(0,1)$ distribution beyond $z(\varepsilon)$ is $\frac{1}{2}\varepsilon$. The interval (4) can be inverted to give a confidence interval for λ_i with confidence $1 - \varepsilon$:

(5) $$\frac{l_i}{1 + \sqrt{2/n}\, z(\varepsilon)} \leq \lambda_i \leq \frac{l_i}{1 - \sqrt{2/n}\, z(\varepsilon)}.$$

Note that the confidence coefficient should be taken large enough so $\sqrt{2/n}\, z(\varepsilon) < 1$. Alternatively, one can use the fact that the limiting distribution of $\sqrt{n}(\log l_i - \log \lambda_i)$ is $N(0, 2)$ by Theorem 4.2.3.

Inference about components of a vector $\boldsymbol{\beta}^{(i)}$ can be based on treating $b^{(i)}$ as being approximately normal with mean $\boldsymbol{\beta}^{(i)}$ and (singular) covariance matrix $1/n$ times (1).

11.6.2. Confidence Region for a Characteristic Vector

We use the asymptotic distribution of the sample characteristic vectors to obtain a large-sample confidence region for the ith characteristic vector of $\boldsymbol{\Sigma}$ [Anderson (1963a)]. The covariance matrix (1) can be written

(6) $$\mathbf{B}\boldsymbol{\Delta}_i^2\mathbf{B}' = \mathbf{B}_i^* \boldsymbol{\Delta}_i^{*2} \mathbf{B}_i^{*\prime},$$

where Δ_i is the $p \times p$ diagonal matrix with 0 as the ith diagonal element and $\sqrt{\lambda_i \lambda_j}/(\lambda_i - \lambda_j)$ as the jth diagonal element, $j \neq i$; Δ_i^* is the $(p-1) \times (p-1)$ diagonal matrix obtained from Δ_i by deleting the ith row and column; and $\boldsymbol{\beta}_i^*$ is the $p \times (p-1)$ matrix formed by deleting the ith column from $\boldsymbol{\beta}$. Then $h^{(i)} = \Delta_i^{*-1} \boldsymbol{\beta}_i^{*\prime} \sqrt{n}\,(b^{(i)} - \boldsymbol{\beta}^{(i)})$ has a limiting normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\mathscr{C}(h^{(i)}) = \Delta_i^{*-1} \boldsymbol{\beta}_i^{*\prime} \left(\boldsymbol{\beta}_i^* \Delta_i^{*2} \boldsymbol{\beta}_i^{*\prime} \right) \boldsymbol{\beta}_i^* \Delta_i^{*-1} = I_{p-1}, \tag{7}$$

and

$$h^{(i)\prime} h^{(i)} = n\left(b^{(i)} - \boldsymbol{\beta}^{(i)}\right)' \boldsymbol{\beta}_i^* \Delta_i^{*-2} \boldsymbol{\beta}_i^{*\prime} \left(b^{(i)} - \boldsymbol{\beta}^{(i)}\right) \tag{8}$$

has a limiting χ^2-distribution with $p-1$ degrees of freedom. The matrix of the quadratic form in $\sqrt{n}\,(b^{(i)} - \boldsymbol{\beta}^{(i)})$ is

$$\begin{aligned} \boldsymbol{\beta}_i^* \Delta_i^{*-2} \boldsymbol{\beta}_i^{*\prime} &= \sum_{j=1}^{p} \boldsymbol{\beta}^{(j)} \left(\frac{\lambda_i}{\lambda_j} - 2 + \frac{\lambda_j}{\lambda_i} \right) \boldsymbol{\beta}^{(j)\prime} - \boldsymbol{\beta}^{(i)} \left(\frac{\lambda_i}{\lambda_i} - 2 + \frac{\lambda_i}{\lambda_i} \right) \boldsymbol{\beta}^{(i)\prime} \\ &= \lambda_i \Sigma^{-1} - 2I + (1/\lambda_i) \Sigma \end{aligned} \tag{9}$$

because $\boldsymbol{\beta} \Lambda^{-1} \boldsymbol{\beta}' = \Sigma^{-1}$, $\boldsymbol{\beta}\boldsymbol{\beta}' = I$, and $\boldsymbol{\beta} \Lambda \boldsymbol{\beta}' = \Sigma$. Then (8) is

$$\begin{aligned} & n\left(b^{(i)} - \boldsymbol{\beta}^{(i)}\right)' \left[\lambda_i \Sigma^{-1} - 2I + (1/\lambda_i) \Sigma \right] \left(b^{(i)} - \boldsymbol{\beta}^{(i)}\right) \\ &= n b^{(i)\prime} \left[\lambda_i \Sigma^{-1} - 2I + (1/\lambda_i) \Sigma \right] b^{(i)} \\ &= n \left[\lambda_i b^{(i)\prime} \Sigma^{-1} b^{(i)} + (1/\lambda_i) b^{(i)\prime} \Sigma b^{(i)} - 2 \right], \end{aligned} \tag{10}$$

because $\boldsymbol{\beta}^{(i)\prime}$ is a characteristic vector of Σ with root λ_i, and of Σ^{-1} with root $1/\lambda_i$. On the left-hand side of (10) we can replace Σ and λ_i by the consistent estimators S and l_i to obtain

$$\begin{aligned} & n\left(b^{(i)} - \boldsymbol{\beta}^{(i)}\right)' \left[l_i S^{-1} - 2I + (1/l_i) S \right] \left(b^{(i)} - \boldsymbol{\beta}^{(i)}\right) \\ &= n \left[l_i \boldsymbol{\beta}^{(i)\prime} S^{-1} \boldsymbol{\beta}^{(i)} + (1/l_i) \boldsymbol{\beta}^{(i)\prime} S \boldsymbol{\beta}^{(i)} - 2 \right], \end{aligned} \tag{11}$$

which has a limiting χ^2-distribution with $p-1$ degrees of freedom.

A confidence region for the ith characteristic vector of Σ with confidence $1-\varepsilon$ consists of the intersection of $\boldsymbol{\beta}^{(i)\prime} \boldsymbol{\beta}^{(i)} = 1$ and the set of $\boldsymbol{\beta}^{(i)}$ such that the right-hand side of (11) is less than $\chi^2_{p-1}(\varepsilon)$, where $\Pr\{\chi^2_{p-1} > \chi^2_{p-1}(\varepsilon)\} = \varepsilon$. Note that the matrix of the quadratic form (9) is positive semidefinite.

This approach also provides a test of the null hypothesis that the ith characteristic vector is a specified $\boldsymbol{\beta}_0^{(i)}$ $(\boldsymbol{\beta}_0^{(i)\prime}\boldsymbol{\beta}_0^{(i)}=1)$. The hypothesis is rejected if the right-hand side of (11) with $\boldsymbol{\beta}^{(i)}$ replaced by $\boldsymbol{\beta}_0^{(i)}$ exceeds $\chi^2_{p-1}(\varepsilon)$.

Mallows (1961) suggested a test of whether *some* characteristic vector of $\boldsymbol{\Sigma}$ is $\boldsymbol{\beta}_0$. Let $\mathbf{B}_0$ be $p\times(p-1)$ matrix such that $\boldsymbol{\beta}_0'\mathbf{B}_0=\mathbf{0}$. If the null hypothesis is true, $\boldsymbol{\beta}_0'X$ and $\mathbf{B}_0'X$ are independent (because $\mathbf{B}_0$ is a nonsingular transform of the set of other characteristic vectors). The test is based on the multiple correlation between $\boldsymbol{\beta}_0'X$ and $\mathbf{B}_0'X$. In principle, the test procedure can be inverted to obtain a confidence region. The usefulness of these procedures is limited by the fact that the hypothesized vector is not attached to a characteristic root; the interpretation depends on the root (e.g., largest versus smallest).

Tyler (1981), (1983b) has generalized the confidence region (11) to include the vectors in a linear subspace. He has also studied easing the restrictions of a normally distributed parent population.

11.6.3. Exact Confidence Limits on the Characteristic Roots

We now consider a confidence interval for the entire set of characteristic roots of $\boldsymbol{\Sigma}$, namely, $\lambda_1\ge\cdots\ge\lambda_p$ [Anderson (1965a)]. We use the facts that $\boldsymbol{\beta}^{(i)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta}^{(i)}=\lambda_i$, $\boldsymbol{\beta}^{(i)\prime}\boldsymbol{\beta}^{(i)}=1$, $i=1,p$, and $\boldsymbol{\beta}^{(1)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta}^{(p)}=0=\boldsymbol{\beta}^{(1)\prime}\boldsymbol{\beta}^{(p)}$. Then $\boldsymbol{\beta}^{(1)\prime}X$ and $\boldsymbol{\beta}^{(p)\prime}X$ are uncorrelated and have variances λ_1 and λ_p, respectively. Hence $n\boldsymbol{\beta}^{(1)\prime}S\boldsymbol{\beta}^{(1)}/\lambda_1$ and $n\boldsymbol{\beta}^{(p)\prime}S\boldsymbol{\beta}^{(p)}/\lambda_p$ are independently distributed as χ^2 with n degrees of freedom. Let l and u be two numbers such that

$$1-\varepsilon=\Pr\{nl\le\chi_n^2\}\Pr\{\chi_n^2\le nu\}. \tag{12}$$

Then

$$\begin{aligned}1-\varepsilon&=\Pr\left\{1\le\frac{\boldsymbol{\beta}^{(1)\prime}S\boldsymbol{\beta}^{(1)}}{\lambda_1},\ \frac{\boldsymbol{\beta}^{(p)\prime}\boldsymbol{\Sigma}\boldsymbol{\beta}^{(p)}}{\lambda_p}\le u\right\}\\&=\Pr\left\{\frac{\boldsymbol{\beta}^{(p)\prime}S\boldsymbol{\beta}^{(p)}}{u}\le\lambda_p,\ \lambda_1\le\frac{\boldsymbol{\beta}^{(1)\prime}S\boldsymbol{\beta}^{(1)}}{l}\right\}\\&\le\Pr\left\{\min_{b'b=1}\frac{b'Sb}{u}\le\lambda_p,\ \lambda_1\le\max_{b'b=1}\frac{b'Sb}{l}\right\}\\&=\Pr\left\{\frac{l_p}{u}\le\lambda_p\le\lambda_1\le\frac{l_1}{l}\right\}.\end{aligned} \tag{13}$$

Theorem 11.6.1. *A confidence interval for the characteristic roots of* Σ *with confidence at least* $1-\varepsilon$ *is*

$$l_p/u \le \lambda_p \le \lambda_1 \le l_1/l, \tag{14}$$

where l *and* u *satisfy* (12).

A tighter inequality can lead to a better lower bound. The matrix $H = n\boldsymbol{\beta}'S\boldsymbol{\beta}$ has characteristic roots $nl_1, \dots, nl_p$ because $\boldsymbol{\beta}$ is orthogonal. We use the following lemma.

Lemma 11.6.1. *For any positive definite matrix* H

$$\operatorname{ch}_p(H) \le \frac{1}{h^{ii}} \le \operatorname{ch}_1(H), \qquad i = 1, \dots, p, \tag{15}$$

where $H^{-1} = (h^{ij})$ *and* $\operatorname{ch}_p(H)$ *and* $\operatorname{ch}_1(H)$ *are the minimum and maximum characteristic roots of* H, *respectively.*

Proof. From Theorem A.2.4 in the Appendix we have $\operatorname{ch}_p(H) \le h_{ii} \le \operatorname{ch}_1(H)$ and

$$\operatorname{ch}_p(H^{-1}) \le h^{ii} \le \operatorname{ch}_1(H^{-1}), \qquad i = 1, \dots, p. \tag{16}$$

Since $\operatorname{ch}_p(H) = 1/\operatorname{ch}_1(H^{-1})$ and $\operatorname{ch}_1(H) = 1/\operatorname{ch}_p(H^{-1})$, the lemma follows. ■

The argument for Theorem 5.2.2 shows that $1/(\lambda_p h^{pp})$ is distributed as χ^2 with $n-p+1$ degrees of freedom, and Theorem 4.3.3 shows that h^{pp} is independent of h_{11}. Let l' and u' be two numbers such that

$$1 - \varepsilon = \Pr\{nl' \le \chi_n^2\} \Pr\{\chi_{n-p+1}^2 \le nu'\}. \tag{17}$$

Then

$$\begin{aligned} 1 - \varepsilon &= \Pr\left\{nl' \le \frac{h_{11}}{\lambda_1}, \frac{1}{\lambda_p h^{pp}} \le nu'\right\} \\ &\le \Pr\left\{\frac{l_p}{u'} \le \lambda_p, \lambda_1 \le \frac{l_1}{l'}\right\} \end{aligned} \tag{18}$$

since $\operatorname{ch}_p(H) = nl_p$ and $\operatorname{ch}_1(H) = nl_1$.

Theorem 11.6.2. *A confidence interval for the characteristic roots of* Σ *with confidence at least* $1 - \varepsilon$ *is*

$$\frac{l_p}{u'} \le \lambda_p \le \lambda_1 \le \frac{l_1}{l'}, \tag{19}$$

where l' *and* u' *satisfy* (17).

Anderson (1965a, 1965b) showed that the above confidence bounds are optimal within the class of bounds

$$f(l_1, \dots, l_p) \le \lambda_p \le \lambda_1 \le g(l_1, \dots, l_p), \tag{20}$$

where f and g are homogeneous of degree 1 and are monotonically nondecreasing in each argument for fixed values of the others. If (20) holds with probability at least $1 - \varepsilon$, then a pair of numbers u' and l' can be found to satisfy (17) and

$$f(l_1, \dots, l_p) \le \frac{l_p}{u'}, \qquad \frac{l_1}{l'} \le g(l_1, \dots, l_p). \tag{21}$$

The homogeneity condition means that the confidence bounds are multiplied by c^2 if the observed vectors are multiplied by c (which is a kind of scale invariance). The monotonicity conditions imply that an increase in the size of S results in an increase in the limits for Σ (which is a kind of consistency).

The confidence bounds given in (31) of Section 10.8 for the roots of Σ based on the distribution of the roots of S when $\Sigma = I$ are greater.

11.7. TESTING HYPOTHESES ABOUT THE CHARACTERISTIC ROOTS OF A COVARIANCE MATRIX

11.7.1. Testing a Hypothesis about the Sum of the Smallest Characteristic Roots

An investigator may raise the question whether the last $p - m$ principal components may be ignored, that is, whether the first m principal components furnish a good approximation to X. He may want to do this if the sum of the variances of the last principal components is less than some specified amount, say γ. Consider the null hypothesis

$$H: \lambda_{m+1} + \cdots + \lambda_p \ge \gamma, \tag{1}$$

where γ is specified, against the alternative that the sum is less than γ. If the characteristic roots of Σ are different, it follows from Theorem 13.5.1 that

$$\sqrt{n}\left(\sum_{i=m+1}^{p} l_i - \sum_{i=m+1}^{p} \lambda_i\right) \tag{2}$$

has a limiting normal distribution with mean 0 and variance $2\sum_{i=m+1}^{p} \lambda_i^2$. The variance can be consistently estimated by $2\sum_{i=m+1}^{p} l_i^2$. Then a rejection region with (large-sample) significance level ε is

$$\sum_{i=m+1}^{p} l_i < \gamma - \frac{\sqrt{2\sum_{i=m+1}^{p} l_i^2}}{\sqrt{n}} z(2\varepsilon), \tag{3}$$

where $z(2\varepsilon)$ is the upper significance point of the standard normal distribution for significance level ε. The (large-sample) probability of rejection is ε if equality holds in (1) and is less than ε if inequality holds.

The investigator may alternatively want an upper confidence interval for $\sum_{i=m+1}^{p} \lambda_i$ with at least approximate confidence level $1-\varepsilon$. It is

$$\sum_{i=m+1}^{p} \lambda_i \leq \sum_{i=m+1}^{p} l_i + \frac{\sqrt{2\sum_{i=m+1}^{p} l_i^2}}{\sqrt{n}} z(2\varepsilon). \tag{4}$$

If the right-hand side is sufficiently small (in particular less than γ), the investigator has confidence that the sum of the variances of the smallest $p-m$ principal components is so small they can be neglected. Anderson (1963a) gave this analysis also in the case that $\lambda_{m+1} = \cdots = \lambda_p$.

11.7.2. Testing a Hypothesis about the Sum of the Smallest Characteristic Roots Relative to the Sum of All the Roots

The investigator may want to ignore the last $p-m$ principal components if their sum is small relative to the sum of all the roots (which is the trace of the covariance matrix). Consider the null hypothesis

$$H: f(\boldsymbol{\lambda}) = \frac{\lambda_{m+1} + \cdots + \lambda_p}{\lambda_1 + \cdots + \lambda_p} \geq \delta, \tag{5}$$

where δ is specified, against the alternative that $f(\boldsymbol{\lambda}) < \delta$. We use the fact that

$$\begin{aligned} \frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_i} &= -\frac{\lambda_{m+1} + \cdots + \lambda_p}{(\lambda_1 + \cdots + \lambda_p)^2}, & i &= 1, \dots, m, \\ \frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_i} &= \frac{\lambda_1 + \cdots + \lambda_m}{(\lambda_1 + \cdots + \lambda_p)^2}, & i &= m+1, \dots, p. \end{aligned} \tag{6}$$

Then the asymptotic variance of $f(l)$ is

$$(7) \qquad 2\left(\frac{\delta}{\operatorname{tr}\Sigma}\right)^2 (\lambda_1^2 + \cdots + \lambda_m^2) + 2\left(\frac{1-\delta}{\operatorname{tr}\Sigma}\right)^2 (\lambda_{m+1}^2 + \cdots + \lambda_p^2)$$

when equality holds in (5), by Theorem 4.2.3. The null hypothesis H is rejected if $\sqrt{n}\,[f(l) - \delta]$ is less than the appropriate significance point of the standard normal distribution times the square root of (7) with λ's replaced by l's and $\operatorname{tr}\Sigma$ by $\operatorname{tr} S$. Alternatively one can construct a large-sample confidence region for $f(\boldsymbol{\lambda})$. A confidence region of approximate confidence $1-\varepsilon$ is $[z = z(2\varepsilon)]$

$$(8)$$

$$\frac{\sum_{i=m+1}^p \lambda_i}{\sum_{i=1}^p \lambda_i} \le \frac{\sum_{i=m+1}^p l_i}{\sum_{i=1}^p l_i} + z\frac{\left[2\left(\sum_{i=m+1}^p l_i\right)^2 \sum_{i=1}^m l_i^2 + 2\left(\sum_{i=1}^m l_i\right)^2 \sum_{i=m+1}^p l_i^2\right]^{\frac{1}{2}}}{\sqrt{n}\left(\sum_{i=1}^p l_i\right)^2}.$$

If the right-hand side is sufficiently small, the investigator may be willing to let the first principal components represent the entire vector of measurements.

11.7.3. Testing Equality of the Smallest Roots

Suppose the observed X is given by $V + U + \boldsymbol{\mu}$, where V and U are unobservable random vectors with means $\mathbf{0}$ and $\boldsymbol{\mu}$ is an unobservable vector of constants. If $\mathscr{E}UU' = \sigma^2 I$, then U can be interpreted as composed of errors of measurement: uncorrelated components with equal variances. (It is assumed that all components of X are in the same units.) Then V can be interpreted as made up of the systematic parts and is supposed to lie in an m-dimensional space. Then $\mathscr{E}VV' = \boldsymbol{\Phi}$ is positive semidefinite of rank m. The observable covariance matrix $\Sigma = \boldsymbol{\Phi} + \sigma^2 I$ has a characteristic root of σ^2 with multiplicity $p - m$ (Problem 11.4).

In this subsection we consider testing the null hypothesis that $\lambda_{m+1} = \cdots = \lambda_p$. That is equivalent to the null hypothesis that $\Sigma = \boldsymbol{\Phi} + \sigma^2 I$, where $\boldsymbol{\Phi}$ is positive semidefinite of rank m. In Section 10.7, we saw that when $m = 0$, the likelihood ratio criterion was the $\frac{1}{2}pN$th power of the ratio of the geometric mean to the arithmetic mean of the sample roots. The analogous criterion here is the $\frac{1}{2}N$th power of

$$(9) \qquad \frac{\prod_{i=m+1}^p l_i}{\left(\sum_{i=m+1}^p l_i\right)^{p-m}} (p-m)^{p-m}.$$

It is also the likelihood ratio criterion, but we shall not derive it. [See Anderson (1963a).] Let $\sqrt{n}(l_i - \lambda_{m+1}) = d_i$, $i = m+1, \dots, p$. The logarithm of (9) multiplied by $-n$ is asymptotically equivalent under the null hypothesis to

$$(10)\quad -n \log \prod_{i=m+1}^{p} l_i + n(p-m)\log \frac{\sum_{i=m+1}^{p} l_i}{p-m}$$

$$= -n \sum_{i=m+1}^{p} \log\left(\lambda_{m+1} + n^{-\frac{1}{2}} d_i\right) + n(p-m)\log \frac{\sum_{i=m+1}^{p}\left(\lambda_{m+1} + n^{-\frac{1}{2}} d_i\right)}{p-m}$$

$$= n\left\{ -\sum_{i=m+1}^{p} \log\left(1 + \frac{d_i}{\lambda_{m+1} n^{\frac{1}{2}}}\right) + (p-m)\log\left(1 + \frac{\sum_{i=m+1}^{p} d_i}{(p-m)\lambda_{m+1} n^{\frac{1}{2}}}\right)\right\}$$

$$= n\left\{ -\sum_{i=m+1}^{p} \left[\frac{d_i}{\lambda_{m+1} n^{\frac{1}{2}}} - \frac{d_i^2}{2\lambda_{m+1}^2 n} + \cdots\right]\right.$$

$$\left. + (p-m)\left[\frac{\sum_{i=m+1}^{p} d_i}{(p-m)\lambda_{m+1} n^{\frac{1}{2}}} - \frac{\left(\sum_{i=m+1}^{p} d_i\right)^2}{2(p-m)^2\lambda_{m+1}^2 n} + \cdots\right]\right\}$$

$$= n\left\{ \sum_{i=m+1}^{p} \frac{d_i^2}{2\lambda_{m+1}^2 n} - \cdots - \frac{\left(\sum_{i=m+1}^{p} d_i\right)^2}{2(p-m)\lambda_{m+1}^2 n} + \cdots\right\}$$

$$= \frac{1}{2\lambda_{m+1}^2}\left[\sum_{i=m+1}^{p} d_i^2 - \frac{1}{p-m}\left(\sum_{i=m+1}^{p} d_i\right)^2\right] + O_p(1).$$

It is shown in Section 13.5.2 that the limiting distribution of $d_{m+1}, \dots, d_p$ is the same as the distribution of the roots of a symmetric matrix $U_{22} = (u_{ij})$, $i, j = m+1, \dots, p$, whose functionally independent elements are independent and normal with mean 0; an off-diagonal element u_{ij}, $i < j$, has variance λ_{m+1}^2, and a diagonal element u_{ii} has variance $2\lambda_{m+1}^2$. See Theorem 13.5.2. Then (10) has the limiting distribution of

$$(11)\quad \frac{1}{2\lambda_{m+1}^2}\left(\operatorname{tr} U_{22}^2 - \frac{1}{p-m}(\operatorname{tr} U_{22})^2\right)$$

$$= \frac{1}{2\lambda_{m+1}^2}\left[\operatorname{tr} U_{22}U_{22}' - \frac{1}{p-m}(\operatorname{tr} U_{22})^2\right]$$

$$= \frac{1}{2\lambda_{m+1}^2}\left[2\sum_{i<j} u_{ij}^2 + \sum_{i=m+1}^{p} u_{ii}^2 - \frac{1}{p-m}\left(\sum_{i=m+1}^{p} u_{ii}\right)^2\right].$$

Thus $\sum_{i<j} u_{ij}^2/\lambda_{m+1}^2$ is asymptotically χ^2 with $\frac{1}{2}(p-m)(p-m-1)$ degrees of freedom; $\frac{1}{2}[\sum_{i=m+1}^{p} u_{ii}^2 - (\sum_{i=m+1}^{p} u_{ii})^2/(p-m)]/\lambda_{m+1}^2$ is asymptotically χ^2 with $p-m-1$ degrees of freedom. Then (10) has a limiting χ^2-distribution with $\frac{1}{2}(p-m+2)(p-m-1)$ degrees of freedom. The hypothesis is rejected if the left-hand side of (10) is greater than the upper-tailed significance point of the χ^2-distribution. If the hypothesis is not rejected, the investigator may consider the last $p-m$ principal components to be composed entirely of error.

When the units of measurement are not all the same, the three hypotheses considered in Section 11.7 have questionable meaning. Corresponding hypotheses for the correlation matrix also have doubtful interpretation. Moreover, the last criterion does not have (usually) a χ^2-distribution. More discussion is given by Anderson (1963a).

The criterion (9) corresponds to the sphericity criterion of Section 10.7, and the number of degrees of freedom of the corresponding χ^2-distribution is $\frac{1}{2}(p-m)(p-m+1)-1$.

11.8. ELLIPTICALLY CONTOURED DISTRIBUTIONS

11.8.1. Observations Elliptically Contoured

Let $x_1, \dots, x_N$ be N observations on a random vector X with density

$$|\Psi|^{-\frac{1}{2}} g\left[(x-\nu)'\Psi^{-1}(x-\nu)\right], \tag{1}$$

where Ψ is a positive definite matrix, $R^2 = (x-\nu)'\Psi^{-1}(x-\nu)$, and $\mathscr{E}R^4 < \infty$. Define $\kappa = p\mathscr{E}R^4/[(\mathscr{E}R^2)^2(p+2)] - 1$. Then $\mathscr{E}X = \nu = \mu$ and $\mathscr{E}(X-\nu)(X-\nu)' = (\mathscr{E}R^2/p)\Psi = \Sigma$.

The maximum likelihood estimators of the principal components of Σ are the characteristic roots and vectors of $\hat{\Sigma} = (\mathscr{E}R^2/p)\hat{\Lambda}$ given by (20) of Section 3.6. Alternative estimators are the characteristic roots and vectors of S, the unbiased estimator of Σ. The asymptotic normal distributions of these estimators are derived in Section 13.7 (Theorem 13.7.1). Let $\Sigma = \boldsymbol{\beta}\Lambda\boldsymbol{\beta}'$ and $S = BLB'$, where Λ and L are diagonal and $\boldsymbol{\beta}$ and B are orthogonal. Let $D = \sqrt{N}(L-\Lambda)$ and $G = \sqrt{N}(B-\boldsymbol{\beta})$. Then the limiting distribution of D and G is normal with G and D independent.

The variance of d_i is $(2+3\kappa)\lambda_i^2$, and the covariance of d_i and d_j $(i \neq j)$ is $\kappa\lambda_i\lambda_j$. The covariance of g_i is

$$\mathscr{C}(g_i) = (1+\kappa)\sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}\boldsymbol{\beta}_k\boldsymbol{\beta}_k'. \tag{2}$$

The covariance of g_i and g_j is

$$\mathscr{AC}(g_i, g_j) = -(1+\kappa)\frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2}\beta_j \beta_i'. \tag{3}$$

For inference about a single ordered root λ_i the limiting standard normal distribution of $\sqrt{N}(l_i - \lambda_i)/(\sqrt{2(2+3\hat{\kappa})}\, l_i)$ can be used.

For inference about a single vector the right-hand side of (11) in Section 11.6.2 can be used with S replaced by $(1+\hat{\kappa})S$ and S^{-1} by $S^{-1}/(1+\hat{\kappa})$.

It is shown in Section 13.7.1 that the limiting distribution of the logarithm of the likelihood ratio criterion for testing the equality of the $q = p - m$ smallest roots is the distribution of $(1+\kappa)\chi^2_{q(q-1)/2-1}$.

11.8.2. Elliptically Contoured Matrix Distributions

Suppose the density of $X = (x_1, \ldots, x_N)$ is

$$|\Psi|^{-N/2} g\left[\operatorname{tr}(X - \varepsilon_N' v)'\Psi^{-1}(X - \varepsilon_N' v)\right]$$
$$= |\Psi|^{-N/2} g\left[\operatorname{tr} A\Psi^{-1} + N(\bar{x} - v)'\Psi^{-1}(\bar{x} - v)\right],$$

where $A = (X - \varepsilon_N'\bar{x})(X - \varepsilon_N'\bar{x})' = nS$ and $n = N - 1$. Thus x and A are a sufficient set of statistics.

Now consider $A = YY'$ having the density $g(\operatorname{tr} A)$. Let $A = BLB'$, where L is diagonal with diagonal elements $l_1 > \cdots > l_p$ and B is orthogonal with $p_{i1} \geq 0$. Then L and B are independent; the roots $l_1, \ldots, l_p$ have the density (18) of Section 13.7, and the matrix B has the conditional Haar invariant distribution.

PROBLEMS

11.1. (Sec. 11.2) Prove that the characteristic vectors of $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ are

$$\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix},$$

corresponding to roots $1+\rho$ and $1-\rho$.

11.2. (Sec. 11.2) Verify that the proof of Theorem 11.2.1 yields a proof of Theorem A.2.1 of the Appendix for any real symmetric matrix.

11.3. (Sec. 11.2) Let $z = y + x$, where $\mathscr{E}y = \mathscr{E}x = 0$, $\mathscr{E}yy' = \Phi$, $\mathscr{E}xx' = \sigma^2 I$, $\mathscr{E}yx' = 0$. The p components of y can be called systematic parts, and the components of x errors.

(a) Find the linear combination $\gamma' z$ of unit variance that has minimum error variance (i.e., $\gamma' x$ has minimum variance).
(b) Suppose $\phi_{ii} + \sigma^2 = 1$, $i = 1, \dots, p$. Find the linear function $\gamma' z$ of unit variance that maximizes the sum of squares of the correlations between z_i and $\gamma' z$, $i = 1, \dots, p$.
(c) Relate these results to principal components.

11.4. (Sec. 11.2) Let $\Sigma = \Phi + \sigma^2 I$, where Φ is positive semidefinite of rank m. Prove that each characteristic vector of Φ is a vector of Σ and each root of Σ is a root of Φ plus σ^2.

11.5. (Sec. 11.2) Let the characteristic roots of Σ be $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$.

(a) What is the form of Σ if $\lambda_1 = \lambda_2 = \cdots = \lambda_p > 0$? What is the shape of an ellipsoid of constant density?
(b) What is the form of Σ if $\lambda_1 > \lambda_2 = \cdots = \lambda_p > 0$? What is the shape of an ellipsoid of constant density?
(c) What is the form of Σ if $\lambda_1 = \cdots = \lambda_{p-1} > \lambda_p > 0$? What is the shape of the ellipsoid of constant density?

11.6. (Sec. 11.2) *Intraclass correlation*. Let

$$\Sigma = \sigma^2[(1-\rho)I + \rho\varepsilon\varepsilon'],$$

where $\varepsilon = (1, \dots, 1)'$. Show that for $\rho > 0$, the largest characteristic root is $\sigma^2[1 + (p-1)\rho]$ and the corresponding characteristic vector is ε. Show that if $\varepsilon' x = 0$, then x is a characteristic vector corresponding to the root $\sigma^2(1-\rho)$. Show that the root $\sigma^2(1-\rho)$ has multiplicity $p-1$.

11.7. (Sec. 11.3) In the example of Section 9.6, consider the three pressing operations (x_2, x_4, x_5). Find the first principal component of this estimated covariance matrix. [*Hint:* Start with the vector $(1,1,1)$ and iterate.)

11.8. (Sec. 11.3) Prove directly the sample analog of Theorem 11.2.1, where $\sum x_\alpha = 0$, $\sum x_\alpha x'_\alpha = A$.

11.9. (Sec. 11.3) Let l_1 and l_p be the largest and smallest characteristic roots of S, respectively. Prove $\mathscr{E} l_1 \ge \lambda_1$ and $\mathscr{E} l_p \le \lambda_p$.

11.10. (Sec. 11.3) Let $U_1 = \beta^{(1)\prime} X$ be the first population principal component with variance $\mathscr{V}(U_1) = \lambda_1$, and let $V_1 = b^{(1)\prime} X$ be the first sample principal component with (sample) variance l_1 (based on S). Let S^* be the covariance matrix of a second (independent) sample. Show $\mathscr{E} b^{(1)\prime} S^* b^{(1)} \le \lambda_1$.

11.11. (Sec. 11.3) Suppose that $\sigma_{ij} > 0$ for every i, j $[\Sigma = (\sigma_{ij})]$. Show that (a) the coefficients of the first principal component are all of the same sign, and (b) the coefficients of each other principal component cannot be all of the same sign.

11.12. (Sec. 11.4) Prove (4) when $\lambda_1 > \lambda_2$.

(a) Show $\Sigma^i = \boldsymbol{\beta}\Lambda^i\boldsymbol{\beta}'$.

(b) Show

$$y_{(i)} = t_i\boldsymbol{\beta}\Lambda^i\boldsymbol{\beta}'x_{(0)} = t_i\lambda_1^i\boldsymbol{\beta}\left(\frac{1}{\lambda_1}\Lambda\right)^i\boldsymbol{\beta}'x_{(0)},$$

where $t_i = \prod_{j=0}^{i} s_j$ and $s_j = 1/\sqrt{x_{(i)}'x_{(i)}}$.

(c) Show

$$\lim_{i\to\infty}\left(\frac{1}{\lambda_1}\Lambda\right)^i = E_{11},$$

where E_{11} has 1 in the upper left-hand position and 0's elsewhere.

(d) Show $\lim_{i\to\infty}(t_i\lambda_1^i)^2 = 1/(\beta^{(1)\prime}x_{(0)})^2$.

(e) Conclude the proof.

11.13. (Sec. 11.4) Let

$$\Sigma = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{(1)}' \\ \boldsymbol{\sigma}_{(1)} & \Sigma_{22} \end{bmatrix}, \qquad K = \begin{bmatrix} 1 & 0 \\ 0 & H \end{bmatrix},$$

where $H = I_{p-1} - 2\boldsymbol{\alpha}\boldsymbol{\alpha}'$ and $\boldsymbol{\alpha}$ has $p-1$ components. Show that $\boldsymbol{\alpha}$ can be chosen so that in

$$K\Sigma K = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{(1)}'H \\ H\boldsymbol{\sigma}_{(1)} & H\Sigma_{22}H \end{bmatrix}$$

$H\boldsymbol{\sigma}_{(1)}$ has all 0 components except the first.

11.14. (Sec. 11.6) Show that

$$\log l_i - \sqrt{\frac{2}{n}}\,z(\varepsilon) < \log\lambda_i < \log l_i + \sqrt{\frac{2}{n}}\,z(\varepsilon)$$

is a confidence interval for $\log\lambda_i$ with approximate confidence $1-\varepsilon$.

11.15. (Sec. 11.6) Prove that $u' < u$ if $l' = l$ and $p > 2$.

11.16. (Sec. 11.6) Prove that $u < u^*$ if $l = l^*$ and $p > 2$, where l^* and u^* are the l and u of Section 10.8.4.

11.17. The lengths, widths, and heights (in millimeters) of 24 male painted turtles [Jolicoeur and Mosimann (1960)] are given below. Find the (sample) principal components and their variances.

Case No.	Length	Width	Height	Case No.	Length	Width	Height
1	93	74	37	13	116	90	43
2	94	78	35	14	117	90	41
3	96	80	35	15	117	91	41
4	101	84	39	16	119	93	41
5	102	85	38	17	120	89	40
6	103	81	37	18	120	93	44
7	104	83	39	19	121	95	42
8	106	83	39	20	125	93	45
9	107	82	38	21	127	96	45
10	112	89	40	22	128	95	46
11	113	88	40	23	131	95	46
12	114	86	40	24	135	106	47

CHAPTER 12

Canonical Correlations and Canonical Variables

12.1. INTRODUCTION

In this section we consider two sets of variates with a joint distribution, and we analyze the correlations between the variables of one set and those of the other set. We find a new coordinate system in the space of each set of variates in such a way that the new coordinates display unambiguously the system of correlation. More precisely, we find linear combinations of variables in the sets that have maximum correlation; these linear combinations are the first coordinates in the new systems. Then a second linear combination in each set is sought such that the correlation between these is the maximum of correlations between such linear combinations as are uncorrelated with the first linear combinations. The procedure is continued until the two new coordinate systems are completely specified.

The statistical method outlined is of particular usefulness in exploratory studies. The investigator may have two large sets of variates and may want to study the interrelations. If the two sets are very large, he may want to consider only a few linear combinations of each set. Then he will want to study those linear combinations most highly correlated. For example, one set of variables may be measurements of physical characteristics, such as various lengths and breadths of skulls; the other variables may be measurements of mental characteristics, such as scores on intelligence tests. If the investigator is interested in relating these, he may find that the interrelation is almost

---

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

completely described by the correlation between the first few canonical variates.

The basic theory was developed by Hotelling (1935), (1936).

In Section 12.2 the canonical correlations and variates in the population are defined; they imply a linear transformation to canonical form. Maximum likelihood estimators are sample analogs. Tests of independence and of the rank of a correlation matrix are developed on the basis of asymptotic theory in Section 12.4.

Another formulation of canonical correlations and variates is made in the case of one set being random and the other set consisting of nonstochastic variables; the expected values of the random variables are linear combinations of the nonstochastic variables (Section 12.6). This is the model of Section 8.2. One set of canonical variables consists of linear combinations of the random variables and the other set consists of the nonstochastic variables; the effect of the regression of a member of the first set on a member of the second is maximized. Linear functional relationships are studied in this framework.

Simultaneous equations models are studied in Section 12.7. Estimation of a single equation in this model is formally identical to estimation of a single linear functional relationship. The limited-information maximum likelihood estimator and the two-stage least squares estimator are developed.

12.2. CANONICAL CORRELATIONS AND VARIATES IN THE POPULATION

Suppose the random vector X of p components has the covariance matrix Σ (which is assumed to be positive definite). Since we are only interested in variances and covariances in this chapter, we shall assume $\mathscr{E}X = \mathbf{0}$ when treating the population. In developing the concepts and algebra we do not need to assume that X is normally distributed, though this latter assumption will be made to develop sampling theory.

We partition X into two subvectors of p_1 and p_2 components, respectively,

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}. \tag{1}$$

For convenience we shall assume $p_1 \le p_2$. The covariance matrix is partitioned similarly into p_1 and p_2 rows and columns,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{2}$$

In the previous chapter we developed a rotation of coordinate axes to a new system in which the variance properties were clearly exhibited. Here we shall develop a transformation of the first p_1 coordinate axes and a transformation of the last p_2 coordinate axes to a new $(p_1 + p_2)$-system that will exhibit clearly the intercorrelations between $X^{(1)}$ and $X^{(2)}$.

Consider an arbitrary linear combination, $U = \boldsymbol{\alpha}' X^{(1)}$, of the components of $X^{(1)}$, and an arbitrary linear function, $V = \boldsymbol{\gamma}' X^{(2)}$, of the components of $X^{(2)}$. We first ask for the linear functions that have maximum correlation. Since the correlation of a multiple of U and a multiple of V is the same as the correlation of U and V, we can make an arbitrary normalization of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. We therefore require $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ to be such that U and V have unit variance, that is,

$$1 = \mathscr{E} U^2 = \mathscr{E} \boldsymbol{\alpha}' X^{(1)} X^{(1)\prime} \boldsymbol{\alpha} = \boldsymbol{\alpha}' \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}, \tag{3}$$

$$1 = \mathscr{E} V^2 = \mathscr{E} \boldsymbol{\gamma}' X^{(2)} X^{(2)\prime} \boldsymbol{\gamma} = \boldsymbol{\gamma}' \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma}. \tag{4}$$

We note that $\mathscr{E} U = \mathscr{E} \boldsymbol{\alpha}' X^{(1)} = \boldsymbol{\alpha}' \mathscr{E} X^{(1)} = 0$ and similarly $\mathscr{E} V = 0$. Then the correlation between U and V is

$$\mathscr{E} UV = \mathscr{E} \boldsymbol{\alpha}' X^{(1)} X^{(2)\prime} \boldsymbol{\gamma} = \boldsymbol{\alpha}' \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}. \tag{5}$$

Thus the algebraic problem is to find $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ to maximize (5) subject to (3) and (4).

Let

$$\psi = \boldsymbol{\alpha}' \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma} - \tfrac{1}{2} \lambda (\boldsymbol{\alpha}' \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} - 1) - \tfrac{1}{2} \mu (\boldsymbol{\gamma}' \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} - 1), \tag{6}$$

where λ and μ are Lagrange multipliers. We differentiate ψ with respect to the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. The vectors of derivatives set equal to zero are

$$\frac{\partial \psi}{\partial \boldsymbol{\alpha}} = \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma} - \lambda \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} = \mathbf{0}, \tag{7}$$

$$\frac{\partial \psi}{\partial \boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{12}' \boldsymbol{\alpha} - \mu \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} = \mathbf{0}. \tag{8}$$

Multiplication of (7) on the left by $\boldsymbol{\alpha}'$ and (8) on the left by $\boldsymbol{\gamma}'$ gives

$$\boldsymbol{\alpha}' \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma} - \lambda \boldsymbol{\alpha}' \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} = 0, \tag{9}$$

$$\boldsymbol{\gamma}' \boldsymbol{\Sigma}_{12}' \boldsymbol{\alpha} - \mu \boldsymbol{\gamma}' \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} = 0. \tag{10}$$

Since $\boldsymbol{\alpha}' \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} = 1$ and $\boldsymbol{\gamma}' \boldsymbol{\Sigma}_{22} \boldsymbol{\gamma} = 1$, this shows that $\lambda = \mu = \boldsymbol{\alpha}' \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}$. Thus (7) and (8) can be written as

$$-\lambda\Sigma_{11}\alpha+\Sigma_{12}\gamma=\mathbf{0}, \tag{11}$$

$$\Sigma_{21}\alpha-\lambda\Sigma_{22}\gamma=\mathbf{0}, \tag{12}$$

since $\Sigma_{12}'=\Sigma_{21}$. In one matrix equation this is

$$\begin{pmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22}\end{pmatrix}\begin{pmatrix}\alpha \\ \gamma\end{pmatrix}=\mathbf{0}. \tag{13}$$

In order that there be a nontrivial solution [which is necessary for a solution satisfying (3) and (4)], the matrix on the left must be singular; that is,

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22}\end{vmatrix}=0. \tag{14}$$

The determinant on the left is a polynomial of degree p. To demonstrate this, consider a Laplace expansion by minors of the first p_1 columns. One term is $|-\lambda\Sigma_{11}|\cdot|-\lambda\Sigma_{22}|=(-\lambda)^{p_1+p_2}|\Sigma_{11}|\cdot|\Sigma_{22}|$. The other terms in the expansion are of lower degree in λ because one or more rows of each minor in the first p_1 columns does not contain λ. Since Σ is positive definite, $|\Sigma_{11}|\cdot|\Sigma_{22}|\neq 0$ (Corollary A.1.3 of the Appendix). This shows that (14) is a polynomial equation of degree p and has p roots, say $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$. [$\alpha'$ and γ' complex conjugate in (9) and (10) prove λ real.]

From (9) we see that $\lambda=\alpha'\Sigma_{12}\gamma$ is the correlation between $U=\alpha'X^{(1)}$ and $V=\gamma'X^{(2)}$ when α and γ satisfy (13) for some value of λ. Since we want the maximum correlation, we take $\lambda=\lambda_1$. Let a solution to (13) for $\lambda=\lambda_1$ be $\alpha^{(1)},\gamma^{(1)}$, and let $U_1=\alpha^{(1)\prime}X^{(1)}$ and $V_1=\gamma^{(1)\prime}X^{(2)}$ Then U_1 and V_1 are normalized linear combinations of $X^{(1)}$ and $X^{(2)}$, respectively, with maximum correlation.

We now consider finding a second linear combination of $X^{(1)}$, say $U=\alpha'X^{(1)}$, and a second linear combination of $X^{(2)}$, say $V=\gamma'X^{(2)}$, such that of all linear combinations uncorrelated with U_1 and V_1 these have maximum correlation. This procedure is continued. At the rth step we have obtained linear combinations $U_1=\alpha^{(1)\prime}X^{(1)}$, $V_1=\gamma^{(1)\prime}X^{(2)},\dots,U_r=\alpha^{(r)\prime}X^{(1)}$, $V_r=\gamma^{(r)\prime}X^{(2)}$ with corresponding correlations [roots of (14)] $\lambda^{(1)}=\lambda_1,\lambda^{(2)},\dots,\lambda^{(r)}$. We ask for a linear combination of $X^{(1)}$, $U=\alpha'X^{(1)}$, and a linear combination of $X^{(2)}$, $V=\gamma'X^{(2)}$, that among all linear combinations uncorrelated with $U_1,V_1,\dots,U_r,V_r$, have maximum correlation. The condition that U be uncorrelated with U_i is

$$0=\mathscr{E}UU_i=\mathscr{E}\alpha'X^{(1)}X^{(1)\prime}\alpha^{(i)}=\alpha'\Sigma_{11}\alpha^{(i)}. \tag{15}$$

Then

$$\mathscr{E}UV_i = \alpha'\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma}^{(i)} = \lambda^{(i)}\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}^{(i)} = 0. \tag{16}$$

The condition that V be uncorrelated with V_i is

$$0 = \mathscr{E}VV_i = \boldsymbol{\gamma}'\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}^{(i)}. \tag{17}$$

By the same argument we have

$$\mathscr{E}VU_i = \boldsymbol{\gamma}'\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}^{(i)} = 0. \tag{18}$$

We now maximize $\mathscr{E}U_{r+1}V_{r+1}$, choosing $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ to satisfy (3), (4), (15), and (17) for $i = 1, 2, \ldots, r$. Consider

$$\begin{aligned}\psi_{r+1} &= \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma} - \tfrac{1}{2}\lambda(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} - 1) - \tfrac{1}{2}\mu(\boldsymbol{\gamma}'\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma} - 1) \\ &\quad + \sum_{i=1}^{r} \nu_i \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}^{(i)} + \sum_{i=1}^{r} \theta_i \boldsymbol{\gamma}'\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}^{(i)},\end{aligned} \tag{19}$$

where $\lambda, \mu, \nu_1, \ldots, \nu_r, \theta_1, \ldots, \theta_r$ are Lagrange multipliers. The vectors of partial derivatives of ψ_{r+1} with respect to the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are set equal to zero, giving

$$\frac{\partial \psi_{r+1}}{\partial \boldsymbol{\alpha}} = \boldsymbol{\Sigma}_{12}\boldsymbol{\gamma} - \lambda\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} + \sum_{i=1}^{r} \nu_i \boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}^{(i)} = \mathbf{0}, \tag{20}$$

$$\frac{\partial \psi_{r+1}}{\partial \boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{21}\boldsymbol{\alpha} - \mu\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma} + \sum_{i=1}^{r} \theta_i \boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}^{(i)} = \mathbf{0}. \tag{21}$$

Multiplication of (20) on the left by $\boldsymbol{\alpha}^{(j)\prime}$ and (21) on the left by $\boldsymbol{\gamma}^{(j)\prime}$ gives

$$0 = \nu_j \boldsymbol{\alpha}^{(j)\prime}\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}^{(j)} = \nu_j, \tag{22}$$

$$0 = \theta_j \boldsymbol{\gamma}^{(j)\prime}\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}^{(j)} = \theta_j. \tag{23}$$

Thus (20) and (21) are simply (11) and (12) or alternatively (13). We therefore take the largest λ_i, say, $\lambda^{(r+1)}$, such that there is a solution to (13) satisfying (3), (4), (15), and (17) for $i = 1, \ldots, r$. Let this solution be $\boldsymbol{\alpha}^{(r+1)}$, $\boldsymbol{\gamma}^{(r+1)}$, and let $U_{r+1} = \boldsymbol{\alpha}^{(r+1)\prime}X^{(1)}$ and $V_{r+1} = \boldsymbol{\gamma}^{(r+1)\prime}X^{(2)}$.

This procedure is continued step by step as long as successive solutions can be found which satisfy the conditions, namely, (13) for some λ_i, (3), (4), (15), and (17). Let m be the number of steps for which this can be done. Now we shall show that $m = p_1$, $(\le p_2)$. Let $\mathbf{A} = (\alpha^{(1)} \cdots \alpha^{(m)})$, $\Gamma_1 = (\gamma^{(1)} \cdots \gamma^{(m)})$, and

$$(24) \qquad \Lambda = \begin{pmatrix} \lambda^{(1)} & 0 & \cdots & 0 \\ 0 & \lambda^{(2)} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda^{(m)} \end{pmatrix}.$$

The conditions (3) and (15) can be summarized as

$$(25) \qquad \mathbf{A}'\boldsymbol{\Sigma}_{11}\mathbf{A} = \boldsymbol{I}.$$

Since $\boldsymbol{\Sigma}_{11}$ is of rank p_1 and $\boldsymbol{I}$ is of rank m, we have $m \le p_1$. Now let us show that $m < p_1$ leads to a contradiction by showing that in this case there is another vector satisfying the conditions. Since $\mathbf{A}'\boldsymbol{\Sigma}_{11}$ is $m \times p_1$, there exists a $p_1 \times (p_1 - m)$ matrix $\boldsymbol{E}$ (of rank $p_1 - m$) such that $\mathbf{A}'\boldsymbol{\Sigma}_{11}\boldsymbol{E} = \mathbf{0}$. Similarly there is a $p_2 \times (p_2 - m)$ matrix $\boldsymbol{F}$ (of rank $p_2 - m$) such that $\boldsymbol{\Gamma}_1'\boldsymbol{\Sigma}_{22}\boldsymbol{F} = \mathbf{0}$. We also have $\boldsymbol{\Gamma}_1'\boldsymbol{\Sigma}_{21}\boldsymbol{E} = \Lambda\mathbf{A}'\boldsymbol{\Sigma}_{11}\boldsymbol{E} = \mathbf{0}$ and $\mathbf{A}'\boldsymbol{\Sigma}_{12}\boldsymbol{F} = \Lambda\boldsymbol{\Gamma}_1'\boldsymbol{\Sigma}_{22}\boldsymbol{F} = \mathbf{0}$. Since $\boldsymbol{E}$ is of rank $p_1 - m$, $\boldsymbol{E}'\boldsymbol{\Sigma}_{11}\boldsymbol{E}$ is nonsingular (if $m < p_1$), and similarly $\boldsymbol{F}'\boldsymbol{\Sigma}_{22}\boldsymbol{F}$ is nonsingular. Thus there is at least one root of

$$(26) \qquad \begin{vmatrix} -\nu\boldsymbol{E}'\boldsymbol{\Sigma}_{11}\boldsymbol{E} & \boldsymbol{E}'\boldsymbol{\Sigma}_{12}\boldsymbol{F} \\ \boldsymbol{F}'\boldsymbol{\Sigma}_{21}\boldsymbol{E} & -\nu\boldsymbol{F}'\boldsymbol{\Sigma}_{22}\boldsymbol{F} \end{vmatrix} = 0,$$

because $|\boldsymbol{E}'\boldsymbol{\Sigma}_{11}\boldsymbol{E}| \cdot |\boldsymbol{F}'\boldsymbol{\Sigma}_{22}\boldsymbol{F}| \neq 0$. From the preceding algebra we see that there exist vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that

$$(27) \qquad \boldsymbol{E}'\boldsymbol{\Sigma}_{12}\boldsymbol{F}\boldsymbol{b} = \nu\boldsymbol{E}'\boldsymbol{\Sigma}_{11}\boldsymbol{E}\boldsymbol{a},$$

$$(28) \qquad \boldsymbol{F}'\boldsymbol{\Sigma}_{21}\boldsymbol{E}\boldsymbol{a} = \nu\boldsymbol{F}'\boldsymbol{\Sigma}_{22}\boldsymbol{F}\boldsymbol{b}.$$

Let $\boldsymbol{E}\boldsymbol{a} = \boldsymbol{g}$ and $\boldsymbol{F}\boldsymbol{b} = \boldsymbol{h}$. We now want to show that ν, $\boldsymbol{g}$, and $\boldsymbol{h}$ form a new solution $\lambda^{(m+1)}, \alpha^{(m+1)}, \gamma^{(m+1)}$. Let $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{h} = \boldsymbol{k}$. Since $\mathbf{A}'\boldsymbol{\Sigma}_{11}\boldsymbol{k} = \mathbf{A}'\boldsymbol{\Sigma}_{12}\boldsymbol{F}\boldsymbol{b} = \mathbf{0}$, $\boldsymbol{k}$ is orthogonal to the rows of $\mathbf{A}'\boldsymbol{\Sigma}_{11}$ and therefore is a linear combination of the columns of $\boldsymbol{E}$, say $\boldsymbol{E}\boldsymbol{c}$. Thus the equation $\boldsymbol{\Sigma}_{12}\boldsymbol{h} = \boldsymbol{\Sigma}_{11}\boldsymbol{k}$ can be written

$$(29) \qquad \boldsymbol{\Sigma}_{12}\boldsymbol{F}\boldsymbol{b} = \boldsymbol{\Sigma}_{11}\boldsymbol{E}\boldsymbol{c}.$$

Multiplication by $\boldsymbol{E}'$ on the left gives

$$(30) \qquad \boldsymbol{E}'\boldsymbol{\Sigma}_{12}\boldsymbol{F}\boldsymbol{b} = \boldsymbol{E}'\boldsymbol{\Sigma}_{11}\boldsymbol{E}\boldsymbol{c}.$$

Since $\boldsymbol{E}'\boldsymbol{\Sigma}_{11}\boldsymbol{E}$ is nonsingular, comparison of (27) and (30) shows that $\boldsymbol{c} = \nu\boldsymbol{a}$, and therefore $k = \nu g$. Thus

$$\Sigma_{12} h = \nu \Sigma_{11} g. \tag{31}$$

In a similar fashion we show that

$$\Sigma_{21} g = \nu \Sigma_{22} h. \tag{32}$$

Therefore $\nu = \lambda^{(m+1)}$, $g = \alpha^{(m+1)}$, $h = \gamma^{(m+1)}$ is another solution. But this is contrary to the assumption that $\lambda^{(m)}, \alpha^{(m)}, \gamma^{(m)}$ was the last possible solution. Thus $m = p_1$.

The conditions on the λ's, α's and γ's can be summarized as

$$\mathrm{A}'\Sigma_{11}\mathrm{A} = I, \tag{33}$$

$$\mathrm{A}'\Sigma_{12}\Gamma_1 = \Lambda, \tag{34}$$

$$\Gamma_1'\Sigma_{22}\Gamma_1 = I. \tag{35}$$

Let $\Gamma_2 = (\gamma^{(p_1+1)} \cdots \gamma^{(p_2)})$ be a $p_2 \times (p_2 - p_1)$ matrix satisfying

$$\Gamma_2'\Sigma_{22}\Gamma_1 = 0, \tag{36}$$

$$\Gamma_2'\Sigma_{22}\Gamma_2 = I. \tag{37}$$

Any Γ_2 can be multiplied on the right by an arbitrary $(p_2 - p_1) \times (p_2 - p_1)$ orthogonal matrix. This matrix can be formed one column at a time: $\gamma^{(p_1+1)}$ is a vector orthogonal to $\Sigma_{22}\Gamma_1$ and normalized so $\gamma^{(p_1+1)'}\Sigma_{22}\gamma^{(p_1+1)} = 1$; $\gamma^{(p_1+2)}$ is a vector orthogonal to $\Sigma_{22}(\Gamma_1 \ \ \gamma^{(p_1+1)})$ and normalized so $\gamma^{(p_1+2)'}\Sigma_{22}\gamma^{(p_1+2)} = 1$; and so forth. Let $\Gamma = (\Gamma_1 \ \ \Gamma_2)$; this square matrix is nonsingular since $\Gamma'\Sigma_{22}\Gamma = I$. Consider the determinant

$$\begin{aligned}
&\begin{vmatrix} \mathrm{A}' & 0 \\ 0 & \Gamma_1' \\ 0 & \Gamma_2' \end{vmatrix} \cdot \begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} \cdot \begin{vmatrix} \mathrm{A} & 0 & 0 \\ 0 & \Gamma_1 & \Gamma_2 \end{vmatrix} \\
&= \begin{vmatrix} -\lambda I & \Lambda & 0 \\ \Lambda & -\lambda I & 0 \\ 0 & 0 & -\lambda I \end{vmatrix} \\
&= (-\lambda)^{p_2-p_1} \begin{vmatrix} -\lambda I & \Lambda \\ \Lambda & -\lambda I \end{vmatrix} \\
&= (-\lambda)^{p_2-p_1} |-\lambda I| \cdot |-\lambda I - \Lambda(-\lambda I)^{-1}\Lambda| \\
&= (-\lambda)^{p_2-p_1} |\lambda^2 I - \Lambda^2| \\
&= (-\lambda)^{p_2-p_1} \Pi(\lambda^2 - \lambda^{(i)2}).
\end{aligned} \tag{38}$$

Except for a constant factor the above polynomial is

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix}. \tag{39}$$

Thus the roots of (14) are the roots of (38) set equal to zero, namely, $\lambda = \pm\lambda^{(i)}$, $i = 1, \dots, p_1$, and $\lambda = 0$ (of multiplicity $p_2 - p_1$). Thus $(\lambda_1, \dots, \lambda_p) = (\lambda_1, \dots, \lambda_{p_1}, 0, \dots, 0, -\lambda_{p_1}, \dots, -\lambda_1)$. The set $\{\lambda^{(i)2}\}$, $i = 1, \dots, p_1$, is the set $\{\lambda_i^2\}$, $i = 1, \dots, p_1$. To show that the set $\{\lambda^{(i)}\}$, $i = 1, \dots, p_1$, is the set $\{\lambda_i\}$, $i = 1, \dots, p_1$, we only need to show that $\lambda^{(i)}$ is nonnegative (and therefore is one of the λ_i, $i = 1, \dots, p_1$). We observe that

$$\Sigma_{12}\gamma^{(r)} = -\lambda^{(r)}\Sigma_{11}(-\alpha^{(r)}), \tag{40}$$

$$\Sigma_{21}(-\alpha^{(r)}) = -\lambda^{(r)}\Sigma_{22}\gamma^{(r)}; \tag{41}$$

thus, if $\lambda^{(r)}, \alpha^{(r)}, \gamma^{(r)}$ is a solution, so is $-\lambda^{(r)}, -\alpha^{(r)}, \gamma^{(r)}$. If $\lambda^{(r)}$ were negative, then $-\lambda^{(r)}$ would be nonnegative and $-\lambda^{(r)} \geq \lambda^{(r)}$. But since $\lambda^{(r)}$ was to be maximum, we must have $\lambda^{(r)} \geq -\lambda^{(r)}$ and therefore $\lambda^{(r)} \geq 0$. Since the set $\{\lambda^{(i)}\}$ is the same as $\{\lambda_i\}$, $i = 1, \dots, p_1$, we must have $\lambda^{(i)} = \lambda_i$.

Let

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_{p_1} \end{pmatrix} = \mathbf{A}'X^{(1)}, \tag{42}$$

$$V^{(1)} = \begin{pmatrix} V_1 \\ \vdots \\ V_{p_1} \end{pmatrix} = \Gamma_1'X^{(2)}, \tag{43}$$

$$V^{(2)} = \begin{pmatrix} V_{p_1+1} \\ \vdots \\ V_{p_2} \end{pmatrix} = \Gamma_2'X^{(2)}. \tag{44}$$

The components of U are one set of canonical variates, and the components of $V = (V^{(1)\prime} \; V^{(2)\prime})'$ are the other set. We have

$$\mathscr{E}\begin{pmatrix} U \\ V^{(1)} \\ V^{(2)} \end{pmatrix}\begin{pmatrix} U' & V^{(1)\prime} & V^{(2)\prime} \end{pmatrix} = \begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_1' \\ \mathbf{0} & \mathbf{\Gamma}_2' \end{pmatrix}\begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_1 & \mathbf{\Gamma}_2 \end{pmatrix}$$
$$= \begin{pmatrix} I_{p_1} & \Lambda & \mathbf{0} \\ \Lambda & I_{p_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{p_2-p_1} \end{pmatrix}, \tag{45}$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_{p_1} \end{pmatrix}. \tag{46}$$

Definition 12.2.1. *Let $X = (X^{(1)\prime} \; X^{(2)\prime})'$, where $X^{(1)}$ has p_1 components and $X^{(2)}$ has p_2 $(= p - p_1 \geq p_1)$ components. The* rth *pair of canonical variates is the pair of linear combinations $U_r = \alpha^{(r)\prime}X^{(1)}$ and $V_r = \gamma^{(r)\prime}X^{(2)}$, each of unit variance and uncorrelated with the first $r-1$ pairs of canonical variates and having maximum correlation. The correlation is the* rth canonical correlation.

Theorem 12.2.1. *Let $X = (X^{(1)\prime} \; X^{(2)\prime})'$ be a random vector with covariance matrix Σ. The* rth *canonical correlation between $X^{(1)}$ and $X^{(2)}$ is the* rth *largest root of* (14). *The coefficients of $\alpha^{(r)\prime}X^{(1)}$ and $\gamma^{(r)\prime}X^{(2)}$ defining the* rth *pair of canonical variates satisfy* (13) *for $\lambda = \lambda_r$ and* (3) *and* (4).

We can now verify (without differentiation) that U_1, V_1 have maximum correlation. The linear combinations $a'U = (a'\mathbf{A}')X^{(1)}$ and $b'V = (b'\mathbf{\Gamma}')X^{(2)}$ are normalized by $a'a = 1$ and $b'b = 1$. Since $\mathbf{A}$ and $\mathbf{\Gamma}$ are nonsingular, any vector α can be written as $\mathbf{A}a$ and any vector γ can be written as $\mathbf{\Gamma}b$, and hence any linear combinations $\alpha'X^{(1)}$ and $\gamma'X^{(2)}$ can be written as $a'U$ and $b'V$. The correlation between them is

$$a'(\Lambda \quad \mathbf{0})b = \sum_{i=1}^{p_1} \lambda_i a_i b_i. \tag{47}$$

Let $\lambda_i a_i / \sqrt{\Sigma(\lambda_i a_i)^2} = c_i$. Then the maximum of

$$a'(\Lambda \; \mathbf{0})b = \sqrt{\Sigma(\lambda_i a_i)^2}\,\Sigma c_i b_i$$

with respect to b is for $b_i = c_i$, since $\Sigma c_i b_i$ is the cosine of the angle between the vector b and $(c_1, \ldots, c_{p_1}, 0, \ldots, 0)$. Then (47) is

$$\sqrt{\Sigma \lambda_i^2 a_i^2} = \sqrt{\Sigma_2^{p_1}(\lambda_i^2 - \lambda_1^2) a_i^2 + \lambda_1^2},$$

and this is maximized by taking $a_i = 0$, $i = 2, \ldots, p_1$. Thus the maximized linear combinations are U_1 and V_1. In verifying that U_2 and V_2 form the second pair of canonical variates we note that lack of correlation between U_1 and a linear combination $a'U$ means $0 = \mathscr{E} U_1 a'U = \mathscr{E} U_1 \Sigma a_i U_i = a_1$ and lack of correlation between V_1 and $b'V$ means $0 = b_1$. The algebra used above gives the desired result with sums starting with $i = 2$.

We can derive a single matrix equation for α or γ. If we multiply (11) by λ and (12) by Σ_{22}^{-1}, we have

$$\lambda \Sigma_{12} \gamma = \lambda^2 \Sigma_{11} \alpha, \tag{48}$$

$$\Sigma_{22}^{-1} \Sigma_{21} \alpha = \lambda \gamma. \tag{49}$$

Substitution from (49) into (48) gives

$$\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \alpha = \lambda^2 \Sigma_{11} \alpha \tag{50}$$

or

$$(\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \lambda^2 \Sigma_{11}) \alpha = \mathbf{0}. \tag{51}$$

The quantities $\lambda_1^2, \ldots, \lambda_{p_1}^2$ satisfy

$$|\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \nu \Sigma_{11}| = 0, \tag{52}$$

and $\alpha^{(1)}, \ldots, \alpha^{(p_1)}$ satisfy (51) for $\lambda^2 = \lambda_1^2, \ldots, \lambda_{p_1}^2$, respectively. The similar equations for $\gamma^{(1)}, \ldots, \gamma^{(p_2)}$ occur when $\lambda^2 = \lambda_1^2, \ldots, \lambda_{p_2}^2$ are substituted with

$$(\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \lambda^2 \Sigma_{22}) \gamma = \mathbf{0}. \tag{53}$$

Theorem 12.2.2. *The canonical correlations are invariant with respect to transformations $X^{(i)*} = C_i X^{(i)}$, where C_i is nonsingular, $i = 1, 2$, and any function of Σ that is invariant is a function of the canonical correlations.*

Proof. Equation (14) is transformed to

(54)
$$0 = \begin{vmatrix} -\lambda C_1 \Sigma_{11} C_1' & C_1 \Sigma_{12} C_2' \\ C_2 \Sigma_{21} C_1' & -\lambda C_2 \Sigma_{22} C_2' \end{vmatrix} = \begin{vmatrix} C_1 & \mathbf{0} \\ \mathbf{0} & C_2 \end{vmatrix} \cdot \begin{vmatrix} -\lambda \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda \Sigma_{22} \end{vmatrix} \cdot \begin{vmatrix} C_1' & \mathbf{0} \\ \mathbf{0} & C_2' \end{vmatrix},$$

and hence the roots are unchanged. Conversely, let $f(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})$ be a vector-valued function of Σ such that $f(C_1\Sigma_{11}C_1', C_1\Sigma_{12}C_2', C_2\Sigma_{22}C_2') = f(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})$ for all nonsingular C_1 and C_2. If $C_1' = \mathbf{A}$ and $C_2 = \Gamma'$, then (54) is (38), which depends only on the canonical correlations. Then $f = f(I, (\Lambda, 0), I)$. ∎

We can make another interpretation of these developments in terms of prediction. Consider two random variables U and V with means 0 and variances σ_u^2 and σ_v^2 and correlation ρ. Consider approximating U by a multiple of V, say bV; then the mean squared error of approximation is

$$\begin{aligned}(55)\qquad \mathscr{E}(U - bV)^2 &= \sigma_u^2 - 2b\sigma_u\sigma_v\rho + b^2\sigma_v^2 \\ &= \sigma_u^2(1-\rho^2) + (b\sigma_v - \rho\sigma_u)^2.\end{aligned}$$

This is minimized by taking $b = \sigma_u \rho/\sigma_v$. We can consider bV as a linear prediction of U from V; then $\sigma_u^2(1-\rho^2)$ is the mean squared error of prediction. The ratio of the mean squared error of prediction to the variance of U is $\sigma_u^2(1-\rho^2)/\sigma_u^2 = 1-\rho^2$; the complement is a measure of the relative effect of V on U or the relative effectiveness of V in predicting U. Thus the greater ρ^2 or $|\rho|$ is, the more effective is V in predicting U.

Now consider the random vector X partitioned according to (1), and consider using a linear combination $V = \gamma' X^{(2)}$ to predict a linear combination $U = \alpha' X^{(1)}$. Then V predicts U best if the correlation between U and V is a maximum. Thus we can say that $\alpha^{(1)\prime} X^{(1)}$ is the linear combination of $X^{(1)}$ that can be predicted best, and $\gamma^{(1)\prime} X^{(2)}$ is the best predictor [Hotelling (1935)].

The mean squared effect of V on U can be measured as

$$(56)\qquad \mathscr{E}(bV)^2 = \rho^2 \frac{\sigma_u^2}{\sigma_v^2} \mathscr{E} V^2 = \rho^2\sigma_u^2,$$

and the relative mean squared effect can be measured by the ratio $\mathscr{E}(bV)^2/\mathscr{E}U^2 = \rho^2$. Thus maximum effect of a linear combination of $X^{(2)}$ on a linear combination of $X^{(1)}$ is made by $\gamma^{(1)\prime} X^{(2)}$ on $\alpha^{(1)\prime} X^{(1)}$.

In the special case of $p_1 = 1$, the one canonical correlation is the multiple correlation between $X^{(1)} = X_1$ and $X^{(2)}$.

The definition of canonical variates and correlations was made in terms of the covariance matrix $\Sigma = \mathscr{E}(X - \mathscr{E}X)(X - \mathscr{E}X)'$. We could extend this treatment by starting with a normally distributed vector Y with $p + p_3$ components and define X as the vector having the conditional distribution of the first p components of Y given the value of the last p_3 components. This would mean treating X_ϕ with mean $\mathscr{E}X_\phi = \Theta y_\phi^{(3)}$; the elements of the covariance matrix would be the partial covariances of the first p elements of Y.

The interpretation of canonical variates may be facilitated by considering the correlations between the canonical variates and the components of the original vectors [e.g., Darlington, Weinberg, and Wahlberg (1973)]. The covariance between the jth canonical variate U_j and X_i is

$$\mathscr{E}U_jX_i = \mathscr{E}\sum_{k=1}^{p_1}\alpha_k^{(j)}X_kX_i = \sum_{k=1}^{p_1}\alpha_k^{(j)}\sigma_{ki}. \tag{57}$$

Since the variance of U_j is 1, the correlation between U_j and X_i is

$$\operatorname{Corr}(U_j, X_i) = \frac{\sum_{k=1}^{p_1}\alpha_k^{(j)}\sigma_{ki}}{\sqrt{\sigma_{ii}}}. \tag{58}$$

An advantage of this measure is that it does not depend on the units of measurement of X_j. However, it is not a scalar multiple of the weight of X_i in U_j (namely, $\alpha_i^{(j)}$).

A special case is $\Sigma_{11} = I$, $\Sigma_{22} = I$. Then

$$\mathbf{A}'\mathbf{A} = I, \qquad \Gamma'\Gamma = I, \qquad \mathbf{A}'\Sigma_{12}\Gamma = (\Lambda \quad \mathbf{0}). \tag{59}$$

From these we obtain

$$\Sigma_{12} = \mathbf{A}(\Lambda \quad \mathbf{0})\Gamma', \tag{60}$$

where $\mathbf{A}$ and Γ are orthogonal and Λ is diagonal. This relationship is known as the *singular value decomposition* of Σ_{12}. The elements of Λ are the square roots of the characteristic roots of $\Sigma_{12}\Sigma_{12}'$, and the columns of $\mathbf{A}$ are characteristic vectors. The diagonal elements of Λ are square roots of the (possibly nonzero) roots of $\Sigma_{12}'\Sigma_{12}$, and the columns of Γ are the characteristic vectors.

12.3. ESTIMATION OF CANONICAL CORRELATIONS AND VARIATES

12.3.1. Estimation

Let $x_1, \dots, x_N$ be N observations from $N(\mu, \Sigma)$. Let x_α be partitioned into two subvectors of p_1 and p_2 components, respectively,

$$x_\alpha = \begin{pmatrix} x_\alpha^{(1)} \\ x_\alpha^{(2)} \end{pmatrix}, \qquad \alpha = 1, \dots, N. \tag{1}$$

The maximum likelihood estimator of Σ [partitioned as in (2) of Section 12.2] is

$$
(2)\quad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix} = \frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'
$$

$$
= \frac{1}{N}\begin{pmatrix} \Sigma\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)' & \Sigma\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)' \\ \Sigma\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)\left(x_\alpha^{(1)} - \bar{x}^{(1)}\right)' & \Sigma\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)\left(x_\alpha^{(2)} - \bar{x}^{(2)}\right)' \end{pmatrix}.
$$

The maximum likelihood estimators of the canonical correlations Λ and the canonical variates defined by $\mathbf{A}$ and Γ involve applying the algebra of the previous section to $\hat{\Sigma}$. The matrices Λ, $\mathbf{A}$, and Γ_1 are uniquely defined if we assume the canonical correlations different and that the first nonzero element of each column of $\mathbf{A}$ is positive. The indeterminacy in Γ_2 allows multiplication on the right by a $(p_2 - p_1)\times(p_2 - p_1)$ orthogonal matrix; this indeterminacy can be removed by various types of requirements, for example, that the submatrix formed by the lower $p_2 - p_1$ rows be upper or lower triangular with positive diagonal elements. Application of Corollary 3.2.1 then shows that the maximum likelihood estimators of $\lambda_1, \dots, \lambda_p$ are the roots of

$$
(3)\qquad \begin{vmatrix} -l\hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & -l\hat{\Sigma}_{22} \end{vmatrix} = 0,
$$

and the jth columns of $\hat{\mathbf{A}}$ and $\hat{\Gamma}_1$ satisfy

$$
(4)\qquad \begin{pmatrix} -l_j\hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & -l_j\hat{\Sigma}_{22} \end{pmatrix}\begin{pmatrix} \hat{\alpha}^{(j)} \\ \hat{\gamma}^{(j)} \end{pmatrix} = \mathbf{0},
$$

$$
(5)\qquad \hat{\alpha}^{(j)\prime}\hat{\Sigma}_{11}\hat{\alpha}^{(j)} = 1, \qquad \hat{\gamma}^{(j)\prime}\hat{\Sigma}_{22}\hat{\gamma}^{(j)} = 1.
$$

$\hat{\Gamma}_2$ satisfies

$$
(6)\qquad \hat{\Gamma}_2'\hat{\Sigma}_{22}\hat{\Gamma}_1 = \mathbf{0},
$$

$$
(7)\qquad \hat{\Gamma}_2'\hat{\Sigma}_{22}\hat{\Gamma}_2 = \boldsymbol{I}.
$$

When the other restrictions on Γ_2 are made, $\hat{\mathbf{A}}$, $\hat{\Gamma}$, and $\hat{\Lambda}$ are uniquely defined.

Theorem 12.3.1. *Let $x_1, \ldots, x_N$ be N observations from $N(\mu, \Sigma)$. Let Σ be partitioned into p_1 and p_2 ($p_1 \le p_2$) rows and columns as in* (2) *in Section* 12.2, *and let x_α be similarly partitioned as in* (1). *The maximum likelihood estimators of the canonical correlations are the roots of* (3), *where $\hat{\Sigma}_{ij}$ are defined by* (2). *The maximum likelihood estimators of the coefficients of the* jth *canonical components satisfy* (4) *and* (5), $j = 1, \ldots, p_1$; *the remaining components satisfy* (6) *and* (7).

In the population the canonical correlations and canonical variates were found in terms of maximizing correlations of linear combinations of two sets of variates. The entire argument can be carried out in terms of the sample. Thus $\hat{\alpha}^{(1)\prime}x_\alpha^{(1)}$ and $\hat{\gamma}^{(1)\prime}x_\alpha^{(2)}$ have maximum sample correlation between any linear combinations of $x_\alpha^{(1)}$ and $x_\alpha^{(2)}$, and this correlation is l_1. Similarly, $\hat{\alpha}^{(2)\prime}x_\alpha^{(1)}$ and $\hat{\gamma}^{(2)\prime}x_\alpha^{(2)}$ have the second maximum sample correlation, and so forth.

It may also be observed that we could define the sample canonical variates and correlations in terms of S, the unbiased estimator of Σ. Then $a^{(j)} = \sqrt{(N-1)/N}\,\hat{\alpha}^{(j)}$, $c^{(j)} = \sqrt{(N-1)/N}\,\hat{\gamma}^{(j)}$, and l_j satisfy

$$S_{12}c^{(j)} = l_j S_{11} a^{(j)}, \tag{8}$$

$$S_{21}a^{(j)} = l_j S_{22} c^{(j)}, \tag{9}$$

$$a^{(j)\prime}S_{11}a^{(j)} = 1, \qquad c^{(j)\prime}S_{22}c^{(j)} = 1. \tag{10}$$

We shall call the linear combinations $a^{(j)\prime}x_\alpha^{(1)}$ and $c^{(j)\prime}x_\alpha^{(2)}$ the sample canonical variates.

We can also derive the sample canonical variates from the sample correlation matrix,

$$R = \left(\frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}}\sqrt{\hat{\sigma}_{jj}}} \right) = \left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \right) = (r_{ij}) = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}. \tag{11}$$

Let

$$S_1 = \begin{pmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{s_{p_1p_1}} \end{pmatrix}, \tag{12}$$

$$(13) \qquad S_2 = \begin{pmatrix} \sqrt{s_{p_1+1,p_1+1}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{p_1+2,p_1+2}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{s_{pp}} \end{pmatrix}.$$

Then we can write (8) through (10) as

$$(14) \qquad R_{12}(S_2 c^{(j)}) = l_j R_{11}(S_1 a^{(j)}),$$

$$(15) \qquad R_{21}(S_1 a^{(j)}) = l_j R_{22}(S_2 c^{(j)}),$$

$$(16) \qquad (S_1 a^{(j)})' R_{11}(S_1 a^{(j)}) = 1, \qquad (S_2 c^{(j)})' R_{22}(S_2 c^{(j)}) = 1.$$

We can give these developments a geometric interpretation. The rows of the matrix $(x_1, \dots, x_N)$ can be interpreted as p vectors in an N-dimensional space, and the rows of $(x_1 - \bar{x}, \dots, x_N - \bar{x})$ are the p vectors projected on the $(N-1)$-dimensional subspace orthogonal to the equiangular line. Denote these as $x_1^*, \dots, x_p^*$. Any vector u^* with components $\alpha'(x_1^{(1)} - \bar{x}^{(1)}, \dots, x_N^{(1)} - \bar{x}^{(1)}) = \alpha_1 x_1^* + \cdots + \alpha_{p_1} x_{p_1}^*$ is in the p_1-space spanned by $x_1^*, \dots, x_{p_1}^*$, and a vector v^* with components $\gamma'(x_1^{(2)} - \bar{x}^{(2)}, \dots, x_N^{(2)} - \bar{x}^{(2)}) = \gamma_1 x_{p_1+1}^* + \cdots + \gamma_{p_2} x_p^*$ is in the p_2-space spanned by $x_{p_1+1}^*, \dots, x_p^*$. The cosine of the angle between these two vectors is the correlation between $u_\alpha = \alpha' x_\alpha^{(1)}$ and $v_\alpha = \gamma' x_\alpha^{(2)}$, $\alpha = 1, \dots, N$. Finding α and γ to maximize the correlation is equivalent to finding the vectors in the p_1-space and the p_2-space such that the angle between them is least (i.e., has the greatest cosine). This gives the first canonical variates, and the first canonical correlation is the cosine of the angle. Similarly, the second canonical variates correspond to vectors orthogonal to the first canonical variates and with the angle minimized.

12.3.2. Computation

We shall discuss briefly computation in terms of the population quantities. Equations (50), (51), or (52) of Section 12.2 can be used. The computation of $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ can be accomplished by solving $\Sigma_{21} = \Sigma_{22} F$ for $\Sigma_{22}^{-1}\Sigma_{21}$ and then multiplying by Σ_{12}. If p_1 is sufficiently small, the determinant $|\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \nu\Sigma_{11}|$ can be expanded into a polynomial in ν, and the polynomial equation may be solved for ν. The solutions are then inserted into (51) to arrive at the vectors α.

In many cases p_1 is too large for this procedure to be efficient. Then one can use an iterative procedure

$$(17) \qquad \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha(i) = \lambda^2(i+1)\alpha(i+1),$$

starting with an initial approximation $\boldsymbol{\alpha}(0)$; the vector $\boldsymbol{\alpha}(i+1)$ may be normalized by

$$\boldsymbol{\alpha}(i+1)'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}(i+1) = 1. \tag{18}$$

The $\lambda^2(i+1)$ converges to λ_1^2 and $\boldsymbol{\alpha}(i+1)$ converges to $\boldsymbol{\alpha}^{(1)}$ (if $\lambda_1 > \lambda_2$). This can be demonstrated in a fashion similar to that used for principal components, using

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \mathbf{A}\boldsymbol{\Lambda}^2\mathbf{A}^{-1} \tag{19}$$

from (45) of Section 12.2. See Problem 12.9.

The right-hand side of (19) is $\sum_{i=1}^{p_1}\boldsymbol{\alpha}^{(i)}\lambda_i^2\tilde{\boldsymbol{\alpha}}^{(i)\prime}$, where $\tilde{\boldsymbol{\alpha}}^{(i)\prime}$ is the ith row of $\mathbf{A}^{-1}$. From the fact that $\mathbf{A}'\boldsymbol{\Sigma}_{11}\mathbf{A} = \boldsymbol{I}$, we find that $\mathbf{A}'\boldsymbol{\Sigma}_{11} = \mathbf{A}^{-1}$ and thus $\boldsymbol{\alpha}^{(i)\prime}\boldsymbol{\Sigma}_{11} = \tilde{\boldsymbol{\alpha}}^{(i)\prime}$. Now

$$\begin{aligned}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \lambda_1^2\boldsymbol{\alpha}^{(1)}\tilde{\boldsymbol{\alpha}}^{(1)\prime} &= \sum_{i=2}^{p_1}\boldsymbol{\alpha}^{(i)}\lambda_i^2\tilde{\boldsymbol{\alpha}}^{(i)\prime} \\ &= \mathbf{A}\begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_{p_1}^2 \end{pmatrix}\mathbf{A}^{-1}.\end{aligned} \tag{20}$$

The maximum characteristic root of this matrix is λ_2^2. If we now use this matrix for iteration, we will obtain λ_2^2 and $\boldsymbol{\alpha}^{(2)}$. The procedure is continued to find as many λ_i^2 and $\boldsymbol{\alpha}^{(i)}$ as desired.

Given λ_i and $\boldsymbol{\alpha}^{(i)}$, we find $\boldsymbol{\gamma}^{(i)}$ from $\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}^{(i)} = \lambda_i\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma}^{(i)}$ or equivalently $(1/\lambda_i)\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}^{(i)} = \boldsymbol{\gamma}^{(i)}$. A check on the computations is provided by comparing $\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma}^{(i)}$ and $\lambda_i\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}^{(i)}$.

For the sample we perform these calculations with $\hat{\boldsymbol{\Sigma}}_{ij}$ or S_{ij} substituted for $\boldsymbol{\Sigma}_{ij}$. It is often convenient to use R_{ij} in the computation (because $-1 < r_{ij} < 1$) to obtain $S_1 a^{(j)}$ and $S_2 c^{(j)}$; from these $a^{(j)}$ and $c^{(j)}$ can be computed.

Modern computational procedures are available for canonical correlations and variates similar to those sketched for principal components. Let

$$\boldsymbol{Z}_1 = \left(x_1^{(1)} - \bar{x}^{(1)}, \dots, x_N^{(1)} - \bar{x}^{(1)}\right), \tag{21}$$

$$Z_2 = \left(x_1^{(2)} - \bar{x}^{(2)}, \dots, x_N^{(2)} - \bar{x}^{(2)}\right). \tag{22}$$

The QR decomposition of the transpose of these matrices (Section 11.4) is $Z_i' = Q_i R_i$, where $Q_i' Q_i = I_{p_i}$ and R_i is upper triangular. Then $S_{ij} = Z_i Z_j' = R_i' Q_i' Q_j R_j$, $i, j = 1, 2$, and $S_{ii} = R_i' R_i$, $i = 1, 2$. The canonical correlations are the singular values of $Q_1' Q_2$ and the square roots of the characteristic roots of $(Q_1' Q_2)(Q_1' Q_2)'$ (by Theorem 12.2.2). Then the singular value decomposition of $Q_1' Q_2$ is $P(L \;\; O)T$, where P and T are orthogonal and L is diagonal. To effect the decomposition Householder transformations are applied to the left and right of $Q_1' Q_2$ to obtain an upper bidiagonal matrix, that is, a matrix with entries on the main diagonal and first superdiagonal. Givens matrices are used to reduce this matrix to a matrix that is diagonal to the degree of approximation required. For more detail see Kennedy and Gentle (1980), Section 7.2 and 12.2, Chambers (1977), Björck and Golub (1973), Golub and Luk (1976), and Golub and Van Loan (1989).

12.4. STATISTICAL INFERENCE

12.4.1. Tests of Independence and of Rank

In Chapter 9 we considered testing the null hypothesis that $X^{(1)}$ and $X^{(2)}$ are independent, which is equivalent to the null hypothesis that $\mathbf{\Sigma}_{12} = \mathbf{0}$. Since $\mathbf{A}' \mathbf{\Sigma}_{12} \mathbf{\Gamma} = (\mathbf{\Lambda} \;\; \mathbf{0})$, it is seen that the hypothesis is equivalent to $\mathbf{\Lambda} = \mathbf{0}$, that is, $\rho_1 = \cdots = \rho_{p_1} = 0$. The likelihood ratio criterion for testing this null hypothesis is the $N/2$ power of

$$(1) \qquad \frac{\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix}}{|A_{11}| \cdot |A_{22}|} = \frac{\begin{vmatrix} \hat{\mathbf{A}}' & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{\Gamma}}' \end{vmatrix}}{|\hat{\mathbf{A}}'| \cdot |\hat{\mathbf{\Gamma}}'|} \cdot \frac{\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix}}{|A_{11}| \cdot |A_{22}|} \cdot \frac{\begin{vmatrix} \hat{\mathbf{A}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{\Gamma}} \end{vmatrix}}{|\hat{\mathbf{A}}| \cdot |\hat{\mathbf{\Gamma}}|}$$

$$= \frac{\begin{vmatrix} I & \hat{\mathbf{\Lambda}} & \mathbf{0} \\ \hat{\mathbf{\Lambda}} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I \end{vmatrix}}{|I| \cdot |I|} = \begin{vmatrix} I & \hat{\mathbf{\Lambda}} \\ \hat{\mathbf{\Lambda}} & I \end{vmatrix} = |I - \hat{\mathbf{\Lambda}}^2| = \prod_{i=1}^{p_1} (1 - r_i^2),$$

where $r_1 = l_1 \geq \cdots \geq r_{p_1} = l_{p_1} \geq 0$ are the p_1 possibly nonzero sample canonical correlations. Under the null hypothesis, the limiting distribution of Bartlett's modification of -2 times the logarithm of the likelihood ratio criterion, namely,

$$(2) \qquad -\left[N - \tfrac{1}{2}(p+3)\right] \sum_{i=1}^{p_1} \log(1 - r_i^2),$$

is χ^2 with $p_1 p_2$ degrees of freedom. (See Section 9.4.) Note that it is approximately

$$N \sum_{i=1}^{p_1} r_i^2 = N \operatorname{tr} A_{11}^{-1} A_{12} A_{21}^{-1} A_{21}, \tag{3}$$

which is N times Nagao's criterion [(2) of Section 9.5].

If $\Sigma_{12} \neq \mathbf{0}$, an interesting question is how many population canonical correlations are different from 0; that is, how many canonical variates are needed to explain the correlations between $X^{(1)}$ and $X^{(2)}$? The number of nonzero canonical correlations is equal to the rank of Σ_{12}. The likelihood ratio criterion for testing the null hypothesis $H_k: \rho_{k+1} = \cdots = \rho_{p_1} = 0$, that is, that the rank of Σ_{12} is not greater than k, is $\prod_{i=k+1}^{p_1}(1-r_i^2)^{\frac{1}{2}N}$ [Fujikoshi (1974)]. Under the null hypothesis

$$-\left[N - \tfrac{1}{2}(p+3)\right] \sum_{i=k+1}^{p_1} \log(1-r_i^2) \tag{4}$$

has approximately the χ^2-distribution with $(p_1-k)(p_2-k)$ degrees of freedom. [Glynn and Muirhead (1978) suggest multiplying the sum in (4) by $N - k - \frac{1}{2}(p+3) + \sum_{i=1}^{k}(1/r_i^2)$; see also Lawley (1959).]

To determine the numbers of nonzero and zero population canonical correlations one can test that all the roots are 0; if that hypothesis is rejected, test that the $p_1 - 1$ smallest roots are 0; etc. Of course, these procedures are not statistically independent, even asymptotically. Alternatively, one could use a sequence of tests in the opposite direction: Test $\rho_{p_1} = 0$, then $\rho_{p_1-1} = \rho_{p_1} = 0$, and so on, until a hypothesis is rejected or until $\Sigma_{12} = \mathbf{0}$ is accepted. Yet another procedure (which can only be carried out for small p_1) is to test $\rho_{p_1} = 0$, then $\rho_{p_1-1} = 0$, and so forth. In this procedure one would use r_j to test the hypothesis $\rho_j = 0$. The relevant asymptotic distribution will be discussed in Section 12.4.2.

12.4.2. Distributions of Canonical Correlations

The density of the canonical correlations is given in Section 13.4 for the case that $\Sigma_{12} = \mathbf{0}$, that is, all the population correlations are 0. The density when some population correlations are different from 0 has been given by Constantine (1963) in terms of a hypergeometric function of two matrix arguments.

The large-sample theory is more manageable. Suppose the first k canonical correlations are positive, less than 1, and different, and suppose that

$p_1 - k$ correlations are 0. Let

$$(5)\qquad \begin{aligned} z_i &= \sqrt{N}\,\frac{r_i^2 - \rho_i^2}{2\rho_i(1-\rho_i^2)}, & i &= 1,\dots,k,\\ z_i &= Nr_i^2, & i &= k+1,\dots,p_1. \end{aligned}$$

Then in the limiting distribution $z_1,\dots,z_k$ and the set $z_{k+1},\dots,z_{p_1}$ are mutually independent, z_i has the limiting distribution $N(0,1)$, $i = 1,\dots,k$, and the density of the limiting distribution of $z_{k+1},\dots,z_{p_1}$ is

$$(6)\qquad \frac{\pi^{\frac{1}{2}(p_1-k)^2}\exp\left(-\frac{1}{2}\sum_{i=k+1}^{p_1} z_i\right)}{2^{\frac{1}{2}(p_1-k)(p_2-k)}\Gamma_{p_1-k}\left[\frac{1}{2}(p_1-k)\right]\Gamma_{p_1-k}\left[\frac{1}{2}(p_2-k)\right]} \cdot \prod_{i=k+1}^{p_1} z_i^{\frac{1}{2}(p_2-p_1-1)} \prod_{\substack{i,j=k+1\\ i<j}}^{p_1} (z_i - z_j).$$

This is the density (11) of Section 13.3 of the characteristic roots of a (p_1-k)-order matrix with distribution $W(I_{p_1-k}, p_2-k)$. Note that the normalizing factor for the squared correlations corresponding to nonzero population correlations is $\sqrt{N}$, while the factor corresponding to zero population correlation is N. See Chapter 13.

In large samples we treat r_i^2 as $N[\rho_i^2,(1/N)4\rho_i^2(1-\rho_i^2)^2]$ or r_i as $N[\rho_i,(1/N)(1-\rho_i^2)^2]$ (by Theorem 4.2.3) to obtain tests of ρ_i or confidence intervals for ρ_i. Lawley (1959) has shown that the transformation $z_i = \tanh^{-1}(r_i)$ [see Section 4.2.3] does not stabilize the variance and has a significant bias in estimating $\zeta_i = \tanh^{-1}(\rho_i)$.

12.5. AN EXAMPLE

In this section we consider a simple illustrative example. Rao [(1952), p. 245] gives some measurements on the first and second adult sons in a sample of 25 families. (These have been used in Problem 3.1 and Problem 4.41.) Let $x_{1\alpha}$ be the head length of the first son in the αth family, $x_{2\alpha}$ be the head breadth of the first son, $x_{3\alpha}$ be the head length of the second son, and $x_{4\alpha}$ be the head breadth of the second son. We shall investigate the relations between the measurements for the first son and for the second. Thus $x_\alpha^{(1)\prime} = (x_{1\alpha}, x_{2\alpha})$

and $x_\alpha^{(2)\prime} = (x_{3\alpha}, x_{4\alpha})$. The data can be summarized as[†]

(1) $\bar{x}' = (185.72, 151.12, 183.84, 149.24)$,

$$S = \begin{pmatrix} 95.2933 & 52.8683 & 69.6617 & 46.1117 \\ 52.8683 & 54.3600 & 51.3117 & 35.0533 \\ 69.6617 & 51.3117 & 100.8067 & 56.5400 \\ 46.1117 & 35.0533 & 56.5400 & 45.0233 \end{pmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

The matrix of correlations is

$$(2) \qquad R = \left(\begin{array}{cc:cc} 1.0000 & 0.7346 & 0.7108 & 0.7040 \\ 0.7346 & 1.0000 & 0.6932 & 0.7086 \\ \hdashline 0.7108 & 0.6932 & 1.0000 & 0.8392 \\ 0.7040 & 0.7086 & 0.8392 & 1.0000 \end{array}\right) = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}.$$

All of the correlations are about 0.7 except for the correlation between the two measurements on second sons. In particular, R_{12} is nearly of rank one, and hence the second canonical will be near zero.

We compute

$$(3) \qquad R_{22}^{-1} R_{21} = \begin{pmatrix} 0.405\,769 & 0.333\,205 \\ 0.363\,480 & 0.428\,976 \end{pmatrix},$$

$$(4) \qquad R_{12} R_{22}^{-1} R_{21} = \begin{pmatrix} 0.544\,311 & 0.538\,841 \\ 0.538\,841 & 0.534\,950 \end{pmatrix}.$$

The determinantal equation is

$$(5) \qquad 0 = \begin{vmatrix} 0.544\,311 - 1.0000\nu & 0.538\,841 - 0.7346\nu \\ 0.538\,841 - 0.7346\nu & 0.534\,950 - 1.0000\nu \end{vmatrix}$$

$$= 0.460\,363\nu^2 - 0.287\,596\nu + 0.000\,830.$$

The roots are 0.621 816 and 0.002 900; thus $l_1 = 0.788\,553$ and $l_2 = 0.053\,852$ Corresponding to these roots are the vectors

$$(6) \qquad S_1 a^{(1)} = \begin{pmatrix} 0.552\,166 \\ 0.521\,548 \end{pmatrix}, \qquad S_1 a^{(2)} = \begin{pmatrix} 1.366\,501 \\ -1.378\,467 \end{pmatrix},$$

where

$$(7) \qquad S_1 = \begin{pmatrix} \sqrt{s_{11}} & 0 \\ 0 & \sqrt{s_{22}} \end{pmatrix} = \begin{pmatrix} 9.7618 & 0 \\ 0 & 7.3729 \end{pmatrix}.$$

[†] Rao's computations are in error; his last "difference" is incorrect.

We apply $(1/l_i)R_{22}^{-1}R_{21}$ to $S_1 a^{(i)}$ to obtain

$$(8) \qquad S_2 c^{(1)} = \begin{pmatrix} 0.504\,511 \\ 0.538\,242 \end{pmatrix}, \qquad S_2 c^{(2)} = \begin{pmatrix} 1.767\,281 \\ -1.757\,288 \end{pmatrix},$$

where

$$(9) \qquad S_2 = \begin{pmatrix} \sqrt{s_{33}} & 0 \\ 0 & \sqrt{s_{44}} \end{pmatrix} = \begin{pmatrix} 10.0402 & 0 \\ 0 & 6.7099 \end{pmatrix}.$$

We check these computations by calculating

$$(10) \quad \frac{1}{l_1} R_{11}^{-1} R_{12}\left(S_2 c^{(1)}\right) = \begin{pmatrix} 0.552\,157 \\ 0.521\,560 \end{pmatrix}, \quad \frac{1}{l_2} R_{11}^{-1} R_{12}\left(S_2 c^{(2)}\right) = \begin{pmatrix} 1.365\,151 \\ -1.376\,741 \end{pmatrix}.$$

The first vector in (10) corresponds closely to the first vector in (6); in fact, it is a slight improvement, for the computation is equivalent to an iteration on $S_1 a^{(1)}$. The second vector in (10) does not correspond as closely to the second vector in (6). One reason is that l_2 is correct to only four or five significant figures (as is $\nu_2 = l_2^2$) and thus the components of $S_2 c^{(2)}$ can be correct to only as many significant figures; secondly, the fact that $S_2 c^{(2)}$ corresponds to the smaller root means that the iteration decreases the accuracy instead of increasing it. Our final results are

$$(11) \qquad \begin{array}{lcc} & (1) & (2) \\ l_i = & 0.789, & 0.054, \\ a^{(i)} = & \begin{pmatrix} 0.0566 \\ 0.0707 \end{pmatrix}, & \begin{pmatrix} 0.1400 \\ -0.1870 \end{pmatrix}, \\ c^{(i)} = & \begin{pmatrix} 0.0502 \\ 0.0802 \end{pmatrix}, & \begin{pmatrix} 0.1760 \\ -0.2619 \end{pmatrix}. \end{array}$$

The larger of the two canonical correlations, 0.789, is larger than any of the individual correlations of a variable of the first set with a variable of the other. The second canonical correlation is very near zero. This means that to study the relation between two head dimensions of first sons and second sons we can confine our attention to the first canonical variates; the second canonical variates are correlated only slightly. The first canonical variate in each set is approximately proportional to the sum of the two measurements divided by their respective standard deviations; the second canonical variate in each set is approximately proportional to the difference of the two standardized measurements.

12.6. LINEARLY RELATED EXPECTED VALUES

12.6.1. Canonical Analysis of Regression Matrices

In this section we develop canonical correlations and variates for one stochastic vector and one nonstochastic vector. The expected value of the stochastic vector is a linear function of the nonstochastic vector (Chapter 8). We find new coordinate systems so that the expected value of each coordinate of the stochastic vector depends on only one coordinate of the nonstochastic vector; the coordinates of the stochastic vector are uncorrelated in the stochastic sense, and the coordinates of the nonstochastic vector are uncorrelated in the sample. The coordinates are ordered according to the effect sum of squares relative to the variance. The algebra is similar to that developed in Section 12.2.

If X has the normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with X, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ partitioned as in (1) and (2) of Section 12.2 and $\boldsymbol{\mu} = (\boldsymbol{\mu}^{(1)\prime}, \boldsymbol{\mu}^{(2)\prime})'$, the conditional distribution of $X^{(1)}$ given $x^{(2)}$ is normal with mean

$$\boldsymbol{\mu}^{(1)} + \boldsymbol{\beta}\left(x^{(2)} - \boldsymbol{\mu}^{(2)}\right), \qquad \boldsymbol{\beta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}, \tag{1}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{2}$$

Since we consider a set of random vectors $X_1^{(1)}, \dots, X_N^{(1)}$ with expected values depending on $x_1^{(2)}, \dots, x_N^{(2)}$ (nonstochastic), we can write the conditional expected value of $X_\phi^{(1)}$ as $\boldsymbol{\tau} + \boldsymbol{\beta}(x_\phi^{(2)} - \bar{x}^{(2)})$, where $\boldsymbol{\tau} = \boldsymbol{\mu}^{(1)} + \boldsymbol{\beta}(\bar{x}^{(2)} - \boldsymbol{\mu}^{(2)})$ can be considered as a parameter vector. This is the model of Section 8.2 with a slight change of notation.

The model of this section is

$$\mathscr{E}X_\phi^{(1)} = \boldsymbol{\tau} + \boldsymbol{\beta}\left(x_\phi^{(2)} - \bar{x}^{(2)}\right), \qquad \phi = 1, \dots, N, \tag{3}$$

where $x_1^{(2)}, \dots, x_N^{(2)}$ are a set of nonstochastic vectors $(q \times 1)$ and $\bar{x}^{(2)} = N^{-1}\sum_{\phi=1}^N x_\phi^{(2)}$. The covariance matrix is

$$\mathscr{E}\left(X_\phi^{(1)} - \mathscr{E}X_\phi^{(1)}\right)\left(X_\phi^{(1)} - \mathscr{E}X_\phi^{(1)}\right)' = \boldsymbol{\Psi}. \tag{4}$$

Consider a linear combination of $X_\phi^{(1)}$, say $U_\phi = \boldsymbol{\alpha}'X_\phi^{(1)}$. Then U_ϕ has variance $\boldsymbol{\alpha}'\boldsymbol{\Psi}\boldsymbol{\alpha}$ and expected value

$$\mathscr{E}U_\phi = \boldsymbol{\alpha}'\boldsymbol{\tau} + \boldsymbol{\alpha}'\boldsymbol{\beta}\left(x_\phi^{(2)} - \bar{x}^{(2)}\right). \tag{5}$$

The mean expected value is $(1/N)\sum_{\phi=1}^{N} \mathscr{E} U_\phi = \boldsymbol{\alpha}'\boldsymbol{\tau}$, and the *mean sum of squares due to* $\boldsymbol{x}^{(2)}$ is

$$(6) \qquad \frac{1}{n}\sum_{\phi=1}^{N}\left(\mathscr{E} U_\phi - \boldsymbol{\alpha}'\boldsymbol{\tau}\right)^2 = \frac{1}{n}\sum_{\phi=1}^{N} \boldsymbol{\alpha}'\boldsymbol{\beta}\left(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)\left(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)'\boldsymbol{\beta}'\boldsymbol{\alpha}$$
$$= \boldsymbol{\alpha}'\boldsymbol{\beta} S_{22}\boldsymbol{\beta}'\boldsymbol{\alpha}.$$

We can ask for the linear combination that maximizes the mean sum of squares relative to its variance; that is, the linear combination of dependent variables on which the independent variables have greatest effect. We want to maximize (6) subject to $\boldsymbol{\alpha}'\boldsymbol{\Psi}\boldsymbol{\alpha} = 1$. That leads to the vector equation

$$(7) \qquad (\boldsymbol{\beta} S_{22}\boldsymbol{\beta}' - \kappa\boldsymbol{\Psi})\boldsymbol{\alpha} = \mathbf{0}$$

for κ satisfying

$$(8) \qquad |\boldsymbol{\beta} S_{22}\boldsymbol{\beta}' - \kappa\boldsymbol{\Psi}| = 0.$$

Multiplication of (7) on the left by $\boldsymbol{\alpha}'$ shows that $\boldsymbol{\alpha}'\boldsymbol{\beta} S_{22}\boldsymbol{\beta}'\boldsymbol{\alpha} = \kappa$ for $\boldsymbol{\alpha}$ and κ satisfying $\boldsymbol{\alpha}'\boldsymbol{\Psi}\boldsymbol{\alpha} = 1$ and (7); to obtain the maximum we take the largest root of (8), say κ_1. Denote this vector by $\boldsymbol{\alpha}^{(1)}$, and the corresponding random variable by $U_{1\phi} = \boldsymbol{\alpha}^{(1)\prime}X_\phi^{(1)}$. The expected value of this first canonical variable is $\mathscr{E} U_{1\phi} = \boldsymbol{\alpha}^{(1)\prime}[\boldsymbol{\beta}(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)}) + \tau]$. Let $\boldsymbol{\alpha}^{(1)\prime}\boldsymbol{\beta} = k\boldsymbol{\gamma}^{(1)\prime}$, where k is determined so

$$(9) \qquad 1 = \frac{1}{n}\sum_{\phi=1}^{N}\left(\boldsymbol{\gamma}^{(1)\prime}\boldsymbol{x}_\phi^{(2)} - \frac{1}{N}\sum_{\eta=1}^{N}\boldsymbol{\gamma}^{(1)\prime}\boldsymbol{x}_\eta^{(2)}\right)^2$$
$$= \frac{1}{n}\sum_{\phi=1}^{N}\boldsymbol{\gamma}^{(1)\prime}\left(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)\left(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)'\boldsymbol{\gamma}^{(1)}$$
$$= \boldsymbol{\gamma}^{(1)\prime} S_{22}\boldsymbol{\gamma}^{(1)}.$$

Then $k = \sqrt{\kappa_1}$. Let $U_{1\phi} = \boldsymbol{\gamma}^{(1)\prime}(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)})$. Then $\mathscr{E} U_{1\phi} = \sqrt{\kappa_1}\, v_\phi^{(1)} + \boldsymbol{\alpha}^{(1)\prime}\boldsymbol{\tau}$.

Next let us obtain a linear combination $U_\phi = \boldsymbol{\alpha}'X_\phi^{(1)}$ that has maximum effect sum of squares among all linear combinations with variance 1 and uncorrelated with $U_{1\phi}$, that is, $0 = \mathscr{E}(U_\phi - \mathscr{E} U_\phi)(U_{1\phi} - \mathscr{E} U_{1\phi})' = \boldsymbol{\alpha}'\boldsymbol{\Psi}\boldsymbol{\alpha}^{(1)}$. As in Section 12.2, we can set up this maximization problem with Lagrange multipliers and find that $\boldsymbol{\alpha}$ satisfies (7) for some κ satisfying (8) and $\boldsymbol{\alpha}'\boldsymbol{\Psi}\boldsymbol{\alpha} = 1$. The process is continued in a manner similar to that in Section 12.2. We summarize the results.

The jth canonical random variable is $U_{j\phi} = \boldsymbol{\alpha}^{(j)\prime}X_\phi^{(1)}$, where $\boldsymbol{\alpha}^{(j)}$ satisfies (7) for $\kappa = \kappa_j$ and $\boldsymbol{\alpha}^{(j)\prime}\boldsymbol{\Psi}\boldsymbol{\alpha}^{(j)} = 1$; $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_{p_1}$ are the roots of (8).

We shall assume the rank of $\boldsymbol{\beta}$ is $p_1 \le p_2$. (Then $\kappa_{p_1} > 0$.) $U_{j\phi}$ has the largest effect sum of squares of linear combinations that have unit variance and are uncorrelated with $U_{1\phi}, \ldots, U_{j-1,\phi}$. Let $\boldsymbol{\gamma}^{(j)} = (1/\sqrt{\kappa_j})\boldsymbol{\beta}'\boldsymbol{\alpha}^{(j)}$, $\nu_j = \boldsymbol{\alpha}^{(j)\prime}\boldsymbol{\tau}$, and $v_{j\phi} = \boldsymbol{\gamma}^{(j)\prime}(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)})$. Then

$$\mathscr{E}U_{j\phi} = \sqrt{\kappa_j}\, v_{j\phi} + \nu_j, \tag{10}$$

$$\frac{1}{n}\sum_{\phi=1}^{N}\left(v_{j\phi} - \frac{1}{N}\sum_{\eta=1}^{N} v_{j\eta}\right)^2 = 1, \tag{11}$$

$$\sum_{\phi=1}^{N}\left(v_{i\phi} - \frac{1}{N}\sum_{\eta=1}^{N} v_{i\eta}\right)\left(v_{j\phi} - \frac{1}{N}\sum_{\eta=1}^{N} v_{j\eta}\right) = 0, \qquad i \ne j. \tag{12}$$

If $p_2 > p_1$, then $\boldsymbol{\gamma}^{(p_1+1)}, \ldots, \boldsymbol{\gamma}^{(p_2)}$ can be chosen so $v_{p_1+1,\phi} = \boldsymbol{\gamma}^{(p_1+1)\prime}(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)}), \ldots, v_{p_2\phi} = \boldsymbol{\gamma}^{(p_2)\prime}(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)})$ satisfy (11) and (12).

Let $\mathbf{A} = (\boldsymbol{\alpha}^{(1)} \cdots \boldsymbol{\alpha}^{(p_1)})$, $\boldsymbol{\Gamma}_1 = (\boldsymbol{\gamma}^{(1)} \cdots \boldsymbol{\gamma}^{(p_1)})$, $\boldsymbol{\Gamma}_2 = (\boldsymbol{\gamma}^{(p_1+1)} \cdots \boldsymbol{\gamma}^{(p_2)})$, $\boldsymbol{\Delta} = \operatorname{diag}(\delta_1, \ldots, \delta_{p_1}) = \operatorname{diag}(\sqrt{\kappa_1}, \ldots, \sqrt{\kappa_{p_1}})$, $U_\phi = \mathbf{A}'X_\phi^{(1)}$, $\boldsymbol{v}_\phi^{(1)} = \boldsymbol{\Gamma}_1'(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)})$, $\boldsymbol{v}_\phi^{(2)} = \boldsymbol{\Gamma}_2'(\boldsymbol{x}_\phi^{(2)} - \bar{\boldsymbol{x}}^{(2)})$, and $\boldsymbol{v}_\phi = (\boldsymbol{v}_\phi^{(1)\prime}\boldsymbol{v}_\phi^{(2)\prime})'$, $\phi = 1, \ldots, N$. Then

$$\mathscr{E}(U_\phi - \mathscr{E}U_\phi)(U_\phi - \mathscr{E}U_\phi)' = \mathbf{A}'\boldsymbol{\Psi}\mathbf{A} = \boldsymbol{I}, \tag{13}$$

$$\mathscr{E}U_\phi = \boldsymbol{\Delta}\boldsymbol{v}_\phi^{(1)} + \boldsymbol{\nu}, \qquad \phi = 1, \ldots, N, \tag{14}$$

$$\frac{1}{n}\sum_{\phi=1}^{N}\left(\boldsymbol{v}_\phi - \frac{1}{N}\sum_{\eta=1}^{N}\boldsymbol{v}_\eta\right)\left(\boldsymbol{v}_\phi - \frac{1}{N}\sum_{\eta=1}^{N}\boldsymbol{v}_\eta\right)' = \boldsymbol{I}. \tag{15}$$

The random canonical variates are uncorrelated and have variance 1. The expected value of each random canonical variate is a multiple of the corresponding nonstochastic canonical variate plus a constant. The nonstochastic canonical variates have sample variance 1 and are uncorrelated in the sample.

If $p_1 > p_2$, the maximum rank of $\boldsymbol{\beta}$ is p_2 and $\kappa_{p_2+1} = \cdots = \kappa_{p_1} = 0$. In that case we define $\mathbf{A}_1 = (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(p_2)})$ and $\mathbf{A}_2 = (\boldsymbol{\alpha}^{(p_2+1)}, \ldots, \boldsymbol{\alpha}^{(p_1)})$, where $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(p_2)}$ (corresponding to positive κ's) are defined as before and $\boldsymbol{\alpha}^{(p_2+1)}, \ldots, \boldsymbol{\alpha}^{(p_1)}$ are any vectors satisfying $\boldsymbol{\alpha}^{(j)\prime}\boldsymbol{\Psi}\boldsymbol{\alpha}^{(j)} = 1$ and $\boldsymbol{\alpha}^{(j)\prime}\boldsymbol{\Psi}\boldsymbol{\alpha}^{(i)} = 0$, $i \ne j$. Then $\mathscr{E}U_\phi^{(i)} = \delta_i v_\phi^{(i)} + \nu_i$, $i = 1, \ldots, p_2$, and $\mathscr{E}U_\phi^{(i)} = \nu_i$, $i = p_2 + 1, \ldots, p_1$.

In either case if the rank of $\boldsymbol{\beta}$ is $r \le \min(p_1, p_2)$, there are r roots of (8) that are nonzero and hence $\mathscr{E}U_\phi^{(i)} = \delta_i v_\phi^{(i)} + \nu_i$ for $i = 1, \ldots, r$.

12.6.2. Estimation

Let $x_1^{(1)}, \dots, x_N^{(1)}$ be a set of observations on $X_1^{(1)}, \dots, X_N^{(1)}$ with the probability structure developed in Section 12.6.1, and let $x_1^{(2)}, \dots, x_N^{(2)}$ be the set of corresponding independent variates. Then we can estimate τ, β, and Ψ by

$$\hat{\tau} = \frac{1}{N} \sum_{\phi=1}^{N} x_\phi^{(1)} = \bar{x}^{(1)}, \tag{16}$$

$$\hat{\beta} = A_{12} A_{22}^{-1} = S_{12} S_{22}^{-1}, \tag{17}$$

$$\begin{aligned} \tilde{\Psi} &= \frac{1}{n} \sum_{\phi=1}^{N} \left[x_\phi^{(1)} - \bar{x}^{(1)} - \hat{\beta}\left(x_\phi^{(2)} - \bar{x}^{(2)}\right)\right]\left[x_\phi^{(1)} - \bar{x}^{(1)} - \hat{\beta}\left(x_\phi^{(2)} - \bar{x}^{(2)}\right)\right]' \\ &= \frac{1}{n}\left(A_{11} - A_{12} A_{22}^{-1} A_{21}\right) = S_{11} - S_{12} S_{22}^{-1} S_{21}, \end{aligned} \tag{18}$$

where the A's and S's are defined as before. (It is convenient to divide by $n = N-1$ instead of by N; the latter would yield maximum likelihood estimators.)

The sample analogs of (7) and (8) are

$$\begin{aligned} 0 &= \left(\hat{\beta} S_{22} \hat{\beta}' - k\tilde{\Psi}\right)\bar{a} \\ &= \left[S_{12} S_{22}^{-1} S_{21} - k\left(S_{11} - S_{12} S_{22}^{-1} S_{21}\right)\right]\bar{a}, \end{aligned} \tag{19}$$

$$\begin{aligned} 0 &= |\hat{\beta} S_{22} \hat{\beta}' - k\tilde{\Psi}| \\ &= |S_{12} S_{22}^{-1} S_{21} - k\left(S_{11} - S_{12} S_{22}^{-1} S_{21}\right)|. \end{aligned} \tag{20}$$

The roots $k_1 \geq \cdots \geq k_{p_1}$ of (20) estimate the roots $\kappa_1 \geq \cdots \geq \kappa_{p_1}$ of (8), and the corresponding solutions $\bar{a}^{(1)}, \dots, \bar{a}^{(p_1)}$ of (19), normalized by $\bar{a}^{(i)\prime}\tilde{\Psi}\bar{a}^{(i)} = 1$, estimate $\alpha^{(1)}, \dots, \alpha^{(p_1)}$. Then $\bar{c}^{(j)} = (1/\sqrt{k_j})\hat{\beta}'\bar{a}^{(j)}$ estimates $\gamma^{(j)}$, and $n_j = \bar{a}^{(j)\prime}\bar{x}^{(1)}$ estimates ν_j. The sample canonical variates are $\bar{a}^{(j)\prime}x_\phi^{(1)}$ and $\bar{c}^{(j)\prime}(x_\phi^{(2)} - \bar{x}^{(2)})$, $j = 1, \dots, p_1$, $\phi = 1, \dots, N$. If $p_1 > p_2$, then $p_1 - p_2$ more $\bar{a}^{(j)}$'s can be defined satisfying $\bar{a}^{(j)\prime}\tilde{\Psi}\bar{a}^{(j)} = 1$ and $\bar{a}^{(j)\prime}\tilde{\Psi}\bar{a}^{(i)} = 0$, $i \neq j$.

12.6.3. Relations Between Canonical Variates

In Section 12.3, the roots $l_1 \geq \cdots \geq l_{p_1}$ were defined to satisfy

$$0 = \begin{vmatrix} -lS_{11} & S_{12} \\ S_{21} & -lS_{22} \end{vmatrix} = (-1)^{p_2} l^{p_2 - p_1} |S_{22}|\,|S_{12} S_{22}^{-1} S_{21} - l^2 S_{11}|. \tag{21}$$

Since (20) can be written

$$0 = \left|(1+k)S_{12}S_{22}^{-1}S_{21} - kS_{11}\right|, \tag{22}$$

we see that $l_i^2 = k_i/(1+k_i)$ and $k_i = l_i^2/(1-l_i^2)$, $i = 1, \dots, p_1$. The vector $a^{(i)}$ in Section 12.3 satisfies

$$\begin{aligned} \mathbf{0} &= \left(S_{12}S_{22}^{-1}S_{21} - l_i^2 S_{11}\right)a^{(i)} \\ &= \left(S_{12}S_{22}^{-1}S_{21} - \frac{k_i}{1+k_i}S_{11}\right)a^{(i)} \\ &= \frac{1}{1+k_i}\left[S_{12}S_{22}^{-1}S_{21} - k_i\left(S_{11} - S_{12}S_{22}^{-1}S_{21}\right)\right]a^{(i)}, \end{aligned} \tag{23}$$

which is equivalent to (19) for $k = k_i$. Comparison of the normalizations $a^{(i)\prime}S_{11}a^{(i)} = 1$ and $\bar{a}^{(i)\prime}(S_{11} - S_{12}S_{22}^{-1}S_{21})\bar{a}^{(i)} = 1$ shows that $\bar{a}^{(i)} = (1/\sqrt{1-l_i^2})a^{(i)}$. Then $\bar{c}^{(j)} = (1/\sqrt{k_j})S_{22}^{-1}S_{21}\bar{a}^{(j)} = c^{(j)}$.

We see that canonical variable analysis can be applied when the two vectors are jointly random and when one vector is random and the other is nonstochastic. The canonical variables defined by the two approaches are the same except for normalization. The measure of relationship between corresponding canonical variables can be the (canonical) correlation or it can be the ratio of "explained" to "unexplained" variance.

12.6.4. Testing Rank

The number of roots κ_j that are different from 0 is the rank of the regression matrix $\boldsymbol{\beta}$. It is the number of linear combinations of the regression variables that are needed to express the expected values of $X_\phi^{(1)}$. We can ask whether the rank is k $(1 \le k \le p_1$ if $p_1 \le p_2)$ against the alternative that the rank is greater than k. The hypothesis is

$$H_k: \kappa_{k+1} = \cdots = \kappa_{p_1} = 0. \tag{24}$$

The likelihood ratio criterion [Anderson (1951b)] is a power of

$$\prod_{i=k+1}^{p_1}(1+k_i)^{-1} = \prod_{i=k+1}^{p_1}(1+l_i^2). \tag{25}$$

Note that this is the same criterion as for the case of both vectors stochastic (Section 12.4). Then

$$-\left[N - \tfrac{1}{2}(p+3)\right]\sum_{i=k+1}^{p_1}\log(1-l_i^2) \tag{26}$$

has approximately the χ^2-distribution with $(p_1-k)(p_2-k)$ degrees of freedom.

The determination of the rank as any number between 0 and p_1 can be done as in Section 12.4.

12.6.5. Linear Functional Relationships

The study of Section 12.6 can be carried out in other terms. For example, the balanced one-way analysis of variance can be set up as

$$Y_{\alpha j} = \boldsymbol{\nu}_\alpha + \boldsymbol{\mu} + U_{\alpha j}, \qquad \alpha = 1,\dots,m, \quad j = 1,\dots,l, \tag{27}$$

where $\mathscr{E}U_\alpha = \mathbf{0}$, $\mathscr{E}U_\alpha U_\alpha' = \boldsymbol{\Psi}$, $\sum_{\alpha=1}^m \boldsymbol{\nu}_\alpha = \mathbf{0}$, and

$$\boldsymbol{\Theta}\boldsymbol{\nu}_\alpha = \mathbf{0}, \qquad \alpha = 1,\dots,m, \tag{28}$$

where $\boldsymbol{\Theta}$ is $q \times p_1$ of rank q $(<p_1)$. This is a special case of the model of Section 12.6.1 with $\phi = 1,\dots,N$, replaced by the pair of indices (α, j), $X_\phi^{(1)} = Y_{\alpha j}$, $\boldsymbol{\tau} = \boldsymbol{\mu}$, and $\boldsymbol{\beta}(x_\phi^{(2)} - \bar{x}^{(2)}) = \boldsymbol{\nu}_\alpha$ by use of dummy variables as in Section 8.8. The rank of $(\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_m)$ is that of $\boldsymbol{\beta}$, namely, $r = p_1 - q$. There are q roots of (8) equal to 0 with

$$\boldsymbol{\beta}S_{22}\boldsymbol{\beta}' = l\sum_{\alpha=1}^m \boldsymbol{\nu}_\alpha\boldsymbol{\nu}_\alpha'. \tag{29}$$

The model (27) can be interpreted as repeated observations on $\boldsymbol{\nu}_\alpha + \boldsymbol{\mu}$ with error. The component equations of (28) are the linear functional relationships.

Let $\bar{y}_\alpha = (1/l)\sum_{j=1}^l y_{\alpha j}$ and $\bar{y} = (1/m)\sum_{\alpha=1}^m \bar{y}_\alpha$. The *sum of squares for effect* is

$$H = l\sum_{\alpha=1}^m (\bar{y}_\alpha - \bar{y})(\bar{y}_\alpha - \bar{y})' = n\hat{\boldsymbol{\beta}}S_{22}\hat{\boldsymbol{\beta}}' \tag{30}$$

with $m-1$ degrees of freedom, and the *sum of squares for error* is

$$G = \sum_{\alpha=1}^m \sum_{j=1}^l (y_{\alpha j} - \bar{y}_\alpha)(y_{\alpha j} - \bar{y}_\alpha)' = n\hat{\boldsymbol{\Psi}} \tag{31}$$

with $m(l-1)$ degrees of freedom. The case $p_1 < p_2$ corresponds to $p_1 < l$ Then a maximum likelihood estimator of $\boldsymbol{\Theta}$ is

$$\hat{\boldsymbol{\Theta}} = (\bar{a}^{(r+1)},\dots,\bar{a}^{(p_1)})', \tag{32}$$

and the maximum likelihood estimators of $\boldsymbol{\nu}_\alpha$ are

$$\hat{\boldsymbol{\nu}}_\alpha = \hat{\boldsymbol{\Psi}}\hat{\Theta}'\hat{\Theta}(\bar{y}_\alpha - \bar{y}), \qquad \alpha = 1, \ldots, n. \tag{33}$$

The estimator (32) can be multiplied by any nonsingular $q \times q$ matrix on the left to obtain another. For a fuller discussion, see Anderson (1984a) and Kendall and Stuart (1973).

12.7. REDUCED RANK REGRESSION

Reduced rank regression involves estimating the regression matrix $\boldsymbol{\beta}$ in $\mathscr{E}X^{(1)}|X^{(2)} = \boldsymbol{\beta}X^{(2)}$ by a matrix $\hat{\boldsymbol{\beta}}$ of preassigned rank k. In the limited-information maximum likelihood method of estimating an equation that is part of a system of simultaneous equations (Section 12.8), the regression matrix is assumed to be of rank one less than the order of the matrix. Anderson (1951a) derived the maximum likelihood estimator of $\boldsymbol{\beta}$ when the model is

$$X_\alpha^{(1)} = \boldsymbol{\tau} + \boldsymbol{\beta}\left(\boldsymbol{x}_\alpha^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right) + Z_\alpha, \qquad \alpha = 1, \ldots, N, \tag{1}$$

the rank of $\boldsymbol{\beta}$ is specified to be k $(\leq p_1)$, the vectors $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_N^{(2)}$ are nonstochastic, and Z_α is normally distributed. On the basis of a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, define $\hat{\Sigma}$ by (2) of Section 12.3 and $\hat{\Lambda}$, $\hat{\mathbf{A}}$, and $\hat{\Gamma}$ by (3), (4), and (5). Partition $\Lambda = \operatorname{diag}(\hat{\Lambda}_1, \hat{\Lambda}_2)$, $\hat{\mathbf{A}} = (\hat{\mathbf{A}}_1, \hat{\mathbf{A}}_2)$, and $\hat{\Gamma} = (\hat{\Gamma}_1, \hat{\Gamma}_2)$, where $\hat{\Lambda}_1$, $\hat{\mathbf{A}}_1$, and $\hat{\Gamma}_1$ have k columns. Let $\hat{\Phi}_1 = \hat{\mathbf{A}}_1(I_k - \hat{\Lambda}_1^2)^{-\frac{1}{2}}$.

Definition 12.7.1 (*Reduced Rank Regression*) The reduced rank regression estimator *in* (1) *is*

$$\hat{B}_k = S_{YX}\hat{\Gamma}_1\hat{\Gamma}_1' = S_{YY}\hat{\mathbf{A}}_1\hat{\Lambda}_1\hat{\Gamma}_1' = \hat{\Sigma}_{\hat{Z}\hat{Z}}\hat{\Phi}_1\hat{\Phi}_1 B, \tag{2}$$

where $B = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}$ *and* $\hat{\Sigma}_{\hat{Z}\hat{Z}} = \hat{\Sigma}_{11} - B\hat{\Sigma}_{22}B'$.

The maximum likelihood estimator of $\boldsymbol{\beta}$ of rank k is the same for $X^{(1)}$ and $X^{(2)}$ normally distributed because the density of $X = (X^{(1)\prime}, X^{(2)\prime})'$ factors as

$$n(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = n\left(\boldsymbol{x}^{(1)}|\boldsymbol{\mu}^{(1)} + \boldsymbol{\beta}(\boldsymbol{x}^{(2)} - \bar{\boldsymbol{x}}^{(2)}), \Sigma_{ZZ}\right)n\left(\boldsymbol{x}^{(2)}|\boldsymbol{\mu}^{(2)}, \Sigma_{22}\right). \tag{3}$$

Reduced rank regression has been applied in many disciplines, including econometrics, time series analysis, and signal processing. See, for example, Johansen (1995) for use of reduced rank regression in estimation of cointegration in economic time series, Tsay and Tiao (1985) and Ahn and Reinsel (1988) for applications in stationary processes, and Stoica and Viberg (1996) for utilization in signal processing. In general the estimated reduced rank regression is a better estimator in a regression model than the unrestricted estimator.

In Section 13.7 the asymptotic distribution of the reduced rank regression estimator is obtained under the assumptions that are sufficient for the asymptotic normality of the least squares estimator $B = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}$. The asymptotic distribution of $\hat{B}_k$ has been obtained by Ryan, Hubert, Carter, Sprague, and Parrott (1992), Schmidli (1996), Stoica and Viberg (1996), and Reinsel and Velu (1998) by use of the expected Fisher information on the assumption that Z_α is normally distributed. Izenman (1975) suggested the term reduced rank regression.

12.8. SIMULTANEOUS EQUATIONS MODELS

12.8.1. The Model

Inference for structural equation models in econometrics is related to canonical correlations. The general model is

$$By_t + \Gamma z_t = u_t, \qquad t = 1, \dots, T, \tag{1}$$

where B is $G \times G$ and Γ is $G \times K$. Here y_t is composed of G jointly dependent variables (endogenous), z_t is composed of K predetermined variables (exogenous and lagged dependent) which are treated as "independent" variables, and u_t consists of G unobservable random variables with

$$\mathscr{E}u_t = \mathbf{0}, \qquad \mathscr{E}u_t u_t' = \Sigma. \tag{2}$$

We require B to be nonsingular. This model was initiated by Haavelmo (1944) and was developed by Koopmans, Marschak, Hurwicz, Anderson, Rubin, Leipnik, et al., 1944–1954, at the Cowles Commission for Research in Economics. Each component equation represents the behavior of some group (such as consumers or producers) and has economic meaning.

The set of structural equations (1) can be solved for y_t (because B is nonsingular):

$$y_t = \Pi z_t + v_t, \tag{3}$$

where

$$\Pi = -B^{-1}\Gamma, \qquad v_t = B^{-1}u_t \tag{4}$$

with

$$\mathscr{E}v_t = \mathbf{0}, \qquad \mathscr{E}v_t v_t' = B^{-1}\Sigma(B')^{-1} = \Omega, \tag{5}$$

say. The equation (3) is called the *reduced form* of the model. It is a multivariate regression model. In principle, it is observable.

12.8.2. Identification by Specified Zeros

The structural equation (1) can be multiplied on the left by an arbitrary nonsingular matrix. To determine component equations that are economically meaningful, restrictions must be imposed. For example, in the case of demand and supply the equation describing demand may be distinguished by the fact that it includes consumer income and excludes cost of raw materials, which is in the supply equation. The exclusion of the latter amounts to specifying that its coefficient in the demand equation is 0.

We consider identification of a structural equation by specifying certain coefficients to be 0. It is convenient to treat the first equation. Suppose the variables are numbered so that the first G_1 jointly dependent variables are included in the first equation and the remaining $G_2 = G - G_1$ are not and the first K_1 predetermined variables are included and $K_2 = K - K_1$ are excluded. Then we can partition the coefficient matrices as

$$(6) \qquad (B \quad \Gamma) = \begin{bmatrix} \boldsymbol{\beta}' & \mathbf{0} & \boldsymbol{\gamma}' & \mathbf{0} \\ \text{—} & \text{—} & \text{—} & \text{—} \end{bmatrix},$$

where the vectors $\boldsymbol{\beta}$, $\mathbf{0}$, $\boldsymbol{\gamma}$, and $\mathbf{0}$ have G_1, G_2, K_1, and K_2 components, respectively. The reduced form is partitioned conformally into G_1 and G_2 sets of rows and K_1 and K_2 sets of columns:

$$(7) \qquad \boldsymbol{\Pi} = \begin{bmatrix} \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ \text{—} & \text{—} \end{bmatrix}.$$

The relation between $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Pi}$ can be expressed as

(8)
$$\begin{bmatrix} \boldsymbol{\gamma}' & \mathbf{0} \\ \text{—} & \text{—} \end{bmatrix} = \boldsymbol{\Gamma} = -\boldsymbol{B}\boldsymbol{\Pi} = -\begin{bmatrix} \boldsymbol{\beta}' & \mathbf{0} \\ \text{—} & \text{—} \end{bmatrix}\begin{bmatrix} \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ \text{—} & \text{—} \end{bmatrix} = -\begin{bmatrix} \boldsymbol{\beta}'\boldsymbol{\Pi}_{11} & \boldsymbol{\beta}'\boldsymbol{\Pi}_{12} \\ \text{—} & \text{—} \end{bmatrix}.$$

The upper right-hand corner of (8) yields

$$(9) \qquad \boldsymbol{\beta}'\boldsymbol{\Pi}_{12} = \mathbf{0}.$$

To determine $\boldsymbol{\beta}$ $(G_1 \times 1)$ uniquely except for a constant of proportionality we need

$$(10) \qquad \operatorname{rank}(\boldsymbol{\Pi}_{12}) = G_1 - 1.$$

This implies

$$K_2 \geq G_1 - 1. \tag{11}$$

Addition of G_2 to (11) gives the *order condition*

$$G_2 + K_2 \geq G_1 + G_2 - 1 = G - 1. \tag{12}$$

The number of specified 0's in an identified equation must be at least equal to 1 less than the number of equations (or jointly dependent variables).

It can be shown that when $\boldsymbol{B}$ is nonsingular (10) holds if and only if the rank of the matrix consisting of the columns of $(\boldsymbol{B}\ \ \boldsymbol{\Gamma})$ with specified 0's in the first row is $G - 1$.

12.8.3. Estimation of the Reduced Form

The model (3) is a typical multivariate regression model. The observations are

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix}, \dots, \begin{pmatrix} y_T \\ z_T \end{pmatrix}. \tag{13}$$

The usual estimators of $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$ (Section 8.2) are

$$\boldsymbol{P} = \sum_{t=1}^{T} y_t z_t' \left(\sum_{t=1}^{T} z_t z_t' \right)^{-1}, \tag{14}$$

$$\hat{\boldsymbol{\Omega}} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \boldsymbol{P} z_t)(y_t - \boldsymbol{P} z_t)'. \tag{15}$$

These are maximum likelihood estimators if the v_t are normal.

If the z_t are exogenous (regardless of normality), then

$$\mathscr{E} \operatorname{vec} \boldsymbol{P} = \operatorname{vec} \boldsymbol{\Pi}, \qquad \mathscr{C}(\operatorname{vec} \boldsymbol{P}) = A^{-1} \otimes \boldsymbol{\Omega}, \tag{16}$$

where

$$A = \sum_{t=1}^{T} z_t z_t' \tag{17}$$

and $\operatorname{vec}(\boldsymbol{d}_1, \dots, \boldsymbol{d}_m) = (\boldsymbol{d}_1', \dots, \boldsymbol{d}_m')'$. If, furthermore, the v_t are normal, then $\boldsymbol{P}$ is normal and $T\hat{\boldsymbol{\Omega}}$ has the Wishart distribution with covariance matrix $\boldsymbol{\Omega}$ and $T - K$ degrees of freedom.

12.8.4. Estimation of the Coefficients of an Equation

First, consider the estimation of the vector of coefficients $\boldsymbol{\beta}$ when $K_2 = G_1 - 1$. Let

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \tag{18}$$

be partitioned as Π. Then the probability is 1 that $\operatorname{rank}(\boldsymbol{P}_{12}) = G_1 - 1$ and the equation

$$\hat{\boldsymbol{\beta}}' P_{12} = \mathbf{0} \tag{19}$$

has a nontrivial solution that is unique except for a constant of proportionality. This is the maximum likelihood estimator when the disturbance terms are normal.

If $K_2 \geq G_1$, then the probability is 1 that $\operatorname{rank}(P_{12}) = G_1$ and (19) has only the trivial solution $\hat{\boldsymbol{\beta}} = \mathbf{0}$, which is unsatisfactory. To obtain a suitable estimator we find $\hat{\boldsymbol{\beta}}$ to minimize $\hat{\boldsymbol{\beta}}'\boldsymbol{P}_{12}$ in som: sense relative to another function of $\hat{\boldsymbol{\beta}}'$.

Let z_t be partitioned into subvectors of K_1 and K_2 components:

$$z_t = \begin{pmatrix} z_t^{(1)} \\ z_t^{(2)} \end{pmatrix}, \tag{20}$$

$$\sum_{t=1}^{T} z_t z_t' = A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \tag{21}$$

$$A_{22\cdot1} = A_{22} - A_{21} A_{11}^{-1} A_{12}. \tag{22}$$

Let y_t and $\boldsymbol{\Omega}$ be partitioned into G_1 and G_2 components:

$$y_t = \begin{pmatrix} y_t^{(1)} \\ y_t^{(2)} \end{pmatrix}, \tag{23}$$

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}. \tag{24}$$

Now set up the multivariate analysis of variance table for $y^{(1)}$:

Source	Sum of Squares
$z_t^{(1)}$	$\sum_{s,t=1}^{T} y_s^{(1)} z_s^{(1)\prime} A_{11}^{-1} z_t^{(1)} y_t^{(1)\prime}$
$z_t^{(2)} \perp z_t^{(1)}$	$P_{12} A_{22\cdot 1} P_{12}'$
Error	$\sum_{t=1}^{T} (y_t^{(1)} - P_{11} z_t^{(1)} - P_{12} z_t^{(2)})(y_t^{(1)} - P_{11} z_t^{(1)} - P_{12} z_t^{(2)})'$
Total	$\sum_{t=1}^{T} y_t^{(1)} y_t^{(1)\prime}$

The first term in the table is the (vector) sum of squares of $y_t^{(1)}$ due to the effect of $z_t^{(1)}$. The second term is due to the effect of $z_t^{(2)}$ beyond the effect of $z_t^{(1)}$. The two add to $(\boldsymbol{PAP}')_{11}$, which is the total effect of z_t, the predetermined variables.

We propose to find the vector $\hat{\beta}$ such that effect of $z_t^{(2)}$ and $\hat{\beta}' y_t^{(1)}$ beyond the effect of $z_t^{(1)}$ is minimized relative to the error sum of squares of $\hat{\beta}' y_t^{(1)}$. We minimize

$$\frac{\hat{\beta}'(P_{12} S_{22\cdot 1} P_{12}')\hat{\beta}}{\hat{\beta}'\hat{\Omega}_{11}\hat{\beta}} = \frac{(\hat{\beta}' P_{12}) S_{22\cdot 1} (\hat{\beta}' P_{12})'}{\hat{\beta}'\hat{\Omega}_{11}\hat{\beta}}, \tag{25}$$

where $T\hat{\Omega} = \sum_{t=1}^{T} y_t y_t - PAP'$. This estimator has been called the *least variance ratio* estimator. Under normality and based only on the 0 restrictions on the coefficients of this single equation, the estimator is maximum likelihood and is known as the *limited-information maximum likelihood* (LIML) estimator [Anderson and Rubin (1949)].

The algebra of minimizing (25) is to find the smallest root, say ν, of

$$\left| P_{12} S_{22\cdot 1} P_{12}' - \lambda \hat{\Omega}_{11} \right| = 0 \tag{26}$$

and the corresponding vector satisfying

$$P_{12} S_{22\cdot 1} P_{12}' \hat{\beta} = \nu \hat{\Omega}_{11} \hat{\beta}. \tag{27}$$

The vector is normalized according to some rule. A frequently used rule is to set one (nonzero) coefficient equal to 1, say the first, $\hat{\beta}_1 = 1$. If we write

$$\boldsymbol{\beta} = \begin{pmatrix} 1 \\ \boldsymbol{\beta}^* \end{pmatrix}, \qquad \hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 \\ \hat{\boldsymbol{\beta}}^* \end{pmatrix}, \tag{28}$$

$$\Pi_{12} = \begin{pmatrix} \pi_{12} \\ \Pi_{12}^* \end{pmatrix}, \qquad P_{11} = \begin{pmatrix} p_{12} \\ P_{12}^* \end{pmatrix}, \tag{29}$$

$$\hat{\Omega}_{11} = \begin{pmatrix} \hat{\omega}_{11} & \hat{\omega}'_{(1)} \\ \hat{\omega}_{(1)} & \hat{\Omega}_{11}^* \end{pmatrix}, \tag{30}$$

then (27) can be replaced by the linear equation

$$\left(P_{12}^* S_{22\cdot 1} P_{12}^{*\prime} - \nu \hat{\Omega}_{11}^*\right)\hat{\boldsymbol{\beta}}^* = -\left(P_{12}^* S_{22\cdot 1} p'_{12} - \nu \hat{\omega}_{(1)}\right). \tag{31}$$

The first component equation in (27) has been dropped because it is linearly dependent on the other equations [because ν is a root of (26)].

12.8.5. Relation to the Linear Functional Relationship

We now show that the model for the single linear functional relationship ($q = 1$) is identical to the model for structural equations in the special case that $G_2 = 0$ ($y_t^{(1)} = y_t$) and $z_t^{(1)} \equiv 1$ ($K_1 = 1$). Write the two models as

$$X_{\alpha j} = \boldsymbol{\mu} + \boldsymbol{\nu}_\alpha + U_{\alpha j}, \qquad \alpha = 1, \dots, n, \quad j = 1, \dots, k, \tag{32}$$

where

$$\sum_{\alpha=1}^{n} \boldsymbol{\nu}_\alpha = \mathbf{0}, \tag{33}$$

and

$$y_t = \Pi_1 + \Pi_2 z_t^{(2)} + v_t, \qquad t = 1, \dots, T, \tag{34}$$

where $\Pi = (\Pi_1 \;\; \Pi_2)$. The correspondence between the models is $p \leftrightarrow G = G_1$,

$$X_{\alpha j} \leftrightarrow y_t, \qquad U_{\alpha j} \leftrightarrow v_t, \tag{35}$$

$$(\alpha, j) \leftrightarrow t, \qquad nk \leftrightarrow T, \tag{36}$$

$$\boldsymbol{\Psi} \leftrightarrow \boldsymbol{\Omega}, \qquad \boldsymbol{\mu} \leftrightarrow \Pi_1. \tag{37}$$

We can write the model for the linear functional relationship with dummy variables. Define

$$(38)\qquad s_{\alpha j} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \alpha\text{th position}, \qquad \alpha = 1, \dots, n-1.$$

$$(39)\qquad s_{nj} = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}.$$

Then

$$(40)\qquad \boldsymbol{\mu} + \boldsymbol{\nu}_\alpha = (\boldsymbol{\mu}, \boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_{n-1}) \begin{pmatrix} 1 \\ s_{\alpha j} \end{pmatrix}, \qquad \alpha = 1, \dots, n,$$

where j may be suppressed. Note

$$(41)\qquad \boldsymbol{\nu}_n = -(\boldsymbol{\nu}_1 + \cdots + \boldsymbol{\nu}_{n-1}).$$

The correspondence is

$$(42)\qquad 1 \leftrightarrow z_t^{(1)}, \qquad s_{\alpha j} \leftrightarrow z_t^{(2)}.$$

$$(43)\qquad \boldsymbol{\mu} \leftrightarrow \boldsymbol{\Pi}_1, \qquad (\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_{n-1}) \leftrightarrow \boldsymbol{\Pi}_2,$$

$$(44)\qquad 1 \leftrightarrow K_1, \qquad n-1 \leftrightarrow K_2,$$

$$(45)\qquad \boldsymbol{B}(\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_{n-1}) = \mathbf{0} \leftrightarrow \boldsymbol{\beta}'\boldsymbol{\Pi}_2 = \mathbf{0}.$$

Let $\boldsymbol{P} = (\boldsymbol{P}_1 \;\; \boldsymbol{P}_2)$. In terms of the statistics we have the correspondence

$$(46)\qquad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} \leftrightarrow \bar{\boldsymbol{y}},$$

$$(47)\qquad \hat{\boldsymbol{\nu}}_\alpha = \bar{\boldsymbol{x}}_\alpha - \bar{\boldsymbol{x}} \leftrightarrow \boldsymbol{P}_2.$$

The effect matrix is

$$(48)\qquad \boldsymbol{H} = k \sum_{\alpha=1}^{n} (\bar{\boldsymbol{x}}_\alpha - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_\alpha - \bar{\boldsymbol{x}})' \leftrightarrow \boldsymbol{P}_2 \boldsymbol{A}_{22\cdot 1} \boldsymbol{P}_2',$$

and the error matrix is

$$(49)\quad G = \sum_{\alpha=1}^{n} \sum_{j=1}^{k} (x_{\alpha j} - \bar{x}_\alpha)(x_{\alpha j} - \bar{x}_\alpha)' \leftrightarrow T\hat{\Omega} = \sum_{t=1}^{T} (y_t - Pz_t)(y_t - Pz_t)'.$$

Then the estimator $\hat{B}$ of the linear functional relationship for $q = 1$ is identical to the LIML estimator [Anderson (1951b), (1976), (1984a)].

12.8.6. Asymptotic Theory as $T \to \infty$

We shall find the limiting distribution of $\sqrt{T}(\hat{\beta}^* - \beta^*)$ defined by (28) and (31) by showing that $\hat{\beta}^*$ is asymptotically equivalent to

$$(50)\quad \hat{\beta}^*_{\mathrm{TSLS}} = -(P^*_{12} S_{22\cdot 1} P^{*\prime}_{12})^{-1} P^*_{22} S_{22\cdot 1} p'_{12}.$$

This derivation is essentially the same as that given in Anderson and Rubin (1950) except for notation. The estimator defined by (50), known as the *two stage least squares* (TSLS) estimator, is an approximation to the LIML estimator obtained by dropping the terms $\nu\hat{\Omega}^*_{11}$ and $\nu\hat{\omega}_{(1)}$ from (31). Let $\hat{\beta}^* = \hat{\beta}^*_{\mathrm{LIML}}$. We assume the conditions for $\sqrt{T}(P - \Pi)$ having a limiting normal distribution. (See Theorem 8.11.1.)

Lemma 12.8.1. *Suppose $(1/T)A \to A^0$, a positive definite matrix, as $T \to \infty$. Then $\nu = O_p(1/T)$, where ν is the smallest root of (26).*

Proof. Let $\tilde{P}_{12} = \sqrt{T}(P_{12} - \Pi_{12})$. Then because $\beta'\Pi_{12} = \mathbf{0}$

$$(51)\quad \frac{\beta' P_{12} S_{22\cdot 1} P_{12}\beta}{\beta'\hat{\Omega}_{11}\beta} = \frac{\beta'\left[\Pi_{12} + (1/\sqrt{T})\tilde{P}_{12}\right] S_{22\cdot 1}\left[\Pi_{12} + (1/\sqrt{T})\tilde{P}_{12}\right]'\beta}{\beta'\hat{\Omega}_{11}\beta}$$

$$= \frac{\beta'\tilde{P}_{12} S_{22\cdot 1}\tilde{P}'_{12}\beta}{T\beta'\hat{\Omega}_{11}\beta} = O_p\left(\frac{1}{T}\right).$$

Since

$$(52)\quad \nu = \min_{\hat{\beta}} \frac{\hat{\beta}' P_{12} S_{22\cdot 1} P'_{12}\hat{\beta}}{\hat{\beta}'\hat{\Omega}_{11}\hat{\beta}} \leq \frac{\beta' P_{12} S_{22\cdot 1} P'_{12}\beta}{\beta'\hat{\Omega}_{11}\beta}$$

the lemma follows. ∎

The import of Lemma 12.8.1 is that the difference between the LIML estimator and the TSLS estimator is $O_p(1/T)$. We have

$$
\begin{aligned}
(53)\quad & \hat{\beta}^*_{\text{LIML}} - \beta^*_{\text{TSLS}} \\
&= (P^*_{12}S_{22\cdot 1}P^{*\prime}_{12})^{-1}P^*_{12}S_{22\cdot 1}p'_{12} \\
&\quad - \left(P^*_{12}S_{22\cdot 1}P^{*\prime}_{12} - \nu\hat{\Omega}_{11}\right)^{-1}\left(P^*_{12}S_{22\cdot 1}p'_{12} - \nu\hat{\omega}_{(1)}\right) \\
&= \left[(P^*_{12}S_{22\cdot 1}P^{*\prime}_{12})^{-1} - \left(P^*_{12}S_{22\cdot 1}P^{*\prime}_{12} - \nu\hat{\Omega}_{11}\right)^{-1}\right]P^*_{12}S_{22\cdot 1}p'_{12} \\
&\quad + \left(P^*_{12}S_{22\cdot 1}P^{*\prime}_{12} - \nu\hat{\Omega}_{11}\right)^{-1}\nu\hat{\omega}_{(1)} \\
&= -\nu(P^*_{12}S_{22\cdot 1}P^{*\prime}_{12})^{-1}\hat{\Omega}_{11}\left(P^*_{12}S_{22\cdot 1}P^{*\prime}_{12} - \nu\hat{\Omega}_{11}\right)^{-1}P^*_{12}S_{22\cdot 1}p'_{12} \\
&\quad + \nu\left(P^*_{12}S_{22\cdot 1}P^*_{12} - \nu\hat{\Omega}_{11}\right)^{-1}\hat{\omega}_{(1)} \\
&= O_p(\nu) = O_p\left(\frac{1}{T}\right).
\end{aligned}
$$

Consider

$$
(54)\qquad p'_{12} + P^{*\prime}_{12}\beta^* = P'_{12}\beta = A^{-1}_{22\cdot 1}\sum_{t=1}^{T} z_t^{(2\cdot 1)}y_t^{(1)}\beta = A^{-1}_{22\cdot 1}\sum_{t=1}^{T} z_t^{(2\cdot 1)}u_{1t},
$$

where $z_t^{(2\cdot 1)} = z_t^{(2)} - A_{21}A_{11}^{-1}z_t^{(1)}$. Thus $\mathscr{E}(p'_{12} + P^{*\prime}_{12}\beta^*) = \mathscr{E}P'_{12}\beta = \mathbf{0}$ and

$$
(55)\qquad \mathscr{E}(p'_{12} + P^{*\prime}_{12}\beta^*)(p'_{12} + P^{*\prime}_{12}\beta^*)' = \mathscr{E}P'_{12}\beta(P'_{12}\beta)' = \sigma_{11}A^{-1}_{22\cdot 1}.
$$

Note that $\hat{\beta}^*_{\text{TSLS}} - \beta^* = -(P^*_{12}S_{22\cdot 1}P^*_{12})^{-1}P^*_{12}S_{22\cdot 1}P'_{12}\beta$ and $(\beta', \mathbf{0})y_t + (\gamma', \mathbf{0})z_t = u_{1t}$.

Theorem 12.8.1. *Under the conditions of Theorem* 8.11.1

$$
(56)\qquad \sqrt{T}\left(\hat{\beta}^*_{\text{LIML}} - \beta^*\right) \xrightarrow{d} N\left[\mathbf{0}, \sigma_{11}\left(\Pi_{12}S^0_{22\cdot 1}\Pi'_{12}\right)^{-1}\right].
$$

Proof. The theorem follows from (55), $S_{22\cdot 1} \to S^0_{22\cdot 1}$, and $P_{12} \xrightarrow{p} \Pi_{12}$. ∎

Because of the correspondence between the LIML estimator and the maximum likelihood estimator for the linear functional relationship as outlined in Section 12.7.5, this asymptotic theory can be translated for the latter.

Suppose the single linear functional relationship is written as

$$(57) \qquad 0 = \boldsymbol{\beta}'\boldsymbol{\nu}_\alpha = \begin{pmatrix} 1 & \boldsymbol{\beta}^{*\prime} \end{pmatrix} \begin{pmatrix} \nu_{1\alpha} \\ \boldsymbol{\nu}_\alpha^* \end{pmatrix} = \nu_{1\alpha} + \boldsymbol{\beta}^{*\prime}\boldsymbol{\nu}_\alpha^*, \qquad \alpha = 1, \dots, n,$$

where

$$(58) \qquad \boldsymbol{\nu}_\alpha = \begin{pmatrix} \nu_{1\alpha} \\ \boldsymbol{\nu}_\alpha^* \end{pmatrix}, \qquad \alpha = 1, \dots, n.$$

Let n ($\leftrightarrow K$) be fixed, and let the number of replications $k \to \infty$ (corresponding to $T/K \to \infty$ for fixed K). Let $\sigma^2 = \boldsymbol{\beta}'\boldsymbol{\Psi}\boldsymbol{\beta}$.

Since $\boldsymbol{\Pi}_{12} A_{22\cdot 1} \boldsymbol{\Pi}_{12}'$ corresponds to $k\sum_{\alpha=1}^n \boldsymbol{\nu}_\alpha^* \boldsymbol{\nu}_\alpha^{*\prime}$, $\hat{\boldsymbol{\beta}}^*$ here has the approximate distribution

$$(59) \qquad N\left[\boldsymbol{\beta}^*, \sigma^2 \left(k \sum_{\alpha=1}^n \boldsymbol{\nu}_\alpha^* \boldsymbol{\nu}_\alpha^{*\prime} \right)^{-1} \right].$$

Although Anderson and Rubin (1950) showed that $\nu\hat{\boldsymbol{\Omega}}_{11}^*$ and $\nu\hat{\boldsymbol{\omega}}_{(1)}$ could be dropped from (31) defining $\hat{\boldsymbol{\beta}}_{\text{LIML}}^*$ and hence that $\hat{\boldsymbol{\beta}}_{\text{TSLS}}^*$ was asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\text{TSLS}}^*$, they did not explicitly propose $\hat{\boldsymbol{\beta}}_{\text{TSLS}}^*$. [As part of the Cowles Commission program, Chernoff and Divinsky (1953) developed a computational program of $\hat{\boldsymbol{\beta}}_{\text{LIML}}$.] The TSLS estimator was proposed by Basmann (1957) and Theil (1961). It corresponds in the linear functional relationship setup to ordinary least squares on the first coordinate. If some other coefficient of $\boldsymbol{\beta}$ were set equal to one, the minimization would be in the direction of that coordinate.

Consider the general linear functional relationship when the error covariance matrix is unknown and there are replications. Constrain $\boldsymbol{B}$ to be

$$(60) \qquad \boldsymbol{B} = \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{B}^* \end{pmatrix}.$$

Partition

$$(61) \qquad \boldsymbol{\nu}_\alpha = \begin{pmatrix} \boldsymbol{\nu}_\alpha^{(1)} \\ \boldsymbol{\nu}_\alpha^{(2)} \end{pmatrix}.$$

Then the least squares estimator of $\boldsymbol{B}^*$ is

$$(62) \qquad \hat{\boldsymbol{B}}_{\text{LS}}^* = -\sum_{\alpha=1}^n \left(\bar{\boldsymbol{x}}_\alpha^{(1)} - \bar{\boldsymbol{x}}^{(1)}\right)\left(\bar{\boldsymbol{x}}_\alpha^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)' \left[\sum_{\alpha=1}^n \left(\bar{\boldsymbol{x}}_\alpha^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)\left(\bar{\boldsymbol{x}}_\alpha^{(2)} - \bar{\boldsymbol{x}}^{(2)}\right)' \right]^{-1}$$

For n fixed and $k \to \infty$ and $\hat{B}^*_{\mathrm{LS}} \xrightarrow{p} B^*$ and

$$(63) \qquad \sqrt{k}\,\mathrm{vec}\left(\hat{B}^*_{\mathrm{LS}} - B^*\right) \to N\left[\mathbf{0}, \left(\sum_{\alpha=1}^{n} \boldsymbol{v}_\alpha^{(2)} \boldsymbol{v}_\alpha^{(2)\prime}\right)^{-1} \otimes B \Psi B'\right].$$

[See Anderson (1984b).] It was shown by Anderson (1951c) that the q smallest sample roots are of such a probability order that the maximum likelihood estimator is asymptotically equivalent, that is, the limiting distribution of $\sqrt{k}\,\mathrm{vec}(\hat{B}^*_{\mathrm{ML}} - B^*)$ is the right-hand side of (63).

12.8.7. Other Asymptotic Theory

In terms of the linear functional relationship it may be more natural to consider $n \to \infty$ and k fixed. When $k = 1$ and the error covariance matrix is $\sigma^2 I_p$, Gleser (1981) has given the asymptotic theory. For the simultaneous equations model, the corresponding conditions are that $K_2 \to \infty$, $T \to \infty$, and K_2/T approaches a positive limit. Kunitomo (1980) has given an asymptotic expansion of the distribution in the case of $p = 2$ and $m = q = 1$.

When $n \to \infty$, the least squares estimator (i.e., minimizing the sum of squares of the residuals in one fixed direction) is not consistent; the LIML and TSLS estimators are not asymptotically equivalent.

12.8.8. Distributions of Estimators

Econometricians have studied intensively the distributions of TSLS and LIML estimator, particularly in the case of two endogenous variables.

Exact distributions have been given by Basmann (1961),(1963), Richardson (1968), Sawa (1969), Mariano and Sawa (1972), Phillips (1980), and Anderson and Sawa (1982). These have not been very informative because they are usually given in terms of infinite series the properties of which are unknown or irrelevant.

A more useful approach is by approximating the distributions. Asymptotic expansions of distributions have been made by Sargan and Mikhail (1971), Anderson and Sawa (1973), Anderson (1974), Kunitomo (1980), and others. Phillips (1982) studied the Padé approach. See also Anderson (1977).

Tables of the distributions of the TSLS and LIML estimators in the case of two endogenous variables have been given by Anderson and Sawa (1977),(1979), and Anderson, Kunitomo, and Sawa (1983a).

Anderson, Kunitomo, and Sawa (1983b) graphed densities of the maximum likelihood estimator and the least squares estimator (minimizing in one direction) for the linear functional relationship (Section 12.6) for the case

$p=2$, $m=q=1$, $\Psi=\sigma^2\Psi_0$ and for various values of β, n, and

$$\delta^2=\frac{1}{\sigma^2}\sum_{\alpha=1}^{n}(\mu_\alpha-\bar{\mu})^2. \tag{64}$$

PROBLEMS

12.1. (Sec. 12.2) Let $z_\alpha=z_{1\alpha}=1$, $\alpha=1,\dots,n$, and $\mathbf{\beta}=\beta$. Verify that $\alpha^{(1)}=\Sigma^{-1}\beta$. Relate this result to the discriminant function (Chapter 6).

12.2. (Sec. 12.2) Prove that the roots of (14) are real.

12.3. (Sec. 12.2)

(a) Let $X'=(X^{(1)\prime}\ X^{(2)\prime})$, $\mathscr{E}X=\mathbf{0}$,

$$\mathscr{E}XX'=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix},$$

$U=\alpha'X^{(1)}$, $V=\gamma'X^{(2)}$, $\mathscr{E}U^2=1=\mathscr{E}V^2$, where α and γ are vectors. Show that choosing α and γ to maximize $\mathscr{E}UV$ is equivalent to choosing α and γ to minimize the generalized variance of $(U\ V)$.

(b) Let $X'=(X^{(1)\prime}\ X^{(2)\prime}\ X^{(3)\prime})$, $\mathscr{E}X=\mathbf{0}$,

$$\mathscr{E}XX'=\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12} & \Sigma_{13}\\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23}\\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33}\end{pmatrix},$$

$U=\alpha'X^{(1)}$, $V=\gamma'X^{(23)}$, $W=\beta'X^{(3)}$, $\mathscr{E}U^2=\mathscr{E}V^2=\mathscr{E}W^2=1$. Consider finding α,γ,β to minimize the generalized variance of (U,V,W). Show that this minimum is invariant with respect to transformations $X^{*(i)}=A_iX^{(i)}$, $|A_i|\neq 0$.

(c) By using such transformations, transform Σ into the simplest possible form.

(d) In the case of $X^{(i)}$ consisting of two components, reduce the problem (of minimizing the generalized variance) to its simplest form.

(e) In this case give the derivative equations.

(f) Show that the minimum generalized variance is 1 if and only if $\Sigma_{12}=\mathbf{0}$, $\Sigma_{13}=\mathbf{0}$, $\Sigma_{23}=\mathbf{0}$. (*Note:* This extension of the notion of canonical variates does not lend itself to a "nice" explicit treatment.)

12.4. (Sec. 12.2) Let

$$X^{(1)}=AZ+Y^{(1)},$$
$$X^{(2)}=BZ+Y^{(2)},$$

where $Y^{(1)}, Y^{(2)}, Z$ are independent with mean zero and covariance matrices I with appropriate dimensionalities. Let $A = (a_1, \dots, a_k)$, $B = (b_1, \dots, b_k)$, and suppose that $A'A, B'B$ are diagonal with positive diagonal elements. Show that the canonical variables for nonzero canonical correlations are proportional to $a_i'X^{(1)}, b_i'X^{(2)}$. Obtain the canonical correlation coefficients and appropriate normalizing coefficients for the canonical variables.

12.5. (Sec. 12.2) Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > 0$ be the positive roots of (14), where Σ_{11} and Σ_{22} are $q \times q$ nonsingular matrices.

(a) What is the rank of Σ_{12}?
(b) Write $\prod_{i=1}^q \lambda_i^2$ as the determinant of a rational function of Σ_{11}, Σ_{12}, Σ_{21}, and Σ_{22}. Justify your answer.
(c) If $\lambda_q = 1$, what is the rank of

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}?$$

12.6. (Sec. 12.2) Let $\Sigma_{11} = (1-g)I_{p_1} + g\varepsilon_{p_1}\varepsilon_{p_1}'$, $\Sigma_{22} = (1-h)I_{p_2} + h\varepsilon_{p_2}\varepsilon_{p_2}'$, $\Sigma_{12} = k\varepsilon_{p_1}\varepsilon_{p_2}'$, where $-1/(p_1-1) < g < 1$, $-1/(p_2-1) < h < 1$, and k is suitably restricted. Find the canonical correlations and variates. What is the appropriate restriction on k?

12.7. (Sec. 12.3) Find the canonical correlations and canonical variates between the first two variables and the last three in Problem 4.42.

12.8. (Sec. 12.3) Prove directly the sample analog of Theorem 12.2.1.

12.9. (Sec. 12.3) Prove that $\lambda_1^2(i+1) \to \lambda_1^2$ and $\alpha(i+1) \to \alpha^{(1)}$ if $\alpha(0)$ is such that $\alpha'(0)\Sigma_{11}\alpha^{(1)} \neq 0$. [*Hint:* Use $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \mathrm{A}\Lambda^2\mathrm{A}^{-1}$.]

12.10. (Sec. 12.6) Prove (9), (10), and (11).

12.11. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$ be the roots of $|\Sigma_1 - \lambda\Sigma_2| = 0$, where Σ_1 and Σ_2 are $q \times q$ positive definite covariance matrices.

(a) What does $\lambda_1 = \lambda_q = 1$ imply about the relationship of Σ_1 and Σ_2?
(b) What does $\lambda_q > 1$ imply about the relationships of the ellipsoids $x'\Sigma_1^{-1}x = c$ and $x'\Sigma_2^{-1}x = c$?
(c) What does $\lambda_1 > 1$ and $\lambda_q < 1$ imply about the relationships of the ellipsoids $x'\Sigma_1^{-1}x = c$ and $x'\Sigma_2^{-1}x = c$?

12.12. (Sec. 12.4) For $q = 2$ express the criterion (2) of Section 9.5 in terms of canonical correlations.

12.13. Find the canonical correlations for the data in Problem 9.11.

CHAPTER 13

The Distributions of Characteristic Roots and Vectors

13.1. INTRODUCTION

In this chapter we find the distribution of the sample principal component vectors and their sample variances when all population variances are 1 (Section 13.3). We also find the distribution of the sample canonical correlations and one set of canonical vectors when the two sets of original variates are independent. This second distribution will be shown to be equivalent to the distribution of roots and vectors obtained in the next section. The distribution of the roots is particularly of interest because many invariant tests are functions of these roots. For example, invariant tests of the general linear hypothesis (Section 8.6) depend on the sample only through the roots of the determinantal equation

$$(1) \qquad \left|(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)A_{11\cdot 2}(\hat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*)' - lN\hat{\Sigma}_\Omega\right| = 0.$$

If the hypothesis is true, the roots have the distribution given in Theorem 13.2.2 or 13.2.3. Thus the significance level of any invariant test of the general linear hypothesis can be obtained from the distribution derived in the next section. If the test criterion is one of the ordered roots (e.g., the largest root), then the desired distribution is a marginal distribution of the joint distribution of roots.

The limiting distributions of the roots are obtained under fairly general conditions. These are needed to obtain other limiting distributions, such as the distribution of the criterion for testing that the smallest variances of

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

principal components are equal. Some limiting distributions are obtained for elliptically contoured distributions.

13.2. THE CASE OF TWO WISHART MATRICES

13.2.1. The Transformation

Let us consider A^* and B^* $(p \times p)$ distributed independently according to $W(\Sigma, m)$ and $W(\Sigma, n)$ respectively $(m, n \geq p)$. We shall call the roots of

$$|A^* - lB^*| = 0 \tag{1}$$

the *characteristic roots of* A^* *in the metric of* B^* and the vectors satisfying

$$(A^* - lB^*)x^* = 0 \tag{2}$$

the *characteristic vectors of* A^* *in the metric of* B^*. In this section we shall consider the distribution of these roots and vectors. Later it will be shown that the squares of canonical correlation coefficients have this distribution if the population canonical correlations are all zero.

First we shall transform A^* and B^* so that the distributions do not involve an arbitrary matrix Σ. Let C be a matrix such that $C\Sigma C' = I$. Let

$$A = CA^*C', \qquad B = CB^*C'. \tag{3}$$

Then A and B are independently distributed according to $W(I, m)$ and $W(I, n)$ respectively (Section 7.3.3). Since

$$\begin{aligned}|A - lB| &= |CA^*C' - lCB^*C'|\\ &= |C(A^* - lB^*)C'| = |C| \cdot |A^* - lB^*| \cdot |C'|,\end{aligned}$$

the roots of (1) are the roots of

$$|A - lB| = 0. \tag{4}$$

The corresponding vectors satisfying

$$(A - lB)x = 0 \tag{5}$$

satisfy

$$\begin{aligned}0 &= C^{-1}(A - lB)x \\ &= C^{-1}(CA^*C' - lCB^*C')x \\ &= (A^* - lB^*)C'x.\end{aligned} \tag{6}$$

Thus the vectors x^* are the vectors $C'x$.

It will be convenient to consider the roots of

$$|A - f(A+B)| = 0 \tag{7}$$

and the vectors y satisfying

$$[A - f(A+B)]y = \mathbf{0}. \tag{8}$$

The latter equation can be written

$$\mathbf{0} = (A - fA - fB)y = [(1-f)A - fB]y. \tag{9}$$

Since the probability that $f = 1$ (i.e., that $|-B| = 0$) is 0, the above equation is

$$\left(A - \frac{f}{1-f}B\right)y = \mathbf{0}. \tag{10}$$

Thus the roots of (4) are related to the roots of (7) by $l = f/(1-f)$ or $f = l/(1+l)$, and the vectors satisfying (5) are equal (or proportional) to those satisfying (8).

We now consider finding the distribution of the roots and vectors satisfying (7) and (8). Let the roots be ordered $f_1 > f_2 > \cdots > f_p > 0$ since the probability of two roots being equal is 0 [Okamoto (1973)]. Let

$$F = \begin{pmatrix} f_1 & 0 & \cdots & 0 \\ 0 & f_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & f_p \end{pmatrix}. \tag{11}$$

Suppose the corresponding vector solutions of (8) normalized by

$$y'(A+B)y = 1 \tag{12}$$

are $y_1, \dots, y_p$. These vectors must satisfy

$$y_i'(A+B)y_j = 0, \tag{13}$$

because $y_i'Ay_j = f_j y_i'(A+B)y_j$ and $y_j'Ay_i = f_i y_j'(A+B)y_i$, and this can be only if (13) holds ($f_i \neq f_j$).

Let the $p \times p$ matrix Y be

$$Y = (y_1, \dots, y_p). \tag{14}$$

Equation (8) can be summarized as

$$AY = (A + B)YF, \tag{15}$$

and (12) and (13) give

$$Y'(A + B)Y = I. \tag{16}$$

From (15) we have

$$Y'AY = Y'(A + B)YF = F. \tag{17}$$

Multiplication of (16) and (17) on the left by $(Y')^{-1}$ and on the right by Y^{-1} gives

$$\begin{aligned} A + B &= (Y')^{-1}Y^{-1}, \\ A &= (Y')^{-1}FY^{-1}. \end{aligned} \tag{18}$$

Now let $Y^{-1} = E$. Then

$$\begin{aligned} A + B &= E'E, \\ A &= E'FE, \\ B &= E'(I - F)E. \end{aligned} \tag{19}$$

We now consider the joint distribution of E and F. From (19) we see that E and F define A and B uniquely. From (7) and (11) and the ordering $f_1 > \cdots > f_p$ we see that A and B define F uniquely. Equations (8) for $f = f_i$ and (12) define y_i uniquely except for multiplication by -1 (i.e., replacing y_i by $-y_i$). Since $YE = I$, this means that E is defined uniquely except that rows of E can be multiplied by -1. To remove this indeterminacy we require that $e_{i1} \geq 0$. (The probability that $e_{i1} = 0$ is 0.) Thus E and F are uniquely defined in terms of A and B.

13.2.2. The Jacobian

To find the density of E and F we substitute in the density of A and B according to (19) and multiply by the Jacobian of the transformation. We devote this subsection to finding the Jacobian

$$\left| \frac{\partial(A, B)}{\partial(E, F)} \right|. \tag{20}$$

Since the transformation from A and B to A and $G = A + B$ has Jacobian unity, we shall find

$$\left| \frac{\partial(A, G)}{\partial(E, F)} \right| = \left| \frac{\partial(A, B)}{\partial(E, F)} \right|. \tag{21}$$

First we notice that if $x_\alpha = f_\alpha(y_1, \dots, y_n)$, $\alpha = 1, \dots, n$, is a one-to-one transformation, the Jacobian is the determinant of the linear transformation

$$dx_\alpha = \sum_\beta \frac{\partial f_\alpha}{\partial y_\beta} dy_\beta, \tag{22}$$

where dx_α and dy_β are only formally differentials (i.e., we write these as a mnemonic device). If $f_\alpha(y_1, \dots, y_n)$ is a polynomial, then $\partial f_\alpha / \partial y_\beta$ is the coefficient of y_β^* in the expansion of $f_\alpha(y_1 + y_1^*, \dots, y_n + y_n^*)$ [in fact the coefficient in the expansion of $f_\alpha(y_1, \dots, y_{\beta-1}, y_\beta + y_\beta^*, y_{\beta+1}, \dots, y_n)$]. The elements of A and G are polynomials in E and F. Thus the derivative of an element of A is the coefficient of an element of E^* and F^* in the expansion of $(E+E^*)'(F+F^*)(E+E^*)$ and the derivative of an element of G is the coefficient of an element of E^* and F^* in the expansion of $(E+E^*)'(E+E^*)$. Thus the Jacobian of the transformation from A, G to E, F is the determinant of the linear transformation

$$dA = (dE)'FE + E'(dF)E + E'F(dE), \tag{23}$$

$$dG = (dE)'E + E'(dE). \tag{24}$$

Since A and G (dA and dG) are symmetric, only the functionally independent component equations above are used.

Multiply (23) and (24) on the left by E'^{-1} and on the right by E^{-1} to obtain

$$E'^{-1}(dA)E^{-1} = E'^{-1}(dE)'F + dF + F(dE)E^{-1}, \tag{25}$$

$$E'^{-1}(dG)E^{-1} = E'^{-1}(dE)' + (dE)E^{-1} \tag{26}$$

It should be kept in mind that (23) and (24) are now considered as a linear transformation without regard to how the equations were obtained.

Let

$$E'^{-1}(dA)E^{-1} = d\bar{A}, \tag{27}$$

$$E'^{-1}(dG)E^{-1} = d\bar{G}, \tag{28}$$

$$(dE)E^{-1} = dW. \tag{29}$$

Then

$$d\bar{A} = (dW)'F + dF + F(dW), \tag{30}$$

$$d\bar{G} = dW' + dW. \tag{31}$$

The linear transformation from dE, dF to dA, dG is considered as the linear transformtion from dE, dF to dW, dF with determinant $|E^{-1}|^p = |E|^{-p}$ (because each row of dE is transformed by E^{-1}), followed by the linear transformation from dW, dF to $d\bar{A}, d\bar{G}$, followed by the linear transformation from $d\bar{A}, d\bar{G}$ to $dA = E'(d\bar{A})E, dG = E'(d\bar{G})E$ with determinant $|E|^{p+1} \cdot |E|^{p+1}$ (from Section 7.3.3); and the determinant of the linear transformation from dE, dF to dA, dG is the product of the determinants of the three component transformations. The transformation (30). (31) is written in components as

$$
(32) \qquad \begin{aligned}
d\bar{a}_{ii} &= df_i + 2f_i\, dw_{ii}, & \\
d\bar{a}_{ij} &= f_j\, dw_{ji} + f_i\, dw_{ij}, & i < j, \\
d\bar{g}_{ii} &= 2dw_{ii}, & \\
d\bar{g}_{ij} &= dw_{ji} + dw_{ij}, & i < j.
\end{aligned}
$$

The determinant is

$$
(33) \qquad \begin{array}{l|cccc|}
 & df_i & dw_{ii} & dw_{ij}\ (i<j) & dw_{ij}\ (i>j) \\
d\bar{a}_{ii} & \boldsymbol{I} & 2\boldsymbol{F} & \boldsymbol{0} & \boldsymbol{0} \\
d\bar{g}_{ii} & \boldsymbol{0} & 2\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\
d\bar{a}_{ij}\ (i<j) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{M} & \boldsymbol{N} \\
d\bar{g}_{ij}\ (i<j) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{I}
\end{array}
$$

$$
= \begin{vmatrix} \boldsymbol{I} & 2\boldsymbol{F} \\ \boldsymbol{0} & 2\boldsymbol{I} \end{vmatrix} \cdot \begin{vmatrix} \boldsymbol{M} & \boldsymbol{N} \\ \boldsymbol{I} & \boldsymbol{I} \end{vmatrix} = 2^p |\boldsymbol{M} - \boldsymbol{N}|,
$$

where

$$
(34) \qquad \boldsymbol{M} = \begin{array}{l} \\ da_{12} \\ \vdots \\ da_{1p} \\ da_{23} \\ \vdots \\ da_{2p} \\ \vdots \\ da_{p-1,p} \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} dw_{12} & \cdots & dw_{1p} & dw_{23} & \cdots & dw_{2p} & \cdots & dw_{p-1,p} \end{array} \\
\left[\begin{array}{ccc|ccc|c|c}
f_1 & \cdots & 0 & 0 & \cdots & 0 & & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & f_1 & 0 & \cdots & 0 & & 0 \\
\hline
0 & \cdots & 0 & f_2 & \cdots & 0 & & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & \cdots & f_2 & & 0 \\
\hline
\vdots & & & & & & & \\
\hline
0 & \cdots & 0 & 0 & \cdots & 0 & & f_{p-1}
\end{array}\right]
\end{array}
$$

and

$$
N = \begin{array}{c} \\ da_{12} \\ \vdots \\ da_{1p} \\ da_{23} \\ \vdots \\ da_{2p} \\ \vdots \\ da_{p-1,p} \end{array}
\begin{array}{c}
\begin{array}{cccccccccc} dw_{21} & \cdots & dw_{p1} & dw_{32} & \cdots & dw_{p2} & & \cdots & & dw_{p,p-1} \end{array} \\
\left[\begin{array}{ccc|ccc|c|c}
f_2 & \cdots & 0 & 0 & \cdots & 0 & & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & f_p & 0 & \cdots & 0 & & 0 \\
\hline
0 & \cdots & 0 & f_3 & \cdots & 0 & & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & \cdots & f_p & & 0 \\
\hline
\vdots & & & & & & & \\
\hline
0 & \cdots & 0 & 0 & \cdots & 0 & & f_p
\end{array}\right]
\end{array}. \tag{35}
$$

Then

$$
|M - N| = \prod_{i<j} (f_i - f_j). \tag{36}
$$

The determinant of the linear transformation (23),(24) is

$$
|E|^{-p}|E|^{p+1}|E|^{p+1}2^p\prod_{i<j}(f_i - f_j) = 2^p|E|^{p+2}\prod_{i<j}(f_i - f_j). \tag{37}
$$

Theorem 13.2.1. *The Jacobian of the transformation* (19) *is the absolute value of* (37).

13.2.3. The Joint Distribution of the Matrix E and the Roots

The joint density of A and B is

$$
w(A|I,m)w(B|I,n) = C_1|A|^{\frac{1}{2}(m-p-1)}|B|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}\operatorname{tr}(A+B)}, \tag{38}
$$

where

$$
C_1 = \left[2^{\frac{1}{2}p(n+m)}\Gamma_p\left(\tfrac{1}{2}n\right)\Gamma_p\left(\tfrac{1}{2}m\right)\right]^{-1}. \tag{39}
$$

Therefore the joint density of E and F is

$$
C_1|E'FE|^{\frac{1}{2}(m-p-1)}|E'(I-F)E|^{\frac{1}{2}(n-p-1)} \cdot e^{-\frac{1}{2}\operatorname{tr} E'E}2^p|E'E|^{\frac{1}{2}(p+2)}\prod_{i<j}(f_i - f_j). \tag{40}
$$

Since $|E'FE| = |E'|\cdot|F|\cdot|E| = |F|\cdot|E'E| = \prod_{i=1}^{p} f_i |E'E|$ and $|E'(I-F)E| = |I-F|\cdot|E'E| = \prod_{i=1}^{p}(1-f_i)|E'E|$, the density of E and F is

(41)
$$2^p C_1 |E'E|^{\frac{1}{2}(m+n-p)} e^{-\frac{1}{2}\operatorname{tr} E'E} \prod_{i=1}^{p} f_i^{\frac{1}{2}(m-p-1)} \prod_{i=1}^{p}(1-f_i)^{\frac{1}{2}(n-p-1)} \prod_{i<j}(f_i - f_j).$$

Clearly, E and F are statistically independent because the density factors into a function of E and a function of F. To determine the marginal densities we have only to find the two normalizing constants (the product of which is $2^p C_1$).

Let us evaluate

$$2^p \int |E'E|^{\frac{1}{2}(m+n-p)} e^{-\frac{1}{2}\operatorname{tr} E'E}\, dE, \tag{42}$$

where the integration is $0 < e_{i1} < \infty$, $-\infty < e_{ij} < \infty$, $j \neq 1$. The value of (42) is unchanged if we let $-\infty < e_{i1} < \infty$ and multiply by 2^{-p}. Thus (42) is

$$(2\pi)^{\frac{1}{2}p^2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |E'E|^{\frac{1}{2}(m+n-p)} \left[\frac{1}{(2\pi)^{\frac{1}{2}p^2}} \exp\left(-\frac{1}{2} \sum_{i,j} e_{ij}^2 \right) \right] \prod de_{ij}. \tag{43}$$

Except for the constant $(2\pi)^{\frac{1}{2}p^2}$, (43) is a definition of the expectation of the $\frac{1}{2}(m+n-p)$th power of $|E'E|$ when the e_{ij} have as density the function within brackets. This expected value is the $\frac{1}{2}(m+n-p)$th moment of the generalized variance $|E'E|$ when $E'E$ has the distribution $W(I, p)$. (See Section 7.5.) Thus (43) is

$$(2\pi)^{\frac{1}{2}p^2} \frac{\Gamma_p\left[\frac{1}{2}(m+n)\right]}{\Gamma_p\left(\frac{1}{2}p\right)} 2^{\frac{1}{2}p(n+m-p)}. \tag{44}$$

Thus the density of E is

$$\frac{\Gamma_p\left(\frac{1}{2}p\right)}{2^{\frac{1}{2}p(m+n-2)} \pi^{\frac{1}{2}p^2} \Gamma_p\left[\frac{1}{2}(m+n)\right]} |E'E|^{\frac{1}{2}(m+n-p)} e^{-\frac{1}{2}\operatorname{tr} E'E}. \tag{45}$$

The density of f_i is (41) divided by (45); that is, the density of f_i is

$$C_2 \prod_{i=1}^{p} f_i^{\frac{1}{2}(m-p-1)} \prod_{i=1}^{p}(1-f_i)^{\frac{1}{2}(n-p-1)} \prod_{i<j}(f_i - f_j) \tag{46}$$

for $0 \leq f_p \leq \cdots \leq f_1 \leq 1$, where

$$C_2 = \frac{\pi^{\frac{1}{2}p^2}\Gamma_p\left[\frac{1}{2}(m+n)\right]}{\Gamma_p(\frac{1}{2}n)\Gamma_p(\frac{1}{2}m)\Gamma_p(\frac{1}{2}p)}. \tag{47}$$

The density of l_i is obtained from (46) by letting

$$f_i = \frac{l_i}{l_i+1}; \tag{48}$$

we have

$$\frac{df_i}{dl_i} = \frac{1}{(l_i+1)^2},$$

$$f_i - f_j = \frac{l_i - l_j}{(l_i+1)(l_j+1)}, \tag{49}$$

$$1 - f_i = \frac{1}{l_i+1}.$$

Thus the density of l_i is

$$C_2 \prod_{i=1}^{p} l_i^{\frac{1}{2}(m-p-1)} \prod_{i=1}^{p}(l_i+1)^{-\frac{1}{2}(m+n)} \prod_{i<j}(l_i - l_j) \tag{50}$$

for $0 \leq l_p \leq \cdots \leq l_1$.

Theorem 13.2.2. *If A and B are distributed independently according to $W(\Sigma, m)$ and $W(\Sigma, n)$ respectively $(m \geq p,\ n \geq p)$, the joint density of the roots of $|A - lB| = 0$ is* (50) *where C_2 is defined by* (47).

The joint density of Y can be found from (45) and the fact that the Jacobian is $|Y|^{-2p}$. (See Theorem A.4.6 of the Appendix.)

13.2.4. The Distribution for A Singular

The matrix A above can be represented as $A = W_1 W_1'$, where the columns of W_1 $(p \times m)$ are independently distributed, each according to $N(\mathbf{0}, \Sigma)$. We now treat the case of $m < p$. If we let $B + W_1 W_1' = G = CC'$ and $W_1 = CU$, then the roots of

$$\begin{aligned} 0 &= |A - f(A+B)| = |W_1 W_1' - fG| \\ &= |CUU'C' - fCC'| = |C| \cdot |UU' - fI_p| \cdot |C| \end{aligned} \tag{51}$$

are the roots of

$$|UU' - fI_p| = 0. \tag{52}$$

We shall show that the nonzero roots $f_1 > \cdots > f_m$ (these roots being distinct with probability 1) are the roots of

$$|U'U - fI_m| = 0. \tag{53}$$

For each root $f \neq 0$ of (52) there is a vector $\boldsymbol{x}$ satisfying

$$(UU' - fI_p)\boldsymbol{x} = \mathbf{0}. \tag{54}$$

Multiplication by U' on the left gives

$$\begin{aligned} \mathbf{0} &= U'(UU' - fI_p)\boldsymbol{x} \\ &= (U'U - fI_m)(U'\boldsymbol{x}). \end{aligned} \tag{55}$$

Thus $U'\boldsymbol{x}$ is a characteristic vector of UU' and f is the corresponding root.

It was shown in Section 8.4 that the density of $U = U_*'$ is (for $I_p - UU'$ positive definite or $I_m - U_* U_*'$ positive definite)

$$K|I_p - UU'|^{\frac{1}{2}(n-p-1)} = K|I_{p^*} - U_* U_*'|^{\frac{1}{2}(n^*-p^*-1)}, \tag{56}$$

where $p^* = m$, $n^* - p^* - 1 = n - p - 1$, and $m^* = p$. Thus $f_1, \ldots, f_m$ must be distributed according to (46) with p replaced by m, m by p, and n by $n + m - p$, that is,

$$\frac{\pi^{\frac{1}{2}m^2}\Gamma_m\left[\frac{1}{2}(m+n)\right]}{\Gamma_m(\frac{1}{2}m)\Gamma_m\left[\frac{1}{2}(m+n-p)\right]\Gamma_m(\frac{1}{2}p)} \cdot \prod_{i=1}^{m}\left[f_i^{\frac{1}{2}(p-m-1)}(1-f_i)^{\frac{1}{2}(n-p-1)}\right]\prod_{i<j}(f_i - f_j). \tag{57}$$

Theorem 13.2.3. *If A is distributed as $W_1 W_1'$, where the m columns of W_1 are independent, each distributed according to $N(\mathbf{0}, \Sigma)$, $m \leq p$, and B is independently distributed according to $W(\Sigma, n)$, $n \geq p$, then the density of the nonzero roots of $|A - f(A+B)| = 0$ is given by* (57).

These distributions of roots were found independently and at about the same time by Fisher (1939), Girshick (1939), Hsu (1939a), Mood (1951), and Roy (1939). The development of the Jacobian in Section 13.2.2 is due mainly to Hsu [as reported by Deemer and Olkin (1951)].

13.3. THE CASE OF ONE NONSINGULAR WISHART MATRIX

In this section we shall find the distribution of the roots of

$$|A - lI| = 0, \tag{1}$$

where the matrix A has the distribution $W(I, n)$. It will be observed that the variances of the principal components of a sample of $n+1$ from $N(\mu, I)$ are $1/n$ times the roots of (1). We shall find the following theorem useful:

Theorem 13.3.1. *If the symmetric matrix B has a density of the form $g(l_1, \dots, l_p)$, where $l_1 > \cdots > l_p$ are the characteristic roots of B, then the density of the roots is*

$$\frac{\pi^{\frac{1}{2}p^2} g(l_1, \dots, l_p) \prod_{i<j}(l_i - l_j)}{\Gamma_p(\frac{1}{2}p)}. \tag{2}$$

Proof. From Theorem A.2.1 of the Appendix we know that there exists an orthogonal matrix C such that

$$B = C'LC, \tag{3}$$

where

$$L = \begin{pmatrix} l_1 & 0 & \cdots & 0 \\ 0 & l_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & l_p \end{pmatrix}. \tag{4}$$

If the l's are numbered in descending order of magnitude and if $c_{i1} \ge 0$, then (with probability 1) the transformation from B to L and C is unique. Let the matrix C be given the coordinates $c_1, \dots, c_{p(p-1)/2}$, and let the Jacobian of the transformation be $f(L, C)$. Then the joint density of L and C is $g(l_1, \dots, l_p) f(L, C)$. To prove the theorem we must show that

$$\int \cdots \int f(L, C)\, dc_1 \cdots dc_{p(p-1)/2} = \frac{\pi^{\frac{1}{2}p^2} \prod_{i<j}(l_i - l_j)}{\Gamma_p(\frac{1}{2}p)}. \tag{5}$$

We show this by taking a special case where $B = UU'$ and U ($p \times m$, $m \ge p$) has the density

$$\pi^{-\frac{1}{2}mp} \frac{\Gamma_p[\frac{1}{2}(m+n)]}{\Gamma_p(\frac{1}{2}n)} |I - UU'|^{\frac{1}{2}(n-p-1)}. \tag{6}$$

Then by Lemma 13.3.1, which will be stated below, $\boldsymbol{B}$ has the density

$$(7)\quad \frac{\Gamma_p[\frac{1}{2}(m+n)]}{\Gamma_p(\frac{1}{2}m)\Gamma_p(\frac{1}{2}n)}|\boldsymbol{I}-\boldsymbol{B}|^{\frac{1}{2}(n-p-1)}|\boldsymbol{B}|^{\frac{1}{2}(m-p-1)}$$

$$= \frac{\Gamma_p[\frac{1}{2}(m+n)]}{\Gamma_p(\frac{1}{2}m)\Gamma_p(\frac{1}{2}n)}\prod_{i=1}^{p}(1-l_i)^{\frac{1}{2}(n-p-1)}\prod_{i=1}^{p}l_i^{\frac{1}{2}(m-p-1)}$$

$$= g^*(l_1,\dots,l_p).$$

The joint density of $\boldsymbol{L}$ and $\boldsymbol{C}$ is $f(\boldsymbol{L},\boldsymbol{C})g^*(l_1,\dots,l_p)$. In the preceding section we proved that the marginal density of $\boldsymbol{L}$ is (50). Thus

$$(8)\quad \int\cdots\int g^*(l_1,\dots,l_p)f(\boldsymbol{L},\boldsymbol{C})\,d\boldsymbol{C} = g^*(l_1,\dots,l_p)\int\cdots\int f(\boldsymbol{L},\boldsymbol{C})\,d\boldsymbol{C}$$

$$= \frac{\pi^{\frac{1}{2}p^2}\prod(l_i-l_j)}{\Gamma_p(\frac{1}{2}p)}g^*(l_1,\dots,l_p).$$

This proves (5) and hence the theorem. ■

The statement above (7) is based on the following lemma:

Lemma 13.3.1. *If the density of Y ($p\times m$) is $f(YY')$, then the density of $\boldsymbol{B}=YY'$ is*

$$(9)\qquad \frac{|\boldsymbol{B}|^{\frac{1}{2}(m-p-1)}f(\boldsymbol{B})\pi^{\frac{1}{2}pm}}{\Gamma_p(\frac{1}{2}m)}.$$

The proof of this, like that of Theorem 13.3.1, depends on exhibiting a special case; let $f(YY')=(2\pi)^{-\frac{1}{2}pm}e^{-\frac{1}{2}\operatorname{tr}YY'}$, then (9) is $w(\boldsymbol{B}|\boldsymbol{I},m)$.

Now let us find the density of the roots of (1). The density of A is

$$(10)\qquad \frac{|A|^{\frac{1}{2}(n-p-1)}e^{-\frac{1}{2}\operatorname{tr}A}}{2^{\frac{1}{2}pn}\Gamma_p(\frac{1}{2}n)} = \frac{\prod_{i=1}^{p}l_i^{\frac{1}{2}(n-p-1)}\exp(-\frac{1}{2}\sum_{i=1}^{p}l_i)}{2^{\frac{1}{2}pn}\Gamma_p(\frac{1}{2}n)}.$$

Thus by the theorem we obtain as the density of the roots of A

$$(11)\qquad \frac{\pi^{\frac{1}{2}p^2}\prod_{i=1}^{p}l_i^{\frac{1}{2}(n-p-1)}\exp(-\frac{1}{2}\sum_{i=1}^{p}l_i)\prod_{i<j}(l_i-l_j)}{2^{\frac{1}{2}pn}\Gamma_p(\frac{1}{2}n)\Gamma_p(\frac{1}{2}p)}.$$

Theorem 13.3.2. *If A $(p \times p)$ has the distribution $W(I, n)$, then the characteristic roots $(l_1 \geq l_2 \geq \cdots \geq l_p \geq 0)$ have the density* (11) *over the range where the density is not* 0.

Corollary 13.3.1. *Let $v_1 \geq \cdots \geq v_p$ be the sample variances of the sample principal components of a sample of size $N = n+1$ from $N(\boldsymbol{\mu}, \sigma^2 I)$. Then $(n/\sigma^2)v_i$ are distributed with density* (11).

The characteristic vectors of A are uniquely defined (except for multiplication by -1) with probability 1 by

$$(A - lI)y = \mathbf{0}, \qquad y'y = 1, \tag{12}$$

since the roots are different with probability 1. Let the vectors with $y_{1i} \geq 0$ be

$$Y = (y_1, \ldots, y_p). \tag{13}$$

Then

$$AY = YL. \tag{14}$$

From Section 11.2 we know that

$$Y'Y = I. \tag{15}$$

Multplication of (14) on the right by $Y^{-1} = Y'$ gives

$$A = YLY'. \tag{16}$$

Thus $Y' = C$, defined above.

Now let us consider the joint distribution of L and C. The matrix A has the distribution of

$$A = \sum_{\alpha=1}^{n} X_\alpha X_\alpha', \tag{17}$$

where the X_α are independently distributed, each according to $N(\mathbf{0}, I)$. Let

$$X_\alpha^* = QX_\alpha, \tag{18}$$

where Q is any orthogonal matrix. Then the X_α^* are independently distributed according to $N(\mathbf{0}, I)$ and

$$A^* = \sum_{\alpha=1}^{n} X_\alpha^* X_\alpha^{*\prime} = QAQ' \tag{19}$$

is distributed according to $W(I, n)$. The roots of A^* are the roots of A; thus

$$A^* = C^{**\prime} L C^{**}, \tag{20}$$

$$C^{**\prime} C^{**} = I \tag{21}$$

define C^{**} if we require $c_{i1}^{**} \geq 0$. Let

$$C^* = CQ'. \tag{22}$$

Let

$$J(C^*) = \begin{pmatrix} \dfrac{c_{11}^*}{|c_{11}^*|} & 0 & \cdots & 0 \\ 0 & \dfrac{c_{21}^*}{|c_{21}^*|} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \dfrac{c_{p1}^*}{|c_{p1}^*|} \end{pmatrix}, \tag{23}$$

with $c_{i1}^*/|c_{i1}^*| = 1$ if $c_{i1}^* = 0$. Thus $J(C^*)$ is a diagonal matrix; the ith diagonal element is 1 if $c_{i1}^* \geq 0$ and is -1 if $c_{i1}^* < 0$. Thus

$$C^{**} = J(C^*)C^* = J(CQ')CQ'. \tag{24}$$

The distribution of C^{**} is the same as that of C. We now shall show that this fact defines the distribution of C.

Definition 13.3.1. *If the random orthogonal matrix E of order p has a distribution such that EQ' has the same distribution for every orthogonal Q, then E is said to have the* Haar invariant *distribution (or normalized measure).*

The definition is possible because it has been proved that there is only one distribution with the required invariance property [Halmos (1950)]. It has also been shown that this distribution is the only one invariant under multiplication on the left by an orthogonal matrix (i.e., the distribution of QE is the same as that of E). From this it follows that the probability is $1/2^p$ that E is such that $e_{i1} \geq 0$. This can be seen as follows. Let $J_1, \ldots, J_{2^p}$ be the 2^p diagonal matrices with elements $+1$ and -1. Since the distribution of $J_i E$ is the same as that of E, the probability that $e_{i1} \geq 0$ is the same as the probability that the elements in the first column of $J_i E$ are nonnegative. These events for $i = 1, \ldots, 2^p$ are mutually exclusive and exhaustive (except for elements being 0, which have probability 0), and thus the probability of any one is $1/2^p$.

The conditional distribution of E given $e_{i1} \geq 0$ is 2^p times the Haar invariant distribution over this part of the space. We shall call it the *conditional Haar invariant* distribution.

Lemma 13.3.2. *If the orthogonal matrix E has a distribution such that $e_{i1} \geq 0$ and if $E^{**} = J(EQ')EQ'$ has the same distribution for every orthogonal Q, then E has the conditional Haar invariant distribution.*

Proof. Let the space V of orthogonal matrices be partitioned into the subspaces $V_1, \dots, V_{2^p}$ so that $J_i V_i = V_1$, say, where $J_1 = I$ and V_1 is the set for which $e_{i1} \geq 0$. Let μ_1 be the measure in V_1 defined by the distribution of E assumed in the lemma. The measure $\mu(W)$ of a (measurable) set W in V_i is defined as $(1/2^p)\mu_1(J_i W)$. Now we want to show that μ is the Haar invariant measure. Let W be any (measurable) set in V_1. The lemma assumes that $2^p\mu(W) = \mu_1(W) = \Pr\{E \in W\} = \Pr\{E^{**} \in W\} = \sum \mu_1(J_i[WQ' \cap V_i]) = 2^p\mu(WQ')$. If U is any (measurable) set in V, then $U = \bigcup_{j=1}^{2^p}(U \cap V_j)$. Since $\mu(U \cap V_j) = (1/2^p)\mu_1[J_j(U \cap V_j)]$, by the above this is $\mu[(U \cap V_j)Q']$. Thus $\mu(U) = \mu(UQ')$. Thus μ is invariant and μ_1 is the conditional invariant distribution. ■

From the lemma we see that the matrix C has the conditional Haar invariant distribution. Since the distribution of C conditional on L is the same, C and L are independent.

Theorem 13.3.3. *If $C = Y'$, where $Y = (y_1, \dots, y_p)$ are the normalized characteristic vectors of A with $y_{1i} \geq 0$ and where A is distributed according to $W(I, n)$, then C has the conditional Haar invariant distribution and C is distributed independently of the characteristic roots.*

From the preceding work we can generalize Theorem 13.3.1.

Theorem 13.3.4. *If the symmetric matrix B has a density of the form $g(l_1, \dots, l_p)$, where $l_1 > \cdots > l_p$ are the characteristic roots of B, then the joint density of the roots is (2) and the matrix of normalized characteristic vectors Y $(y_{1i} \geq 0)$ is independently distributed according to the conditional Haar invariant distribution.*

Proof. The density of QBQ', where $QQ' = I$, is the same as that of B (for the roots are invariant), and therefore the distribution of $J(Y'Q')Y'Q'$ is the same as that of Y'. Then Theorem 13.3.4 follows from Lemma 13.3.2. ■

We shall give an application of this theorem to the case where $B = B'$ is normally distributed with the (functionally independent) components of B independent with means 0 and variances $\mathscr{E}b_{ii}^2 = 1$ and $\mathscr{E}b_{ij}^2 = \frac{1}{2}$ $(i < j)$.

Theorem 13.3.5. *Let $B = B'$ have the density*

$$\pi^{-p(p+1)/4} 2^{-\frac{1}{2}p} e^{-\frac{1}{4}\operatorname{tr} B^2}. \tag{25}$$

Then the characteristic roots $l_1 > \cdots > l_p$ of B have the density

$$2^{-\frac{1}{2}p} \pi^{p(p-1)/4} \Gamma_p^{-1}(\tfrac{1}{2}p) \exp\left(-\frac{1}{2}\sum_{i=1}^{p} l_i^2\right) \prod_{i<j} (l_i - l_j), \tag{26}$$

and the matrix Y of the normalized characteristic vectors $(y_{1i} \geq 0)$ is independently distributed according to the conditional Haar invariant distribution.

Proof. Since the characteristic roots of B^2 are $l_1^2, \dots, l_p^2$ and $\operatorname{tr} B^2 = \Sigma l_i^2$, the theorem follows directly. ■

Corollary 13.3.2. *Let nS be distributed according to $W(I, n)$, and define the diagonal matrix L and B by $S = C'LC$, $C'C = I$, $l_1 > \cdots > l_p$, and $c_{i1} \geq 0$, $i = 1, \dots, p$. Then the density of the limiting distribution of $\sqrt{n}(L - I) = D$ diagonal is (26) with l_i replaced by d_i, and the matrix C is independently distributed according to the conditional Haar measure.*

Proof. The density of the limiting distribution of $\sqrt{n}(S - I)$ is (25), and the diagonal elements of D are the characteristic roots of $\sqrt{n}(S - I)$ and the columns of C' are the characteristic vectors. ■

13.4. CANONICAL CORRELATIONS

The sample canonical correlations were shown in Section 12.3 to be the square roots of the roots of

$$|A_{12} A_{22}^{-1} A_{21} - f A_{11}| = 0, \tag{1}$$

where

$$A_{ij} = \sum_{\alpha=1}^{N} (X_\alpha^{(i)} - \bar{X}^{(i)})(X_\alpha^{(j)} - \bar{X}^{(j)})', \qquad i, j = 1, 2, \tag{2}$$

and the distribution of

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \tag{3}$$

is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \tag{4}$$

From Section 3.3 we know that the distribution of A_{ij} is the same as that of

$$A_{ij} = \sum_{\alpha=1}^{n} Y_\alpha^{(i)} Y_\alpha^{(j)\prime}, \qquad i, j = 1, 2, \tag{5}$$

where $n = N - 1$ and

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} \tag{6}$$

is distributed according to $N(\mathbf{0}, \boldsymbol{\Sigma})$. Let us assume that the dimensionality of $Y^{(1)}$, say p_1, is not greater than the dimensionality of $Y^{(2)}$, say p_2. Then there are p_1 nonzero roots of (1), say

$$f_1 > f_2 > \cdots > f_{p_1}. \tag{7}$$

Now we shall find the distribution of $\{f_i\}$ when

$$\boldsymbol{\Sigma}_{12} = \mathbf{0}. \tag{8}$$

For the moment assume $\{Y_\alpha^{(2)}\}$ to be fixed. Then A_{22} is fixed, and

$$B = A_{12} A_{22}^{-1} \tag{9}$$

is the matrix of regression coefficients of $Y^{(1)}$ on $Y^{(2)}$. From Section 4.3 we know that

$$\begin{aligned} A_{11\cdot 2} &= \sum_{\alpha=1}^{n} \left(Y_\alpha^{(1)} - BY_\alpha^{(2)}\right)\left(Y_\alpha^{(1)} - BY_\alpha^{(2)}\right)' = A_{11} - BA_{22}B' \\ &= A_{11} - A_{12}A_{22}^{-1}A_{21} \end{aligned} \tag{10}$$

and

$$Q = BA_{22}B' = A_{12}A_{22}^{-1}A_{21} \tag{11}$$

($\boldsymbol{\beta} = \mathbf{0}$) are independently distributed according to $W(\boldsymbol{\Sigma}_{11}, n - p_2)$ and $W(\boldsymbol{\Sigma}_{11}, p_2)$, respectively. In terms of Q the equation (1) defining f is

$$|Q - f(A_{11\cdot 2} + Q)| = 0. \tag{12}$$

The distribution of f_i, $i = 1, \ldots, p_1$, is the distribution of the nonzero roots of (12), and the density is given by (see Section 13.2)

$$\text{(13)} \quad \pi^{\frac{1}{2}p_1^2} \frac{\Gamma_{p_1}(\frac{1}{2}n)}{\Gamma_{p_1}[\frac{1}{2}(n-p_2)]\Gamma_{p_1}(\frac{1}{2}p_1)\Gamma_p(\frac{1}{2}p_2)} \cdot \prod_{i=1}^{p_1} \left\{ f_i^{\frac{1}{2}(p_2-p_1-1)} (1-f_i)^{\frac{1}{2}(N-p_2-p_1-2)} \right\} \prod_{i<j}^{p_1} (f_i - f_j).$$

Since the conditional density (13) does not depend upon $Y^{(2)}$, (13) is the unconditional density of the squares of the sample canonical correlation coefficients of the two sets $X_\alpha^{(1)}$ and $X_\alpha^{(2)}$, $\alpha = 1, \ldots, N$. The density (13) also holds when the $X^{(2)}$ are actually fixed variate vectors or have any distribution, so long as $X^{(1)}$ and $X^{(2)}$ are independently distributed and $X^{(1)}$ has a multivariate normal distribution.

In the special case when $p_1 = 1$, $p_2 = p - 1$, (13) reduces to

$$\text{(14)} \qquad \frac{\Gamma[\frac{1}{2}(N-1)]}{\Gamma[\frac{1}{2}(N-p)]\Gamma[\frac{1}{2}(p-1)]} f^{\frac{1}{2}(p-3)} (1-f)^{\frac{1}{2}(N-p-2)},$$

which is the density of the square of the sample multiple correlation coefficient between $X^{(1)}$ $(p_1 = 1)$ and $X^{(2)}$ $(p_2 = p - 1)$.

13.5. ASYMPTOTIC DISTRIBUTIONS IN THE CASE OF ONE WISHART MATRIX

13.5.1. All Population Roots Different

In Section 13.3 we found the density of the diagonal matrix L and the orthogonal matrix B defined by $S = BLB'$, $l_1 \geq \cdots \geq l_p$, and $b_{1i} \geq 0$, $i = 1, \ldots, p$, when nS is distributed according to $W(I, n)$. In this section we find the asymptotic distribution of L and B when nS is distributed according to $W(\Sigma, n)$ and the characteristic roots of Σ are different. (Corollary 13.3.2 gave the asymptotic distribution when $\Sigma = I$.)

Theorem 13.5.1. *Suppose nS has the distribution $W(\Sigma, n)$. Define diagonal Λ and L and orthogonal $\boldsymbol{\beta}$ and B by*

$$\text{(1)} \qquad \Sigma = \boldsymbol{\beta}\Lambda\boldsymbol{\beta}', \qquad S = BLB',$$

$\lambda_1 > \lambda_2 > \cdots > \lambda_p$, $l_1 \geq l_2 \geq \cdots \geq l_p$, $\beta_{1i} \geq 0$, $b_{1i} \geq 0$, $i = 1, \ldots, p$. *Define $G = \sqrt{n}(B - \boldsymbol{\beta})$ and diagonal $D = \sqrt{n}(L - \Lambda)$. Then the limiting distribution of*

D and G is normal with D and G independent, and the diagonal elements of D are independent. The diagonal element d_i has the limiting distribution $N(0, 2\lambda_i^2)$. The covariance matrix of g_i in the limiting distribution of $G = (g_1, \dots, g_p)$ is

$$\mathscr{C}(g_i) = \sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \beta_k \beta_k', \tag{2}$$

where $\beta = (\beta_1, \dots, \beta_p)$. The covariance matrix of g_i and g_j in the limiting distribution is

$$\mathscr{C}(g_i, g_j) = -\frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \beta_j \beta_i', \qquad i \neq j. \tag{3}$$

Proof. The matrix $nT = n\beta' S \beta$ is distributed according to $W(\Lambda, n)$. Let

$$T = YLY', \tag{4}$$

where Y is orthogonal. In order that (4) determine Y uniquely, we require $y_{ii} \geq 0$. Let $\sqrt{n}(T - \Lambda) = U$ and $\sqrt{n}(Y - I) = W$. Then (4) can be written

$$\Lambda + \frac{1}{\sqrt{n}} U = \left(I + \frac{1}{\sqrt{n}} W\right)\left(\Lambda + \frac{1}{\sqrt{n}} D\right)\left(I + \frac{1}{\sqrt{n}} W\right)', \tag{5}$$

which is equivalent to

$$U = W\Lambda + D + \Lambda W' + \frac{1}{\sqrt{n}}(WD + W\Lambda W' + DW') + \frac{1}{n} WDW'. \tag{6}$$

From $I = YY' = [I + (1/\sqrt{n})W][I + (1/\sqrt{n})W']$, we have

$$0 = W + W' + \frac{1}{\sqrt{n}} WW'. \tag{7}$$

We shall proceed heuristically and justify the method later. If we neglect terms of order $1/\sqrt{n}$ and $1/n$ (6) and (7), we obtain

$$U = W\Lambda + D + \Lambda W', \tag{8}$$

$$0 = W + W'. \tag{9}$$

When we substitute $W' = -W$ from (9) into (8) and write the result in components, we obtain $w_{ii} = 0$,

$$d_i = u_{ii}, \qquad i = 1, \dots, p, \tag{10}$$

$$w_{ij} = \frac{u_{ij}}{\lambda_j - \lambda_i}, \qquad i \neq j, \quad i, j = 1, \dots, p. \tag{11}$$

(Note $w_{ij} = -w_{ji}$.) From Theorem 3.4.4 we know that in the limiting normal distribution of U the functionally independent elements are statistically independent with means 0 and variances $\mathscr{AV}(u_{ii}) = 2\lambda_i^2$ and $\mathscr{AV}(u_{ij}) = \lambda_i\lambda_j$, $i \neq j$. Then the limiting distribution of D and W is normal, and $d_1, \ldots, d_p, w_{12}, w_{13}, \ldots, w_{p-1,p}$ are independent with means 0 and variances $\mathscr{AV}(d_i) = 2\lambda_i^2$, $i = 1, \ldots, p$, and $\mathscr{AV}(w_{ij}) = \lambda_i\lambda_j/(\lambda_j - \lambda_i)^2$, $j = i+1, \ldots, p$, $i = 1, \ldots, p-1$. Each column of B is $\pm$ the corresponding column of $\boldsymbol{\beta}Y$; since $Y \xrightarrow{p} I$, we have $\boldsymbol{\beta}Y \xrightarrow{p} \boldsymbol{\beta}$, and with arbitrarily high probability each column of B is nearly identical to the corresponding column of $\boldsymbol{\beta}Y$. Then $G = \sqrt{n}(B - \boldsymbol{\beta})$ has the limiting distribution of $\boldsymbol{\beta}\sqrt{n}(Y - I) = \boldsymbol{\beta}W$. The asymptotic variances and covariances follow.

Now we justify the limiting distribution of D and W. The equations $T = YLY'$ and $I = YY'$ and conditions $l_1 > \cdots > l_p$, $y_{ii} > 0$, $i = 1, \ldots, p$, define a 1–1 transformation of T to Y, L except for a set of measure 0. The transformation from Y, L to T is continuously differentiable. The inverse is continuously differentiable in a neighborhood of $Y = I$ and $L = \Lambda$, since the equations (8) and (9) can be solved uniquely. Hence Y, L as a function of T satisfies the conditions of Theorem 4.2.3. ■

13.5.2. One Root of Higher Multiplicity

In Section 11.7.3 we used the asymptotic distribution of the q smallest sample roots when the q smallest population roots are equal. We shall now derive that distribution. Let

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \lambda^* I_q \end{pmatrix}, \tag{12}$$

where the diagonal elements of the diagonal matrix Λ_1 are different and are larger than λ^* (> 0). Let

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \qquad Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}, \qquad L = \begin{bmatrix} L_1 & 0 \\ 0 & L_2 \end{bmatrix}. \tag{13}$$

Then $T \xrightarrow{p} \Lambda$, which implies $L \xrightarrow{p} \Lambda$, $Y_{11} \xrightarrow{p} I$, $Y_{12} \xrightarrow{p} 0$, $Y_{21} \xrightarrow{p} 0$, but Y_{22} does not have a probability limit. However, $Y_{22}Y_{22}' \xrightarrow{p} I_q$. Let the singular value decomposition of Y_{22} be EJF, where J is diagonal and E and F are orthogonal. Define $C_2 = EF$, which is orthogonal. Let $U = \sqrt{n}(I - \Lambda)$ and $D = \sqrt{n}(L - \Lambda)$ be partitioned similarly to T and L. Define $W_{11} = \sqrt{n}(Y_{11} - I)$, $W_{12} = \sqrt{n}\,Y_{12}$, $W_{21} = \sqrt{n}\,Y_{21}$, and $W_{22} = \sqrt{n}(Y_{22} = C_2) = \sqrt{n}\,E(J -$

$I_q)F$. Then (4) can be written

$$(14)\quad \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \lambda^* I_q \end{pmatrix} + \frac{1}{\sqrt{n}} \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}$$

$$= \left[\begin{pmatrix} I_{p-q} & 0 \\ 0 & C_2 \end{pmatrix} + \frac{1}{\sqrt{n}} \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \right]$$

$$\cdot \left[\begin{pmatrix} \Lambda_1 & 0 \\ 0 & \lambda^* I_q \end{pmatrix} + \frac{1}{\sqrt{n}} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \right]$$

$$\cdot \left[\begin{pmatrix} I_{p-q} & 0 \\ 0 & C_2' \end{pmatrix} + \frac{1}{\sqrt{n}} \begin{pmatrix} W_{11}' & W_{21}' \\ W_{12}' & W_{22}' \end{pmatrix} \right]$$

$$= \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \lambda^* I_q \end{pmatrix} + \frac{1}{\sqrt{n}} \left[\begin{pmatrix} D_1 & 0 \\ 0 & C_2 D_2 C_2' \end{pmatrix} \right.$$

$$\left. + \begin{pmatrix} W_{11}\Lambda_1 & \lambda^* W_{12} C_2' \\ W_{21}\Lambda_1 & \lambda^* W_{22} C_2' \end{pmatrix} + \begin{pmatrix} \Lambda_1 W_{11}' & \Lambda_1 W_{21}' \\ \lambda^* C_2 W_{12}' & \lambda^* C_2 W_{22}' \end{pmatrix} \right] + \frac{1}{n} M,$$

where the submatrices of M are sums of products of C_2, Λ_1, $\lambda^* I_q$, D_k, W_{kl}, and $1/\sqrt{n}$. The orthogonality of Y ($I_p = YY'$) implies

$$(15)\quad I_p = \begin{pmatrix} I_{p-q} & 0 \\ 0 & I_q \end{pmatrix} + \frac{1}{\sqrt{n}} \left[\begin{pmatrix} W_{11} & W_{12} C_2' \\ W_{21} & W_{22} C_2' \end{pmatrix} + \begin{pmatrix} W_{11}' & W_{21}' \\ C_2' W_{12}' & C_2 W_{22}' \end{pmatrix} \right] + \frac{1}{n} N,$$

where the submatrices of N are sums of products of W_{kl}. From (14) and (15) we find that

$$(16)\qquad U_{22} = C_2 D_2 C_2' + O_p(1/\sqrt{n}).$$

The limiting distribution of $(1/\lambda^*)U_{22}$ has the density (25) of Section 13.3 with p replaced by q. Then the limiting distribution of D_2 and C_2 is the distribution of D_2^* and Y_{22}^* defined by $U_{22}^* = Y_{22}^* D_2^* Y_{22}^{*\prime}$, where $(1/\lambda^*)U_{22}^*$ has the density (25) of Section 13.3.

Theorem 13.5.2. *Under the conditions of Theorem* 13.5.1 *and* $\Lambda = \operatorname{diag}(\Lambda_1, \lambda^* I_q)$, *the density of the limiting distribution of* $d_{p-q+1}, \ldots, d_p$ *is*

$$(17)\quad 2^{-\frac{1}{2}q}(\lambda^*\pi)^{q(q-1)/4}\Gamma_p^{-\frac{1}{2}}(\tfrac{1}{2}p)\exp\left(-\frac{1}{2\lambda^*}\sum_{i=p-q+1}^{p} d_i^2\right)\prod_{i<j}(d_i - d_j).$$

To justify the preceding derivation we note that D_2 and Y_{22} are functions of U depending on n that converge to the solution of $U_{22}^* = Y_{22}^* D_{22}^* Y_{22}^{*\prime}$. We can use the following theorem given by Anderson (1963a) and due to Rubin.

Theorem 13.5.3. *Let $F_n(u)$ be the cumulative distribution function of a random matrix U_n. Let V_n be a matrix-valued function of U_n, $V_n = f_n(u_n)$, and Let $G_n(v)$ be the (induced) distribution of V_n. Suppose $F_n(u) \to F(u)$ in every continuity point of $F(u)$, and suppose for every continuity point u of $f(u)$, $f_n(u_n) \to f(u)$ when $u_n \to u$. Let $G(v)$ be the distribution of the random matrix $V = f(U)$, where U has the distribution $F(u)$ If the probability of the set of discontinuities of $f(u)$ according to $F(u)$ is* 0, *then*

$$(18)\qquad \lim_{n\to\infty} G_n(v) = G(v)$$

in every continuity point of $G(v)$.

The details of verifying that $U(n)$ and

$$(19)\qquad (D_2(n), Y_{22}(n)) = f_n(U(nn))$$

satisfy the conditions of the theorem have been given by Anderson (1963a).

13.6. ASYMPTOTIC DISTRIBUTIONS IN THE CASE OF TWO WISHART MATRICES

13.6.1. All Population Roots Different

In Section 13.2 we studied the distributions of the roots $l_1 \geq l_2 \geq \cdots \geq l_p$ of

$$(1)\qquad |S^* - lT^*| = 0$$

and the vectors satisfying

$$(2)\qquad (S^* - lT^*)x^* = \mathbf{0}$$

and $x^{*\prime}T^*x^* = 1$ when $A^* = mS^*$ and $B^* = nT^*$ are distributed independently according to $W(\Sigma, m)$ and $W(\Sigma, n)$, respectively. In this section we study the asymptotic distributions of the roots and vectors as $n \to \infty$ when A^* and B^* are distributed independently according to $W(\Phi, m)$ and $W(\Sigma, n)$, respectively, and $m/n \to \eta > 0$. We shall assume that the roots of

$$|\Phi - \lambda\Sigma| = 0 \tag{3}$$

are distinct. (In Section 13.2 $\lambda_1 = \cdots = \lambda_p = 1$.)

Theorem 13.6.1. *Let mS^* and nT^* be independently distributed according to $W(\Phi, m)$ and $W(\Sigma, n)$, respectively. Let $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ (> 0) be the roots of* (3), *and let Λ be the diagonal matrix with the roots as diagonal elements in descending order; let $\gamma_1, \dots, \gamma_p$ be the solutions to*

$$(\Phi - \lambda_i\Sigma)\gamma = \mathbf{0}, \qquad i = 1, \dots, p, \tag{4}$$

$\gamma'\Sigma\gamma = 1$, and $\gamma_{1i} \geq 0$, and let $\Gamma = (\gamma_1, \dots, \gamma_p)$. Let $l_1 \geq \cdots \geq l_p$ (> 0) be the roots of (1), *and let L be the diagonal matrix with the roots as diagonal elements in descending order; let $x_1^*, \dots, x_p^*$ be the solutions to* (2) *for $l = l_i$, $i = 1, \dots, p$, $x^{*\prime}T^*x^* = 1$, and $x_{1i}^* > 0$, and let $X^* = (x_1^*, \dots, x_p^*)$. Define $Z^* = \sqrt{n}\,(X^* - \Gamma)$ and diagonal $D = \sqrt{n}\,(L - \Lambda)$. Then the limiting distribution of D and Z^* is normal with means $\mathbf{0}$ as $n \to \infty$, $m \to \infty$, and $m/n \to \eta$ (> 0). The asymptotic variances and covariances that are not* 0 *are*

$$\mathscr{AV}(d_i) = 2\frac{\lambda_i^2(1+\eta)}{\eta}, \tag{5}$$

$$\mathscr{AC}(z_i^*) = \sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i(\lambda_k + \eta\lambda_i)}{\eta(\lambda_k - \lambda_i)^2}\gamma_k\gamma_k' + \tfrac{1}{2}\gamma_i\gamma_i', \tag{6}$$

$$\mathscr{AC}(d_i, z_i^*) = \lambda_i\gamma_i, \tag{7}$$

$$\mathscr{AC}(z_i^*, z_j^*) = -\frac{\lambda_i\lambda_j(1+\eta)}{\eta(\lambda_j - \lambda_i)^2}\gamma_j\gamma_i', \qquad i \neq j. \tag{8}$$

Proof. Let

$$S = \Gamma'S^*\Gamma, \qquad T = \Gamma'T^*\Gamma. \tag{9}$$

Then mS and nT are distributed independently according to $W(\Lambda, m)$ and $W(I, n)$, respectively (Section 7.3.3). Then $l_1, \dots, l_p$ are the roots of

$$|S - lT| = 0. \tag{10}$$

Let $x_1, \dots, x_p$ be the solutions to

$$(S - l_i T)x = 0, \qquad i = 1, \dots, p, \tag{11}$$

and $x'Tx = 1$, and let $X = (x_1, \dots, x_p)$. Then $x_i^* = \Gamma x_i$ and $X^* = \Gamma X$ except possibly for multiplication of columns of X (or X^*) by -1. If $Z = \sqrt{n}(X - I)$, then $Z^* = \Gamma Z$ (except possibly for multiplication of columns by -1).

We shall now find the limiting distribution of D and Z. Let $\sqrt{n}(S - \Lambda) = U$ and $\sqrt{n}(T - I) = V$. Then U and V have independent limiting normal distributions with means 0. The functionally independent elements of U and V are statistically independent in the limiting distribution. The variances are $\mathscr{E}u_{ii}^2 \doteq 2(n/m)\lambda_i^2 \to 2\lambda_i^2/\eta$; $\mathscr{E}u_{ij}^2 = (r/m)\lambda_i\lambda_j \to \lambda_i\lambda_j/\eta$, $i \neq j$; $\mathscr{E}v_{ii}^2 = 2$; $\mathscr{E}v_{ij}^2 = 1$, $i \neq j$.

From the definition of L and X we have $SX = TXL$, $X'TX = I$, and $X'SX = L$. If we let $X^{-1} = G$, we obtain

$$S = G'LG, \qquad T = G'G. \tag{12}$$

We require $g_{ii} > 0$, $i = 1, \dots, p$. Since $S \xrightarrow{p} \Lambda$ and $T \xrightarrow{p} I$, we have $L \xrightarrow{p} \Lambda$ and $G \xrightarrow{p} I$. Let $\sqrt{n}(G - I) = H$. Then we write (12) as

$$\Lambda + \frac{1}{\sqrt{n}}U = \left(I + \frac{1}{\sqrt{n}}H'\right)\left(\Lambda + \frac{1}{\sqrt{n}}D\right)\left(I + \frac{1}{\sqrt{n}}H\right), \tag{13}$$

$$I + \frac{1}{\sqrt{n}}V = \left(I + \frac{1}{\sqrt{n}}H'\right)\left(I + \frac{1}{\sqrt{n}}H\right). \tag{14}$$

These can be rewritten

$$U = D + \Lambda H + H'\Lambda + \frac{1}{\sqrt{n}}(DH + H'D + H'\Lambda H) + \frac{1}{n}H'DH, \tag{15}$$

$$V = H + H' + \frac{1}{\sqrt{n}}H'H. \tag{16}$$

If we neglect the terms of order $1/\sqrt{n}$ and $1/n$ (as in Section 13.5), we can write

$$U = D + \Lambda H + H'\Lambda, \tag{17}$$

$$V = H + H', \tag{18}$$

$$U - V\Lambda = D + \Lambda H - H\Lambda. \tag{19}$$

The diagonal elements of (18) and the components of (19) are

$$v_{ii} = 2h_{ii}, \tag{20}$$

$$u_{ii} - \lambda_i v_{ii} = d_i, \tag{21}$$

$$u_{ij} - v_{ij}\lambda_j = (\lambda_i - \lambda_j)h_{ij}, \qquad i \neq j. \tag{22}$$

The limiting distribution of $\boldsymbol{H}$ and $\boldsymbol{D}$ is normal with means 0. The pairs (h_{ij}, h_{ji}) of off-diagonal elements of $\boldsymbol{H}$ are independent with variances

$$\mathscr{AV}(h_{ij}) = \frac{\lambda_j(\lambda_i + \eta\lambda_j)}{\eta(\lambda_i - \lambda_j)^2}, \qquad i \neq j, \tag{23}$$

and covariances

$$\mathscr{AC}(h_{ij}, h_{ji}) = -\frac{\lambda_i\lambda_j(1+\eta)}{\eta(\lambda_i - \lambda_j)^2}, \qquad i \neq j. \tag{24}$$

The pairs (d_i, h_{ii}) of diagonal elements of $\boldsymbol{D}$ and $\boldsymbol{H}$ are independent with variances (5),

$$\mathscr{AV}(h_{ii}) = \tfrac{1}{2}, \tag{25}$$

and covariance

$$\mathscr{AC}(d_i, h_{ii}) = -\lambda_i. \tag{26}$$

The diagonal elements of $\boldsymbol{D}$ and $\boldsymbol{H}$ are independent of the off-diagonal elements of $\boldsymbol{H}$.

That the limiting distribution of $\boldsymbol{D}$ and $\boldsymbol{H}$ is normal is justified by Theorem 4.2.3. $\boldsymbol{S}$ and $\boldsymbol{T}$ are polynomials in $\boldsymbol{L}$ and $\boldsymbol{G}$, and their derivatives are polynomials and hence continuous. Since the equations (12) with auxiliary conditions can be solved uniquely for $\boldsymbol{L}$ and $\boldsymbol{G}$, the inverse function is also continuously differentiable at $\boldsymbol{L} = \boldsymbol{\Lambda}$ and $\boldsymbol{G} = \boldsymbol{I}$. By Theorem 4.2.3, $\boldsymbol{D} = \sqrt{n}(\boldsymbol{L} - \boldsymbol{\Lambda})$ and $\boldsymbol{H} = \sqrt{n}(\boldsymbol{G} - \boldsymbol{I})$ have a limiting normal distribution. In turn, $\boldsymbol{X} = \boldsymbol{G}^{-1}$ is continuously differentiable at $\boldsymbol{G} = \boldsymbol{I}$, and $\boldsymbol{Z} = \sqrt{n}(\boldsymbol{X} - \boldsymbol{I}) = \sqrt{n}(\boldsymbol{G}^{-1} - \boldsymbol{I})$ has the limiting distribution of $-\boldsymbol{H}$. (Expand $\sqrt{n}\{[\boldsymbol{I} + (1/\sqrt{n})\boldsymbol{H}]^{-1} - \boldsymbol{I}\}$.) Since $\boldsymbol{G} \xrightarrow{p} \boldsymbol{I}$, $\boldsymbol{X} \xrightarrow{p} \boldsymbol{I}$, and $x_{ii} > 0$, $i = 1, \dots, p$ with probability approaching 1. Then $\boldsymbol{Z}^* = \sqrt{n}(\boldsymbol{X}^* - \boldsymbol{\Gamma})$ has the limiting distribution of $\boldsymbol{\Gamma Z}$. (Since $\boldsymbol{X} \xrightarrow{p} \boldsymbol{I}$, we have $\boldsymbol{X}^* = \boldsymbol{\Gamma X} \xrightarrow{p} \boldsymbol{\Gamma}$ and $x_{1i} > 0$, $i = 1, \dots, p$, with probability approaching 1.) The asymptotic variances and covariances (6) to (8) are obtained from (23) to (26). ∎

Anderson (1989b), has derived the limiting distribution of the characteristic roots and vectors of one sample covariance matrix in the metric of another with population roots of arbitrary multiplicities.

13.6.2. One Root of Higher Multiplicity

In Section 13.6.1 it was assumed that mS^* and nT^* were distributed independently according to $W(\Phi, m)$ and $W(\Sigma, n)$, respectively, and that the roots of $|\Phi - \lambda\Sigma| = 0$ were distinct. In this section we assume that the k larger roots are distinct and greater than the $p-k$ smaller roots, which are assumed equal. Let the diagonal matrix Λ of characteristic roots be $\Lambda = \operatorname{diag}(\Lambda_1, \lambda^* I_{p-k})$, and let Γ be a matrix satisfying

$$\Phi\Gamma = \Sigma\Gamma\Lambda, \qquad \Gamma'\Sigma\Gamma = I. \tag{27}$$

Define S and T by (9) and diagonal L and G by (12). Then $S \xrightarrow{p} \Lambda$, $T \xrightarrow{p} I_p$, and $L \xrightarrow{p} \Lambda$. Partition S, T, L, and G as

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \qquad T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \tag{28}$$

$$L = \begin{bmatrix} L_1 & 0 \\ 0 & L_2 \end{bmatrix}, \qquad G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

where S_{11}, T_{11}, L_1, and G_{11} are $k \times k$. Then $G_{11} \xrightarrow{p} I_k$, $G_{12} \xrightarrow{p} 0$, and $G_{21} \xrightarrow{p} 0$, but G_{22} does not have a probability limit. Instead $G_{22}'G_{22} \xrightarrow{p} I_{p-k}$. Let the singular value decomposition of G_{22} be EJF, where E and F are orthogonal and J is diagonal. Let $C_2 = EF$.

The limiting distribution of $U = \sqrt{n}(S - \Lambda)$ and $V = \sqrt{n}(T - I)$ is normal with the covariance structure given above (12) with $\lambda_{k+1} = \cdots = \lambda_p = \lambda^*$. Define $D = \sqrt{n}(L - \Lambda)$, $H_{11} = \sqrt{n}(G_{11} - I)$, $H_{12} = \sqrt{n}\, G_{12}$, $H_{21} = \sqrt{n}\, G_{21}$, and $H_{22} = \sqrt{n}(G_{22} - C_2) = \sqrt{n}\, E(J - I_{p-k})F$. Then (13) and (15) are replaced by

$$\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \lambda^* I_{p-k} \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \tag{29}$$

$$= \begin{bmatrix} I_k + \frac{1}{\sqrt{n}} H_{11}' & \frac{1}{\sqrt{n}} H_{21}' \\ \frac{1}{\sqrt{n}} H_{12}' & C_2' + \frac{1}{\sqrt{n}} H_{22}' \end{bmatrix} \begin{bmatrix} \Lambda_1 + \frac{1}{\sqrt{n}} D_1 & 0 \\ 0 & \lambda^* I_{p-k} + \frac{1}{\sqrt{n}} D_2 \end{bmatrix}$$

$$\cdot\begin{bmatrix} I_k + \frac{1}{\sqrt{n}}H_{11} & \frac{1}{\sqrt{n}}H_{12} \\ \frac{1}{\sqrt{n}}H_{21} & C_2 + \frac{1}{\sqrt{n}}H_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \lambda^* I_{p-k} \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} D_1 & 0 \\ 0 & C_2' D_2 C_2 \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} \Lambda_1 H_{11} & \Lambda_1 H_{12} \\ \lambda^* C_2' H_{21} & \lambda^* C_2' H_{22} \end{bmatrix}$$

$$+ \frac{1}{\sqrt{n}} \begin{bmatrix} H_{11}'\Lambda_1 & \lambda^* H_{21}' C_2 \\ H_{12}'\Lambda_1 & \lambda^* H_{22}' C_2 \end{bmatrix} + O_p\left(\frac{1}{\sqrt{n}}\right),$$

and (14) and (16) are replaced by

$$\begin{bmatrix} I_k & 0 \\ 0 & I_{p-k} \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \tag{30}$$

$$= \begin{bmatrix} I + \frac{1}{\sqrt{n}}H_{11}' & \frac{1}{\sqrt{n}}H_{21}' \\ \frac{1}{\sqrt{n}}H_{12}' & C_2' + \frac{1}{\sqrt{n}}H_{22}' \end{bmatrix} \begin{bmatrix} I + \frac{1}{\sqrt{n}}H_{11} & \frac{1}{\sqrt{n}}H_{12} \\ \frac{1}{\sqrt{n}}H_{21} & C_2 + \frac{1}{\sqrt{n}}H_{22} \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 \\ 0 & I_{p-k} \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & C_2' H_{22} \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} H_{11}' & H_{21}' \\ H_{12}' & H_{22}' C_2 \end{bmatrix} + \frac{1}{n} H'H.$$

If we neglect the terms of order $1/\sqrt{n}$ and $1/n$, instead of (19) we can write

(31)

$$\begin{bmatrix} U_{11} - V_{11}\Lambda_1 & U_{12} - \lambda^* V_{12} \\ U_{21} - V_{21}\Lambda_1 & U_{22} - \lambda^* V_{22} \end{bmatrix} = \begin{bmatrix} D_1 + \Lambda_1 H_{11} - H_{11}\Lambda_1 & (\lambda^* I - \Lambda_1) H_{12} C_2 \\ C_2' H_{21}(\lambda^* I - V_1) & C_2' D_2 C_2 \end{bmatrix}$$

Then $v_{ii} = 2h_{ii}$, $i = 1, \dots, k$; $u_{ii} - \lambda_i v_{ii} = d_i$, $i = 1, \dots, k$; $u_{ij} - v_{ij}\lambda_j = (\lambda_i - \lambda_j)h_{ij}$, $i \neq j$, $i, j = 1, \dots, k$; $U_{22} - \lambda^* V_{22} = C_2' D_2 C_2$; $C_2(U_{21} - V_{21}\Lambda_1) = H_{21}(\lambda^* I - \Lambda_1)$; and $(U_{12} - \lambda^* V_{12})C_2' = (\lambda^* I - \Lambda_1)H_{12}$. The limiting distribution of $U_{22} - \lambda^* V_{22}$ is normal with mean $\mathbf{0}$; $\mathscr{E}(u_{ii} - \lambda^* v_{ii})^2 = \mathscr{E} d_i^2 = 2\lambda^{*2}(1 + \eta)/\eta$, $i = k+1, \dots, p$; and $\mathscr{E}(u_{ij} - \lambda^* v_{ij})^2 = \lambda^{*2}(1 + \eta)/\eta$, $i \neq j$, $i, j = k+1, \dots, p$. The limiting distribution of D_2 and C_2 is the distribution of D_2 and C_2 defined by $U_{22} - \lambda^* V_{22} = C_2' D_2 C_2$ where $(1/\lambda^*)(U_{22} - V_{22})$ has the density of (25) of Section 13.3.

13.7. ASYMPTOTIC DISTRIBUTION IN A REGRESSION MODEL

13.7.1. Both Sets of Variates Stochastic

The sample canonical correlations $l_1, \dots, l_{p_2}$ and vectors $\hat{\alpha}_1, \dots, \hat{\alpha}_{p_1}$, and $\hat{\gamma}_1, \dots, \hat{\gamma}_{p_2}$ are defined in Section 12.3. The set $\hat{\gamma}_1, \dots, \hat{\gamma}_{p_2}$ and $l_1, \dots, l_{p_2}$ are defined by

$$S_{21} S_{11}^{-1} S_{12} \hat{\gamma} = S_{22} \hat{\gamma} l^2, \qquad \hat{\gamma}' S_{22} \hat{\gamma} = 1. \tag{1}$$

The asymptotic distribution of these quantities was given by Anderson (1999a) when $X = (X^{(1)\prime}, X^{(2)\prime})'$ has a normal distribution and also when $X^{(1)}$ is normally distributed with a linear function of nonstochastic $X^{(2)}$ as expected value. We shall now find the asymptotic distribution when X has a normal distribution. The model in regression form is

$$X^{(1)} = \mathbf{\beta} X^{(2)} + Z, \tag{2}$$

where $X^{(2)}$ and Z are independently normally distributed with expected values $\mathscr{E} X^{(2)} = \mathbf{0}$ and $\mathscr{E} Z = \mathbf{0}$ and covariances $\mathscr{E} X^{(2)} X^{(2)} = \Sigma_{22}$, $\mathscr{E} ZZ' = \Sigma_{ZZ}$ ($\mathscr{E} X^{(2)} Z' = \mathbf{0}$). Then $\mathscr{E} X^{(1)} = \mathbf{0}$ and $\mathscr{E} X^{(1)} X^{(1)\prime} = \Sigma_{11} = \Sigma_{ZZ} + \mathbf{\beta} \Sigma_{22} \mathbf{\beta}'$ and $\mathscr{E} X^{(1)} X^{(2)\prime} = \mathbf{\beta} \Sigma_{22}$. Inference is based on a sample of X of n observations.

First we transform to canonical variables $U = \mathrm{A}' X^{(1)}$, $V = \Gamma' X^{(2)}$, and $W = \mathrm{A}' Z$. Then (1) is transformed to

$$U = \Theta V + W, \tag{3}$$

where $\Theta = A' \mathbf{\beta} (\Gamma')^{-1}$, $\mathscr{E} UU' = \Sigma_{UU} = I_{p_1}$, $\mathscr{E} VV' = \Sigma_{VV} = I_{p_2}$, $\mathscr{E} UV' = \Sigma_{UV} = (\Lambda, \mathbf{0}) = \overline{\Lambda}$, $\mathscr{E} WW' = \Sigma_{WW} = I_{p_1} - \Lambda^2$, and $\mathscr{E} VW' = \mathbf{0}$. [See (33) to (37) and (45) of Section 12.2.] Let the sample covariance matrices be $S_{UU} = \mathrm{A}' S_{11} \mathrm{A}'$, $S_{UV} = \mathrm{A}' S_{12} \Gamma$, and $S_{VV} = \Gamma' S_{22} \Gamma$. Let the sample vectors constitute $H = \Gamma^{-1} \hat{\Gamma} = \Gamma^{-1} (\hat{\gamma}_1, \dots, \hat{\gamma}_{p_2})$. Then H satisfies

$$S_{VU} S_{UU}^{-1} S_{UV} H = S_{VV} H \hat{\Lambda}^{+2}, \qquad H' S_{VV} H = I_{p_2}, \tag{4}$$

where $\hat{\Lambda}^{+} = \operatorname{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_{p_1}, 0, \dots, 0)$; if $p_1 < p_2$, there are $p_2 - p_1$ 0's in $\hat{\Lambda}^{+}$.

We have $S_{UU} \xrightarrow{p} I_{p_1}$, $S_{VV} \xrightarrow{p} I_{p_2}$, $S_{UV} \xrightarrow{p} \overline{\Lambda}$. Then $\hat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_{p_1}) \xrightarrow{p} \Lambda$. Let

$$H = (H_1, H_2) = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}, \tag{5}$$

where H_{11} is $p_1 \times p_1$ and H_{22} is $(p_2 - p_1) \times (p_2 - p_1)$. The first p_1 columns of (4) are

$$S_{VU} S_{UU}^{-1} S_{UV} H_1 = S_{VV} H_1 \hat{\Lambda}^2; \tag{6}$$

the last $p_2 - p_1$ columns of (4) are $S_{UV} H_2 = \mathbf{0}$. Then $H_{11} \xrightarrow{p} I_{p_1}$, $H_{12} \xrightarrow{p} \mathbf{0}$, and $H_{21} \xrightarrow{p} \mathbf{0}$, but the probability limit of (4) only implies $H_{22}' H_{22} \xrightarrow{p} I_{p_2 - p_1}$. Let the singular value decomposition of H_{22} be $H_{22} = EJF$.

Define $S_{UU}^* = \sqrt{n}(S_{UU} - I_{p_1})$, $S_{VV}^* = \sqrt{n}(S_{VV} - I_{p_2})$, $S_{UV}^* = \sqrt{n}(S_{UV} - \overline{\Lambda})$, $H_1^* = \sqrt{n}(H_1 - I_{(p_1)})$, and $\overline{\Lambda}^* = [\sqrt{n}(\hat{\Lambda} - \Lambda), \mathbf{0}]$, where $I_{(p_1)} = (I_{p_1}, \mathbf{0})'$. Then expansion of (6) yields

$$\left(\overline{\Lambda}' + \frac{1}{\sqrt{n}} S_{VU}^*\right)\left(I_{p_1} + \frac{1}{\sqrt{n}} S_{UU}^*\right)^{-1}\left(\overline{\Lambda} + \frac{1}{\sqrt{n}} S_{UV}^*\right)\left(I_{(p_1)} + \frac{1}{\sqrt{n}} H_1^*\right) \tag{7}$$
$$= \left(I_{p_2} + \frac{1}{\sqrt{n}} S_{VV}^*\right)\left(I_{(p_1)} + \frac{1}{\sqrt{n}} H_1^*\right)\left(\Lambda + \frac{1}{\sqrt{n}} \Lambda^*\right)^2.$$

From (7) we obtain

$$S_{VU}^* \overline{\Lambda} I_{(p_1)} - \overline{\Lambda}' S_{UU}^* \overline{\Lambda} I_{(p_1)} + \overline{\Lambda}' S_{UV}^* I_{(p_1)} + \overline{\Lambda}' \overline{\Lambda} H_1^* \tag{8}$$
$$= S_{VV}^* I_{(p_1)} \Lambda^2 + H_1^* \Lambda^2 + 2 I_{(p_1)} \Lambda \Lambda^* + o_p(1)$$

From $H_1' S_{VV} H_1 = I_{p_1}$ we derive

$$H_{11}^* + H_{11}^{*\prime} = -S_{VV}^{*11} + o_p(1). \tag{9}$$

In terms of partitioned matrices (8) is

$$\begin{bmatrix} S_{UV}^{*11}\Lambda - \Lambda S_{UU}^{*11}\Lambda + \Lambda S_{UV}^{*11} - S_{VV}^{*11}\Lambda^2 \\ S_{VU}^{*21}\Lambda - S_{VV}^{*21}\Lambda^2 \end{bmatrix} \tag{10}$$
$$= \begin{bmatrix} 2\Lambda\Lambda^* + H_{11}^*\Lambda^2 - \Lambda^2 H_{11}^* \\ H_{21}^*\Lambda^2 \end{bmatrix} + o_p(1).$$

The lower submatrix equation $[(p_2 - p_1) \times p_1]$ of (10) is

$$H_{21}^* \Lambda = S_{VU}^{*21} - S_{VV}^{*21}\Lambda + o_p(1) = S_{VW}^{*21} + o_p(1) = \left(S_{WV}^{*12}\right)' + o_p(1). \tag{11}$$

A diagonal element of the upper submatrix equation of (10) is

$$(12)\quad \lambda_i^* = \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} \left[(u_{i\alpha} v_{i\alpha} - \lambda_i) - \tfrac{1}{2}\lambda_i (u_{i\alpha}^2 - 1) - \tfrac{1}{2}\lambda_i (v_{i\alpha}^2 - 1) \right] + o_p(1).$$

The right-hand side of (12) is the expansion of the sample correlation coefficient of $u_{i\alpha}$ and $v_{i\alpha}$. See Section 4.2.3. The limiting distribution of λ_i^* is $N[0,(1-\lambda_i^2)^2]$.

The (i,j)th component of H_{11}^* in (10) is

(13)

$$(\lambda_j^2 - \lambda_i^2) h_{ij}^* = \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} \left(\lambda_j v_{i\alpha} u_{j\alpha} + \lambda_i u_{i\alpha} v_{j\alpha} - \lambda_i \lambda_j u_{i\alpha} u_{j\alpha} - \lambda_j^2 v_{i\alpha} v_{j\alpha} \right) + o_p(1).$$

$$i \neq j \quad i,j = 1,\ldots,p_1.$$

The asymptotic covariance of $(\lambda_j^2 - \lambda_i^2) h_{ij}^*$ and $(\lambda_i^2 - \lambda_j^2) h_{ji}^*$ is

$$(14)\quad \begin{bmatrix} (1-\lambda_j^2)(\lambda_i^2 + \lambda_j^2 - 2\lambda_i^2\lambda_j^2) & (1-\lambda_i^2)(1-\lambda_j^2)(\lambda_i^2+\lambda_j^2) \\ (1-\lambda_i^2)(1-\lambda_j^2)(\lambda_i^2+\lambda_j^2) & (1-\lambda_i^2)(\lambda_i^2+\lambda_j^2 - 2\lambda_i^2\lambda_j^2) \end{bmatrix}.$$

The pair (h_{ij}^*, h_{ji}^2) is uncorrelated with other pairs.

Suppose $p_1 = p_2$. Then $\hat{\Gamma}^* = \Gamma H_{11}^* = \Gamma H^*$. Let $\Gamma = (\gamma_1, \ldots, \gamma_{p_1})$, $\hat{\Gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_{p_1})$. Then $\hat{\gamma}_j^* = \sum_{i=1}^{p_1} \gamma_i h_{ij}^*$, where h_{ij}^*, $i \neq j$, is obtained from (13) and h_{ii}^* from (9). We obtain

$$(15)\quad n\mathscr{E}(\hat{\gamma}_j - \gamma_j)(\hat{\gamma}_j - \gamma_j)' = \tfrac{1}{2}\gamma_j\gamma_j' + (1-\lambda_j^2) \sum_{\substack{k=1 \\ k \neq j}}^{p_1} \frac{\lambda_k^2 + \lambda_j^2 - 2\lambda_k^2\lambda_j^2}{(\lambda_j^2 - \lambda_k^2)^2},$$

$$(16)\quad n\mathscr{E}(\hat{\gamma}_j - \gamma_j)(\hat{\gamma}_l - \gamma_l)' = \frac{(1-\lambda_j^2)(1-\lambda_l^2)(\lambda_j^2 + \lambda_l^2)}{(\lambda_j^2 - \lambda_l^2)^2} \gamma_l \gamma_j', \qquad j \neq l.$$

Anderson (1999a), has also given the asymptotic covariances of $\hat{\alpha}_j$ and of $\hat{\gamma}_j$ and $\hat{\alpha}_l$. Note that h_{ij}^* depends linearly on $(u_{i\alpha}, v_{i\alpha})$ and that the pairs $(u_{i\alpha}, v_{i\alpha})$ and $(u_{j\alpha}, v_{j\alpha})$, $i \neq j$, are uncorrelated. The covariances (14) do not depend on (U, V) being normal.

Now suppose that the rank of Γ_{12} is $k < p_1$. Define $H_1 = (H_{11}', H_{21}')'$ as the first k columns of H satisfying (4), and define $\hat{\Lambda}_1 = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$. Then H_1 satisfies (6), and H_1^* satisfies (8), (9), (10), and (11). Then λ_i^* is given by (12) for $i = 1, \ldots, k$.

The last $p_1 - k$ columns of (4) are

$$\begin{bmatrix} \Lambda_1 S_{UV}^{*12} C_2 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} -\Lambda_1^2 H_{12}^* \\ \mathbf{0} \end{bmatrix} + o_p(1). \tag{17}$$

Hence

$$\Lambda_1 H_{12}^* = -S_{UV}^{*12} C_2 + o_p(1) = -S_{WV}^{*12} C_2 + o_p(1). \tag{18}$$

13.7.2. One Set of Variates Stochastic and the Other Set Nonstochastic

Now consider the case that $X^{(2)}$ in (2) is nonstochastic, where $\mathscr{E} Z_\alpha = \mathbf{0}$ and $\mathscr{E} Z_\alpha Z_\alpha' = \Sigma_{ZZ}$. We observe $X = x_1, \dots, x_n$. We assume

$$S_{22} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha^{(2)} x_\alpha^{(2)\prime} \to \Sigma_{22}, \tag{19}$$

and Σ_{22} is nonsingular. Then

$$S_{11} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha^{(1)} x_\alpha^{(1)\prime} = \beta S_{22} \beta + S_{Z2} \beta' + \beta S_{2Z} + S_{ZZ} \xrightarrow{p} \beta \Sigma_{22} \beta' + \Sigma_{ZZ}, \tag{20}$$

$$S_{12} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha^{(1)} x_\alpha^{(2)\prime} = \beta S_{22} + S_{Z2} \xrightarrow{p} \beta \Sigma_{22}. \tag{21}$$

Define λ, α, and γ by solutions to

$$\begin{bmatrix} -\lambda(\Sigma_{ZZ} + \beta S_{22} \beta') & \beta S_{22} \\ S_{22} \beta & -\lambda S_{22} \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = \mathbf{0}, \tag{22}$$

$$\alpha'(\Sigma_{ZZ} + \beta S_{22} \beta')\alpha = 1, \qquad \gamma' S_{22} \gamma - 1. \tag{23}$$

We shall first assume $p_1 = p_2$ and $\lambda_1 > \cdots > \lambda_{p_1} > 0$. Then (22) and (23) and $\alpha_{ii} > 0$ define

$$\operatorname{diag}(\lambda_1, \dots, \lambda_{p_1}) = \Lambda_n, \qquad (\alpha_1, \dots, \alpha_{p_1}) = \mathrm{A}_n, \qquad (\gamma_1, \dots, \gamma_{p_1}) = \Gamma_n. \tag{24}$$

Let $U = \mathrm{A}_n' X^{(1)}$, $v_\alpha = \Gamma_n' x_\alpha$, $\alpha = 1, \dots, n$, $W = \mathrm{A}_n' Z$, $\Theta = \mathrm{A}_n' \beta (\Gamma_n')^{-1} = \Lambda$, $H = \Gamma_n^{-1} \hat{\Gamma}$. Then H and $\hat{\Lambda}$ satisfy (4). Then $S_{VV} = I$,

$$S_{UV} = \Theta S_{VV} + S_{WV} = \Theta + S_{WV} \xrightarrow{p} \Theta, \tag{25}$$

$$S_{UU} = \Theta S_{VV} \Theta + \Theta S_{VW} + S_{WV} \Theta + S_{WW} \xrightarrow{p} I. \tag{26}$$

Then (4) can be written

$$(27)\quad (\Lambda + S_{VW})(\Lambda^2 + \Lambda S_{VW} + S_{WV}\Lambda + S_{WW})^{-1}(\Lambda + S_{WV})H = H\hat{\Lambda}^2.$$

Note that $S_{VW} \xrightarrow{p} \mathbf{0}$, $S_{WW} \xrightarrow{p} I - \Lambda^2$, and hence $H \xrightarrow{p} I$, $\hat{\Lambda} \xrightarrow{p} \Lambda$.

Let $S_{VW}^* = \sqrt{n}\, S_{VW}$, $S_{WW}^* = \sqrt{n}\,[S_{WW} - (I - \Lambda^2)]$. Then (27) leads to

$$(28)\quad \begin{aligned}(I - \Lambda^2) S_{VW}^* \Lambda + \Lambda S_{WV}^*(I - \Lambda^2) - \Lambda S_{WW}^* \Lambda + \Lambda^2 H^* \\ = H^* \Lambda^2 + 2\Lambda\Lambda^* + o_p(1).\end{aligned}$$

A diagonal term of (28) gives

$$(29)\quad \lambda_i^* = (1 - \lambda_i^2)\frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} v_{i\alpha} w_{i\alpha} - \tfrac{1}{2}\lambda_i \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} \left[w_{i\alpha}^2 - (1 - \lambda_i^2)\right] + o_p(1).$$

Since

$$(30)\quad \mathscr{E}(v_{i\alpha} w_{i\alpha})^2 = v_{i\alpha}^2 (1 - \lambda_i^2),$$

$$(31)\quad \mathscr{E}\left[w_{i\alpha}^2 - (1 - \lambda_i^2)\right]^2 = 2(1 - \lambda_i^2)^2$$

under the assumption that W is normally distributed, the limiting distribution of $\sqrt{n}(\hat{\lambda}_i - \lambda_i)$ is $N[0, (1 - \lambda_i^2)^2(1 - \frac{1}{2}\lambda_i^2)]$. Note that this variance is smaller than in the case of $X^{(2)}$ stochastic.

From (28) we find

(32)
$$(\lambda_j^2 - \lambda_i^2) h_{ij}^* = \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} \left[(1 - \lambda_i^2) v_{i\alpha} w_{j\alpha} \lambda_j + \lambda_i w_{i\alpha} v_{j\alpha}(1 - \lambda_j^2) - \lambda_i w_{i\alpha} w_{j\alpha} \lambda_j\right] + o_p(1).$$

Then

$$(33)\quad (\lambda_j^2 - \lambda_i^2)^2 \mathscr{E}(h_{ij}^*)^2 \to (1 - \lambda_i^2)(1 - \lambda_j^2)(\lambda_i^2 - \lambda_j^2 - \lambda_i^2\lambda_j^2).$$

The equation $H' S_{VV} H = I$ implies $H'H = I$, leading to $H^* = -H^{*\prime} + o_p(1)$, that is, $h_{ij}^* = -h_{ji}^* + o_p(1)$.

Now suppose that the rank of $\boldsymbol{\beta}$ is $k < p_1 = p_2$. Then $\Lambda = \operatorname{diag}(\Lambda_1, \mathbf{0})$, where $\Lambda_1 = \operatorname{diag}(\lambda_1, \dots, \lambda_k)$. Let $\Gamma = (\Gamma_1, \Gamma_2)$, where Γ_1 has k columns and Γ_2 has $p_1 - k$ columns. Define the partition (5) to be made into k and $p_1 - k$ rows and columns. The probability limit of (4) implies $H_{11} \xrightarrow{p} I_k$, $H_{12} \xrightarrow{p} \mathbf{0}$, $H_{21} \xrightarrow{p} \mathbf{0}$, and $H_{22}' H_{22} \xrightarrow{p} I$. Let the singular value decomposition of H_{22} be

EJF, where J is a diagonal matrix of order $p_1 - k$ and E and F are orthogonal matrices of order $p_1 - k$. Define $C_2 = EF$. The expansion of (4) in terms of $S_{VW}^* = \sqrt{n}(S_{VW} - \Lambda)$, $S_{WV}^* = \sqrt{n}(S_{UV}^* - \Lambda)$, $S_{WW}^* = \sqrt{n}[S_{WW} - (I - \Lambda^2)]$, $H_{11}^* = \sqrt{n}(H_{11} - I)$, $H_{12}^* = \sqrt{n}H_{12}$, $H_{21}^* = \sqrt{n}H_{21}$, and $H_{22}^* = \sqrt{n}(H_{22} - C_2) = \sqrt{n}E(J - I)F$ yields

$$\text{(34)} \quad \begin{bmatrix} \Lambda_1 S_{WV}^{*11}(I - \Lambda_1^2) + (I - \Lambda_1^2)S_{VW}^{*11}\Lambda_1 - \Lambda_1 S_{WV}^{*11}\Lambda_1 & \Lambda_1 S_{WV}^{*12}C_2 \\ S_{VW}^{*21}\Lambda_1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2\Lambda_1\Lambda_1^* + H_{11}^*\Lambda_1^2 - \Lambda_1^2 H_{11}^* & -\Lambda_1^2 H_{12}^* \\ H_{21}^*\Lambda_1^2 & 0 \end{bmatrix} + o_p(1).$$

The ith diagonal term of (34) is (29) for $i = 1, \ldots, k$. The i, jth element of the upper left-hand submatrix is (32) for $i \neq j$ and $i, j = 1, \ldots, k$. Two other submatrix equations of (34) are

$$\text{(35)} \qquad \Lambda_1 H_{12}^* = -S_{WV}^{*12}C_2 + o_p(1),$$

$$\text{(36)} \qquad H_{21}^*\Lambda_1 = S_{VW}^{*21} + o_p(1).$$

The equation $I = H'S_{VV}H = H'H$ yields

$$\text{(37)} \qquad \begin{bmatrix} H_{11}^* + H_{11}^* & H_{12}^* + H_{21}^{*\prime}C_2 \\ C_2'H_{21}^* + H_{12}^{*\prime} & C_2 H_{22}^* + H_{22}^{*\prime}C_2 \end{bmatrix} = 0 + o_p(1).$$

The off-diagonal submatrices of (37) agree with (35) and (36).

13.7.3. Reduced Rank Regression Estimator

When the rank of $\boldsymbol{\beta}$ is specified to be k $(< p_1)$, the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\text{(38)} \qquad \hat{B}_k = S_{12}\hat{\Gamma}_1\hat{\Gamma}_1'.$$

See Section 12.7. In terms of (3) the reduced-rank regression estimator of Θ is

$$\text{(39)} \qquad \hat{\Theta}_k = S_{UV}H_1H_1'.$$

Suppose $X^{(2)}$ is stochastic and $\Theta = \operatorname{diag}(\Theta_1, 0) = \operatorname{diag}(\Lambda_1, 0)$. We define $\Theta_k^* = \sqrt{n}(\hat{\Theta}_k - \Theta)$, $H_1^* = \sqrt{n}(H_1 - I_{(k)})$, $S_{UV}^* = \sqrt{n}(S_{UV} - \Lambda)$, and $S_{VV}^* = \sqrt{n}(S_{VV} - I)$. From $H_1'S_{VV}H_1 = I$ we find $H_{11}^* + H_{11}^{*\prime} = -S_{VV}^{*11} + o_p(1)$.

From (39) and (9) we obtain

$$(40)\qquad \hat{\Theta}_k^* = \begin{bmatrix} S_{UV}^{*11} + \Lambda_1(H_{11}^* + H_{11}^{*\prime}) & \Lambda_1 H_{21}^{*\prime} \\ S_{UV}^{*21} & \mathbf{0} \end{bmatrix} + o_p(1)$$

$$= \begin{bmatrix} S_{WV}^{*11} & S_{WV}^{*12} \\ S_{WV}^{*21} & \mathbf{0} \end{bmatrix} + o_p(1).$$

We can compare $\hat{\Theta}_k$ with the maximum likelihood estimator unrestricted by a rank condition $\hat{\Theta} = S_{UV}S_{VV}^{-1}$. Then

$$(41)\qquad \hat{\Theta}^* = \sqrt{n}\,(\hat{\Theta} - \Theta) = (S_{UV} - \Theta S_{VV})S_{VV}^{-1}$$

$$= S_{WV}^* + o_p(1) = \begin{bmatrix} S_{WV}^{*11} & S_{WV}^{*12} \\ S_{WV}^{*21} & S_{WV}^{*22} \end{bmatrix} + o_p(1),$$

since $S_{VV} \xrightarrow{p} I$. The effect of the rank restriction is to replace the lower right-hand submatrix of S_{WV}^* by $\mathbf{0}$ (the parameter value).

Since $S_{WV}^* = (1/\sqrt{n})\sum_{\alpha=1}^n W_\alpha V_\alpha'$, we have $\operatorname{vec} S_{WV}^* = (1/\sqrt{n})\sum_{\alpha=1}^n (V_\alpha \otimes W_\alpha)$. Because V_α and W_α are independent,

$$(42)\qquad \mathscr{E} \operatorname{vec} S_{WV}^* (\operatorname{vec} S_{WV}^*)'$$

$$= \mathscr{E}VV' \otimes \mathscr{E}WW' = I \otimes (I - \Lambda^2) = \operatorname{diag}(I - \Lambda^2, \ldots, I - \Lambda^2).$$

where $\Lambda = \operatorname{diag}(\Lambda_1, \mathbf{0})$ and $I - \Lambda^2 = \operatorname{diag}(I - \Lambda_1^2, I)$. On the other hand

$$(43)\qquad \operatorname{vec} \hat{\Theta}_k^* = \operatorname{vec} \frac{1}{\sqrt{n}} \sum_{\alpha=1}^n \left[W_\alpha V_\alpha^{(1)\prime}, \begin{pmatrix} W_\alpha^{(1)} \\ \mathbf{0} \end{pmatrix} V_\alpha^{(2)\prime} \right] + o_p(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{\alpha=1}^n \begin{bmatrix} V_\alpha^{(1)} & \otimes & W_\alpha \\ V_\alpha^{(2)} & \otimes & \begin{pmatrix} W_\alpha^{(1)} \\ \mathbf{0} \end{pmatrix} \end{bmatrix} + o_p(1),$$

where $V_\alpha = (V_\alpha^{(1)\prime}, V_\alpha^{(2)\prime})'$ and $W_\alpha = (W_\alpha^{(1)\prime}, W_\alpha^{(2)\prime})'$. Then

(44)

$$\mathscr{E} \operatorname{vec} \hat{\Theta}_k^* (\operatorname{vec} \hat{\Theta}_k^*)' \to \begin{bmatrix} I_k \otimes (I_{p_1} - \Lambda^2) & \mathbf{0} \\ \mathbf{0} & I_{p_2 - k} \otimes \begin{bmatrix} I_k - \Lambda_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \end{bmatrix}$$

$$= \operatorname{diag}\left(I_{p_1} - \Lambda^2, \ldots, I_{p_1} - \Lambda^2, I_k - \Lambda_1^2, \mathbf{0}, \ldots, I_k - \Lambda_1^2, \mathbf{0}\right)$$

where there are k blocks of $I_{p_1} - \Lambda^2$ and $p_1 - k$ blocks of $\operatorname{diag}(I_k - \Lambda_1^2, \mathbf{0})$

In the original coordinate system

$$(45)\quad \begin{aligned} \operatorname{vec}(\hat{\boldsymbol{B}}_k - \boldsymbol{\beta}) &= \operatorname{vec}\left[(\boldsymbol{\Lambda}')^{-1}(\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta})\boldsymbol{\Gamma}'\right] \\ &= \left[\boldsymbol{\Gamma} \otimes (\boldsymbol{\Lambda}')^{-1}\right]\operatorname{vec}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \\ &= \left[(\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2) \otimes \boldsymbol{\Sigma}_{ZZ}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2)(\boldsymbol{I} - \boldsymbol{\Lambda}^2)^{-1}\right]\operatorname{vec}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}). \end{aligned}$$

From (44) and (45) we obtain

$$(46)\quad \begin{aligned} &\mathscr{E} \operatorname{vec} n(\hat{\boldsymbol{B}}_k - \boldsymbol{\beta})\left[\operatorname{vec}(\hat{\boldsymbol{B}}_k - \boldsymbol{\beta})\right]' \\ &\to \left[(\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1') \otimes \boldsymbol{\Sigma}_{ZZ}\boldsymbol{\Lambda}(\boldsymbol{I}_p - \boldsymbol{\Lambda}^2)^{-1}\right] \\ &\qquad + \left[\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2' \otimes \boldsymbol{\Sigma}_{ZZ}\boldsymbol{\Lambda}_1(\boldsymbol{I}_k - \boldsymbol{\Lambda}_1^2)^{-1}\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{ZZ}\right] \\ &= [\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1' \otimes \boldsymbol{\Sigma}_{ZZ}] + \left[\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2' \otimes \boldsymbol{\Sigma}_{ZZ}\boldsymbol{\Lambda}_1(\boldsymbol{I}_k - \boldsymbol{\Lambda}_1^2)^{-1}\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{ZZ}\right] \\ &= \boldsymbol{\Sigma}_{XX}^{-1} \otimes \boldsymbol{\Sigma}_{ZZ} - (\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2' \otimes \boldsymbol{\Sigma}_{ZZ}\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2'\boldsymbol{\Sigma}_{ZZ}). \end{aligned}$$

If we define $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Gamma}_1 = \boldsymbol{\Sigma}_{ZZ}\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1(\boldsymbol{I} - \boldsymbol{\Lambda}_1^2)^{-1}$ and $\boldsymbol{\Pi} = \boldsymbol{\Gamma}_1$, then $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\Pi}'$. We have

$$(47)\qquad \boldsymbol{\Omega}(\boldsymbol{\Omega}'\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}' = \boldsymbol{\Sigma}_{ZZ} - \boldsymbol{\Sigma}_{ZZ}\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2'\boldsymbol{\Sigma}_{ZZ},$$

$$(48)\qquad \boldsymbol{\Pi}(\boldsymbol{\Pi}'\boldsymbol{\Sigma}_{XX}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}' = \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1' = \boldsymbol{\Sigma}_{XX}^{-1} - \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2'.$$

Thus (46) can be written

$$(49)\quad \begin{aligned} \mathscr{E} \operatorname{vec}\hat{\boldsymbol{B}}_k^* (\operatorname{vec}\hat{\boldsymbol{B}}_k^*)' \to \boldsymbol{\Sigma}_{XX}^{-1} \otimes \boldsymbol{\Sigma}_{ZZ} - \left[\boldsymbol{\Sigma}_{XX}^{-1} - \boldsymbol{\Pi}(\boldsymbol{\Pi}'\boldsymbol{\Sigma}_{XX}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}'\right] \\ \otimes \left[\boldsymbol{\Sigma}_{ZZ} - \boldsymbol{\Omega}(\boldsymbol{\Omega}'\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}'\right]. \end{aligned}$$

Theorem 13.7.1. *Let* $(\boldsymbol{X}^{(1)\prime}, \boldsymbol{X}^{(2)\prime})'$, $\alpha = 1, \dots, n$, *be observations on the random vector* $\boldsymbol{x}_\alpha$ *with mean* $\boldsymbol{0}$ *and covariance matrix* $\boldsymbol{\Sigma}$. *Let* $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. *Let the columns of* $\hat{\boldsymbol{\Gamma}}_1$ *satisfy* (1) *and* $\hat{\gamma}_{ii} > 0$. *Suppose that* $\boldsymbol{X}^{(1)} - \boldsymbol{\beta}\boldsymbol{X}^{(2)} = \boldsymbol{Z}$ *is independent of* $\boldsymbol{X}^{(2)}$. *Then the limiting distribution of* $\operatorname{vec}\hat{\boldsymbol{B}}_k^* = \sqrt{n}\operatorname{vec}(\hat{\boldsymbol{B}}_k - \boldsymbol{\beta})$, *with* $\hat{\boldsymbol{B}}_k = \boldsymbol{S}_{12}\hat{\boldsymbol{\Gamma}}_1\hat{\boldsymbol{\Gamma}}_1'$, *is normal with mean* $\boldsymbol{0}$ *and covariance matrix* (46) *or* (49).

Note that $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\Pi}' = \boldsymbol{\Omega}\boldsymbol{M}'(\boldsymbol{\Pi}\boldsymbol{M}^{-1})'$ for arbitrary nonsingular $\boldsymbol{M}$; however, (47) and (48) are invariant with respect to the transformation $\boldsymbol{\Omega} \to \boldsymbol{\Omega}\boldsymbol{M}$ and $\boldsymbol{\Pi} \to \boldsymbol{\Pi}\boldsymbol{M}^{-1}$. Thus (49) holds for any factorization $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\Pi}'$.

The limiting distribution of $\hat{B}_k^*$ only depends on $\sqrt{n}\, S_{Z2} S_{22}^{-1} = (A')^{-1} S_{WV}^* S_{VV}^{-1} \Gamma'$ and hence holds under the same conditions as the asymptotic normality of the least squares estimator $\hat{B}$.

Now suppose that $X_\alpha^{(2)} = x_\alpha^{(2)}$, $\alpha = 1, \dots, n$, is nonstochastic and that (19) holds. The model is (2); in the transformed coordinates $[U = \mathrm{A}'_n x^{(1)},\ v_\alpha = \Gamma'_n x_\alpha^{(2)},\ W = \mathrm{A}'_n Z,\ \Theta = \mathrm{A}'_n \boldsymbol{\beta} (\Gamma'_n)^{-1} = \Lambda]$ the model is (3). $H_1 = \Gamma^{-1}\Gamma_1$ satisfies (34) and (37). Again (39) holds. Further, (42) and (43) hold with $V_\alpha = v_\alpha$ nonstochastic.

Corollary 13.7.1. *Let $x_1^{(2)}, \dots, x_n^{(2)}$ be a set of vectors such that* (19) *holds. Let $x_\alpha^{(1)} = \boldsymbol{\beta} x_\alpha^{(2)} + z_\alpha$, $\alpha = 1, \dots, n$, where z_α is an observation on a random vector Z with $\mathscr{E} Z = \mathbf{0}$ and $\mathscr{E} ZZ' = \Sigma_{ZZ}$. Suppose $\boldsymbol{\beta}$ has rank k. Then the limiting distribution of $\sqrt{n}\,\mathrm{vec}(\hat{B}_k - \boldsymbol{\beta})$ is normal with mean $\mathbf{0}$ and covariance* (46) *or* (49).

13.8. ELLIPTICALLY CONTOURED DISTRIBUTIONS

13.8.1. Observations Elliptically Contoured

Let $x_1, \dots, x_N$ be N observations on a random vector X with density

$$(1) \qquad |\Psi|^{-\frac{1}{2}} g\left[(x - \nu)' \Psi^{-1} (x - \nu)\right],$$

where Ψ is a positive definite matrix, $R^2 = (x - \nu)' \Psi^{-1} (x - \nu)$, and $\mathscr{E} R^2 < \infty$. Define $\kappa = p\, \mathscr{E} R^4 / [(\mathscr{E} R^2)^2 (p+2)] - 1$. Then $\mathscr{E} X = \nu = \mu$ and $\mathscr{E}(X - \nu)(X - \nu)' = (\mathscr{E} R^2 / p)\Psi = \Sigma$. Define $\bar{x}$ and S as the sample mean and covariance matrix. Define the orthogonal matrices $\boldsymbol{\beta}$ and B and the diagonal matrices Λ and L by

$$(2) \qquad \Sigma = \boldsymbol{\beta} \Lambda \boldsymbol{\beta}', \qquad S = BLB',$$

$\lambda_1 > \cdots > \lambda_p$, $l_1 > \cdots > l_p$, $\beta_{i1} \geq 0$, $b_{i1} \geq 0$, $i = 1, \dots, p$. As in Section 13.5.1, define $T = \boldsymbol{\beta}' S \boldsymbol{\beta} = YLY'$, where $Y = \boldsymbol{\beta}' B$ is orthogonal and $y_{i1} \geq 0$. Then $\mathscr{E} T = \boldsymbol{\beta}' \Sigma \boldsymbol{\beta} = \Lambda$.

The limiting covariances of $\sqrt{N}\,\mathrm{vec}(S - \Sigma)$ and $\sqrt{N}\,\mathrm{vec}(T - \Lambda)$ are

$$(3) \qquad \lim_{N \to \infty} N \mathscr{E}\, \mathrm{vec}(S - \Sigma)[\mathrm{vec}(S - \Sigma)]' = (\kappa + 1)(I_{p^2} + K_{pp})(\Sigma \otimes \Sigma) + \kappa\, \mathrm{vec}\, \Sigma\, (\mathrm{vec}\, \Sigma)',$$

$$(4) \qquad \lim_{N \to \infty} N \mathscr{E}\, \mathrm{vec}(T - \Lambda)[\mathrm{vec}(T - \Lambda)]' = (\kappa + 1)(I_{p^2} + K_{pp}) + \kappa\, \mathrm{vec}\, I_p\, (\mathrm{vec}\, I_p)'.$$

In terms of components $\mathscr{E}t_{ij} = \lambda_i \delta_{ij}$ and

$$\lim_{N\to\infty} N\mathscr{E}(t_{ij} - \lambda_i\delta_{ij})(t_{kl} - \lambda_k\delta_{kl}) \tag{5}$$
$$= (\kappa+1)(\lambda_i\lambda_j\delta_{ik}\delta_{jl} + \lambda_i\lambda_k\delta_{il}\delta_{jk}) + \kappa\lambda_i\lambda_k\delta_{ij}\delta_{kl}.$$

Let $\sqrt{N}(T-\Lambda) = U$, $\sqrt{N}(L-\Lambda) = D$, and $\sqrt{N}(Y - I_p) = W$. The set $u_{11}, \dots, u_{pp}$ are asymptotically independent of the set $(u_{12}, \dots, u_{p-1,p})$; the covariances u_{ij}, $i \neq j$, are mutually independent with variances $(\kappa+1)\lambda_i\lambda_j$; the variance of $u_{ii} = d_i$ converges to $(3\kappa+2)\lambda_i^2$; the covariance of $u_{ii} = d_i$ and $u_{kk} = d_k$, $i \neq k$, converges to $\kappa\lambda_i\lambda_k$. The limiting distribution of w_{ij}, $i \neq j$, is the limiting distribution of $u_{ij}/(\lambda_j - \lambda_i)$. Thus the w_{ij}, $i<j$, are asymptotically mutually independent with $\mathscr{E}w_{ij}^2 = (\kappa+1)\lambda_i\lambda_j/(\lambda_j-\lambda_i)^2$. These variances and covariances for the normal case hold for $\kappa = 0$.

Theorem 13.8.1. *Define diagonal Λ and L and orthogonal $\boldsymbol{\beta}$ and B by* (2), $\lambda_1 > \cdots > \lambda_p$, $l_1 > \cdots > l_p$, $\beta_{1i} \geq 0$, $b_{1i} \geq 0$, $i = 1, \dots, p$. *Define $G = \sqrt{N}(B - \boldsymbol{\beta})$ and diagonal $D = \sqrt{N}(L - \Lambda)$. Then the limiting distribution of G and D is normal with G and D independent. The variance of d_i is $(2+3\kappa)\lambda_i^2$, and the covariance of d_i and d_k is $\kappa\lambda_i\lambda_k$. The covariance of g_i is*

$$\mathscr{A}\mathscr{C}(g_i) = (1+\kappa)\sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i\lambda_k}{(\lambda_i - \lambda_k)^2}\boldsymbol{\beta}_k\boldsymbol{\beta}_k'. \tag{6}$$

The covariance matrix of g_i and g_j is

$$\mathscr{A}\mathscr{C}(g_i, g_j) = -(1+\kappa)\frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}\boldsymbol{\beta}_j\boldsymbol{\beta}_i', \qquad i \neq j. \tag{7}$$

Proof. The proof is the same as for Theorem 13.5.1 except that (4) is used instead of (4) with $\kappa = 0$. ■

In Section 11.7.3 we used the asymptotic distribution of the smallest q sample roots when the smallest q population roots are equal. Let $\Lambda = \operatorname{diag}(\Lambda_1, \lambda^* I_q)$, where the diagonal elements of (diagonal) Λ_1 are different and are larger than λ^*. As before, let $U = \sqrt{N}(T - \Lambda)$, and let U_{22} be the lower right-hand $q \times q$ submatrix of U. Let D_2 and Y_{22} be the lower right-hand $q \times q$ submatrices of D and Y. It was shown in Section 13.5.2 that $U_{22} = Y_{22}D_2Y_{22}' + o_p(1)$.

The criterion for testing the null hypothesis $\lambda_{p-q+1} = \cdots = \lambda_p$ is

$$(8) \qquad \frac{\prod_{i=p-q+1}^{p} l_i}{\left(\sum_{i=p-q+1}^{p} l_i/q\right)^q}.$$

In Section 11.7.3 it was shown that $-N$ times the logarithm of (8) has the limiting distribution of

$$(9) \quad \frac{1}{2\lambda^{*2}}\left[\operatorname{tr} U_{22}U_{22}' - \frac{1}{q}(\operatorname{tr} U_{22})^2\right]$$

$$= \frac{1}{2\lambda^{*2}}\left[2\sum_{\substack{i=p-q+1\\ i<j}}^{p} u_{ij}^2 + \sum_{i=p-q+1}^{p} u_{ii}^2 - \frac{1}{q}\left(\sum_{i=p-q+1}^{p} u_{ii}\right)^2\right].$$

The term $\sum_{i<j} u_{ij}^2$ has the limiting distribution of $(1+\kappa)\lambda^{*2}\chi^2_{q(q-1)/2}$. The limiting distribution of $(u_{p-q-1,\,p-q+1},\ldots,u_{pp})$ is normal with mean $\mathbf{0}$ and covariance matrix $\lambda^{*2}[2(1+\kappa)I_q + \kappa\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\lambda^{*2}$. The limiting distribution of $[\sum u_{ii}^2 - (\sum u_{ii})^2/q]\lambda^{*2}$ is $2(1+\kappa)\lambda^{*2}\chi^2_{q-1}$. Hence, the limiting distribution of (9) is the distribution of $(1+\kappa)\chi^2_{q(q+1)/2-1}$.

We are also interested in the characteristic roots and vectors of one covariance matrix in the metric of another covariance matrix.

Theorem 13.8.2. *Let S^* be the sample covariance matrix of a sample of size M from* (1), *and let T^* be the sample covariance matrix of a sample of size N from* (1) *with $\boldsymbol{\Psi}$ replaced by $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Lambda}$ be the diagonal matrix with $\lambda_1 > \cdots > \lambda_p$ (>0) as the diagonal elements, where $\lambda_1,\ldots,\lambda_p$ are the roots of $|\boldsymbol{\Psi} - \lambda\boldsymbol{\Sigma}| = 0$. Let $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_p)$ be the matrix with $\boldsymbol{\gamma}_i$ the solution of $(\boldsymbol{\Psi} - \lambda_i\boldsymbol{\Sigma})\boldsymbol{\gamma} = \mathbf{0}$, $\boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma} = 1$, and $\gamma_{1i} \geq 0$. Let $X^* = (x_1^*,\ldots,x_p^*)$ and diagonal L^* consist of the solutions to*

$$(10) \qquad (S^* - lT^*)x^* = \mathbf{0},$$

$x^{*\prime}T^*x^* = 1$, and $x_1^* \geq 0$. As $M \to \infty$, $N \to \infty$, $M/N \to \eta$, the limiting distribution of $Z^* = \sqrt{N}(X^* - \boldsymbol{\Gamma})$ and diagonal $D^* = \sqrt{N}(L - \boldsymbol{\Lambda})$ is normal with the following covariances:

$$(11) \qquad \mathscr{AV}(d_i) = (2+3\kappa)\lambda_i^2\frac{1+\eta}{\eta},$$

$$(12) \qquad \mathscr{AC}(d_i,d_j) = \kappa\lambda_i\lambda_j\frac{1+\eta}{\eta},$$

$$(13)\qquad \mathscr{AC}(z_i) = (1+\kappa)\sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i(\lambda_k + \eta\lambda_i)}{(\lambda_k - \lambda_i)^2}\gamma_k\gamma_k' + \frac{2+3\kappa}{4}\gamma_i\gamma_i',$$

$$(14)\qquad \mathscr{AC}(d_i, z_i) = \frac{2+3\kappa}{2}\lambda_i\gamma_i,$$

$$(15)\qquad \mathscr{AC}(z_i^*, z_j^*) = -(1+\kappa)\frac{\lambda_i\lambda_j(1+\eta)}{\eta(\lambda_j-\lambda_i)^2}\gamma_j\gamma_i' + \frac{\kappa}{4}\gamma_i\gamma_j, \qquad i\neq j,$$

$$(16)\qquad \mathscr{AC}(d_i, z_j) = \frac{\kappa}{2}\lambda_i\gamma_j.$$

Proof. Transform S^* and T^* to $S=\Gamma'S^*\Gamma$ and $T=\Gamma'T^*\Gamma$, Φ and Σ to $\Lambda=\Gamma'\Phi\Gamma$ and $I=\Gamma'\Sigma\Gamma$, and X^* to $X=\Gamma^{-1}X^*=G^{-1}$. Let $D=\sqrt{N}(L-\Lambda)$, $H=\sqrt{N}(G-I)$, $U=\sqrt{N}(S-\Lambda)$, and $V=\sqrt{N}(T-I)$. The matrices U and V and D and H have limiting normal distributions; they are related by (20), (21), and (22) of Section 13.6. From there and the covariances of the limiting distributions we derive (11) to (16). ■

13.8.2. Elliptically Contoured Matrix Distributions

Let Y $(p\times N)$ have the density $g(\operatorname{tr} YY')$. Then $A=YY'$ has the density (Lemma 13.3.1)

$$(17)\qquad \frac{\pi^{\frac{1}{2}pN}|A|^{\frac{1}{2}(N-p-1)}}{\Gamma_p(\frac{1}{2}N)}g(\operatorname{tr} A).$$

Let $A=BLB'$, where L is diagonal with diagonal elements $l_1>\cdots>l_p$ and B is orthogonal with $b_{i1}\geq 0$. Since $g(\operatorname{tr} A)=g(\sum_{i=1}^p l_i)$, the density of $l_1,\dots,l_p$ is (Theorem 13.3.4)

$$(18)\qquad \frac{\pi^{\frac{1}{2}p^2}g(\sum_{i=1}^p l_i)\prod_{i<j}(l_i-l_j)}{\Gamma_p(\frac{1}{2}p)},$$

and the matrix B is independently distributed according to the conditional Haar invariant distribution.

Suppose Y^* $(p\times m)$ and Z^* $(p\times n)$ have the density

$$\Psi^{(m+n)/2}g\left[tr(Y^{*\prime}\Psi^{-1}Y^*+Z^{*\prime}\Psi^{-1}Z^*)\right]\qquad (m,n>p).$$

Let C be a matrix such that $C\Psi C'=I$. Then $Y=CY^*$ and $Z=CZ^*$ have the density $g[\operatorname{tr}(YY'+ZZ')]$. Let $A^*=Y^*Y^{*\prime}$, $B^*=Z^*Z^{*\prime}$, $A=YY'$, and $B=ZZ'$. The roots of $|A^*-lB^*|=0$ are the roots of $(A-lB|=0$. Let the roots of $|A - f(A+B)| = 0$ be $f_1 > \cdots > f_p$, and let $F = \operatorname{diag}(f_1, \ldots, f_p)$. Define E $(p \times p)$ by $A + B = E'E$, and $A = E'FE$, and $e_{i1} \geq 0$, $i = 1, \ldots, p$.

Theorem 13.8.3. *The matrices E and F are independent. The density of F is*

$$(19) \qquad \frac{\pi^{\frac{1}{2}p^2}\Gamma_p\left[\frac{1}{2}(m+n)\right]}{\Gamma_p(\frac{1}{2}n)\Gamma_p(\frac{1}{2}m)\Gamma_p(\frac{1}{2}p)} \prod_{i=1}^{p} f_i^{\frac{1}{2}(m-p-1)} \prod_{i=1}^{p} (1-f_i)^{\frac{1}{2}(n-p-1)} \prod_{i<j} (f_i - f_j);$$

the density of E is

$$(20) \qquad \frac{2^p\Gamma_p(\frac{1}{2}p)\pi^{\frac{1}{2}p(n+m-p)}}{2^{\frac{1}{2}p(m+n-2)}\pi^{\frac{1}{2}p^2}\Gamma_p\left[\frac{1}{2}(m+n)\right]} |E'E|^{\frac{1}{2}p(m+n-p)} g(\operatorname{tr} E'E).$$

In the development in Section 13.2 the observations Y, Z have the density

$$(21) \qquad (2\pi)^{-\frac{1}{2}p(n+m)} e^{-\frac{1}{2}\operatorname{tr}(Y'Y+Z'Z)} = (2\pi)^{-\frac{1}{2}p(n+m)} e^{-\frac{1}{2}\operatorname{tr}(A+B)},$$

and in Section 13.7 $g[\operatorname{tr}(Y'Y + Z'Z)] = g[\operatorname{tr}(A+B)]$. The distribution of the roots does not depend on the form of $g(\cdot)$; the distribution of E depends only on $E'E = A + B$. The algebra in Section 13.2 carries over to this more general case.

PROBLEMS

13.1. (Sec. 13.2) Prove Theorem 13.2.1 for $p = 2$ by calculating the Jacobian directly.

13.2. (Sec. 13.2) Prove Theorem 13.3.2 for $p = 2$ directly by representing the orthogonal matrix C in terms of the cosine and sine of an angle.

13.3. (Sec. 13.2) Consider the distribution of the roots of $|A - lB| = 0$ when A and B are of order two and are distributed according to $W(\Sigma, m)$ and $W(\Sigma, n)$, respectively.

(a) Find the distribution of the larger root.
(b) Find the distribution of the smaller root.
(c) Find the distribution of the sum of the roots.

13.4. (Sec. 13.2) Prove that the Jacobian $|\partial(G, A)/\partial(E, F)|$ is $\prod(f_i - f_j)$ times a function of E by showing that the Jacobian vanishes for $f_i = f_j$ and that its degree in f_i is the same as that of $\prod(f_i - f_j)$.

13.5. (Sec. 13.3) Give the Haar invariant distribution explicitly for the 2×2 orthogonal matrix represented in terms of the cosine and sine of an angle.

13.6. (Sec. 13.3) Let A and B be distributed according to $W(\Sigma, m)$ and $W(\Sigma, n)$ respectively. Let $l_1 > \cdots > l_p$ be the roots of $|A - lB| = 0$ and $m_1 > \cdots > m_p$ be the roots of $|A - m\Sigma| = 0$. Find the distribution of the m's from that of the l's by letting $n \to \infty$.

13.7. (Sec. 13.3) Prove Lemma 13.3.1 in as much detail as Theorem 13.3.1.

13.8. Let A be distributed according to $W(\Sigma, n)$. In case of $p = 2$ find the distribution of the characteristic roots of A. [*Hint:* Transform so that Σ goes into a diagonal matrix.]

13.9. From the result in Problem 13.6 find the distribution of the sphericity criterion (when the null hypothesis is not true).

13.10. (Sec. 13.3) Show that X $(p \times n)$ has the density $f_X(X'X)$ if and only if T has the density

$$\frac{2^p \pi^{pn/2}}{\Gamma_p(n/2)} \prod_{i=1}^{p} t_{ii}^{n-i} f_X(TT'),$$

where T is the lower triangular matrix with positive diagonal elements such that $TT' = X'X$. [Srivastava and Khatri (1979)]. [*Hint:* Compare Lemma 13.3.1 with Corollary 7.2.1.]

13.11. (Sec. 13.5.2) In the case that the covariance matrix is (12) find the limiting distribution of D_1, W_{11}, W_{12}, and W_{21}.

13.12. (Sec. 13.3) Prove (6) of Section 12.4.

CHAPTER 14

Factor Analysis

14.1. INTRODUCTION

Factor analysis is based on a model in which the observed vector is partitioned into an unobserved *systematic* part and an unobserved *error* part. The components of the error vector are considered as uncorrelated or independent, while the systematic part is taken as a linear combination of a relatively small number of unobserved factor variables. The analysis separates the effects of the factors, which are of basic interest, from the errors. From another point of view the analysis gives a description or explanation of the interdependence of a set of variables in terms of the factors without regard to the observed variability. This approach is to be compared with principal component analysis, which describes or "explains" the *variability* observed. Factor analysis was developed originally for the analysis of scores on mental tests; however, the methods are useful in a much wider range of situations, for example, analyzing sets of tests of attitudes, sets of physical measurements, and sets of economic quantities. When a battery of tests is given to a group of individuals, it is observed that the score of an individual on a given test is more related to his scores on other tests than to the scores of other individuals on the other tests; that is, usually the scores for any particular individual are interrelated to some degree. This interrelation is "explained" by considering a test score of an individual as made up of a part which is peculiar to this particular test (called error) and a part which is a function of more fundamental quantities called *scores of primary abilities* or *factor scores*. Since they enter several test scores, it is their effect that connects the various

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

test scores. Roughly, the idea is that a person who is more intelligent in some respects will do better on many tests than someone who is less intelligent.

The model for factor analysis is defined and discussed in Section 14.2. Maximum likelihood estimators of the parameters are derived in the case that the factor scores and errors are normally distributed, and a test that the model fits is developed. The large-sample distribution theory is given for the estimators and test criterion (Section 14.3). Maximum likelihood estimators for fixed factors do not exist, but alternative estimation procedures are suggested (Section 14.4). Some aspects of interpretation are treated in Section 14.5. The maximum likelihood estimators are derived when the factors are normal and identification is effected by specified zero loadings. Finally the estimation of factor scores is considered. Anderson (1984a) discusses the relationship of factor analysis to principal components and linear functional and structural relationships.

14.2. THE MODEL

14.2.1. Definition of the Model

Let the observable vector X be written as

$$X = \Lambda f + U + \mu, \tag{1}$$

where X, U, and μ are column vectors of p components, f is a column vector of m $(\leq p)$ components, and Λ is a $p \times m$ matrix. We assume that U is distributed independently of f and with mean $\mathscr{E}U = \mathbf{0}$ and covariance matrix $\mathscr{E}UU' = \Psi$, which is diagonal. The vector f will be treated alternatively as a random vector and as a vector of parameters that varies from observation to observation.

In terms of mental tests each component of X is a score on a test or battery of tests. The corresponding component of μ is the average score of this test in the population. The components of f are the scores of the mental factors; linear combinations of these enter into the test scores. The coefficients of these linear combinations are the elements of Λ, and these are called *factor loadings*. Sometimes the elements of f are called *common* factors because they are common to several different tests; in the first presentation of this kind of model [Spearman (1904)] f consisted of one component and was termed the *general* factor. A component of U is the part of the test score not "explained" by the common factors. This is considered as made up of the error of measurement in the test plus a *specific factor*, having to do only with this particular test. Since in our model (with one set of observations on each individual) we cannot distinguish between these two components of the coordinate of U, we shall simply term the element of U the error of measurement.

The specification of a given component of X is similar to that in regression theory (or analysis of variance) in that it is a linear combination of other variables. Here, however, f, which plays the role of the independent variable, is not observed.

We can distinguish between two kinds of models. In one we consider the vector f to be a random vector, and in the other we consider f to be a vector of nonrandom quantities that varies from one individual to another. In the second case, it is more accurate to write $X_\alpha = \Lambda f_\alpha + U + \mu$. The nonrandom factor score vector may seem a better description of the systematic part, but it poses problems of inference because the likelihood function may not have a maximum. In principle, the model with random factors is appropriate when different samples consist of different individuals; the nonrandom factor model is suitable when the specific individuals involved and not just the structure are of interest.

When f is taken as random, we assume $\mathscr{E}f = \mathbf{0}$. (Otherwise, $\mathscr{E}X = \Lambda\,\mathscr{E}f + \mu$, and μ can be redefined to absorb $\Lambda\,\mathscr{E}f$.) Let $\mathscr{E}ff' = \Phi$. Our analysis will be made in terms of first and second moments. Usually, we shall consider f and U to have normal distributions. If f is not random, then $f = f_\alpha$ for the αth individual. Then we shall assume usually $(1/N)\Sigma_{\alpha=1}^N f_\alpha = \mathbf{0}$ and $(1/N)\Sigma_{\alpha=1}^N f_\alpha f_\alpha' = \Phi$.

There is a fundamental indeterminacy in this model. Let $f = Cf^*$ $(f^* = C^{-1}f)$ and $\Lambda^* = \Lambda C$, where C is a nonsingular $m \times m$ matrix. Then (1) can be written as

$$X = \Lambda^* f^* + U + \mu. \tag{2}$$

When f is random, $\mathscr{E}f^*f^{*\prime} = C^{-1}\Phi(C^{-1})' = \Phi^*$; when f is nonrandom, $(1/N)\Sigma_{\alpha=1}^N f_\alpha^* f_\alpha^{*\prime} = \Phi^*$. The model with Λ and f is equivalent to the model with Λ^* and f^*; that is, by observing X we cannot distinguish between these two models.

Some of the indeterminacy in the model can be eliminated by requiring that $\mathscr{E}ff' = I$ if f is random, or $\Sigma_{\alpha=1}^N f_\alpha f_\alpha' = NI$ if f is not random. In this case the factors are said to be *orthogonal*; if Φ is not diagonal, the factors are said to be *oblique*. When we assume $\Phi = I$, then $\mathscr{E}f^*f^{*\prime} = C^{-1}(C^{-1})' = I$ $(I = CC')$. The indeterminacy is equivalent to multiplication by an orthogonal matrix; this is called the problem of *rotation*. Requiring that Φ be diagonal means that the components of f are independently distributed when f is assumed normal. This has an appeal to psychologists because one idea of common mental factors is (by definition) that they are independent or uncorrelated quantities.

A crucial assumption is that the components of U are uncorrelated. Our viewpoint is that the errors of observation and the specific factors are by definition uncorrelated. That is, the interrelationships of the test scores are caused by the common factors, and that is what we want to investigate. There is another point of view on factor analysis that is fundamentally quite different; that is, that the common factors are supposed to explain or account for as much of the variance of the test scores as possible. To follow this point of view, we should use a different model.

A geometric picture helps the intuition. Consider a p-dimensional space. The columns of Λ can be considered as m vectors in this space. They span some m-dimensional subspace; in fact, they can be considered as coordinate axes in the m-dimensional space, and f can be considered as coordinates of a point in that space referred to this particular axis system. This subspace is called the *factor space*. Multiplying Λ on the right by a matrix corresponds to taking a new set of coordinate axes in the factor space.

If the factors are random, the covariance matrix of the observed X is

$$\Sigma = \mathscr{E}(X-\mu)(X-\mu)' = \mathscr{E}(\Lambda f+U)(\Lambda f+U)' = \Lambda\Phi\Lambda' + \Psi. \tag{3}$$

If the factors are orthogonal ($\mathscr{E}ff' = I$), then (3) is

$$\Sigma = \Lambda\Lambda' + \Psi. \tag{4}$$

If f and U are normal, all the information about the structure comes from (3) [or (4)] and $\mathscr{E}X = \mu$.

14.2.2. Identification

Given a covariance matrix Σ and a number m of factors, we can ask whether there exist a triplet Λ, Φ positive definite, and Ψ positive definite and diagonal to satisfy (3); if so, is the triplet unique? Since any triplet can be transformed into an equivalent structure ΛC, $C^{-1}\Phi C'^{-1}$, and Ψ, we can put m^2 independent conditions on Λ and Φ to rule out this indeterminacy. The number of components in the observable Σ and the number of conditions (for uniqueness) is $\frac{1}{2}p(p+1)+m^2$; the numbers of parameters in Λ, Φ, and Ψ are pm, $\frac{1}{2}m(m+1)$, and p, respectively. If the excess of observed quantities and conditions over number of parameters, namely, $\frac{1}{2}[(p-m)^2 - p - m]$, is positive, we can expect a problem of existence but can anticipate uniqueness if a set of parameters does exist. If the excess is negative, we can expect existence but possibly not uniqueness; if the excess is 0, we can hope for both existence and uniqueness (or at least a finite number of solutions). The question of existence of a solution is whether there exists a diagonal matrix Ψ with nonnegative diagonal entries such that $\Sigma - \Psi$ is positive semidefinite of rank m. Anderson and Rubin (1956) include most of the known results on this problem.

If a solution exists and is unique, the model is said to be *identified*. As noted above, some m^2 conditions have to be put on Λ and Φ to eliminate a transformation $\Lambda^* = \Lambda C$ and $\Phi^* = C^{-1}\Phi C'^{-1}$. We have referred above to the condition $\Phi = I$, which forces a transformation C to be orthogonal. [There are $\frac{1}{2}m(m+1)$ component equations in $\Phi = I$.] For some purposes, it is convenient to add the restrictions that

$$\Gamma = \Lambda' \Psi^{-1} \Lambda \tag{5}$$

is diagonal. If the diagonal elements of Γ are ordered and different ($\gamma_{11} > \gamma_{22} > \cdots > \gamma_{mm}$), Λ is uniquely determined. Alternative conditions are that the first m rows of Λ form a lower triangular matrix. A generalization of this condition is to require that the first m rows of $B\Lambda$ form a lower triangular matrix, where B is given in advance. (This condition is implied by the so-called centroid method.)

Simple Structure

These are conditions proposed by Thurstone (1947, p. 335) for choosing a matrix out of the class ΛC that will have particular psychological meaning. If $\lambda_{i\alpha} = 0$, then the αth factor does not enter into the ith test. The general idea of *simple structure* is that many tests should not depend on all the factors when the factors have real psychological meaning. This suggests that, given a Λ, one should consider all rotations, that is, all matrices ΛC where C is orthogonal, and choose the one giving most 0 coefficients. This matrix can be considered as giving the simplest structure and presumably the one with most meaningful psychological interpretation. It should be remembered that the psychologist can construct his or her tests so that they depend on the assumed factors in different ways.

The positions of the 0's are not chosen in advance, but rotations C are tried until a Λ is found satisfying these conditions. It is not clear that these conditions effect identification. Reiersøl (1950) modified Thurstone's conditions so that there is only one rotation that satisfies the conditions, thus effecting identification.

Zero Elements in Specified Positions

Here we consider a set of conditions that requires of the investigator more a priori information. He or she must know that some particular tests do not depend on some specific factors. In this case, the conditions are that $\lambda_{i\alpha} = 0$ for specified pairs (i, α); that is, that the αth factor does not affect the ith test score. Then we do not assume that $\mathscr{E}ff' = I$. These conditions are similar to some used in econometric models. The coefficients of the αth column are identified except for multiplication by a scale factor if (a) there are at least $m-1$ zero elements in that column and if (b) the rank of $\Lambda^{(\alpha)}$ is $m-1$, where $\Lambda^{(\alpha)}$ is the matrix composed of the rows containing the assigned 0's in the αth column with those assigned 0's deleted (i.e., the αth column deleted). (See Problem 14.1.) The multiplication of a column by a scale constant can be eliminated by a normalization, such as $\phi_{\alpha\alpha} = 1$ or $\lambda_{i\alpha} = 1$ for *some* i for *each* α. If $\phi_{\alpha\alpha} = 1$, $\alpha = 1, \dots, m$, then Φ is a correlation matrix.

It will be seen that there are m normalizations and a minimum of $m(m-1)$ zero conditions. This is equal to the number of elements of C. If there are more than $m-1$ zero elements specified in one or more columns of Λ, then there may be more conditions than are required to take out the indeterminacy in ΛC; in this case the conditions may restrict $\Lambda \Phi \Lambda'$.

As an example, consider the model

$$X = \mu + \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & \lambda_{32} \\ 0 & \lambda_{42} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v \\ a \end{bmatrix} + U \tag{6}$$

$$= \mu + \begin{bmatrix} v \\ \lambda_{21} v \\ \lambda_{31} v + \lambda_{32} a \\ \lambda_{42} a \\ a \end{bmatrix} + U$$

for the scores on five tests, where v and a are measures of verbal and arithmetic ability. The first two tests are specified to depend only on verbal ability while the last two tests depend only on arithmetic ability. The normalizations put verbal ability into the scale of the first test and arithmetic ability into the scale of the fifth test.

Koopmans and Reiersøl (1950), Anderson and Rubin (1956), and Howe (1955) suggested the use of preassigned 0's for identification and developed maximum likelihood estimation under normality for this case. [See also Lawley (1958).] Jöreskog (1969) called factor analysis under these identification conditions *confirmatory* factor analysis; with arbitrary conditions or with rotation to simple structure, it has been called *exploratory* factor analysis.

Other Conditions

A convenient set of conditions is to require the upper square submatrix of Λ to be the identity. This assumes that the upper square matrix without this condition is nonsingular. In fact, if $\Lambda^* = (\Lambda_1^{*\prime}, \Lambda_2^{*\prime})'$ is an arbitrary $p \times m$ matrix with Λ_1^* square and nonsingular, then $\Lambda = \Lambda^* \Lambda_1^{*-1} = (I_m, \Lambda_2')'$ satisfies the condition. (This specification of the leading $m \times m$ submatrix of Λ as I_m is a convenient identification condition and does not imply any substantive meaning.)

14.2.3. Units of Measurement

We have considered factor analysis methods applied to covariance matrices. In many cases the unit of measurement of each component of X is arbitrary. For instance, in psychological tests the unit of scoring has no intrinsic meaning.

Changing the units of measurement means multiplying each component of X by a constant; these constants are not necessarily equal. When a given test score is multiplied by a constant, the factor loadings for the test are multiplied by the same constant and the error variance is multiplied by square of the constant. Suppose $DX = X^*$, where D is a diagonal matrix with positive diagonal elements. Then (1) becomes

$$X^* = \Lambda^* f + U^* + \mu^*, \tag{7}$$

where $\mu^* = \mathscr{E}X^* = D\mu$, $\Lambda^* = D\Lambda$, and $U^* = DU$ has covariance matrix $\Psi^* = D\Psi D$. Then

$$\mathscr{E}(X^* - \mu^*)(X^* - \mu^*)' = \Lambda^* \Phi \Lambda^{*\prime} + \Psi^* = \Sigma^*, \tag{8}$$

where $\Sigma^* = D\Sigma D$. Note that if the identification conditions are $\Phi = I$ and $\Lambda'\Psi^{-1}\Lambda$ diagonal, then Λ^* satisfies the latter condition. If Λ is identified by specified 0's and the normalization is by $\phi_{\alpha\alpha} = 1$, $\alpha = 1, \dots, m$ (i.e., Φ is a correlation matrix), then $\Lambda^* = D\Lambda$ is similarly identified. (If the normalization is $\lambda_{i\alpha} = 1$ for specified i for each α, each column of $D\Lambda$ has to be renormalized.)

A particular diagonal matrix D consists of the reciprocals of the observable standard deviations $d_{ii} = 1/\sqrt{\sigma_{ii}}$. Then $\Sigma^* = D\Sigma D$ is the correlation matrix.

We shall see later that the maximum likelihood estimators with identification conditions Γ diagonal or specified 0's transform in the above fashion; that is, the transformation $x_\alpha^* = Dx_\alpha$, $\alpha = 1, \dots, N$, induces $\hat{\Lambda}^* = D\hat{\Lambda}$ and $\hat{\Psi}^* = D\hat{\Psi}D$.

14.3. MAXIMUM LIKELIHOOD ESTIMATORS FOR RANDOM ORTHOGONAL FACTORS

14.3.1. Maximum Likelihood Estimators

In this section we find the maximum likelihood estimators of the parameters when the observations are normally distributed, that is, the factor scores and errors are normal [Lawley (1940)]. Then $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. We impose conditions on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ to make them just identified. These do not restrict $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$; it is a positive definite matrix of rank m. For convenience we suppose that $\boldsymbol{\Phi} = \boldsymbol{I}$ (i.e., the factors are orthogonal or uncorrelated) and that $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is diagonal. Then the likelihood depends on the mean $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. The maximum likelihood estimators of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ under some other conditions effecting just identification [e.g., $\boldsymbol{\Lambda} = (\boldsymbol{I}_m, \boldsymbol{\Lambda}_2')'$] are transformations of the maximum likelihood estimators of $\boldsymbol{\Lambda}$ under the preceding conditions. If $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ are a set of N observations on $\boldsymbol{X}$, the likelihood function for this sample is

$$(1) \qquad L = (2\pi)^{-\frac{1}{2}pN} |\boldsymbol{\Sigma}|^{-\frac{1}{2}N} \exp\left[-\tfrac{1}{2}\sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_\alpha - \boldsymbol{\mu})\right].$$

The maximum likelihood estimator of the mean $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = (1/N)\sum_{\alpha=1}^{N} \boldsymbol{x}_\alpha$.

Let

$$(2) \qquad \boldsymbol{A} = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'.$$

Next we shall maximize the logarithm of (1) with $\boldsymbol{\mu}$ replaced by $\hat{\boldsymbol{\mu}}$; this is†

$$(3) \qquad -\tfrac{1}{2}pN\log 2\pi - \tfrac{1}{2}N\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\operatorname{tr}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}.$$

(This is the logarithm of the *concentrated likelihood*.) From $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}$, we obtain for any parameter θ

$$(4) \qquad \frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\theta} = -\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta}\boldsymbol{\Sigma}^{-1}.$$

Then the partial derivative of (3) with regard to ψ_{ii}, a diagonal element of $\boldsymbol{\Psi}$, is $-N/2$ times

$$(5) \qquad \sigma^{ii} - \sum_{k,j=0}^{p} c_{kj}\sigma^{ji}\sigma^{ik},$$

†We could add the restriction that the off-diagonal elements of $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ are 0 with Lagrange multipliers, but then the Lagrange multipliers become 0 when the derivatives are set equal to 0. Such restrictions do not affect the maximum.

where $\Sigma^{-1} = (\sigma^{ij})$ and $(c_{ij}) = C = (1/N)A$. In matrix notation, (5) set equal to 0 yields

$$\operatorname{diag} \Sigma^{-1} = \operatorname{diag} \Sigma^{-1} C \Sigma^{-1}, \tag{6}$$

where $\operatorname{diag} H$ indicates the diagonal terms of the matrix H. Equivalently $\operatorname{diag}\Sigma^{-1}(\Sigma - C)\Sigma^{-1} = \operatorname{diag}\mathbf{0}$. The derivative of (3) with respect to $\lambda_{k\tau}$ is $-N$ times

$$\sum_{j=1}^{p} \sigma^{kj}\lambda_{j\tau} - \sum_{h,g,j=1}^{p} \sigma^{kh}c_{hg}\sigma^{gj}\lambda_{j\tau}, \qquad k = 1,\dots,p, \quad \tau = 1,\dots,m. \tag{7}$$

In matrix notation (7) set equal to 0 yields

$$\Sigma^{-1}\Lambda = \Sigma^{-1}C\Sigma^{-1}\Lambda. \tag{8}$$

We have

$$\Sigma\Psi^{-1}\Lambda = (\Lambda\Lambda' + \Psi)\Psi^{-1}\Lambda = \Lambda\Gamma + \Lambda = \Lambda(\Gamma + I). \tag{9}$$

From this we obtain $\Psi^{-1}\Lambda(\Gamma + I)^{-1} = \Sigma^{-1}\Lambda$. Multiply (8) by Σ and use the above to obtain

$$\Lambda(\Gamma + I) = C\Psi^{-1}\Lambda, \tag{10}$$

or

$$(C - \Psi)\Psi^{-1}\Lambda = \Lambda\Gamma. \tag{11}$$

Next we want to show that $\Sigma^{-1} - \Sigma^{-1}C\Sigma^{-1} = \Sigma^{-1}(\Sigma - C)\Sigma^{-1}$ is $\Psi^{-1}(\Sigma - C)\Psi^{-1}$ when (8) holds. Multiply the latter by Σ on the left and on the right to obtain

(12)

$$\begin{aligned}\Sigma\Psi^{-1}(\Sigma - C)\Psi^{-1}\Sigma &= (\Lambda\Lambda' + \Psi)\Psi^{-1}(\Psi + \Lambda\Lambda' - C)\Psi^{-1}(\Lambda\Lambda' + \Psi)\\ &= \Psi + \Lambda\Lambda' - C\end{aligned}$$

because

$$\begin{aligned}\Lambda\Lambda'\Psi^{-1}(\Psi + \Lambda\Lambda' - C) &= \Lambda\Lambda' + \Lambda\Gamma\Lambda' - \Lambda\Lambda'\Psi^{-1}C\\ &= \Lambda\left[(I + \Gamma)\Lambda' - \Lambda'\Psi^{-1}C\right]\\ &= \mathbf{0}\end{aligned} \tag{13}$$

by virtue of (10). Thus

$$\Sigma^{-1}(\Sigma - C)\Sigma^{-1} = \Psi^{-1}(\Sigma - C)\Psi^{-1}. \tag{14}$$

Then (6) is equivalent to $\operatorname{diag} \boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{C})\boldsymbol{\Psi}^{-1} = \operatorname{diag} \mathbf{0}$. Since $\boldsymbol{\Psi}$ is diagonal, this equation is equivalent to

$$\operatorname{diag}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) = \operatorname{diag} \boldsymbol{C}. \tag{15}$$

The estimators $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ are determined by (10), (15), and the requirement that $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is diagonal.

We can multiply (11) on the left by $\boldsymbol{\Psi}^{-\frac{1}{2}}$ to obtain

$$\boldsymbol{\Psi}^{-\frac{1}{2}}(\boldsymbol{C} - \boldsymbol{\Psi})\boldsymbol{\Psi}^{-\frac{1}{2}}(\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda}) = (\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda})\boldsymbol{\Gamma}, \tag{16}$$

which shows that the columns of $\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda}$ are characteristic vectors of $\boldsymbol{\Psi}^{-\frac{1}{2}}(\boldsymbol{C} - \boldsymbol{\Psi})\boldsymbol{\Psi}^{-\frac{1}{2}} = \boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{\Psi}^{-\frac{1}{2}} - \boldsymbol{I}$ and the corresponding diagonal elements of $\boldsymbol{\Gamma}$ are the characteristic roots. [In fact, the characteristic vectors of $\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{\Psi}^{-\frac{1}{2}} - \boldsymbol{I}$ are the characteristic vectors of $\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{\Psi}^{-\frac{1}{2}}$ because $(\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{\Psi}^{-\frac{1}{2}} - \boldsymbol{I})\boldsymbol{x} = \gamma\boldsymbol{x}$ is equivalent to $\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{C}\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{x} = (1+\gamma)\boldsymbol{x}$.] The vectors are normalized by $(\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda})'(\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda}) = \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Gamma}$. The characteristic roots are chosen to maximize the likelihood. To evaluate the maximized likelihood function we calculate

$$\begin{aligned} \operatorname{tr} \boldsymbol{C}\hat{\boldsymbol{\Sigma}}^{-1} &= \operatorname{tr} \boldsymbol{C}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}')\hat{\boldsymbol{\Psi}}^{-1} \\ &= \operatorname{tr}\left[\boldsymbol{C}\hat{\boldsymbol{\Psi}}^{-1} - (\boldsymbol{C}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Lambda}})\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\right] \\ &= \operatorname{tr}\left[\boldsymbol{C}\hat{\boldsymbol{\Psi}}^{-1} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\right] \\ &= \operatorname{tr}\left[(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}})\hat{\boldsymbol{\Psi}}^{-1} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\right] \\ &= p. \end{aligned} \tag{17}$$

The third equality follows from (8) multiplied on the left by $\hat{\boldsymbol{\Sigma}}$; the fourth equality follows from (15) and the fact that $\hat{\boldsymbol{\Psi}}$ is diagonal. Next we find

$$\begin{aligned} |\hat{\boldsymbol{\Sigma}}| &= |\hat{\boldsymbol{\Psi}}^{\frac{1}{2}}| \cdot \left|\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}} + \boldsymbol{I}_p\right| \cdot |\hat{\boldsymbol{\Psi}}^{\frac{1}{2}}| \\ &= |\hat{\boldsymbol{\Psi}}| \cdot \left|\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}\boldsymbol{\Lambda} + \boldsymbol{I}_m\right| \\ &= |\hat{\boldsymbol{\Psi}}| \cdot |\hat{\boldsymbol{\Gamma}} + \boldsymbol{I}_m| \\ &= \prod_{i=1}^{p} \hat{\psi}_{ii} \prod_{j=1}^{m} (\hat{\gamma}_j + 1). \end{aligned} \tag{18}$$

The second equality is $|\boldsymbol{U}\boldsymbol{U}' + \boldsymbol{I}_p| = |\boldsymbol{U}'\boldsymbol{U} + \boldsymbol{I}_m|$ for $\boldsymbol{U}$ $p \times m$, which is proved as in (14) of Section 8.4. From the fact that the characteristic roots of

$\Psi^{-\frac{1}{2}}(C-\Psi)\Psi^{-\frac{1}{2}}$ are the roots $\gamma_1 > \gamma_2 > \cdots > \gamma_p$ of $0 = |C - \Psi - \gamma\Psi| = |C-(1+\gamma)\Psi|$,

$$\frac{|C|}{|\hat{\Psi}|} = \prod_{i=1}^{p} (1+\hat{\gamma}_i). \tag{19}$$

[Note that the roots $1+\gamma_i$ of $\Psi^{-\frac{1}{2}}C\Psi^{-\frac{1}{2}}$ are positive. The roots γ_i of $\Psi^{-\frac{1}{2}}(C-\Psi)\Psi^{-\frac{1}{2}}$ are not necessarily positive; usually some will be negative.]

Then

$$|\hat{\Sigma}| = \frac{|C|\prod_{j\in S}(1+\hat{\gamma}_j)}{\prod_{i=1}^{p}(1+\hat{\gamma}_i)} = \frac{|C|}{\prod_{j\notin S}(1+\hat{\gamma}_j)}, \tag{20}$$

where S is the set of indices corresponding to the roots in $\hat{\Gamma}$. The logarithm of the maximized likelihood function is

$$-\tfrac{1}{2}pN\log 2\pi - \tfrac{1}{2}N\log|C| - \tfrac{1}{2}N\sum_{j\notin S}\log(1+\hat{\gamma}_j) - \tfrac{1}{2}Np. \tag{21}$$

The largest roots $\hat{\gamma}_1 > \cdots > \hat{\gamma}_m$ should be selected for diagonal elements of $\hat{\Gamma}$. Then $S = \{1, \ldots, m\}$. The logarithm of the concentrated likelihood (3) is a function of $\Sigma = \Lambda\Lambda' + \Psi$. This matrix is positive definite for every Λ and every diagonal Ψ that is positive definite; it is also positive definite for some diagonal Ψ's that are not positive definite. Hence there is not necessarily a relative maximum for Ψ positive definite. The concentrated likelihood function may increase as one or more diagonal elements of Ψ approaches 0. In that case the derivative equations may not be satisfied for Ψ positive definite.

The equations for the estimators (11) and (15) can be written as polynomial equations [multiplying (11) by $|\Psi|$], but cannot be solved directly. There are various iterative procedures for finding a maximum of the likelihood function, including steepest descent, Newton–Raphson, scoring (using the information matrix), and Fletcher–Powell. [See Lawley and Maxwell (1971), Appendix II, for a discussion.]

Since there may not be a relative maximum in the region for which $\psi_{ii} > 0$, $i = 1, \ldots, p$, an iterative procedure may define a sequence of values of $\hat{\Lambda}$ and $\hat{\Psi}$ that includes $\hat{\psi}_{ii} < 0$ for some indices i. Such negative values are inadmissible because ψ_{ii} is interpreted as the variance of an error. One may impose the condition that $\psi_{ii} \geq 0$, $i = 1, \ldots, p$. Then the maximum may occur on the boundary (and not all of the derivative equations will be satisfied). For some indices i the estimated variance of the error is 0; that is, some test scores are exactly linear combinations of factor scores. If the identification conditions

$\Phi = I$ and $\Lambda'\Psi^{-1}\Lambda$ diagonal are dropped, we can find a coordinate system for the factors such that the test scores with 0 error variance can be interpreted as (transformed) factor scores. That interpretation does not seem useful. [See Lawley and Maxwell (1971) for further discussion.]

An alternative to requiring ψ_{ii} to be positive is to require ψ_{ii} to be bounded away from 0. A possibility is $\psi_{ii} \geq \varepsilon\sigma_{ii}$ for some small ε, such as 0.005. Of course, the value of ε is arbitrary; increasing ε will decrease the value of the maximum if the maximum is not in the interior of the restricted region, and the derivative equations will not all be satisfied.

The nature of the concentrated likelihood is such that more than one relative maximum may be possible. Which maximum an iterative procedure approaches will depend on the initial values. Rubin and Thayer (1982) have given an example of three sets of estimates from three different initial estimates using the EM algorithm.

The EM (expectation–maximization) algorithm is a possible computational device for maximum likelihood estimation [Dempster, Laird, and Rubin (1977), Rubin and Thayer (1982)]. The idea is to treat the unobservable f's as missing data. Under the assumption that f and U have a joint normal distribution, the sufficient statistics are the means and covariances of the X's and f's. The E-step of the algorithm is to obtain the expectation of the covariances on the basis of trial values of the parameters. The M-step is to maximize the likelihood function on the basis of these covariances; this step provides updated values of the parameters. The steps alternate, and the procedure usually converges to the maximum likelihood estimators. (See Problem 14.3.)

As noted in Section 14.2, the structure is equivariant and the factor scores are invariant under changes in the units of measurement of the observed variables $X \to DX$, where D is a diagonal matrix with positive diagonal elements and Λ is identified by $\Lambda'\Psi^{-1}\Lambda$ is diagonal. If we let $D\Lambda = \Lambda^*$, $D\Psi D = \Psi^*$, and $DCD = C^*$, then the logarithm of the likelihood function is a constant plus a constant times

$$(22)\qquad -\log|\Psi^* + \Lambda^*\Lambda^{*\prime}| - \operatorname{tr} C^*(\Psi^* + \Lambda^*\Lambda^{*\prime})^{-1}$$
$$= -\log|\Psi + \Lambda\Lambda'| - \operatorname{tr} C(\Psi + \Lambda\Lambda')^{-1} - 2\log|D|.$$

The maximum likelihood estimators of Λ^* and Ψ^* are $\hat{\Lambda}^* = D\hat{\Lambda}$ and $\hat{\Psi}^* = D\hat{\Psi}D$, and $\hat{\Lambda}^{*\prime}\hat{\Psi}^{*-1}\hat{\Lambda}^* = \hat{\Lambda}\hat{\Psi}^{-1}\hat{\Lambda}$ is diagonal. That is, the estimated factor loadings and error variances are merely changed by the units of measurement.

It is often convenient to use $d_{ii} = 1/\sqrt{c_{ii}}$, so $DCD = (r_{ij})$ is made up of the sample correlation coefficients. The analysis is independent of the units of measurement. This fact is related to the fact that psychological test scores do not have natural units.

The fact that the factors do not depend on the location and scale factors is one reason for considering factor analysis as an analysis of interdependence.

It is convenient to give some rules of thumb for initial estimates of the communalities, $\sum_{j=1}^{m}\hat{\lambda}_{ij}^2 = 1 - \hat{\psi}_{ii}$, in terms of observed correlations. One rule is to use the $R^2_{i\cdot 1,\ldots,i-1,i+1,\ldots,p}$. Another is to use $\max_{h\neq i}|r_{ih}|$.

14.3.2. Test of the Hypothesis That the Model Fits

We shall derive the likelihood ratio test that the model fits; that is, that for a specified m the covariance matrix can be written as $\mathbf{\Sigma} = \mathbf{\Psi} + \mathbf{\Lambda}\mathbf{\Lambda}'$ for some diagonal positive definite $\mathbf{\Psi}$ and some $p \times m$ matrix $\mathbf{\Lambda}$. The likelihood ratio criterion is

$$(23)\qquad \frac{\max_{\boldsymbol{\mu},\mathbf{\Lambda},\mathbf{\Psi}} L(\boldsymbol{\mu}, \mathbf{\Psi} + \mathbf{\Lambda}\mathbf{\Lambda}')}{\max_{\boldsymbol{\mu},\mathbf{\Sigma}} L(\boldsymbol{\mu}, \mathbf{\Sigma})} = \frac{|C|^{\frac{1}{2}N}}{|\hat{\mathbf{\Psi}} + \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'|^{\frac{1}{2}N}} = \prod_{j=m+1}^{p}(1+\hat{\gamma}_j)^{-\frac{1}{2}N}$$

because the unrestricted maximum likelihood estimator of $\mathbf{\Sigma}$ is C, $\operatorname{tr} C(\hat{\mathbf{\Psi}} + \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}')^{-1} = p$ by (17), and $|C|/|\hat{\mathbf{\Sigma}}| = \prod_{j=m+1}^{k}(1+\hat{\gamma}_j)^{\frac{1}{2}N}$ from (20). The null hypothesis is rejected if (23) is too small. We can use -2 times the logarithm of the likelihood ratio criterion:

$$(24)\qquad -N\sum_{j=m+1}^{p}\log(1+\hat{\gamma}_j)$$

and reject the null hypothesis if (24) is too large.

If the regularity conditions for $\hat{\mathbf{\Psi}}$ and $\hat{\mathbf{\Lambda}}$ to be asymptotically normally distributed hold, the limiting distribution of (24) under the null hypothesis is χ^2 with degrees of freedom $\frac{1}{2}[(p-m)^2 - p - m]$, which is the number of elements of $\mathbf{\Sigma}$ plus the number of identifying restrictions minus the number of parameters in $\mathbf{\Psi}$ and $\mathbf{\Lambda}$. Bartlett (1950) suggested replacing N by[†] $N - (2p+11)/6 - 2m/3$. See also Amemiya and Anderson (1990).

From (15) and the fact that $\hat{\gamma}_1, \ldots, \hat{\gamma}_p$ are the characteristic roots of $\hat{\mathbf{\Psi}}^{-\frac{1}{2}}(C - \hat{\mathbf{\Psi}})\hat{\mathbf{\Psi}}^{-\frac{1}{2}}$ we have

$$(25)\qquad \begin{aligned} 0 &= \operatorname{tr}\hat{\mathbf{\Psi}}^{-\frac{1}{2}}(C - \hat{\mathbf{\Psi}} - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}')\hat{\mathbf{\Psi}}^{-\frac{1}{2}} \\ &= \operatorname{tr}\hat{\mathbf{\Psi}}^{-\frac{1}{2}}(C - \hat{\mathbf{\Psi}})\hat{\mathbf{\Psi}}^{-\frac{1}{2}} - \operatorname{tr}\hat{\mathbf{\Psi}}^{-\frac{1}{2}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Psi}}^{-\frac{1}{2}} \\ &= \operatorname{tr}\hat{\mathbf{\Psi}}^{-\frac{1}{2}}(C - \hat{\mathbf{\Psi}})\hat{\mathbf{\Psi}}^{-\frac{1}{2}} - \operatorname{tr}\hat{\mathbf{\Gamma}} \\ &= \sum_{i=1}^{p}\hat{\gamma}_i - \sum_{i=1}^{m}\hat{\gamma}_i = \sum_{i=m+1}^{p}\hat{\gamma}_i. \end{aligned}$$

[†]This factor is heuristic. If $m = 0$, the factor from Chapter 9 is $N - (2p+11)/6$; Bartlett suggested replacing N and p by $N - m$ and $p - m$, respectively.

If $|\hat{\gamma}_j| < 1$ for $j = m+1, \ldots, p$, we can expand (24) using (25) as

$$(26) \quad -N \sum_{j=m+1}^{p} \left(\hat{\gamma}_j - \tfrac{1}{2}\hat{\gamma}_j^2 + \tfrac{1}{3}\hat{\gamma}_j^3 - \cdots \right) = \tfrac{1}{2}N \sum_{j=m+1}^{p} \left(\hat{\gamma}_j^2 - \tfrac{2}{3}\hat{\gamma}_j^3 + \cdots \right).$$

The criterion is approximately $\frac{1}{2}N\sum_{j=m+1}^{p}\hat{\gamma}_j^2$. The estimators $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Lambda}}$ are found so that $\boldsymbol{C} - \hat{\boldsymbol{\Psi}} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$ is small in a statistical sense or, equivalently, so $\boldsymbol{C} - \hat{\boldsymbol{\Psi}}$ is approximately of rank m. Then the smallest $p-m$ roots of $\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}(\boldsymbol{C} - \hat{\boldsymbol{\Psi}})\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}$ should be near 0. The criterion measures the deviations of these roots from 0. Since $\hat{\gamma}_{m+1}, \ldots, \hat{\gamma}_p$ are the nonzero roots of $\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}(\boldsymbol{C} - \hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}$, we see that

$$\begin{aligned}(27) \quad \frac{1}{2}\sum_{j=m+1}^{p}\hat{\gamma}_j^2 &= \tfrac{1}{2}\operatorname{tr}\left[\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}(\boldsymbol{C} - \hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{\Psi}}^{-\frac{1}{2}}\right]^2 \\ &= \tfrac{1}{2}\operatorname{tr}\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{C} - \hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{C} - \hat{\boldsymbol{\Sigma}}) \\ &= \sum_{i<j}\frac{(c_{ij} - \hat{\sigma}_{ij})^2}{\hat{\psi}_{ii}\hat{\psi}_{jj}}\end{aligned}$$

because the diagonal elements of $\boldsymbol{C} - \hat{\boldsymbol{\Sigma}}$ are 0.

In many situations the investigator does not know a value of m to hypothesize. He or she wants to determine the smallest number of factors such that the model is consistent with the data. It is customary to test successive values of m. The investigator starts with a test that the number of factors is a specified m_0 (possibly 0 or 1). If that hypothesis is rejected, one proceeds to test that the number is $m_0 + 1$. One continues in that fashion until a hypothesis is accepted or until $\frac{1}{2}[(p-m)^2 - p - m) \le 0$. In the last event one concludes that no nontrivial factor model fits. Unfortunately, the probabilities of errors under this procedure are unknown, even asymptotically.

14.3.3. Asymptotic Distributions of the Estimators

The maximum likelihood estimators $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ maximize the average concentrated log likelihood functions $L^*(\boldsymbol{C}, \boldsymbol{\Lambda}^*, \boldsymbol{\Psi}^*)$ given by (3) divided by N for $\boldsymbol{\Sigma}^* = \boldsymbol{\Psi}^* + \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime}$, subject to $\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Psi}^{*-1}\boldsymbol{\Lambda}^*$ being diagonal. If $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$ (the "true" covariance matrix), then $L^*(\boldsymbol{C}, \boldsymbol{\Lambda}^*, \boldsymbol{\Psi}^*) \to L^*(\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}', \boldsymbol{\Lambda}^*, \boldsymbol{\Psi}^*)$ uniformly in probability in a neighborhood of $\boldsymbol{\Lambda}, \boldsymbol{\Psi}$, and $L^*(\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}', \boldsymbol{\Lambda}^*, \boldsymbol{\Psi}^*)$ has a unique maximum at $\boldsymbol{\Psi}^* = \boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}$. Because the function is continuous, the $\boldsymbol{\Lambda}^*, \boldsymbol{\Psi}^*$ that maximize $L^*(\boldsymbol{C}, \boldsymbol{\Lambda}^*, \boldsymbol{\Psi}^*)$ must converge stochastically to $\boldsymbol{\Lambda}, \boldsymbol{\Psi}$.

Theorem 14.3.1. *If Λ and Ψ are identified by $\Lambda'\Psi^{-1}\Lambda$ being diagonal, if the diagonal elements are different and ordered, and if $C \xrightarrow{p} \Psi + \Lambda\Lambda'$, then $\hat{\Psi} \xrightarrow{p} \Psi$ and $\hat{\Lambda} \xrightarrow{p} \Lambda$.*

A sufficient condition for $C \xrightarrow{p} \Sigma$ is that $(f' \; U')'$ has a distribution with finite second-order moments.

The estimators $\hat{\Lambda}$ and $\hat{\Psi}$ are the solutions to the equations (10), (15), and the requirement that $\Lambda'\Psi^{-1}\Lambda$ is diagonal. These equations are polynomial equations. The derivatives of $\hat{\Lambda}$ and $\hat{\Psi}$ as functions of C are continuous unless they become infinite. Anderson and Rubin (1956) investigated conditions for the derivative to be finite and proved the following theorem:

Theorem 14.3.2. *Let*

$$(\theta_{ij}) = \Theta = \Psi - \Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'. \tag{28}$$

If (θ_{ij}^2) is nonsingular, if Λ and Ψ are identified by the condition that $\Lambda'\Psi^{-1}\Lambda$ is diagonal and the diagonal elements are different and ordered, if $C \xrightarrow{p} \Psi + \Lambda\Lambda'$, and if $\sqrt{N}(C - \Sigma)$ has a limiting normal distribution, then $\sqrt{N}(\hat{\Lambda} - \Lambda)$ and $\sqrt{N}(\hat{\Psi} - \Psi)$ have a limiting normal distribution.

For example, $\sqrt{N}(C - \Sigma)$ will have a limiting distribution if $(f' \; U')'$ has a distribution with finite fourth moments.

The covariance matrix of the limiting distribution of $\sqrt{N}(\hat{\Lambda} - \Lambda)$ and $\sqrt{N}(\hat{\Psi} - \Psi)$ is too complicated to derive or even present here. Lawley (1953) found covariances for $\sqrt{N}(\hat{\Lambda} - \Lambda)$ appropriate for Ψ known, and Lawley (1967) extended his work to the case of Ψ estimated. [See also Lawley and Maxwell (1971).] Jennrich and Thayer (1973) corrected an error in his work.

The covariance of $\sqrt{N}(\hat{\psi}_{ii} - \psi_{ii})$ and $\sqrt{N}(\hat{\psi}_{jj} - \psi_{jj})$ in the limiting distribution is

$$2\psi_{ii}^2\psi_{jj}^2\xi^{ij}, \qquad i, j = 1, \dots, p, \tag{29}$$

where $(\xi^{ij}) = (\theta_{ij}^2)^{-1}$. The other covariances are too involved to give here.

While the asymptotic covariances are too complicated to give insight into the sampling variability, they can be programmed for computation. In that case the parameters are replaced by their consistent estimators.

14.3.4. Minimum-Distance Methods

An alternative to maximum likelihood is generalized least squares. The estimators are the values of Ψ and Λ that minimize

$$\operatorname{tr}(C - \Sigma)H(C - \Sigma)H, \tag{30}$$

where $\Sigma = \Psi + \Lambda\Lambda'$ and $H = \Sigma^{-1}$ or some consistent estimator of Σ^{-1}. When $H = \Sigma^{-1}$, the objective function is of the form

$$[c - \sigma(\Psi, \Lambda)]'[\operatorname{cov} c]^{-1}[c - \sigma(\Psi, \Lambda)], \tag{31}$$

where c represents the elements of C arranged in a vector, $\sigma(\Psi, \Lambda)$ is $\Psi + \Lambda\Lambda'$ arranged in a corresponding vector, and $\operatorname{cov} c$ is the covariance matrix of c under normality [Anderson (1973a)]. Jöreskog and Goldberger (1972) use C^{-1} for H and minimize

$$\operatorname{tr}(C - \Sigma)C^{-1}(C - \Sigma)C^{-1} = \operatorname{tr}(I - \Sigma C^{-1})^2. \tag{32}$$

The matrix of derivatives with respect to the elements of Λ set equal to 0 forms the matrix equation

$$C^{-1}(C - \Sigma)C^{-1}\Lambda = \mathbf{0}. \tag{33}$$

This can be rewritten as

$$\Lambda = \Sigma C^{-1}\Lambda. \tag{34}$$

Multiplication on the left by $\Sigma^{-1}C\Sigma^{-1}$ yields (8), which leads to (10). This estimator of Λ given Ψ is the same as the maximum likelihood estimator except for normalization of columns. The equation obtained by setting the derivatives of (32) with respect to Ψ equal to 0 is

$$\operatorname{diag} C^{-1}[(\Psi + \Lambda\Lambda') - C]C^{-1} = \operatorname{diag}\mathbf{0}. \tag{35}$$

An alternative is to minimize

$$\tfrac{1}{2}\operatorname{tr}\{(\Psi + \Lambda\Lambda')^{-1}[C - (\Psi + \Lambda\Lambda')]\}^2. \tag{36}$$

This leads to (8) or (10) and

$$\operatorname{diag}\Sigma^{-1}C\Sigma^{-1}(C - \Sigma)\Sigma^{-1} = \operatorname{diag}\mathbf{0}. \tag{37}$$

Browne (1974) showed that the generalized least squares estimator of Ψ has the same asymptotic distribution as the maximum likelihood estimator. Dahm and Fuller (1981) showed that if $\operatorname{cov} c$ in (31) is replaced by a matrix converging to $\operatorname{cov} c$ and Ψ, Λ, and Φ depend on some parameters, then the asymptotic distributions are the same as for maximum likelihood.

14.3.5. Relation to Principal Component Analysis

What is the relation of maximum likelihood to the principal component analysis proposed by Hotelling (1933)? As explained in Chapter 11, the vector of sample principal components is the orthogonal transformation $B'X$, where the columns of B are the characteristic vectors of C normalized by $B'B = I$. Then

$$C = BTB' = \sum_{i=1}^{p} b_i t_i b_i', \tag{38}$$

where T is the diagonal matrix with diagonal elements $t_1, \dots, t_p$, the characteristic roots of C. If $t_{m+1}, \dots, t_p$ are small, C can be approximated by

$$B_1 T_1 B_i' = \sum_{i=1}^{m} b_i t_i b_i', \tag{39}$$

where T_1 is the diagonal matrix with diagonal elements $t_1, \dots, t_m$, and X is approximated by

$$B_1 B_1' X = \sum_{i=1}^{m} b_i (b_i' X). \tag{40}$$

Then the sample covariance of the difference between X and the approximation (40) is the sample covariance of

$$X - B_1 B_1' X = B_2 B_2' X, \tag{41}$$

which is $B_2 T_2 B_2' = \sum_{i=m+1}^{p} b_i t_i b_i'$, and the sum of the variances of the components is $\sum_{i=m+1}^{p} t_i$. Here T_2 is the diagonal matrix with $t_{m+1}, \dots, t_p$ as diagonal elements.

This analysis is in terms of some common unit of measurement. The first m components "explain" a large proportion of the "variance," $\operatorname{tr} C$. When the units of measurement are not the same (e.g., when the units are arbitrary), it is customary to standardize each measurement to (sample) variance 1. However, then the principal components do not have the interpretation in terms of variance.

Another difference between principal component analysis and factor analysis is that the former does not separate the error from the systematic part. This fault is easily remedied, however. Thomson (1934) proposed the following estimation procedure for the factor analysis model. A diagonal matrix Ψ is subtracted from C, and the principal component analysis is carried out on $C - \Psi$. However, Ψ is determined so $C - \Psi$ is close to rank m. The equations are

$$(C - \Psi)\Lambda = \Lambda L, \tag{42}$$

$$\operatorname{diag}(\Psi + \Lambda\Lambda) = \operatorname{diag} C, \tag{43}$$

$$\Lambda'\Lambda = L \text{ diagonal}. \tag{44}$$

The last equation is a normalization and takes out the indeterminacy in Λ. This method allows for the error terms, but still depends on the units of measurement. The estimators are consistent but not (asymptotically) efficient in the usual factor analysis model.

14.3.6. The Centroid Method

Before the availability of high-speed computers, the centroid method was used almost exclusively because of its computational ease. For the sake of history we give a sketch of the method. Let R^* be the correlation reduced matrix, that is, the matrix consisting of r_{ij}, $i \neq j$, and $1-\hat{\psi}_{ii}^*$, where $\hat{\psi}_{ii}^*$ is an initial estimate of the error variance in standard deviation units. Thomson's principal components approach is first to find the m characteristic vectors of $R_0 = R^*$ corresponding to the m largest characteristic roots. As indicated in Chapter 11, one computational method involves starting with an initial estimate of the first vector, say $x^{(0)}$, calculating $x^{(1)} = R_0 x^{(0)}$, and iterating. At the rth step $x^{(r)}$ is approximately $\gamma_1 x^{(r-1)}$, where γ_1 is the largest root and $x^{(r)\prime}x^{(r)} \sim \gamma_1^2 x^{(r-1)\prime}x^{(r-1)}$. Then $y_1 = x^{(r)}/\sqrt{\gamma_1 x^{(r-1)\prime}x^{(r-1)}}$ is approximately the first characteristic vector normalized so $y_1'y_1 = \gamma_1$. To obtain the second vector, apply the same procedure to $R_1 = R^* - y_1 y_1'$.

The centroid method can be considered as a very rough approximation to the principal component approach. With psychological tests the correlation matrix usually consists of positive entries, and the first characteristic vector has all positive components, often of about the same value. The centroid method uses $\varepsilon = (1,\dots,1)'$ as the initial estimate of the first vector. Then $R^*\varepsilon = x^{(1)}$ is the first iterate and should be an approximation to the first characteristic vector. An approximation to the first characteristic root is $\varepsilon' R^* \varepsilon/\varepsilon'\varepsilon$. Then $y_1 = x^{(1)}/\sqrt{\varepsilon' R^* \varepsilon}$ is an approximation to the first characteristic vector of R^* normalized to have length squared γ_1. The operations can be carried out on an adding machine or on a desk calculator because $R^*\varepsilon$ amounts to adding across rows and $\varepsilon' R^* \varepsilon$ is the sum of those row totals.

The second characteristic vector is orthogonal to the first. A vector orthogonal to ε is ε^* consisting of $p/2$ 1's and $p/2$ -1's. Then $R_1\varepsilon^* = x_2$ is an approximation to the second characteristic vector, and $\varepsilon^{*\prime} R_1 \varepsilon^*/\varepsilon^{*\prime}\varepsilon^*$ approximates the second characteristic root. These operations involve changing signs of entries of R_1 and adding. The positions of the -1's in ε^* are selected to maximize $\varepsilon^{*\prime} R_1 \varepsilon^*$. The procedure can be continued.

14.4. ESTIMATION FOR FIXED FACTORS

Let $x_\alpha = (x_{1\alpha},\dots,x_{p\alpha})'$ be an observation on X_α given by

$$X_\alpha = \Lambda f_\alpha + \mu + U_\alpha \tag{1}$$

with f_α being a nonstochastic vector (an incidental parameter), $\alpha = 1, \dots, N$, satisfying $\sum_{\alpha=1}^{N} f_\alpha = \mathbf{0}$. The likelihood function is

$$(2)\quad L = \frac{1}{\left[(2\pi)^p \prod_{i=1}^{p} \psi_{ii}\right]^{N/2}} \prod_{i=1}^{p} \exp\left\{ -\frac{1}{2} \sum_{\alpha=1}^{N} \frac{\left(x_{i\alpha} - \mu_i - \sum_{j=1}^{m} \lambda_{ij} f_{j\alpha}\right)^2}{\psi_{ii}} \right\}.$$

This likelihood function does not have a maximum. To show this fact, let $\mu_1 = 0$, $\lambda_{11} = 1$, $\lambda_{1j} = 0$ $(j \neq 1)$, $f_{1\alpha} = x_{1\alpha}$. Then $x_{1\alpha} - \mu_1 - \sum_{j=1}^{m} \lambda_{ij} f_{j\alpha} = 0$, and ψ_{11} does not appear in the exponent but appears only in the constant. As $\psi_{11} \to 0$, $L \to \infty$. Thus the likelihood does not have a maximum, and the maximum likelihood estimators do not exist [Anderson and Rubin (1956)]. Lawley (1941) set the partial derivatives of the likelihood equal to 0, but Solari (1969) showed that the solution is only a stationary value, not a maximum.

Since maximum likelihood estimators do not exist in the case of fixed factors, what estimation methods can be used? One possibility is to use the maximum likelihood method appropriate for random factors. It was stated by Anderson and Rubin (1956) and proved by Fuller, Pantula, and Amemiya (1982) in the case of identification by 0's that the asymptotic normal distribution of the maximum likelihood estimators for the random case is the same as for fixed factors.

The sample covariance matrix under normality has the noncentral Wishart distribution [Anderson (1946a)] depending on $\mathbf{\Psi}$, $\mathbf{\Lambda \Phi \Lambda}'$, and $N-1$. Anderson and Rubin (1956) proposed maximizing this likelihood function. However, one of the equations is difficult to solve. Again the estimators are asymptotically equivalent to the maximum likelihood estimators for the random-factor case.

14.5. FACTOR INTERPRETATION AND TRANSFORMATION

14.5.1. Interpretation

The identification restrictions of $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ diagonal or the first m rows of $\mathbf{\Lambda}$ being $\mathbf{I}_m$ may be convenient for computing the maximum likelihood estimators, but the components of the factor score vector may not have any intrinsic meaning. We saw in Section 14.2 that 0 coefficients may give meaning to a factor by the fact that this factor does not affect certain tests. Similarly, large factor loadings may help in interpreting a factor. The coefficient of verbal ability, for example, should be large on tests that look like they are verbal.

In psychology each variable or factor usually has a natural positive direction: more answers right on a test and more of the ability represented by the factor. It is usually expected that more ability leads to higher performance; that is, the factor loading should be positive if it is not 0. Therefore, roughly

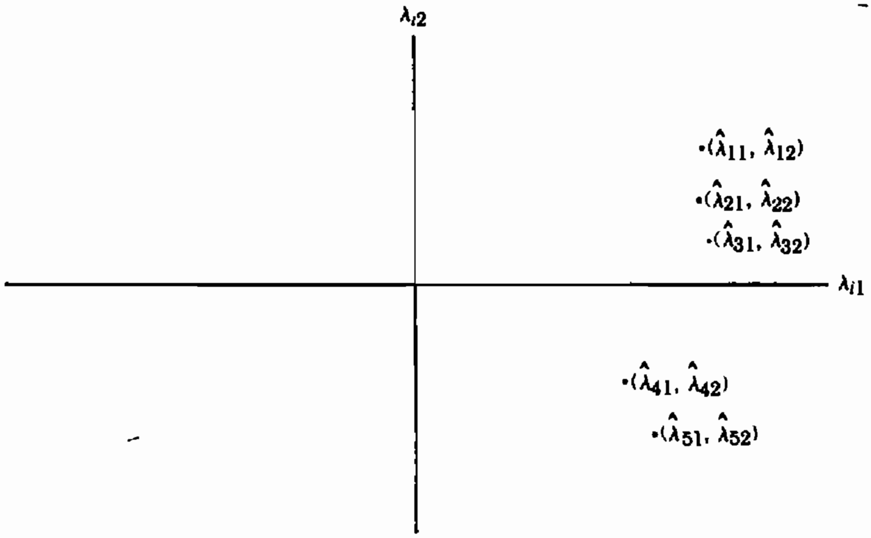

Figure 14.1. Rows of $\hat{\Lambda}$.

speaking, for the sake of interpretation, one may look for factor loadings that are either 0 or positive and large.

14.5.2. Transformations

The maximum likelihood estimators on the basis of some arbitrary identification conditions including $\Phi = I$ are $\hat{\Lambda}$ and $\hat{\Psi}$. We consider transformations

$$(1) \qquad \hat{\Lambda}^* = \hat{\Lambda} P, \qquad \hat{\Phi}^* = P^{-1}(P^{-1})' = (P'P)^{-1}.$$

If the factors are to be orthogonal, then $\Phi^* = I$ and P is orthogonal. If the factors are permitted to be oblique, P can be an arbitrary nonsingular matrix and $\hat{\Phi}^*$ an arbitrary positive definite matrix.

The rows of $\hat{\Lambda}$ can be plotted in an m-dimensional space. Figure 14.1 is a plot of the rows of a 5×2 matrix $\hat{\Lambda}$. The coordinates refer to factors and the points refer to tests. If Φ^* is required to be I_m, we are seeking a rotation of coordinate axes in this space. In the example that is graphed, a rotation of 45° would put all of the points into the positive quadrant, that is, $\lambda_{ij}^* \geq 0$. One of the new coordinates would be large for each of the first three points and small for the other two points, and the other coordinate would be small for the first three and large for the last two. The first factor is representative of what is common to the first three tests, and the second factor of what is common to the last two tests.

If $m > 2$, a general rotation can be approximated manually by a sequence of two-dimensional rotations.

If Φ^* is not required to be I_m, the transformation P is simply nonsingular. If the normalization of the jth column of Λ is $\lambda_{i(j),j} = 1$, then

$$(2) \qquad 1 = \hat{\lambda}^*_{i(j),j} = \sum_{k=1}^{m} \hat{\lambda}_{i(j),k} p_{kj};$$

each column of P satisfies such a constraint. If the normalization is $\phi_{jj} = 1$, then

$$(3) \qquad 1 = \phi^*_{jj} = \sum_{k=1}^{m} (p^{jk})^2,$$

where $(p^{jk}) = P^{-1}$.

Of the various computational procedures that are based on optimizing an objective function, we describe the varimax method proposed by Kaiser (1958) to be carried out on pairs of factors. Horst (1965), Chapter 18, extended the method to be done on all factors simultaneously. A modified criterion is

$$(4) \qquad \sum_{j=1}^{m}\sum_{i=1}^{p}\left(\lambda^{*2}_{ij} - \frac{\sum_{h=1}^{p}\lambda^{*2}_{hj}}{p}\right)^2 = \sum_{j=1}^{m}\left[\sum_{i=1}^{p}\lambda^{*4}_{ij} - \frac{\left(\sum_{h=1}^{p}\lambda^{*2}_{hj}\right)^2}{p}\right],$$

which is proportional to the sum of the column variances of the squares of the transformed factor loadings. The orthogonal matrix P is selected so as to maximize (4). The procedure tends to maximize the scatter of λ^{*2}_{ij} within columns. Since $\lambda^{*2}_{ij} \geq 0$, there is a tendency to obtain some large loadings and some near 0. Kaiser's original criterion was (4) with λ^{*2}_{ij} replaced by $\lambda^{*2}_{ij}/\sum_{h=1}^{m}\lambda^{*2}_{ih}$.

Lawley and Maxwell (1971) describe other criteria. One of them is a measure of similarity to a predetermined $p \times m$ matrix of 1's and 0's.

14.5.3. Orthogonal versus Oblique Factors

In the case of orthogonal factors the components are uncorrelated in the population or in the sample according to whether the factors are considered random or fixed. The idea of uncorrelated factor scores has appeal. Some psychologists claim that the orthogonality of the factor scores is essential if one is to consider the factor scores more basic than the test scores. Considerable debate has gone on among psychologists concerning this point. On the other side, Thurstone (1947), page vii, says "it seems just as unnecessary to require that mental traits shall be uncorrelated in the general population as to require that height and weight be uncorrelated in the general population."

As we have seen, given a pair of matrices Λ, Φ, equivalent pairs are given by $\Lambda P, P^{-1}\Phi P'^{-1}$ for nonsingular P's. The pair may be selected (i.e., the P

given $\Lambda\Phi$) as the one with the most meaningful interpretation in terms of the subject matter of the tests. The idea of simple structure is that with 0 factor loadings in certain patterns the component factor scores can be given meaning regardless of the moment matrix. Permitting Φ to be an arbitrary positive definite matrix allows more 0's in Λ.

Another consideration in selecting transformations or identification conditions is autonomy, or permanence, or invariance with regard to certain changes. For example, what happens if a selection of the constituents of a population is made? In case of intelligence tests, suppose a selection is made, such as college admittees out of high school seniors, that can be assumed to involve the primary abilities. One can envisage that the relation between unobserved factor scores f and observed test scores x is unaffected by the selection, that is, that the matrix of factor loadings Λ is unchanged. The variance of the errors (and specific factors), the diagonal elements of Ψ, may also be considered as unchanged by the selection because the errors are uncorrelated with the factors (primary abilities).

Suppose there is a *true* model, Λ, Φ, Ψ, and the investigator applies identification conditions that permit him to discover it. Next, suppose there is a selection that results in a new population of factor scores so that their covariance matrix is Φ^*. When the investigator analyzes the new observed covariance matrix $\Psi + \Lambda\Phi^*\Lambda'$, will he find Λ again? If part of the identification conditions are that the factor moment matrix is I, then he will obtain a different factor loading matrix. On the other hand, if the identification conditions are entirely on the factor loadings (specified 0's and 1's), the factor loading matrix from the analysis is the same as before.

The same consideration is relevant in comparing two populations. It may be reasonable to consider that $\Psi_1 = \Psi_2$, $\Lambda_1 = \Lambda_2$, but $\Phi_1 \neq \Phi_2$. To test the hypothesis that $\Phi_1 = \Phi_2$, one wants to use identification conditions that agree with $\Lambda_1 = \Lambda_2$ (rather than $\Lambda_1 = \Lambda_2 C$). The condition should be on the factor loadings.

What happens if more tests are added (or deleted)? In addition to observing $X = \Lambda f + \mu + U$, suppose one observes $X^* = \Lambda^* f + \mu^* + U^*$, where U^* is uncorrelated with U. Since the common factors f are unchanged, Φ is unchanged. However, the (arbitrary) condition that $\Lambda'\Psi^{-1}\Lambda$ is diagonal is changed; use of this type of condition would lead to a rotation of $(\Lambda' \ \Lambda^{*\prime})$.

14.6. ESTIMATION FOR IDENTIFICATION BY SPECIFIED ZEROS

We now consider estimation of Λ, Ψ, and Φ when Φ is unrestricted and Λ is identified by specified 0's and 1's. We assume that each column of Λ has at least $m + 1$ 0's in specified positions and that the submatrix consisting of the rows of Λ containing the 0's specified for a given column is of rank $m - 1$. (See Section 14.2.2.) We further assume that each column of Λ has 1 in a specified position or, alternatively, that the diagonal element of Φ corresponding to that column is 1. Then the model is identified.

The likelihood function is given by (1) of Section 14.3. The derivatives of the likelihood function set equal to 0 are

$$\operatorname{diag} \Sigma^{-1}[C - (\Psi + \Lambda\Phi\Lambda')]\Sigma^{-1} = \operatorname{diag} \mathbf{0}, \tag{1}$$

$$\Lambda'\Sigma^{-1}[C - (\Psi + \Lambda\Phi\Lambda')]\Sigma^{-1}\Lambda = \mathbf{0} \tag{2}$$

for positions in Φ that are not specified, and

$$\Sigma^{-1}[C - (\Psi + \Lambda\Phi\Lambda')]\Sigma^{-1}\Lambda = \mathbf{0} \tag{3}$$

for positions in Λ not specified, where

$$\Sigma = \Psi + \Lambda\Phi\Lambda. \tag{4}$$

These equations cannot be simplified as in Section 14.3.1 because (3) holds only for unspecified positions in Λ, and hence one cannot multiply by Σ on the left. [See Howe (1955), Anderson and Rubin (1956), and Lawley (1958).]

These equations are not useful for computation. The likelihood function, however, can be maximized numerically.

As noted before, a change in units of measurement, $X^* = DX$, results in a corresponding change in the parameters Λ and Ψ if identification is by 0 in specified positions of Λ and normalization is by $\phi_{jj} = 1$, $j = 1, \dots, m$. It is readily verified that the derivative equations (1), (2), (3), and (4) are changed in a corresponding manner.

Anderson and Amemiya (1988a) have derived the asymptotic distribution of the estimators under general conditions. Normality of the observations is not required. See also Anderson and Amemiya (1988b).

14.7. ESTIMATION OF FACTOR SCORES

It is frequently of interest to estimate the factor scores of the individuals in the group being studied. In the model with nonstochastic factors the factor scores are incidental parameters that characterize the individuals. As we have seen (Section 14.4), the maximum likelihood estimators of the parameters $(\Psi, \Lambda, \mu, f_1, \dots, f_N)$ do not exist. We shall therefore study the estimation of the factor scores on the basis that the structural parameters (Ψ, Λ, μ) are known.

When f_α is considered as an incidental parameter, $x_\alpha - \mu$ is an observation from a distribution with mean Λf_α and covariance matrix Ψ. The weighted least squares estimator of f_α is

$$\hat{f}_\alpha = (\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}(x_\alpha - \mu) \tag{1}$$
$$= \Gamma^{-1}\Lambda'\Psi^{-1}(x_\alpha - \mu),$$

where $\Gamma = \Lambda'\Psi^{-1}\Lambda$ (not necessarily diagonal). This estimator is unbiased and its covariance matrix is

$$\mathscr{E}(\hat{f}_\alpha - f_\alpha)(\hat{f}_\alpha - f_\alpha)' = (\Lambda'\Psi^{-1}\Lambda)^{-1} = \Gamma^{-1} \tag{2}$$

by the usual generalized least squares theory [Bartlett (1937b),(1938)]. It is the minimum variance unbiased linear estimator of f_α. If x_α is normal, the estimator is also maximum likelihood.

When f_α is considered random [Thomson (1951)], we suppose X_α and f_α have a joint normal distribution with mean vector $(\mu', 0')'$ and covariance matrix

$$\mathscr{C}\begin{pmatrix} X \\ f \end{pmatrix} = \begin{pmatrix} \Psi + \Lambda\Phi\Lambda' & \Lambda\Phi \\ \Phi\Lambda' & \Phi \end{pmatrix}. \tag{3}$$

Then the regression of f on X (Section 2.5) is

$$\mathscr{E}(f|X) = \Phi\Lambda'(\Psi + \Lambda\Phi\Lambda')^{-1}(x - \mu) \tag{4}$$
$$= \Phi(\Phi + \Phi\Gamma\Phi)^{-1}\Phi\Lambda'\Psi^{-1}(x - \mu).$$

The estimator or predictor of f_α is

$$\hat{f}^*_\alpha = \Phi(\Phi + \Phi\Gamma\Phi)^{-1}\Phi\Lambda'\Psi^{-1}(x_\alpha - \mu). \tag{5}$$

If $\Phi = I$, the predictor is

$$\hat{f}^*_\alpha = (I + \Gamma)^{-1}\Lambda'\Psi^{-1}(x_\alpha - \mu). \tag{6}$$

When Γ is also diagonal, the jth element of (6) is $\gamma_j/(1 + \gamma_j)$ times the jth element of (1). In the conditional distribution of x_α given f_α (for $\Phi = I$)

$$\mathscr{E}(\hat{f}^*_\alpha | f_\alpha) = (I + \Gamma)^{-1}\Gamma f_\alpha, \tag{7}$$

$$\mathscr{C}(\hat{f}^*_\alpha | f_\alpha) = (I + \Gamma)^{-1}\Gamma(I + \Gamma)^{-1}, \tag{8}$$

(9) $\mathscr{E}\left[(\hat{f}_\alpha^* - f_\alpha)(\hat{f}_\alpha^* - f_\alpha)'|f_\alpha\right] = (I+\Gamma)^{-1}(\Gamma + f_\alpha f_\alpha')(I+\Gamma)^{-1}.$

(10) $\mathscr{E}(\hat{f}_\alpha^* - f_\alpha)(\hat{f}_\alpha^* - f_\alpha)' = (I+\Gamma)^{-1}.$

This last matrix, describing the mean squared error, is smaller than (2) describing the unbiased estimator. The estimator (5) or (6) is a Bayes estimator and is appropriate when f_α is treated as random.

PROBLEMS

14.1. (Sec. 14.2) *Identification by 0's.* Let

$$\Lambda = \begin{pmatrix} 0 & \Lambda^{(1)} \\ \lambda_{(1)} & \Lambda_{(1)} \end{pmatrix}, \qquad C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & C_{22} \end{pmatrix},$$

where C is nonsingular. Show that

$$\Lambda C = \begin{pmatrix} 0 & \Lambda^{*(1)} \\ \lambda^*_{(1)} & \Lambda^*_{(1)} \end{pmatrix}$$

implies

$$C = \begin{pmatrix} c_{11} & c_{12} \\ 0 & C_{22} \end{pmatrix}$$

if and only if $\Lambda^{(1)}$ is of rank $m-1$.

14.2. (Sec. 14.3) For $p=3$, $m=1$, and $\Lambda = \lambda$, prove $|\theta_{ij}^2| = \prod_{i=1}^3(\lambda_i^2/\psi_{ii})$.

14.3. (Sec. 14.3) *The EM algorithm.*

(a) If f and U are normal and f and X are observed, show that the likelihood function based on $(x_1, f_1), \dots, (x_N, f_N)$ is

$$\prod_{\alpha=1}^N \left\{ \frac{1}{(2\pi)^{\frac{1}{2}p}\prod_{i=1}^p \psi_{ii}} \exp\left[-\frac{1}{2}\sum_{i=1}^p \frac{\left(x_{i\alpha} - \mu_i - \sum_{j=1}^m \lambda_{ij} f_{j\alpha}\right)^2}{\psi_{ii}} \right] \cdot \frac{1}{(2\pi)^{\frac{1}{2}m}|\Phi|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2} f_\alpha' \Phi^{-1} f_\alpha\right] \right\}.$$

(b) Show that when the factor scores are included as data the sufficient set of statistics is $\bar{x}, \bar{f}, C_{xx} = C$,

$$C_{xf} = \frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(f_\alpha - \bar{f})',$$

$$C_{ff} = \frac{1}{N}\sum_{\alpha=1}^{N}(f_\alpha - \bar{f})(f_\alpha - \bar{f})'.$$

(c) Show that the conditional expectations of the covariances in (b) given $X = (x_1, \dots, x_N)$, Λ, Φ, and Ψ are

$$C_{xx}^* = \mathscr{E}(C_{xx}|X, \Lambda, \Phi, \Psi) = C_{xx},$$

$$C_{xf}^* = \mathscr{E}(C_{xf}|X, \Lambda, \Phi, \Psi) = C_{xx}(\Psi + \Lambda\Phi\Lambda')^{-1}\Lambda\Phi,$$

$$C_{ff}^* = \mathscr{E}(C_{ff}|X, \Lambda, \Phi, \Psi) = \Phi\Lambda'(\Psi + \Lambda\Phi\Lambda')^{-1}C_{xx}(\Psi + \Lambda\Phi\Lambda')^{-1}\Lambda\Phi$$
$$+ \Phi - \Phi\Lambda'(\Psi + \Lambda\Phi\Lambda')^{-1}\Lambda\Phi.$$

(d) Show that the maximum likelihood estimators of Λ and Ψ given $\Phi = I$ are

$$\hat{\Lambda} = C_{xf}^* C_{ff}^{*-1},$$

$$\hat{\Psi} = C_{xx}^* - C_{xf}^* C_{ff}^{*-1} C_{xf}^{*\prime}.$$

CHAPTER 15

Patterns of Dependence; Graphical Models

15.1. INTRODUCTION

An emphasis in multivariate statistical analysis is that several measurements on a number in individuals or objects may be correlated, and the methods developed in this book take account of that dependence. The amount of association between two variables may be measured by the (Pearson) correlation of them (a symmetric measure); the association between one variable and a set may be quantified by a multiple correlation; and the dependence between one set and another set may be studied by criteria of independence such as studied in Chapter 9 or by canonical correlations. Similar measures can be applied in conditional distributions. Another kind of dependence (asymmetrical) is characterized by regression coefficients and related measures. In this chapter we study models which involve several kinds of dependence or more intricate patterns of dependence.

A *graphical model* in statistics is a visual diagram in which observable variables are identified with points (*vertices* or *nodes*) connected by *edges* and an associated family of probability distributions satisfying some independences specified by the visual pattern. Edges may be *undirected* (drawn as line segments) or *directed* (drawn as arrows). Undirected edges have to do with symmetrical dependence and independence, while directed edges may reflect a possible direction of action or sequence in time. These independences may come from a priori knowledge of the subject matter or may derive from these or other data. Advantages of the graphical display include

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

ease of comprehension, particularly of complicated patterns, ease of elicitation of expert opinion, and ease of comparing probabilities.

Use of such diagrams goes back at least to the work of the geneticist Sewall Wright (1921), (1934), who used the term "path analysis." An elaborate algebra has been developed for graphical models. Specification of independences reduces the number of parameters to be determined. Some of these independences are known as *Markov properties*. In a time series analysis of a Markov process (or order 1), for example, the future of the process is considered independent of the past when the present is given; in such a model the correlation between a variable in the past and a variable in the future is determined by the correlation between the present variable and the variable of the immediate future. This idea is expanded in several ways.

The family of probability distributions associated with a given diagram depends on the properties of the distribution that are represented by the graph. These properties for diagrams consisting of undirected edges (known as *undirected graphs*) will be described in Section 15.2; the properties for diagrams consisting entirely of directed edges (known as *directed graphs*) in Section 15.3; and properties of diagrams with both types of edges in Section 15.4. The methods of statistical inference will be given in Section 15.5.

In this chapter we assume that the variables have a joint nonsingular normal distribution; hence, the characterization of a model is in terms of the covariance matrix and its inverse, and functions of them. This assumption implies that the variables are quantitative and have a positive density. The mathematics of graphical models may apply to discrete variables (contingency tables) and to nonnormal quantitative variables, but we shall not develop the theory necessary to include them.

There is a considerable social science literature that has followed Wright's original work. For recent reviews of this writing see, for example, Pearl (2000) and McDonald (2002).

15.2. UNDIRECTED GRAPHS

A *graph* is a set of vertices and edges, $G \equiv (V, E)$. Each vertex is identified with a random vector. In this chapter the random variables have a joint normal distribution. Each undirected edge is a line connecting two vertices. It is designated by its two end points; (u, v) is the same as (v, u) in an undirected graph (but not in directed graphs).

Two vertices connected by an edge are called *adjacent*; if not connected by an edge, they are called *nonadjacent*. In Figure 15.1(a) all vertices are

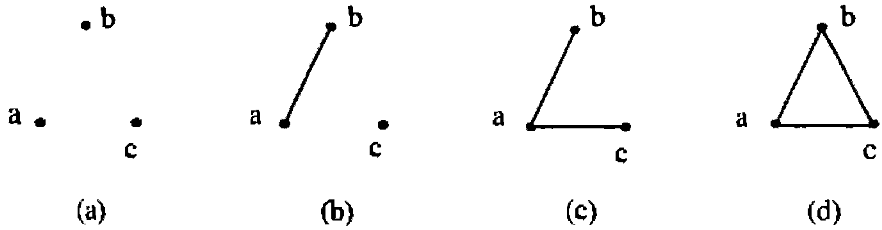

Figure 15.1

nonadjacent; in (b) a and b are adjacent; in (c) the pair a and b and the pair a and c are adjacent; in (d) every pair of vertices are adjacent.

The family of (normal) distributions associated with G is defined by a set of requirements on conditional distributions, known as *Markov properties.* Since the distributions considered here are normal, the conditions have to do with the covariance matrix Σ and its inverse $\Lambda = \Sigma^{-1}$, which is known as the *concentration matrix.* However, many of the lemmas and theorems hold for nonnormal distributions. We shall consider three definitions of Markov and then show that they are equivalent.

Definition 15.2.1. *The probability distribution on a graph is* pairwise Markov with respect to G *if for every pair of vertices* (u, v) *that are not adjacent* X_u *and* X_v *are independent conditional on all the other variables in the graph.*

In symbols

$$X_u \perp\!\!\!\perp X_v | X_{V\setminus(u,v)}, \tag{1}$$

where $\perp\!\!\!\perp$ means independence and $V\setminus(u, v)$ indicates the set V with u and v deleted. The definition of pairwise Markov is that $\rho_{uv\cdot V\setminus(u,v)} = 0$ for all pairs for which $(u, v) \notin E$. We may also write $u \perp\!\!\!\perp v | V\setminus(u, v)$.

Let Σ and $\Lambda = \Sigma^{-1}$ be partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \Lambda_{AA} & \Lambda_{AB} \\ \Lambda_{BA} & \Lambda_{BB} \end{bmatrix}, \tag{2}$$

where A and B are disjoint sets of vertices. The conditional distribution of X_A given X_B is

$$N\left(\Sigma_{AB}\Sigma_{BB}^{-1}X_B, \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right). \tag{3}$$

The conditional covariance matrix is

$$\Sigma_{AA\cdot B} = \Sigma_{AA} - X_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} = \Lambda_{AA}^{-1}. \tag{4}$$

If $A = (1, 2)$ and $B = (3, \ldots, p)$, the covariance of X_1 and X_2 given $X_3, \ldots, X_p$ is $\sigma_{12 \cdot 3 \cdots p}$ in $\Sigma_{A \cdot B} = (\sigma_{ij \cdot 3 \cdots p})$. This is 0 if and only if $\lambda_{12} = 0$; that is, $\Sigma_{A \cdot B}$ is diagonal if and only if Λ_{AA} is diagonal.

Theorem 15.2.1. *If a distribution on a graph is pairwise Markov, $\lambda_{ij} = 0$ for $(i, j) \notin V$.*

Definition 15.2.2. *The* boundary *of a set A, termed* bd(A), *consists of those vertices not in A that are adjacent to A. The closure of A, termed* cl(A), *is $A \cup$* bd(A).

Definition 15.2.3. *A distribution on a graph is* locally Markov *if for every vertex v the variable X_v is independent of the variables not in* cl(B) *conditional on the boundary of v: in notation,*

$$X_v \perp\!\!\!\perp X_{V \setminus \mathrm{cl}(v)} | X_{\mathrm{bd}(v)}. \tag{5}$$

Theorem 15.2.2. *The conditional independences*

$$X \perp\!\!\!\perp Y | Z, \qquad X \perp\!\!\!\perp Z | Y \tag{6}$$

hold if and only if

$$X \perp\!\!\!\perp (Y, Z). \tag{7}$$

Proof. The relations (6) imply that the density of X, Y, and Z can be written as

$$\begin{aligned} f(x, y, z) &= f(x|z) g(y|z) h(z) \\ &= k(x|y) l(z|y) m(y). \end{aligned} \tag{8}$$

Since $g(y|z)h(z) = n(y, z) = l(z|y)m(y)$, (8) implies $f(x|z) = k(x|y)$, which in turn implies $f(x|z) = k(x|y) = p(x)$. Hence

$$f(x, y, z) = p(x) n(y, z), \tag{9}$$

which is the density generating (7). Conversely, (9) can be written as either form in (8), implying (7). ■

Corollary 15.2.1. *The relations*

$$X \perp\!\!\!\perp Y | Z, W, \qquad X \perp\!\!\!\perp Z | Y, W \tag{10}$$

hold if and only if

$$X \perp\!\!\!\perp (Y, Z) | W. \tag{11}$$

The relations in Theorem 15.2.2 and Corollary 15.2.1 are sometimes called the *block independence theorem*. They are based on positive densities, that is, nonsingular normal distributions.

Theorem 15.2.3. *A locally Markov distribution on a graph is pairwise Markov.*

Proof. Suppose the graph is locally Markov (Definition 15.2.3). Let u and v be nonadjacent vertices. Because v is not adjacent to u, it is not in bd(u); hence,

$$X_u \perp\!\!\!\perp X_{V \setminus \mathrm{cl}(u)} | X_{\mathrm{bd}(u)}. \tag{12}$$

The relation (12) can be written

$$X_u \perp\!\!\!\perp \{X_v, X_{V \setminus [u, v, \mathrm{bd}(u)]}\} | \mathrm{bd}(u). \tag{13}$$

Then Corollary 15.2.1 ($X = X_u$, $Y = X_v$, $Z = Z_{V \setminus [\mathrm{cl}(u), v]}$, $W = X_{\mathrm{bd}(u)}$) implies

$$X_u \perp\!\!\!\perp X_v | X_{V \setminus (u, v)}. \qquad \blacksquare \tag{14}$$

Theorem 15.2.4. *A pairwise Markov distribution on a graph is locally Markov.*

Proof. Let $V \setminus \mathrm{cl}(u) = v_1 \cup \cdots \cup v_n$. Then

$$u \perp\!\!\!\perp v_1 | \mathrm{bd}(u) \cup v_2 \cup \cdots \cup v_n, \qquad u \perp\!\!\!\perp v_2 | \mathrm{bd}(u) \cup v_1 \cup v_3 \cup \cdots \cup v_n, \tag{15}$$

which by Corollary 15.2.1 implies

$$u \perp\!\!\!\perp v_1 \cup v_2 | \mathrm{bd}(u) \cup v_3 \cup \cdots \cup v_n. \tag{16}$$

Further, (16) and

$$u \perp\!\!\!\perp v_3 | \mathrm{bd}(u) \cup v_1 \cup v_2 \cup v_4 \cup \cdots \cup v_n \tag{17}$$

imply

$$u \perp\!\!\!\perp v_1 \cup v_2 \cup v_3 | \mathrm{bd}(u) \cup v_4 \cup \cdots \cup v_n. \tag{18}$$

This procedure leads to

$$u \perp\!\!\!\perp v_1 \cup \cdots \cup v_n | \mathrm{bd}(u). \qquad \blacksquare \tag{19}$$

A third notion of Markov, namely, global, requires some definitions.

Definition 15.2.4. *A* path *from B to C is a sequence $v_0, v_1, v_2, \dots, v_n$ of adjacent vertices with $v_0 \in B$ and $v_n \in C$.*

Definition 15.2.5. *A set S* separates *sets B and C if S, B, and C are disjoint and every path from B to C intersects S.*

Thus S separates B and C if for every sequence of vertices $v_0, v_1, \dots, v_n$ with $v_0 \in B$ and $v_n \in C$ at least one of $v_1, \dots, v_{n-1}$ is a vertex in S. Here B and/or C are nonempty, but S can be empty.

Definition 15.2.6. *A distribution on a graph is* globally Markov *if for every triplet of disjoint sets S, B, and C such that S separates B and C the vector variables X_B and X_C are independent conditional on X_S.*

In the example of Figure 15.1(c), a separates b and c. If $\rho_{bc\cdot a} = 0$, that is, $\rho_{bc} - \rho_{ba}\rho_{ac} = 0$, the distribution is globally Markov. Note that a set of vertices is identified with a vector of variables.

The global Markov property puts restrictions on the possible (normal) distributions, and that implies fewer parameters about which to make inferences.

Suppose $V = A \cup B \cup S$, where A, B, and S are disjoint. Partition Σ and $\Lambda = \Sigma^{-1}$, the *concentration* matrix, as

$$(20) \qquad \Lambda = \begin{bmatrix} \Lambda_{AA} & \Lambda_{AB} & \Lambda_{AS} \\ \Lambda_{BA} & \Lambda_{BB} & \Lambda_{BS} \\ \Lambda_{SA} & \Lambda_{SB} & \Lambda_{SS} \end{bmatrix} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} & \Sigma_{AS} \\ \Sigma_{BA} & \Sigma_{BB} & \Sigma_{BS} \\ \Sigma_{SA} & \Sigma_{SB} & \Sigma_{SS} \end{bmatrix}^{-1}.$$

The conditional distribution of $(X_A', X_B')'$ given X_S is normal with covariance matrix

$$(21) \qquad \Sigma_{(A,B)\cdot S} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} - \begin{bmatrix} \Sigma_{AS} \\ \Sigma_{BS} \end{bmatrix} \Sigma_{SS}^{-1} \begin{bmatrix} \Sigma_{SA} & \Sigma_{SB} \end{bmatrix}$$
$$= \begin{bmatrix} \Lambda_{AA} & \Lambda_{AB} \\ \Lambda_{BA} & \Lambda_{BB} \end{bmatrix}^{-1}.$$

Theorem 15.2.5. *If S separates A and B in a graph with a globally Markov distribution,* $\Lambda_{AB} = 0$.

Proof. Because S separates A and B, every element u of A and every element v of B are nonadjacent, for otherwise the path (u, v) would connect

A and B without intersecting S. The globally Markov property is that X_A and X_B are uncorrelated in the conditional distribution, implying that $\Sigma_{(A,B)\cdot S}$ is block diagonal and hence that $\Lambda_{AB} = \mathbf{0}$. ∎

Theorem 15.2.6. *A distribution on a globally Markov graph is pairwise Markov.*

Proof. Let the set B be i, the set C be j not adjacent to i, and the set A the rest of the variables. Any path from B to C must include elements of A. Hence i is independent of j in the distribution conditioned on the other variables. ∎

Theorem 15.2.7. *A globally Markov family of distributions on a graph is locally Markov.*

Proof. The boundary of a set B separates B and $V \backslash \mathrm{cl}(B)$. ∎

Theorem 15.2.8. *A pairwise Markov family of distributions on a graph is globally Markov.*

Proof. Let A, B, and S be disjoint sets in a pairwise Markov graph such that S separates A and B. Let $\#(S)$ and $\#(V)$ denote the numbers of vertices in S and V, respectively. If $\#(V) = \#(S) + 2$, that is, $V = A \cup B \cup S$, then there must be one vertex in each of A and B, and the pairwise Markov property is exactly the globally Markov property. The rest of the proof is a backward induction on $\#(S)$. Suppose $\#(V) - \#(S) > 2$ and $V = A \cup B \cup S$. Then either A or B or both have more than one vertex. Suppose A has more than one vertex, and let $u \in A$. Then $S \cup u$ separates $A \backslash u$ and B, and $S \cup A$ separates u and B. By the induction hypothesis

$$(22) \qquad X_{A\backslash u} \perp\!\!\!\perp X_B \mid (X_S, X_u), \qquad X_u \perp\!\!\!\perp X_B \mid (X_S, X_{A\backslash u}).$$

By Corollary 15.2.1

$$(23) \qquad X_A \perp\!\!\!\perp X_B \mid X_S.$$

Now suppose $A \cup B \cup S \subset V$. Let $u \in V \backslash (A \cup B \cup S)$. Then $S \cup u$ separates A and B. By the induction hypothesis

$$(24) \qquad X_A \perp\!\!\!\perp X_B \mid (X_S, X_u).$$

Also, either $A \cup S$ separates u and B or $B \cup S$ separates A and u. (Otherwise there would be a path from B to u and from u to A that would not intersect S.) If $A \cup S$ separates u and B,

(25) $$X_u \perp\!\!\!\perp X_B | (X_S, X_A).$$

Then Corollary 15.2.1 applied to (19) and (20) implies

(26) $$(X_A, X_u) \perp\!\!\!\perp X_B | X_S,$$

from which we derive $X_A \perp\!\!\!\perp X_B | X_S$. ■

Theorems 15.2.3, 15.2.5, and 15.2.6 show that the three Markov properties are equivalent: any one implies the other two. The proofs here hold fairly generally, but in this chapter a nonsingular multivariate normal distribution is assumed: thus all densities are positive.

Definition 15.2.7. *A graph* $G = (V, E)$ *is* complete *if and only if every two vertices in* V *are adjacent.*

The definition implies that the graph specifies no restriction on the covariance matrix of the multivariate normal distribution.

A subset $A \subseteq V$ induces a subgraph $G_A = (A, E_A)$, where the edge set E_A includes all edges (u, v) of G with $(u, v) \in E$, where $u \in A$ and $v \in A$. A subset of a graph is complete if and only if every two vertices in A are adjacent in E_A.

Definition 15.2.8. *A* clique *is a maximal complete set of vertices.*

"Maximal" means that if another vertex from V is added to the set, the set will no longer be complete. A clique can be constructed by starting with one vertex, say v_1. If it is not adjacent to any other vertex, v_1 alone constitutes a clique. If v_2 is adjacent to v_1 $[(v_1, v_2) \in E]$, continue constructing a clique with v_1 and v_2 in it until a maximal complete subset is obtained. Thus every vertex is a member of at least one clique, and every edge is included in at least one clique.

Lemma 15.2.1. *If the distribution of* X_V *is Markov, it is determined by the set of marginal distributions of all cliques.*

In Figure 15.1(a) each of a, b, c is a clique; in (b) each of (a, b) and c is a clique; in (c) each of (a, b) and (a, c) is a clique; in (d) (a, b, c) is a clique.

Definition 15.2.9. *The density* $f(X_V)$ factorizes with respect to G *if there are nonnegative functions* $g_C(X_C)$ *depending on the complete subgraphs such that*

(27) $$f(X_V) = \prod_{C \text{ complete}} g_C(X_C).$$

Since it suffices to consider only cliques, an alternative factorization is

$$f(X_V) = \prod_{C^* \text{ cliques}} g_{C^*}(X_{C^*}). \tag{28}$$

These functions $g_C(X_C)$ and $g_{C^*}(X_{C^*})$ are not necessarily densities or conditional densities. The problems of statistical inference may be reduced to the problems of the complete subgraphs or cliques.

Definition 15.2.10. *A* decomposition *of a graph is formed by three disjoint sets A, B, S if $V = A \cup B \cup S$, S separates A and B, and S is complete.*

In this definition one or more of the sets A, B, and S may be empty. If both A and B are nonempty, the decomposition is termed *proper*.

Definition 15.2.11. *A graph is* decomposable *if it is complete or if there is a proper decomposition (A, B, S) into decomposable subgraphs $G_{A \cup S}$ and $G_{B \cup S}$.*

Theorem 15.2.9. *Suppose A, B, S decomposes $G = (V, E)$. Then the density of X_V factorizes with respect to G if and only if its marginal densities $f_{A \cup S}(\boldsymbol{x}_{A \cup S})$ and $f_{B \cup S}(\boldsymbol{x}_{B \cup S})$ factorize and the densities satisfy*

$$f(\boldsymbol{x}_V) = \frac{f_{A \cup S}(\boldsymbol{x}_{A \cup S}) f_{B \cup S}(\boldsymbol{x}_{B \cup S})}{f_S(\boldsymbol{x}_S)}. \tag{29}$$

Proof. Suppose that $f_V(\boldsymbol{x}_V)$ factorizes as

$$f_V(\boldsymbol{x}_V) = \prod_{c \in C} g_c(\boldsymbol{x}_c). \tag{30}$$

Because A, B, S decomposes G, every clique is either a subset of $A \cup S$ or a subset of $B \cup S$. Let $\mathscr{A}$ denote the cliques that are subsets of $A \cup S$, and $\mathscr{B}$ those that are subsets of B. Then $f_V(\boldsymbol{x}_V) = h(\boldsymbol{x}_{A \cup S}) k(\boldsymbol{x}_{B \cup S})$, where

$$h(\boldsymbol{x}_{A \cup S}) = \prod_{C \in \mathscr{A}} g_C(\boldsymbol{x}_C), \tag{31}$$

$$k(\boldsymbol{x}_{B \cup S}) = \prod_{C \in \mathscr{B} \setminus \mathscr{A}} g_C(\boldsymbol{x}_C). \tag{32}$$

Integration of (30) with respect to x_C for $C \in \mathscr{B} \setminus \mathscr{A}$ gives

$$f_{A \cup S}(\boldsymbol{x}_{A \cup S}) = h(\boldsymbol{x}_{A \cup S}) \bar{k}(\boldsymbol{x}_S), \tag{33}$$

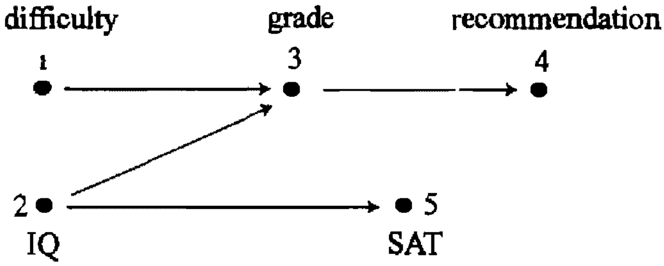

Figure 15.2

where

(34) $$\bar{k}(x_S) = \int k(x_{B \cup S})\, dx_B. \qquad \blacksquare$$

In turn $f_{A \cup S}(x_{A \cup S})$ and $f_{B \cup S}(x_{B \cup S})$ can be factorized, leading to (28).

15.3. DIRECTED GRAPHS

We now include relations with a direction; the measurement represented by one vertex u may precede the measurement represented by another vertex v. In the graph this *directed edge* is displayed as an arrow pointing from u to v; in notation it appears as (u, v), which is now distinguished from (v, u). The precedence may indicate the times of measurement, for example, the precipitation on two successive days, or may indicate possible causation.

The difficulty of an examination x_1 may affect the grade of a student x_3; the grade is also affected by his/her IQ x_2. In turn the grade of the student influences the quality of a letter of recommendation x_4; the IQ is a factor in performance on the SAT, x_5. See Figure 15.2. (We shall draw figures so that the action proceeds from left to right.)

A graph composed entirely of directed edges is called a *directed graph*. A cycle, such as $1 \to 2$, $2 \to 3$, $3 \to 1$, is hard to interpret and hence is usually ruled out. A directed graph without a cycle is an *acyclic directed graph* (ADG or DAG), also known as an acyclic digraph. All directed graphs in this chapter are acyclic.

An acyclic directed graph may represent a *recursive linear system*. For example, Figure 15.2 could represent

(1) $$X_1 = u_1,$$

(2) $$X_2 = u_2,$$

$$X_3 = \beta_{31} X_1 + \beta_{32} X_2 + u_3, \tag{3}$$

$$X_4 = \beta_{43} X_3 + u_4, \tag{4}$$

$$X_5 = \beta_{52} X_2 + u_5, \tag{5}$$

where u_1, u_2, u_3, u_4, u_5 are mutually independent unobserved variables.

Wold (1960) called such models *causal chains*. Note that the matrix of coefficients is lower triangular. In general X_i may depend on $X_1, \ldots, X_{i-1}$.

The recursive linear system (1) to (5) generates the recursive factorization

$$\begin{aligned} f_{12345}(x_1, x_2, x_3, x_4, x_5) \\ = f_1(x_1) f_2(x_2) f_{3|12}(x_3|x_1, x_2) f_{4|123}(x_4|x_3) f_{5|1234}(x_5|x_2). \end{aligned} \tag{6}$$

A directed graph induces a partial order.

Definition 15.3.1. *A* partial ordering *of an acyclic directed graph $u \le v$ is defined by the existence of a directed path*

$$u = v_0 \to v_1 \to \cdots \to v_n = v. \tag{7}$$

The partial ordering satisfies the conditions (i) reflexive: $v \le v$; (ii) transitive: $u \le v$ and $v \le w$ imply $u \le w$; and (iii) antisymmetric: $u \le v$ and $v \le u$ imply $u = v$. Further, $u \le v$ and $u \neq v$ defines $u < v$.

Definition 15.3.2. *If $u \to v$, then u is a* parent *of v, termed $u = \mathrm{pa}(v)$, and v is a* child *of u, termed $v = \mathrm{ch}(u)$. In symbols*

$$\mathrm{pa}(v) = \{w \in V \setminus v | w \to v\}, \tag{8}$$

$$\mathrm{ch}(u) = \{w \in V \setminus u | u \to w\}. \tag{9}$$

In the graph displayed in Figure 15.2 we have $(1, 2) = \mathrm{pa}(3)$, $3 = \mathrm{pa}(4)$, $2 = \mathrm{pa}(5)$, $3 = \mathrm{ch}(1, 2)$, $4 = \mathrm{ch}(3)$, and $5 = \mathrm{ch}(2)$.

Definition 15.3.3. *If $u < v$, then v is a* descendant *of u,*

$$\mathrm{de}(u) = \{v | u < v\}, \tag{10}$$

and u is an ancestor *of v,*

$$\mathrm{an}(v) = \{u | u < v\}. \tag{11}$$

The set of *nondescendants* of u is $\mathrm{Nd}(u) = V \setminus \mathrm{de}(u)$, and the set of *strict nondescendants* is $\mathrm{nd}(u) = \mathrm{Nd}(u) \setminus u$. Define $\mathrm{An}(A) = \mathrm{an}(A) \cup A$.

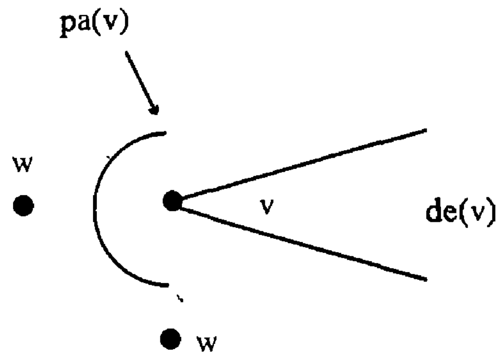

Figure 15.3

Note that

$$\mathrm{pa}(v) \subseteq \mathrm{an}(v) \subseteq \mathrm{nd}(v). \tag{12}$$

In our study of undirected graphs we considered three Markov properties independently defined and then showed that a graph with one Markov property also has the other two. In the case of acyclic directed graphs we shall define three similar Markov properties, but the definitions are different because they take account of the direction of action or influence.

Definition 15.3.4. *A distribution on an acyclic directed graph G is* pairwise Markov *if for every $v \in V$ and $w \in \mathrm{nd}(v)\setminus \mathrm{pa}(v)$*

$$v \perp\!\!\!\perp w \mid \mathrm{nd}(v)\setminus w. \tag{13}$$

In comparison with Definition 15.2.1 for undirected graphs, note that attention is paid only to vertices in $\mathrm{nd}(v)$; since $\mathrm{pa}(v)$ is the effective boundary of v, the vertices w and v are nonadjacent. (See Figure 15.3.) Note also that the conditioning set inclues the parents of v, but not the children (which are descendants).

Definition 15.3.5. *A distribution on an acyclic directed graph is* locally Markov *if*

$$v \perp\!\!\!\perp [\mathrm{nd}(v)\setminus \mathrm{pa}(v)] \mid \mathrm{pa}(v). \tag{14}$$

In the definition of locally Markov the conditioning is only on the parents of v, but in the definition of pairwise Markov the conditioning is on all of the other nondescendants. These features correspond to Definitions 15.2.1 and 15.2.3 for undirected graphs.

In Figure 15.2, we have $1 \perp\!\!\!\perp 2,5$, $3 \perp\!\!\!\perp 5 \mid 2$, $4 \perp\!\!\!\perp 1,2,5 \mid 3$, and $5 \perp\!\!\!\perp 1,3,4 \mid 2$. In an undirected graph constructed by replacing arrows in Figure 15.2 by lines (directed edges by undirected edges), a locally Markov distribution on the graph would include the conditional independences $1 \perp\!\!\!\perp 2 | 3$, $1, 2 \perp\!\!\!\perp 4 | 3$, $1, 3, 4 \perp\!\!\!\perp 5$. In the interpretation of the arrow indicating time sequence X_4 relates to the future of (X_2, X_3); the future cannot be conditioned on.

As another example, consider an autoregressive time series $y_0, y_1, \ldots, y_T$ defined by

$$y_t = \rho y_{t-1} + u_t, \qquad i = 1, 2, \ldots, T, \tag{15}$$

where $u_1, \ldots, u_T$ are independent $N(0, \sigma^2)$ variables and y_0 has distribution $N[0, \sigma^2/(1-\rho)^2)]$. In this case given y_t, the future $y_{t+1}, \ldots, y_T$ is independent of the past $y_0, \ldots, y_{t-1}$.

Theorem 15.3.1. *A locally Markov distribution on an acyclic directed graph is pairwise Markov.*

Proof. The proof is the same as the proof of Theorem 15.2.3 for undirected graphs. ■

Theorem 15.3.2. *A pairwise Markov distribution on an acyclic directed graph is locally Markov.*

Proof. The proof is the same as the proof of Theorem 15.2.4. ■

Another Markov property is based on numbering the vertices in an order reflecting the direction of the action or the partial ordering induced.

Definition 15.3.6. *An enumeration of the elements of V is called* well-numbered *if $i < j \Rightarrow v_j \nless v_i$, or equivalently $v_j < v_i \Rightarrow j < i$.*

Theorem 15.3.3. *A finite ordered set $(V, \leq)$ admits at least one well-numbering.*

Definition 15.3.7. *An element $a^* \in V$ is* maximal (*or* terminal) *if $a^* \leq b \Rightarrow a^* = b$.*

Lemma 15.3.1. *A finite, partially ordered set $(V, \leq)$ has at least one maximal element a^*.*

Proof of Lemma. The proof is by induction with $a^* = a$ if $\#(V) = 1$. Assume the lemma holds for $\#(V) = n$, and consider $\#(V) = n + 1$. Then $V = a \cup (V \backslash a)$ for any $a \in V$. Since $\#(V \backslash a) = n$, $V \backslash a$ has a maximal element, say $\bar{a}$. Then either $\bar{a} \leq a$ and so a is maximal, or $\bar{a} \nless a$ and so $\bar{a}$ is maximal. ■

Proof of Theorem 15.3.3. We shall construct a well-numbering. Let v^* be a maximal element; define $v_n = v^*$. In $V \setminus v_n$ let v^{**} be a maximal element; define $v_{n-1} = v^{**}$. At the jth stage let v^{***} be a maximal element in $V \setminus (v_n, \dots, v_{n-j+1})$; define $v_{n-j} = v^{***}$, $j = 3, \dots, n-1$. Then $v_1 = V \setminus (v_1, \dots, v_{n-1})$. This construction satisfies Definition 15.3.6. ■

The well-numbering of V as $v^{(1)}, \dots, v^{(n)}$ implies that in any directed path $u = v^{(i_0)} \to v^{(i_1)} \to \cdots \to v^{(i_n)} = v$ the indices satisfy $i_0 \leq i_1 \leq \cdots \leq i_n$. The well-numbering is not necessarily unique. Since V is finite, a maximal element can be found by comparing v_i and v_j for at most $n(n-1)/2$ pairs.

Definition 15.3.8. *Let $\{v_1, \dots, v_n\}$ be a well-numbering of the acyclic directed graph G. A distribution on G is* well-numbered Markov *with respect to this well-numbering if*

$$v_i \perp\!\!\!\perp (v_1, \dots, v_{i-1}) \setminus \mathrm{pa}(v_i) \mid \mathrm{pa}(v_i), \qquad i = 3, \dots, n. \tag{16}$$

Apparently the definition depends on the choice of well-numbering, but this is not the case, by Theorem 15.3.4.

Theorem 15.3.4. *A distribution on an acyclic directed graph that is well-numbered Markov is locally Markov.*

Proof. $(v_1, \dots, v_{i-1}) \in \mathrm{nd}(v_i) \setminus \mathrm{pa}(v_i)$. ■

The definition of the global Markov property depends on relating the directed graph to a corresponding undirected graph.

Definition 15.3.9. *The* moral graph G^m *of an acyclic directed graph $G = (V, E)$ is the undirected graph constructed by adding (undirected) edges between parents of each vertex $v \in V$ and replacing every directed edge by an undirected edge.*

In the jargon of graph theory, the parents of a vertex are "married."

Definition 15.3.10. *A distribution on an acyclic directed graph is* globally Markov *if $A \perp\!\!\!\perp B \mid S$ for every A, B, and S such that S separates A and B in $[G_{\mathrm{An}(A \cup B \cup S)}]^m$.*

Theorem 15.3.5. *A distribution on an acyclic directed graph that is globally Markov is locally Markov.*

Proof. For any $v \in V$ let $\mathrm{pa}(v) = S$ in the definition of globally Markov. Let $v = A$ and $\mathrm{nd}(v)\backslash\mathrm{pa}(v) = B$. A vertex $w \in \mathrm{nd}(v)\backslash\mathrm{pa}(v)$ is a vertex in $\mathrm{An}(A \cup B \cup S)$. Let $\pi = w = v_0, v_1, \ldots, v_n = v$ be a path from w to v in $[G_{\mathrm{An}(A \cup B \cup S)}]^m = [G_{\mathrm{Nd}(v)}]^m$. If (v_{n-1}, v_n) corresponds to a directed edge $(v_{n-1} \to v_n)$ in $[G_{\mathrm{Nd}(v)}]^m$, then $v_{n-1} \in \mathrm{pa}(v) = S$ and $\mathrm{pa}(v)$ separates $\mathrm{nd}(v)\backslash\mathrm{pa}(v)$ and v. [The directed edge $(v_{n-1} \leftarrow v_n)$ implies $v_{n-1} \in \mathrm{de}(v)$.] ∎

Theorem 15.3.6. *A distribution on an acyclic directed graph that is locally Markov is globally Markov.*

The proof is very lengthy and is omitted.

Recursive Factorization

The recursive aspect of the acyclic directed graph permits a systematic factorization of the density. Use the construction of Theorem 15.3.4. Let $n = |V|$; then v_n is a maximal element of V. Then

$$X_{V\backslash \mathrm{cl}(v_n)} \perp\!\!\!\perp X_{v_n} | \mathrm{pa}(v_n). \tag{17}$$

Thus (in the normal case)

$$\mathscr{E} X_{v_n} | \mathrm{pa}(v_n) = \boldsymbol{\alpha}_n + \boldsymbol{B}_n X_{\mathrm{pa}(v_n)}, \tag{18}$$

$$\mathscr{E}(X_{v_n} - \mathscr{E}X_{v_n})(X_{v_n} - \mathscr{E}X_{v_n})' = \Sigma_n. \tag{19}$$

At the jth step let v_{n-j+1} be a maximal element of $V\backslash(v_n, \ldots, v_{n-j+2})$. Then

$$X_{V\backslash[v_n, \ldots, v_{n-j+2}, \mathrm{cl}(v_{n-j+1})]} \perp\!\!\!\perp X_{v_{n-j+1}} | \mathrm{pa}(v_{n-j+1}). \tag{20}$$

Thus

$$\mathscr{E} X_{v_{n-j+1}} | \mathrm{pa}(v_{n-j+1}) = \boldsymbol{\alpha}_{n-j+1} + \boldsymbol{B}_{n-j+1} \boldsymbol{x}_{\mathrm{pa}(v_{n-j+1})}, \tag{21}$$

$$\mathscr{E}(X_{v_{n-j+1}} - \mathscr{E}X_{v_{n-j+1}})(X_{v_{n-j+1}} - \mathscr{E}X_{v_{n-j+1}})' = \Sigma_{n-j+1}, \quad j = 1, \ldots, n-1. \tag{22}$$

The vector $X_{v_{n-j+1}}$ is independent of $\mathrm{pa}(v_{n-j+1})$. The relations (18) to (22) can be written as generating equations. Let

$$\boldsymbol{x}_1 = \boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}_1, \tag{23}$$

$$\boldsymbol{x}_2 = \boldsymbol{\alpha}_2 + \boldsymbol{B}_2 \boldsymbol{x}_1 + \boldsymbol{\varepsilon}_2, \tag{24}$$

$$\vdots$$

$$\boldsymbol{x}_{n-1} = \boldsymbol{\alpha}_{n-1} + \boldsymbol{B}_{n-1}(\boldsymbol{x}_1', \ldots, \boldsymbol{x}_{n-2}')' + \boldsymbol{\varepsilon}_{n-1}, \tag{25}$$

$$\boldsymbol{x}_n = \boldsymbol{\alpha}_n + \boldsymbol{B}_n(\boldsymbol{x}_1', \ldots, \boldsymbol{x}_{n-1}')' + \boldsymbol{\varepsilon}_n, \tag{26}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random vectors with $\mathscr{E}\varepsilon_j\varepsilon_j' = \Sigma_j$. In matrix form (23) to (26) are

$$Bx = \alpha + \varepsilon, \tag{27}$$

where

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \cdots \\ \alpha_n \end{bmatrix}, \quad B = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ -B_{21} & I & 0 & \cdots & 0 \\ -B_{31} & -B_{32} & I & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -B_{n1} & -B_{n2} & -B_{n3} & \cdots & I \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdots \\ \varepsilon_n \end{bmatrix}, \tag{28}$$

and $B_{ji} = 0$ if $i < j - k_j$. Because the determinant of B is 1, (27) can be solved for

$$x = \Gamma^{-1}\alpha + \Gamma^{-1}\varepsilon. \tag{29}$$

The matrix Γ^{-1} is also lower triangular.

15.4. CHAIN GRAPHS

A *chain graph* includes both directed and undirected edges; however, only certain patterns of vertices and edges are permitted. Suppose the set of vertices V of the graph $G = (V, E)$ can be partitioned into subsets $V = V(1) \cup \cdots \cup V(T)$ so that within a subset the vertices are joined by undirected edges and directed edges join vertices in different subsets. Let $\mathscr{T}(G)$ be the set of vertices $1, \ldots, T$ and let $\mathscr{E}(G)$ be the (directed) edge set such that $\tau \to \sigma$ if and only if there is at least one element $u \in V(\tau)$ and at least one element $v \in V(\sigma)$ such that $u \to v$ is in E, the edge set of G. Then $\mathscr{D}(G) = [\mathscr{T}(G), \mathscr{E}(G)]$ is an acyclic directed graph; we can define $\mathrm{pa}_{\mathscr{D}}(\tau)$, etc., for $\mathscr{D}(G)$.

Let $X_\tau = \{X_u | u \in V(\tau)\}$. Within a set the vertices form an undirected graph relative to the probability distribution conditional on the past (that is, earlier sets). See Figure 15.4 [Lauritzen (1996)] and Figure 15.5.

We now define the Markov properties as specified by Lauritzen and Wermuth (1989) and Frydenberg (1990):

(C1) The distribution of X_τ, $\tau = 1, \ldots, T$, is locally Markov with respect to the acyclic directed graph $\mathscr{D}(G)$; that is,

$$X_\tau \perp\!\!\!\perp X_\sigma | X_{\mathrm{pa}_{\mathscr{D}}(\tau)}, \qquad \sigma \in \mathrm{nd}_{\mathscr{D}}(\tau) \setminus \mathrm{pa}_{\mathscr{D}}(\tau). \tag{1}$$

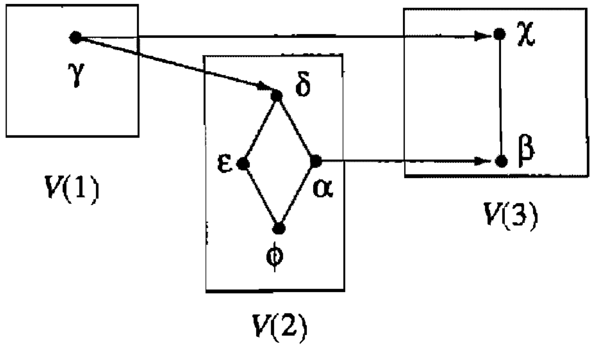

Figure 15.4. A chain graph.

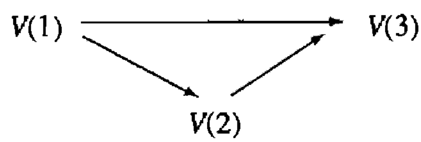

Figure 15.5. The corresponding induced acyclic directed graph on $V = V(1) \cup V(2) \cup V(3)$.

(C2) For each τ the conditional distribution of X_τ given $X_{\mathrm{pa}_{\mathscr{D}}}(\tau)$ is globally Markov with respect to the undirected graph on $V(\tau)$.

(C3)

$$(2) \qquad X_u \perp\!\!\!\perp X_v | X_{\mathrm{bd}_G}(U), \; u \in U \subseteq (\tau), \; v \in \mathrm{pa}_{\mathscr{D}}(\tau) \setminus \mathrm{pa}_G(U).$$

Here $\mathrm{bd}_G(U) = \mathrm{pa}_G(U) \cup \mathrm{nb}_G(U)$. A distribution on the chain graph G that satisfies (C1), (C2), (C3) is *LWF block recursive Markov.*

In Figure 15.6 $\mathrm{pa}_{\mathscr{D}}(\tau) = \{\tau - 1, \tau - 2\}$ and $\mathrm{nd}_{\mathscr{D}}(\tau) \setminus \mathrm{pa}_{\mathscr{D}}(\tau) = \{\tau - 3, \tau - 4, \ldots, 1\}$. The set $U = \{u, w\}$ is a set in $V(\tau)$, and $\mathrm{pa}_G(U)$ is the set in $V(\tau - 1) \cup V(\tau - 2)$ that includes $\mathrm{pa}_G(u)$ for $u \in U$; that is, $\mathrm{pa}_G(U) = \{x, y\}$.

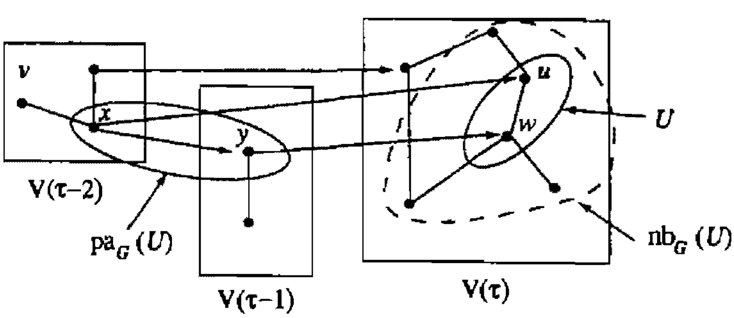

Figure 15.6. A chain graph.

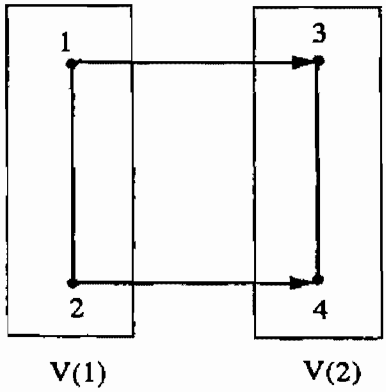

Figure 15.7. A chain graph.

Andersson, Madigan, and Perlman (2001) have proposed an alternative Markov property (AMP), replacing (C3) by

(C3*)

$$(3) \qquad X_u \perp\!\!\!\perp X_v | X_{\mathrm{pa}_G(U)}, \qquad u \in U \subseteq V(\tau), \quad v \in \mathrm{pa}_{\mathscr{D}}(\tau) \setminus \mathrm{pa}_G(U).$$

In Figure 15.6, X_v for a vertex v in $V(\tau-2) \cup V(\tau-1)$ is conditionally independent of X_u $[u \in U \subseteq V(\tau)]$ when regressed on $X_{\mathrm{pa}_G(U)} = (X_x, X_y)$. The difference between (C3) and (C3*) is that the conditioning in (C3) is on $\mathrm{bd}_G(U) = \mathrm{pa}_G(U) \cup \mathrm{nb}_G(U)$, but the conditioning in (C3*) is on $\mathrm{pa}_G(U)$ only. See Figure 15.6. The conditioning in (C3*) is on variables in the past. Figure 15.7 [Andersson, Madigan, and Perlman (2001)] illustrates the difference between the LWF and AMP Markov properties:

$$(4) \qquad \text{LWF:} \qquad X_1 \perp\!\!\!\perp X_4 | X_2, X_3, \quad X_2 \perp\!\!\!\perp X_3 | X_1, X_4,$$

$$(5) \qquad \text{AMP:} \qquad X_1 \perp\!\!\!\perp X_4 | X_2, \qquad X_2 \perp\!\!\!\perp X_3 | X_1.$$

Note that in (5) X_1 and X_4 are conditionally independent given X_2; the conditional distribution of X_4 depends on $\mathrm{pa}(v_2)$, but not X_3.

The AMP specification allows a block recursive equation formulation. In the example in Figure 15.7 the distribution of scalars X_1 and X_2 $[v_1, v_2 \in V(1)]$ can be specified as

$$(6) \qquad X_1 = \varepsilon_1,$$

$$(7) \qquad X_2 = \varepsilon_2,$$

where $(\varepsilon_1, \varepsilon_2)$ has an arbitrary (normal) distribution. Since X_3 depends directly on X_1 and X_4 depends directly on X_2, we write

$$X_3 = \beta_{31} X_1 + \varepsilon_3, \tag{8}$$

$$X_4 = \beta_{42} X_2 + \varepsilon_4, \tag{9}$$

where $(\varepsilon_3, \varepsilon_4)$ has an arbitrary distribution independent of $(\varepsilon_1, \varepsilon_2)$, and hence independent of (X_1, X_2).

In general the AMP model can be expressed as (26) of Section 15.3.

15.5. STATISTICAL INFERENCE

15.5.1. Normal Distribution

Let $x_1, \dots, x_N$ be N observations on X with distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\bar{x} = N^{-1}\sum_{\alpha=1}^N x_\alpha$ and $S = (N-1)^{-1}\sum_{\alpha=1}^N (x_\alpha - \bar{x})(x_\alpha - \bar{x})' = (N-1)^{-1}[\sum_{\alpha=1}^N x_\alpha x_\alpha' - N\bar{x}\bar{x}']$. The likelihood is

$$\begin{aligned}(1)\qquad & (2\pi)^{-Np/2}|\boldsymbol{\Sigma}|^{-N/2} e^{-\frac{1}{2}\sum_{\alpha=1}^N (x_\alpha - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x_\alpha - \boldsymbol{\mu})} \\ &= (2\pi)^{-Np/2}|\boldsymbol{\Sigma}|^{-N/2} e^{-\frac{1}{2}[(N-1)\operatorname{tr} S\boldsymbol{\Sigma}^{-1} + N(\bar{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{x}-\boldsymbol{\mu})]}.\end{aligned}$$

The above form shows that $\bar{x}$ and S are a pair of sufficient statistics for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and they are independently distributed. The interest in this chapter is on the dependences, which depend only on the covariance matrix $\boldsymbol{\Sigma}$, not $\boldsymbol{\mu}$. For the rest of this chapter we shall suppress the mean. Accordingly, we suppose that the parent distribution is $N(\mathbf{0}, \boldsymbol{\Sigma})$ and the sample is $x_1, \dots, x_n$), and $S = (1/n)\sum_{\alpha=1}^n x_\alpha x_\alpha'$. The likelihood function can be written

$$\begin{aligned}(2)\qquad & (2\pi)^{-pn/2}|\boldsymbol{\Lambda}|^{pn/2} e^{-\frac{1}{2}\operatorname{tr} n\boldsymbol{\Lambda} S} \\ &= \exp\left[-\Psi(\boldsymbol{\Lambda}) - \frac{1}{2}\sum_{i=1}^{p}\lambda_{ii}t_{ii} - \sum_{i<j}\lambda_{ij}t_{ij}\right],\end{aligned}$$

where $\boldsymbol{\Lambda} = (\lambda_{ij}) = \boldsymbol{\Sigma}^{-1}$, $T = (t_{ij}) = \sum_{\alpha=1}^N x_\alpha x_\alpha'$, and $\Psi(\boldsymbol{\Lambda}) = \frac{1}{2}pn\log(2\pi) - \frac{1}{2}n\log|\boldsymbol{\Lambda}|$.

The likelihood is in the exponential family with canonical parameter $\boldsymbol{\Lambda}$ and statistic T. The maximum likelihood estimator of $\boldsymbol{\Sigma}$ with no restriction is $\hat{\boldsymbol{\Sigma}} = S = (1/n)T$. Since $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is a 1-to-1 transformation of $\boldsymbol{\Sigma}$, the maximum likelihood estimator of $\boldsymbol{\Lambda}$ of $\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Sigma}}^{-1}$.

15.5.2. Covariance Selection Models

In undirected graphs many of the models involve zero restrictions on elements of $\mathbf{\Lambda}$. Dempster (1972) studied such models and introduced the term *covariance selection*. When the (directed) graph satisfies the pairwise Markov condition, $\lambda_{ij} = 0$ for $(i, j) \notin E$. We assume here that the graph satisfies this Markov condition. Further we assume $n \geq p$; then $\boldsymbol{S}$ is positive definite with probability 1.

The likelihood function is

$$(3) \qquad (2\pi)^{-p/2} |\mathbf{\Lambda}|^{-n/2} \exp\left(\tfrac{1}{2} \sum_{i=1}^{p} \lambda_{ii} + \sum_{(i,j)\in E} \lambda_{ij} n s_{ij} \right),$$

where $\mathbf{\Lambda}$ satisfies the condition $\lambda_{ij} = 0$, $(i, j) \notin E$. In this form the *canonical parameters* are $\lambda_{11}, \dots, \lambda_{pp}$ and λ_{ij}, $(i, j) \in E$. The *canonical variables* are $s_{11}, \dots, s_{pp}$ and s_{ij}, $(i, j) \in E$; these form a sufficient set of statistics. To maximize the likelihood function we differentiate (3) with respect to λ_{ii}, $i = 1, \dots, p$, and λ_{ij}, $(i, j) \in E$, to obtain the equations (4) and (5).

Theorem 15.5.1. *The maximum likelihood estimator of* $\mathbf{\Sigma}$ *in the model* (3) *is given by*

$$(4) \qquad \hat{\sigma}_{ij} = s_{ij}, \qquad i = j \text{ or } (i, j) \in E,$$

$$(5) \qquad \lambda_{ij} = 0, \qquad i \neq j \text{ and } (i, j) \notin E,$$

where $\mathbf{\Lambda} = \hat{\mathbf{\Sigma}}^{-1}$.

This result follows from the general theory of exponential families. See Lauritzen (1996), Theorem 5.3 and Appendix D.1.

Here we shall show that for a decomposable graph the equations (4) and (5) have a unique positive definite solution by developing an algorithm for its computation. We follow Speed and Kiiveri (1986).

Theorem 15.5.2. *Let* $\boldsymbol{L}$ *and* $\boldsymbol{M}$ *be* $p \times p$ *positive definite matrices. There exists a unique positive definite matrix* $\boldsymbol{K}$ *such that*

$$(6) \qquad k_{ij} = l_{ij}, \qquad i = j \text{ or } (i, j) \in E,$$

$$(7) \qquad k^{ij} = m^{ij}, \qquad i \neq j \text{ and } (i, j) \notin E,$$

where $(k^{ij}) = \boldsymbol{K}^{-1}$ *and* $(m^{ij}) = \boldsymbol{M}^{-1}$.

The proof of Theorem 15.5.2 depends on several lemmas. In the maximum likelihood estimation $\boldsymbol{L} = \boldsymbol{S}$, $\boldsymbol{M} = \boldsymbol{I}$ or any other diagonal matrix, and $\boldsymbol{K} = \hat{\mathbf{\Sigma}}$.

To develop this subject we use the Kullback information. For a pair of multivariate normal distributions $N(\mathbf{0}, \mathbf{P})$ and $N(\mathbf{0}, \mathbf{R})$ define

$$(8) \qquad I(\mathbf{P}|\mathbf{R}) = \mathscr{E}_P \log \frac{n(x|\mathbf{0},\mathbf{P})}{n(x|\mathbf{0},\mathbf{R})}$$

$$= -\tfrac{1}{2}\left[\log|\mathbf{P}\mathbf{R}^{-1}| + \operatorname{tr}(\mathbf{I}-\mathbf{P}\mathbf{R}^{-1})\right].$$

Lemma 15.5.1. *Suppose $\mathbf{P}$ and $\mathbf{R}$ are positive definite. Then:*

(i) *$I(\mathbf{P}|\mathbf{R}) > 0$, $\mathbf{P} \neq \mathbf{R}$, and $I(\mathbf{P}|\mathbf{P}) = 0$.*
(ii) *If $\{\mathbf{P}_n\}$ and $\{\mathbf{R}_n\}$ are sequences of positive definite matrices such that $I(\mathbf{P}_n|\mathbf{R}_n) \to 0$, then $\mathbf{P}_n\mathbf{R}_n^{-1} \to \mathbf{I}$.*

Proof. (i) Let the roots of $|\mathbf{P} - s\mathbf{R}| = 0$ be $s_1 \leq \cdots \leq s_p$. Then

$$(9) \qquad \log|\mathbf{P}\mathbf{R}^{-1}| + \operatorname{tr}(\mathbf{I}-\mathbf{P}\mathbf{R}^{-1}) = \sum_{i=1}^{p}(\log s_i + 1 - s_i) \geq 0,$$

and (9) is 0 if and only if $s_1 = \cdots = s_p = 1$.

(ii) Let the roots of $|\mathbf{P}_n - s\mathbf{R}_n| = 0$ be $s_1(n) \leq \cdots \leq s_p(n)$. Then $I(\mathbf{P}_n|\mathbf{R}_n) \to 0$ implies $[s_1(n), \dots, s_p(n)] \to (1, \dots, 1)$, which implies that $\mathbf{P}_n\mathbf{R}_n^{-1} \to \mathbf{I}$. ∎

Lemma 15.5.2. *Let*

$$(10) \qquad \mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}.$$

Then

(i) *The matrix*

$$(11) \qquad \mathbf{Q} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{11}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{P}_{11} & \mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} + \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{P}_{11}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \end{bmatrix}$$

satisfies $\mathbf{Q}_{11} = \mathbf{P}_{11}$, $\mathbf{Q}^{12} = \mathbf{R}^{12}$, and $\mathbf{Q}^{22} = \mathbf{R}^{22}$, where

$$(12) \qquad \mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{P}_{11}^{-1} + \mathbf{R}^{12}(\mathbf{R}^{22})^{-1}\mathbf{R}^{21} & \mathbf{R}^{12} \\ \mathbf{R}^{21} & \mathbf{R}^{22} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{11}^{-1} - \mathbf{R}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \mathbf{R}^{-1}.$$

(ii) *$I(\mathbf{P}|\mathbf{R}) = I(\mathbf{P}|\mathbf{Q}) + I(\mathbf{Q}|\mathbf{R})$.*

Proof. (i) Let

$$Q^{-1} = \begin{bmatrix} S & R^{21} \\ R^{21} & R^{22} \end{bmatrix}. \tag{13}$$

Then $I = Q^{-1}Q$ can be solved for $S = P_{11}^{-1} + R^{12}(R^{22})^{-1}R^{21}$; $Q = (Q^{-1})^{-1}$ follows from Theorem A.3.3. Then (ii) follows from

$$PQ^{-1} = PR^{-1} + \begin{pmatrix} I - P_{11}R_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \qquad QR^{-1} = \begin{pmatrix} P_{11}R_{11}^{-1} & 0 \\ R_{21}R_{11}^{-1}P_{11} & I \end{pmatrix}, \tag{14}$$

and $|PQ^{-1}| \cdot |QR^{-1}| = |PR^{-1}|$. From (13) and (14)

$$\operatorname{tr} PQ^{-1} + \operatorname{tr} QR^{-1} = \operatorname{tr} PR^{-1} + \operatorname{tr} I_p. \qquad \blacksquare \tag{15}$$

Lemma 15.5.2 provides the solution to the problem of finding a matrix Q, given positive definite matrices P and R, such that

$$q_{ij} = p_{ij}, \qquad (i,j) \in \{1,\ldots,t\}, \tag{16}$$

$$q^{ij} = r^{ij}, \qquad (i,j) \notin \{1,\ldots,t\}. \tag{17}$$

We now develop an iterative method to find K to satisfy (6) and (7), thus proving Theorem 15.5.2. Suppose $E = c_1 \cup \cdots \cup c_m$, where $c_1,\ldots,c_m$ are the m cliques of a decomposable graph $G = (V,E)$. Let $K_0^{-1} = M^{-1}$. Define recursively $K_n = (k_{ij}(n))$ such that

$$k_{ij}(n) = l_{ij}, \qquad i,j \in c_{n \bmod m}, \tag{18}$$

$$k^{ij}(n) = k^{ij}(n-1), \qquad i,j \notin c_{n \bmod m}. \tag{19}$$

By Lemma 15.5.2, K_n is uniquely determined. (The algorithm cycles through the cliques.) By construction

$$I(L|K_{n-1}) = I(L|K_n) + I(K_n|K_{n-1}). \tag{20}$$

Summation of (20) from 1 to q gives

$$I(L|K_0) = I(L|K_q) + \sum_{j=1}^{q} I(K_j|K_{j-1}). \tag{21}$$

Since $I(L|K_q| \geq 0$, $\sum_{j=1}^{q} I(K_j|K_{j-1})$ is bounded and $I(K_j|K_{j-1}) \to 0$ as $n \to \infty$. The set $\{K^{-1}|I(L|K) \leq I(L|K_0)\}$ is strictly convex.

Consider the vector sequence $(K_{rm+1},\ldots,K_{rm+m})$ with index r $(n = rm)$. It has a convergent subsequence $\{r(i)\}$; that is, $(K_{mr(i)+1},\ldots,K_{mr(i)+m})$ converges to $(K_1^*,\ldots,K_m^*)$, say. Since $I(K_j|K_{j-1}) \to 0$, $K_jK_{j-1}^{-1} \to I$. Then the

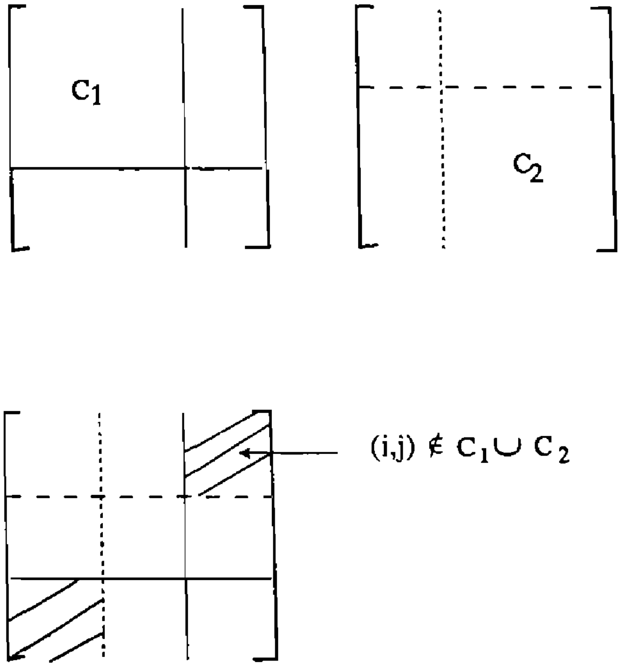

Figure 15.8. Diagram of c_1 and c_2 for $c_1 \cup c_2 = E$.

matrix $K_{mr(i)+j}K^{-1}_{mr(i)+j-1} \to I$, $j = 2, \ldots, m$, which implies $K_1^* = \cdots = K_m^* = K$, say. Note that $\{i,j|i,j \notin E\}$ satisfies $i,j \notin c_i$, $i = 1, \ldots, m$. Hence K_n satisfies (7), $n = 0, 1, \ldots,$ and K does too. Further, $k_{ij}(mr(i)+t)$ satisfies (18) $i,j \in c_t$ and K does, too. Figure 15.8 diagrams the sets for $c_1 = (i,j)$, $i,j = 1, \ldots, t$, and $c_2 = (i,j)$, $i,j = u, u+1, \ldots, p$, $u < t$.

The procedure allows for construction of a multivariate normal distribution with arbitrary marginal distributions over the cliques $c_1, \ldots, c_m$, provided that the specified marginal distributions are consistent.

Theorem 15.5.2 provides a proof of the existence and uniqueness of the maximum likelihood estimators.

The equation (12) is an updating equation. When $Q^{-1} = K_n^{-1}$ and $R^{-1} = K_{n-1}^{-1}$, the entries in K_{n-1}^{-1} not in $c_{n \bmod m}$ remain unchanged.

Dempster (1972) also proposes some iterative methods for finding K satisfying $k_{ij} = l_{ij}$, $(i,j) \in D$, and $k^{ij} = m^{ij}$, $(i,j) \notin D$. The entropy of $n(x|\mathbf{0}, P)$ is

$$(22) \qquad \mathcal{E}_p \log n(x|\mathbf{0}, P) = -\tfrac{1}{2}(p \log 2\pi - p - \log|P|).$$

Note that $|\Sigma| = \prod_{i=1}^{p} \sigma_{ii}|R|$, where $R = (\rho_{ij})$. Given that $\hat{\sigma}_{ii} = s_{ii}$, the selection of $\hat{\rho}_{ij}$ to maximize the entropy of the fitted normal distribution satisfying the requirements also minimizes $|R|$ [Demspter (1972)].

15.5.3. Decomposition of Covariance Selection Models

An undirected graph is decomposable if the graph is formed by three disjoint sets A, B, C, where $V = A \cup B \cup C$, A and B are nonempty, C separates A and B, and C is complete. Then if X_V is globally Markov with respect to G, we have $X_A \perp\!\!\!\perp X_B | X_C$,

$$(23) \qquad \Sigma^{-1} = \Lambda = \begin{bmatrix} \Lambda_{AA} & 0 & \Lambda_{AC} \\ 0 & \Lambda_{BB} & \Lambda_{BC} \\ \Lambda_{CA} & \Lambda_{CB} & \Lambda_{CC} \end{bmatrix},$$

$$(24) \qquad \Sigma_{(AB)\cdot C} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} - \begin{bmatrix} \Sigma_{AC} \\ \Sigma_{BC} \end{bmatrix} \Sigma_{CC}^{-1} (\Sigma_{CA}, \Sigma_{CB}),$$

and

$$(25) \qquad \Sigma_{AB} - \Sigma_{AC} \Sigma_{CC}^{-1} \Sigma_{CB} = 0.$$

The maximum likelihood estimator of Σ can be constructed from the maximum likelihood estimators of $\Sigma_{AA\cdot C}$, $\Sigma_{AB\cdot C}$, $\Sigma_{BB\cdot C}$, $\beta_{(AB)\cdot C}$, and Σ_{CC}.

If there is no restriction on Σ, the maximum likelihood estimator of Σ is

$$(26) \qquad \hat{\Sigma}_\Omega = \begin{bmatrix} S_{(AB)\cdot C} + S_{(AB)C} S_{CC}^{-1} S_{C(AB)} & S_{(AB)C} \\ S_{C(AB)} & S_{CC} \end{bmatrix} = S_{ABC},$$

where

$$(27) \qquad S_{(AB)\cdot C} = \begin{bmatrix} S_{AA\cdot C} & S_{AB\cdot C} \\ S_{BA\cdot C} & S_{BB\cdot C} \end{bmatrix}, \qquad S_{(AB)C} = (S_{AC}, S_{BC}).$$

If the restriction (25) is imposed, the maximum likelihood estimator is (26) with $S_{AB\cdot C}$ replaced by $\mathbf{0}$ to obtain

$$(28) \qquad \hat{\Sigma}_\omega = \begin{bmatrix} \begin{pmatrix} S_{AA\cdot C} & 0 \\ 0 & S_{BB\cdot C} \end{pmatrix} + \begin{pmatrix} S_{AC} \\ S_{BC} \end{pmatrix} S_{CC}^{-1} (S_{CA}, S_{CB}) & \begin{pmatrix} S_{AC} \\ S_{BC} \end{pmatrix} \\ (S_{CA}, S_{CB}) & S_{CC} \end{bmatrix}.$$

The matrix $S_{(AB)C}$ has the Wishart distribution $W[\Sigma_{(AB)\cdot C}, n - (p_A + p_B)]$, where p_A and p_B are the number of components of X_A and X_B, respectively

(Section 8.3). The matrix $B_{(AB)\cdot C} = S_{(AB)C} S_{CC}^{-1}$ conditional on $(X_{C1}, \dots, X_{Cn}) = X_C$ has a normal distribution, the covariance of which is given by

$$\mathscr{E} \operatorname{vec}\begin{bmatrix} B_{A\cdot C} \\ B_{B\cdot C} \end{bmatrix}\left(\operatorname{vec}\begin{bmatrix} B_{A\cdot C} \\ B_{B\cdot C} \end{bmatrix}\right)' \Bigg| S_{CC} = S_{CC}^{-1} \otimes \begin{bmatrix} \Sigma_{AA} & 0 \\ 0 & \Sigma_{BB} \end{bmatrix}, \tag{29}$$

and S_{CC} has the Wishart distribution $W(\Sigma_{CC}, n)$. The matrix $S_{(AB)\cdot C}$ and the matrix $B_{(AB)\cdot C}$, are independent (Chapter 8).

Consider testing the null hypothesis (25). This is testing the null hypothesis $\Sigma_{AB\cdot C} = 0$ against the alternative $\Sigma_{AB\cdot C} \neq 0$. The determinant of (26) is $|\hat{\Sigma}_\Omega| = |S_{AB\cdot C}| \cdot |S_{CC}|$; the determinant of (28) is $|\hat{\Sigma}_\omega| = |S_{AA}| \cdot |S_{BB}| \cdot |S_{CC}|$. The likelihood ratio criterion is

$$\left(\frac{|\hat{\Sigma}_\omega|}{|\hat{\Sigma}_\Omega|}\right)^{n/2} = \left(\frac{|S_{(AB)\cdot C}|}{|S_{AA\cdot C}| \cdot |S_{BB\cdot C}|}\right)^{n/2}. \tag{30}$$

Since the sample covariance matrix $S_{(AB)\cdot C}$ has the Wishart distribution $W[\Sigma_{(AB)\cdot C}, n - (p_A + p_B)]$, where p_A and p_B are the numbers of components of X_A and X_B (Section 8.2), the criterion is, in effect, $u_{p_A, p_B, n-(p_A+p_B)}$, studied in Sections 8.4 and 8.5.

As another example, consider the graph in Figure 15.9. Note that node 4 separates (1, 2, 3) and (5, 6); nodes 1, 4 separate 2 and 3; and node 4 separates 5 and 6. These separations imply three conditional independences: $(X_1, X_2, X_3) \perp\!\!\!\perp (X_5, X_6) | X_4$, $X_2 \perp\!\!\!\perp X_3 | (X_1, X_4)$, and $X_5 \perp\!\!\!\perp X_6 | X_4$. In terms of covariances these conditional independences are

$$\Sigma_{(123)(56)\cdot 4} = \Sigma_{(123)(56)} - \Sigma_{(123)4} \Sigma_{44}^{-1} \Sigma_{4(56)} = 0, \tag{31}$$

$$\Sigma_{23\cdot(14)} = \Sigma_{23} - \Sigma_{2(14)} \Sigma_{(14)(14)}^{-1} \Sigma_{(14)3} = 0, \tag{32}$$

$$\Sigma_{56\cdot 4} = \Sigma_{56} - \Sigma_{54} \Sigma_{44}^{-1} \Sigma_{46} = 0. \tag{33}$$

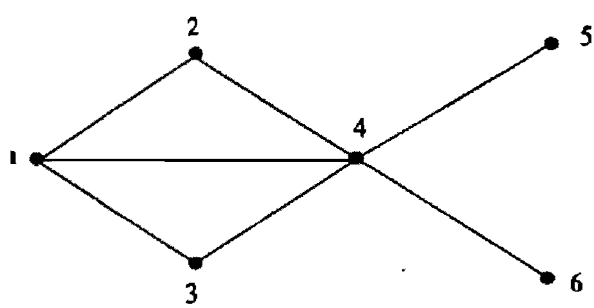

Figure 15.9

In view of (31) the restriction (32) can be written as

$$(34) \qquad \boldsymbol{\Sigma}_{23\cdot 14} = \boldsymbol{\Sigma}_{23\cdot 4} - \boldsymbol{\Sigma}_{21\cdot 4}\boldsymbol{\Sigma}_{11\cdot 4}^{-1}\boldsymbol{\Sigma}_{12\cdot 4} = \mathbf{0}.$$

It will be convenient to reorder the subvectors as $X_2, X_3, X_1, X_5, X_6, X_4$ to write

$$(35) \quad S = \begin{bmatrix} S_{22} & \cdots & S_{26} & S_{24} \\ \vdots & & \vdots & \vdots \\ S_{62} & \cdots & S_{66} & S_{64} \\ S_{42} & \cdots & S_{46} & S_{44} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{bmatrix} S_{22\cdot 4} & \cdots & S_{26\cdot 4} \\ \vdots & & \vdots \\ S_{22\cdot 4} & \cdots & S_{66\cdot 4} \end{bmatrix} + \begin{bmatrix} S_{24} \\ \vdots \\ S_{64} \end{bmatrix} S_{44}^{-1} \begin{bmatrix} S_{42} & \cdots & S_{46} \end{bmatrix} & \begin{bmatrix} S_{24} \\ \vdots \\ S_{64} \end{bmatrix} \\ \begin{bmatrix} S_{42} & \cdots & S_{46} \end{bmatrix} & S_{44} \end{bmatrix}$$

$$= \begin{bmatrix} S_{(2\ldots 6)(2\ldots 6)\cdot 4} + S_{(2\ldots 6)4}S_{44}^{-1}S_{4\cdot(2\ldots 6)} & S_{(2\ldots 6)4} \\ S_{4(2\ldots 6)} & S_{44} \end{bmatrix}.$$

The determinant of S is

$$(36) \qquad |S| = |S_{(2\ldots 6)(2\ldots 6)4}| \cdot |S_{44}|.$$

If the condition $(X_1, X_2, X_3) \perp\!\!\!\perp (X_5, X_6) | X_4$ is imposed, the maximum likelihood estimator is (35) with $S_{(125)(56)\cdot 4}$ replaced by $\mathbf{0}$ to obtain

$$(37) \qquad \begin{bmatrix} \begin{pmatrix} S_{(231)(231)\cdot 4} & \mathbf{0} \\ \mathbf{0} & S_{(56)(56)\cdot 4} \end{pmatrix} + S_{(2\ldots 6)4}S_{44}^{-1}S_{4(2\ldots 6)} & S_{(2\ldots 6)4} \\ S_{4(2\ldots 6)} & S_{44} \end{bmatrix}.$$

The determinant of (37) is

$$(38) \qquad |S_{(231)(231)\cdot 4}| \cdot |S_{(56)(56)\cdot 4}| \cdot |S_{44}|.$$

The likelihood ratio criterion for $\boldsymbol{\Sigma}_{(123)(56)\cdot 4} = \mathbf{0}$ is

$$(39) \qquad \left(\frac{|S_{(2\ldots 6)(2\ldots 6)\cdot 4}|}{|S_{(231)(231)\cdot 4}| \cdot |S_{(56)(56)\cdot 4}|} \right)^{n/2} = U_{(231)(56)\cdot 4}^{n/2}.$$

Here $U_{(231)(56)\cdot 4}$ has the distribution of $U_{p_1+p_2+p_3, p_5+p_6, n-p}$ (Section 8.4) since the distribution of $S_{(2\ldots 6)(2\ldots 6)\cdot 4}$ is $W(\boldsymbol{\Sigma}_{(2\ldots 6)(2\ldots 6)\cdot 4}, n-p_4)$, independent of S_{44}.

The first three rows and columns of $S_{(2\ldots 6)(2\ldots 6)\cdot 4}$ constitute the matrix

(40)

$$S_{(231)(231)\cdot 4} = \begin{bmatrix} S_{22\cdot 4} & S_{23\cdot 4} & S_{21\cdot 4} \\ S_{32\cdot 4} & S_{33\cdot 4} & S_{31\cdot 4} \\ S_{12\cdot 4} & S_{13\cdot 4} & S_{11\cdot 4} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{pmatrix} S_{22\cdot 14} & S_{23\cdot 14} \\ S_{32\cdot 14} & S_{33\cdot 14} \end{pmatrix} + \begin{pmatrix} S_{21\cdot 4} \\ S_{31\cdot 4} \end{pmatrix} S_{11\cdot 4}^{-1} (S_{12\cdot 4}, S_{13\cdot 4}) & \begin{pmatrix} S_{21\cdot 4} \\ S_{31\cdot 4} \end{pmatrix} \\ (S_{12\cdot 4}, S_{13\cdot 4}) & S_{11\cdot 4} \end{bmatrix},$$

$$= \begin{bmatrix} \begin{pmatrix} S_{22\cdot 14} & S_{23\cdot 14} \\ S_{32\cdot 14} & S_{33\cdot 14} \end{pmatrix} + \begin{pmatrix} S_{2(14)} \\ S_{3(14)} \end{pmatrix} S_{(14)(14)}^{-1} (S_{(14)2}, S_{(14)3}) & \begin{pmatrix} S_{21\cdot 4} \\ S_{31\cdot 4} \end{pmatrix} \\ (S_{12\cdot 4}, S_{13\cdot 4}) & S_{11\cdot 4} \end{bmatrix},$$

The determinant of (40) is

$$(41) \qquad |S_{(231)(231)\cdot 4}| = |S_{(23)(23)\cdot 14}| \cdot |S_{11\cdot 4}|.$$

The estimator of $S_{(231)(231)\cdot 4}$ with $X_2 \perp\!\!\!\perp X_3 | X_1, X_4$ imposed is (40) with $S_{23\cdot 14}$ replace by $\mathbf{0}$ to obtain

$$(42) \qquad \begin{bmatrix} \begin{pmatrix} S_{22\cdot 14} & \mathbf{0} \\ \mathbf{0} & S_{33\cdot 14} \end{pmatrix} + \begin{pmatrix} S_{21\cdot 4} \\ S_{31\cdot 4} \end{pmatrix} S_{11\cdot 4}^{-1} (S_{12\cdot 4}, S_{13\cdot 4}) & \begin{pmatrix} S_{21\cdot 4} \\ S_{31\cdot 4} \end{pmatrix} \\ (S_{12\cdot 4}, S_{13\cdot 4}) & S_{11\cdot 4} \end{bmatrix},$$

the determinant of which is

$$(43) \qquad |S_{22\cdot 14}| \cdot |S_{33\cdot 14}| \cdot |S_{11\cdot 4}|.$$

The likelihood ratio criterion for $\Sigma_{23\cdot 14} = \mathbf{0}$ is

$$\left(\frac{|S_{(23)(23)\cdot 14}|}{|S_{22\cdot 14}| \cdot |S_{33\cdot 14}|} \right)^{n/2} = U_{23\cdot 14}^{n/2}. \tag{44}$$

The statistic $U_{23\cdot 14}$ has the distribution of $U_{p_2, p_3, n-(p_1+p_2+p_3+p_4)}$ (Section 8.4) since $S_{(23)(23)\cdot 14}$ has the distribution $W[\Sigma_{(23)(23)\cdot 14}, n-(p_1+p_4)]$ independent of $S_{(14)(14)}$.

The estimator of $\Sigma_{(56)(56)\cdot 4}$ with $\Sigma_{56\cdot 4} = \mathbf{0}$ imposed is $S_{(56)(56)\cdot 4}$ with $S_{56\cdot 4}$ replaced by $\mathbf{0}$ to obtain

$$\begin{bmatrix} S_{55\cdot 4} & \mathbf{0} \\ \mathbf{0} & S_{66\cdot 4} \end{bmatrix}. \tag{45}$$

The likelihood ratio criterion for testing $\Sigma_{56\cdot 4} = \mathbf{0}$ is

$$\left(\frac{|S_{(56)(56)\cdot 4}|}{|S_{55\cdot 4}| \cdot |S_{66\cdot 4}|} \right)^{n/2} = U_{56\cdot 4}^{n/2}. \tag{46}$$

The statistic $u_{56\cdot 4}$ has the distribution of $u_{p_5, p_6, n-(p_4+p_5+p_6)}$ since $S_{(56)(56)\cdot 4}$ has the distribution $W(\Sigma_{(56)(56)\cdot 4}, n-p_4)$ independent of S_{44}.

The estimator of Σ under three null hypotheses is (37) with $S_{(231)(231)\cdot 4}$ replaced by (42) and $S_{(56)(56)\cdot 4}$ replaced by (45). The determinant of this matrix is

$$|\hat{\Sigma}_\omega| = |S_{22\cdot 14}| \cdot |S_{33\cdot 14}| \cdot |S_{11\cdot 4}| \cdot |S_{55\cdot 4}| \cdot |S_{66\cdot 4}| \cdot |S_{44}|. \tag{47}$$

The likelihood ratio criterion for testing the three null hypotheses is

$$\left(\frac{|\hat{\Sigma}_\Omega|}{|\hat{\Sigma}_\omega|} \right)^{n/2} = \left(U_{(231)(56)\cdot 4} U_{23\cdot 14} U_{56\cdot 4} \right)^{n/2}. \tag{48}$$

When the null hypotheses are true, the factors $U_{(231)(56)\cdot 4}$, $U_{23\cdot 14}$, and $U_{56\cdot 4}$ are independent. Their distributions are discussed in Sections 8.4 and 8.5. In particular the moments of these factors are given and asymptotic expansions of distributions are described.

15.5.4. Directed Graphs

We suppose that the vertices are well-numbered, $1, \dots, n$; the N observations $x_{(1)}, \dots, x_{(N)}$ are made on $X = (X_1', \dots, X_p')'$. The model is (22) to (25) or (26) of Section 15.3. Let $\bar{x} = N^{-1}\sum_{\alpha=1}^N x_{(\alpha)}$ and $S = (N-1)^{-1}\sum_{\alpha=1}^N (x_{(\alpha)} - \bar{x})(x_{(\alpha)} - \bar{x})'$. The model (26) consists of $x_1 = \alpha_1 + \varepsilon_1$ and $n-1$ regressions

(23) to (25). The vector $\boldsymbol{\alpha}_1$ in $x_1 = \boldsymbol{\alpha}_1 + \varepsilon_1$ is estimated by $\bar{x}_1$. If $\mathrm{pa}(v_2)$ is vacuous, $\boldsymbol{\alpha}_2$ is estimated by $\bar{x}_2$; if $\mathrm{pa}(v_2)$ is not vacuous and $\boldsymbol{X}_{\mathrm{pa}(v_2)} = \boldsymbol{X}_1$, then $\boldsymbol{B}_2$ and $\boldsymbol{\alpha}_2$ are estimated by

$$(49)\quad \hat{\boldsymbol{B}}_2 = \sum_{\alpha=1}^{N} (x_{2(\alpha)} - \bar{x}_2)(x_{1(\alpha)} - \bar{x}_1)' \left[\sum_{\alpha=1}^{N} (x_{1(\alpha)} - \bar{x}_1)(x_{1(\alpha)} - \bar{x}_1)' \right]^{-1},$$

$$(50)\quad \hat{\boldsymbol{\alpha}}_2 = \bar{x}_2 + \hat{\boldsymbol{B}}_2 \bar{x}_1.$$

In general

$$(51)\quad \hat{\boldsymbol{B}}_j = \sum_{\alpha=1}^{N} \left[x_{j(\alpha)} - \bar{x}_j \right] \left[x_{\mathrm{pa}(v_j)(\alpha)} - \bar{x}_{\mathrm{pa}(v_j)} \right]' \cdot \left\{ \sum_{\alpha=1}^{N} \left[x_{\mathrm{pa}(v_j)(\alpha)} - \bar{x}_{\mathrm{pa}(v_j)} \right] \left[x_{\mathrm{pa}(v_j)(\alpha)} - \bar{x}_{\mathrm{pa}(v_j)} \right]' \right\}^{-1},$$

$$(52)\quad \hat{\boldsymbol{\alpha}}_j = \bar{x}_j + \hat{\boldsymbol{B}}_j \bar{x}_{\mathrm{pa}(v_j)}.$$

Conditional on $x_{\mathrm{pa}(v_j)(\alpha)}$, the distribution of these estimators is normal.

15.5.5. Chain Graphs

The condition (C1) of Section 15.4 specifies that $\boldsymbol{X}_u \perp\!\!\!\perp \boldsymbol{X}_v | \boldsymbol{X}_{\mathrm{pa}_{\mathscr{D}}\tau}$ for $u \in V(\tau)$ and for $v \in V(\sigma)$, where $\sigma \in \mathrm{nd}_{\mathscr{D}}(\tau) \setminus \mathrm{pa}_{\mathscr{D}}(\tau)$; that is, the past earlier than $\mathrm{pa}_{\mathscr{D}}(\tau)$ is independent of the present. This condition corresponds to the Markov property in time series analysis. Thus $\boldsymbol{X}_u$ is in terms of deviations from the regression of $\boldsymbol{X}_\tau$ on $\boldsymbol{X}_{\mathrm{pa}_{\mathscr{D}}(\tau)}$

$$\mathscr{E} \boldsymbol{X}_\tau | \boldsymbol{X}_{\mathrm{pa}_{\mathscr{D}}(\tau)} = \boldsymbol{\alpha}_\tau + \boldsymbol{B}_\tau \boldsymbol{X}_{\mathrm{pa}_{\mathscr{D}}(\tau)}.$$

The vector $\boldsymbol{\alpha}_\tau$ and the matrix $\boldsymbol{B}_\tau$ are estimated as for directed graphs.

The Markov property (C2) indicates the analysis in terms of deviations $\boldsymbol{X}_\tau - \boldsymbol{\alpha}_\tau - \boldsymbol{B}_\tau \boldsymbol{X}_{\mathrm{pa}_{\mathscr{D}}(\tau)}$. The estimation of the structure of dependence within $V(\tau)$ is carried out as in Section 15.5.2.

The Markov property (C3*) specifies $\boldsymbol{X}_u \perp\!\!\!\perp \boldsymbol{X}_v | \boldsymbol{X}_{\mathrm{pa}_G(U)}$ for $u \in U \subseteq V(\tau)$ and $v \in \mathrm{pa}_{\mathscr{D}}(U) \cup \mathrm{nb}_G(U)$. The property is a restriction on the regression of $\boldsymbol{X}_\tau$ on $\boldsymbol{X}_{\mathrm{pa}_{\mathscr{D}}(\tau)}$.

ACKNOWLEDGMENTS

The preparation of this chapter drew heavily on lecture notes of Michael Perlman. Thanks are also due Michael Perlman and Ingram Olkin for reading drafts of the manuscript.

APPENDIX A

Matrix Theory

A.1. DEFINITION OF A MATRIX AND OPERATIONS ON MATRICES

In this appendix we summarize some of the well-known definitions and theorems of matrix algebra. A number of results that are not always contained in books on matrix algebra are proved here.

An $m \times n$ matrix $\boldsymbol{A}$ is a rectangular array of real numbers

$$(1) \qquad \boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix},$$

which may be abbreviated (a_{ij}), $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. Capital boldface letters will be used to denote matrices whose elements are the corresponding lowercase letters with appropriate subscripts. The sum of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same numbers of rows and columns, respectively, is defined by

$$(2) \qquad \boldsymbol{A} + \boldsymbol{B} = (a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij}).$$

The product of a matrix by a real number λ is defined by

$$(3) \qquad \lambda \boldsymbol{A} = \boldsymbol{A} \lambda = (\lambda a_{ij}).$$

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

These operations have the algebraic properties

$$A + B = B + A, \tag{4}$$

$$(A + B) + C = A + (B + C). \tag{5}$$

$$A + (-1)A = (0), \tag{6}$$

$$(\lambda + \mu)A = \lambda A + \mu A, \tag{7}$$

$$\lambda(A + B) = \lambda A + \lambda B, \tag{8}$$

$$\lambda(\mu A) = (\lambda\mu)A. \tag{9}$$

The matrix (0) with all elements 0 is denoted as $\mathbf{0}$. The operation $A + (-1B)$ is denoted as $A - B$.

If A has the same number of columns as B has rows, that is, $A = (a_{ij})$, $i = 1, \dots, l$, $j = 1, \dots, m$, $B = (b_{jk})$, $j = 1, \dots, m$, $k = 1, \dots, n$, then A and B can be multiplied according to the rule

$$AB = (a_{ij})(b_{jk}) = \left(\sum_{j=1}^{m} a_{ij} b_{jk} \right), \qquad i = 1, \dots, l, \quad k = 1, \dots, n; \tag{10}$$

that is, AB is a matrix with l rows and n columns, the element in the ith row and kth column being $\sum_{j=1}^{m} a_{ij} b_{jk}$. The matrix product has the properties

$$(AB)C = A(BC), \tag{11}$$

$$A(B + C) = AB + AC, \tag{12}$$

$$(A + B)C = AC + BC. \tag{13}$$

The relationships (11)–(13) hold provided one side is meaningful (i.e., the numbers of rows and columns are such that the operations can be performed); it follows then that the other side is also meaningful. Because of (11) we can write

$$(AB)C = A(BC) = ABC. \tag{14}$$

The product BA may be meaningless even if AB is meaningful, and even when both are meaningful they are not necessarily equal.

The *transpos*· of the $l \times m$ matrix $A = (a_{ij})$ is defined to be the $m \times l$ matrix A' which has in the jth row and ith column the element that A has in the ith row and jth column. The operation of transposition has the properties

$$(A')' = A, \tag{15}$$

$$(A+B)' = A' + B', \tag{16}$$

$$(AB)' = B'A', \tag{17}$$

again with the restriction (which is understood throughout this book) that at least one side is meaningful.

A vector x with m components can be treated as a matrix with m rows and one column. Therefore, the above operations hold for vectors.

We shall now be concerned with square matrices of the same size, which can be added and multiplied at will. The number of rows and columns will be taken to be p. A is called *symmetric* if $A = A'$. A particular matrix of considerable interest is the *identity* matrix

$$I = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = (\delta_{ij}), \tag{18}$$

where δ_{ij}, the Kronecker delta, is defined by

$$\begin{aligned} \delta_{ij} &= 1, & i &= j, \\ &= 0, & i &\neq j. \end{aligned} \tag{19}$$

The identity matrix satisfies

$$IA = AI = A. \tag{20}$$

We shall write the identity as I_p when we wish to emphasize that it is of order p. Associated with any square matrix A is the determinant $|A|$, defined by

$$|A| = \sum (-1)^{f(j_1, \ldots, j_p)} \prod_{i=1}^{p} a_{ij_i}, \tag{21}$$

where the summation is taken over all permutations $(j_1, \ldots, j_p)$ of the set of integers $(1, \ldots, p)$, and $f(j_1, \ldots, j_p)$ is the number of transpositions required to change $(1, \ldots, p)$ into $(j_1, \ldots, j_p)$. A transposition consists of interchanging two numbers, and it can be shown that, although one can transform $(1, \ldots, p)$ into $(j_1, \ldots, j_p)$ by transpositions in many different ways, the number of transpositions required is always even or always odd, so that $(-1)^{f(j_1,\ldots,j_p)}$ is consistently defined. Then

(22) $$|AB| = |A| \cdot |B|.$$

Also

(23) $$|A| = |A'|.$$

A *submatrix* of A is a rectangular array obtained from A by deleting rows and columns. A *minor* is the determinant of a square submatrix of A. The minor of an element a_{ij} is the determinant of the submatrix of a square matrix A obtained by deleting the ith row and jth column. The *cofactor* of a_{ij}, say A_{ij}, is $(-1)^{i+j}$ times the minor of a_{ij}. It follows from (21) that

(24) $$|A| = \sum_{j=1}^{p} a_{ij} A_{ij} = \sum_{j=1}^{p} a_{jk} A_{jk}.$$

If $|A| \neq 0$, there exists a unique matrix B such that $AB = I$. Then B is called the *inverse* of A and is denoted by A^{-1}. Let a^{hk} be the element of A^{-1} in the hth row and kth column. Then

(25) $$a^{hk} = \frac{A_{kh}}{|A|}.$$

The operation of taking the inverse satisfies

(26) $$(AC)^{-1} = C^{-1}A^{-1},$$

since

(27) $$(AC)(C^{-1}A^{-1}) = A(CC^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I.$$

Also $I^{-1} = I$ and $A^{-1}A = I$. Furthermore, since the transposition of (27) gives $(A^{-1})'A' = I$, we have $(A^{-1})' = (A')^{-1}$.

A matrix whose determinant is not zero is called *nonsingular*. If $|A| \neq 0$, then the only solution to

(28) $$Az = \mathbf{0}$$

is the trivial one $z = \mathbf{0}$ [by multiplication of (28) on the left by A^{-1}]. If $|A| = 0$, there is at least one nontrivial solution (that is, $z \neq \mathbf{0}$). Thus an equivalent definition of A being nonsingular is that (28) have only the trivial solution.

A set of vectors $z_1, \ldots, z_r$ is said to be *linearly independent* if there exists no set of scalars $c_1, \ldots, c_r$, not all zero, such that $\sum_{i=1}^{r} c_i z_i = \mathbf{0}$. A $q \times p$

matrix D is said to be of *rank* r if the maximum number of linearly independent columns is r. Then every minor of order $r+1$ must be zero (from the remarks in the preceding paragraph applied to the relevant square matrix of order $r+1$), and at least one minor of order r must be nonzero. Conversely, if there is at least one minor of order r that is nonzero, there is at least one set of r columns (or rows) which is linearly independent. If all minors of order $r+1$ are zero, there cannot be any set of $r+1$ columns (or rows) that are linearly independent, for such linear independence would imply a nonzero minor of order $r+1$, but this contradicts the assumption. Thus rank r is equivalently defined by the maximum number of linearly independent rows, by the maximum number of linearly independent columns, or by the maximum order of nonzero minors.

We now consider the quadratic form

$$x'Ax = \sum_{i,j=1}^{p} a_{ij}x_i x_j, \tag{29}$$

where $x' = (x_1, \ldots, x_p)$ and $A = (a_{ij})$ is a symmetric matrix. This matrix A and the quadratic form are called *positive semidefinite* if $x'Ax \geq 0$ for all x. If $x'Ax > 0$ for all $x \neq \mathbf{0}$, then A and the quadratic form are called *positive definite*. In this book *positive definite* implies the matrix is symmetric.

Theorem A.1.1. *If C with p rows and columns is positive definite, and if B with p rows and q columns, $q \leq p$, is of rank q, then $B'CB$ is positive definite.*

Proof. Given a vector $y \neq \mathbf{0}$, let $x = By$. Since B is of rank q, $By = x \neq \mathbf{0}$. Then

$$\begin{aligned} y'(B'CB)y &= (By)'C(By) \\ &= x'Cx > 0. \end{aligned} \tag{30}$$

The proof is completed by observing that $B'CB$ is symmetric. As a converse, we observe that $B'CB$ is positive definite only if B is of rank q, for otherwise there exists $y \neq \mathbf{0}$ such that $By = \mathbf{0}$. ■

Corollary A.1.1. *If C is positive definite and B is nonsingular, then $B'CB$ is positive definite.*

Corollary A.1.2. *If C is positive definite, then C^{-1} is positive definite.*

Proof. C must be nonsingular; for if $Cx = \mathbf{0}$ for $x \neq \mathbf{0}$, then $x'Cx = \mathbf{0}$ for this x, but that is contrary to the assumption that C is positive definite. Let

B in Theorem A.1.1 be C^{-1}. Then $B'CB = (C^{-1})'CC^{-1} = (C^{-1})'$. Transposing $CC^{-1} = I$, we have $(C^{-1})'C' = (C^{-1})'C = I$. Thus $C^{-1} = (C^{-1})'$. ■

Corollary A.1.3. *The $q \times q$ matrix formed by deleting $p - q$ rows of a positive definite matrix C and the corresponding $p - q$ columns of C is positive definite.*

Proof. This follows from Theorem A.1.1 by forming B by taking the $p \times p$ identity matrix and deleting the columns corresponding to those deleted from C. ■

The *trace* of a square matrix A is defined as $\operatorname{tr} A = \sum_{i=1}^{p} a_{ii}$. The following properties are verified directly:

$$\operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B, \tag{31}$$

$$\operatorname{tr} AB = \operatorname{tr} BA. \tag{32}$$

A square matrix A is said to be *diagonal* if $a_{ij} = 0$, $i \neq j$. Then $|A| = \prod_{i=1}^{p} a_{ii}$, for in (24) $|A| = a_{11}A_{11}$, and in turn A_{11} is evaluated similarly.

A square matrix A is said to be *triangular* if $a_{ij} = 0$ for $i > j$ or alternatively for $i < j$. If $a_{ij} = 0$ for $i > j$, the matrix is *upper* triangular, and, if $a_{ij} = 0$ for $i < j$, it is *lower* triangular. The product of two upper triangular matrices A, B is upper triangular, for the i,jth term $(i > j)$ of AB is $\sum_{k=1}^{p} a_{ik}b_{kj} = 0$ since $a_{ik} = 0$ for $k < i$ and $b_{kj} = 0$ for $k > j$. Similarly, the product of two lower triangular matrices is lower triangular. The determinant of a triangular matrix is the product of the diagonal elements. The inverse of a nonsingular triangular matrix is triangular in the same way.

Theorem A.1.2. *If A is nonsingular, there exists a nonsingular lower triangular matrix F such that $FA = A^*$ is nonsingular upper triangular.*

Proof. Let $A = A_1$. Define recursively $A_g = (a_{ij}^{(g)}) = F_{g-1}A_{g-1}$, $g = 2, \ldots, p$, where $F_{g-1} = (f_{ij}^{(g-1)})$ has elements

$$f_{jj}^{(g-1)} = 1, \qquad j = 1, \ldots, p, \tag{33}$$

$$f_{i,g-1}^{(g-1)} = -\frac{a_{i,g-1}^{(g-1)}}{a_{g-1,g-1}^{(g-1)}}, \qquad i = g, \ldots, p, \tag{34}$$

$$f_{ij}^{(g-1)} = 0, \qquad \text{otherwise}. \tag{35}$$

Then

$$(36)\quad a_{ij}^{(g)} = 0, \qquad i = j+1,\dots,p, \quad j = 1,\dots,g-1,$$

$$(37)\quad a_{ij}^{(g)} = a_{ij}^{(g-1)}, \qquad i = 1,\dots,g-1, \quad j = 1,\dots,p,$$

$$(38)\quad a_{ij}^{(g)} = a_{ij}^{(g-1)} + f_{i,g-1}^{(g-1)} a_{g-1,j}^{(g-1)} = a_{ij}^{(g-1)} - \frac{a_{i,g-1}^{(g-1)} a_{g-1,j}^{(g-1)}}{a_{g-1,g-1}^{(g-1)}}, \qquad i, j = g,\dots,p.$$

Note that $F = F_{p-1},\dots,F_1$ is lower triangular and the elements of A_g in the first $g-1$ columns below the diagonal are 0; in particular $A^* = FA$ is upper triangular. From $|A| \neq 0$ and $|F_{g-1}| = 1$, we have $|A_{g-1}| \neq 0$. Hence $a_{11}^{(1)},\dots,a_{g-2,g-2}^{(g-2)}$ are different from 0 and the last $p-g$ columns of A_{g-1} can be numbered so $a_{g-1,g-1}^{(g-1)} \neq 0$; then $f_{i,g-1}^{(g-1)}$ is well defined. ∎

The equation $FA = A^*$ can be solved to obtain $A = LR$, where $R = A^*$ is upper triangular and $L = F^{-1}$ is lower triangular and has 1's on the main diagonal (because F is lower triangular and has 1's on the main diagonal). This is known as the *LR decomposition.*

Corollary A.1.4. *If A is positive definite, there exists a lower triangular nonsingular matrix F such that FAF' is diagonal and positive definite.*

Proof. From Theorem A.1.2, there exists a lower triangular nonsingular matrix F such that FA is upper triangular and nonsingular. Then FAF' is upper triangular and symmetric; hence it is diagonal. ∎

Corollary A.1.5. *The determinant of a positive definite matrix A is positive.*

Proof. From the construction of FAF',

$$(39)\qquad FAF' = \begin{bmatrix} a_{11}^{(1)} & 0 & 0 & \cdots & 0 \\ 0 & a_{22}^{(2)} & 0 & \cdots & 0 \\ 0 & 0 & a_{33}^{(3)} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{pp}^{(p)} \end{bmatrix}$$

is positive definite, and hence $a_{gg}^{(g)} > 0$, $g = 1,\dots,p$, and $0 < |FAF'| = |F| \cdot |A| \cdot |F| = |A|$. ∎

Corollary A.1.6. *If A is positive definite, there exists a lower triangular matrix G such that $GAG' = I$.*

Proof. Let $FAF' = D^2$, and let D be the diagonal matrix whose diagonal elements are the positive square roots of the diagonal elements of D^2. Then $C = D^{-1}F$ serves the purpose. ■

Corollary A.1.7 (Cholsky Decomposition). *If A is positive definite, there exists a unique lower triangular matrix T $(t_{ij} = 0,\ i < j)$ with positive diagonal elements such that $A = TT'$.*

Proof. From Corollary A.1.6, $A = G^{-1}(G')^{-1}$, where G is lower triangular. Then $T = G^{-1}$ is lower triangular. ■

In effect this theorem was proved in Section 7.2 for $A = VV'$.

A.2. CHARACTERISTIC ROOTS AND VECTORS

The *characteristic roots* of a square matrix B are defined as the roots of the *characteristic equation*

$$(1) \qquad |B - \lambda I| = 0.$$

Alternative terms are *latent roots* and *eigenvalues*. For example, with

$$B = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix},$$

we have

$$(2) \qquad |B - \lambda I| = \begin{vmatrix} 5-\lambda & 2 \\ 2 & 5-\lambda \end{vmatrix} = 25 - 4 - 10\lambda + \lambda^2 = \lambda^2 - 10\lambda + 21.$$

The degree of the polynomial equation (1) is the order of the matrix B and the constant term is $|B|$.

A matrix C is said to be *orthogonal* if $C'C = I$; it follows that $CC' = I$. Let the vectors $x' = (x_1, \dots, x_p)$ and $y' = (y_1, \dots, y_p)$ represent two points in a p-dimensional Euclidean space. The distance squared between them is $D(x, y) = (x - y)'(x - y)$. The transformation $z = Cx$ can be thought of as a change of coordinate axes in the p-dimensional space. If C is orthogonal, the transformation is distance-preserving, for

$$(3)\qquad \begin{aligned} D(Cx, Cy) &= (Cy - Cx)'(Cy - Cx) \\ &= (y - x)'C'C(y - x) = (y - x)'(y - x) = D(x, y). \end{aligned}$$

Since the angles of a triangle are determined by the lengths of its sides, the transformation $z = Cx$ also preserves angles. It consists of a rotation together with a possible reflection of one or more axes. We shall denote $\sqrt{x'x}$ by $\|x\|$.

Theorem A.2.1. *Given any symmetric matrix B, there exists an orthogonal matrix C such that*

$$(4)\qquad C'BC = D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix}.$$

If B is positive semidefinite, then $d_i \geq 0$, $i = 1, \ldots, p$; if B is positive definite, then $d_i > 0$, $i = 1, \ldots, p$.

The proof is given in the discussion of principal components in Section 11.2 for the case of B positive semidefinite and holds for B symmetric. The characteristic equation (1) under transformation by C becomes

$$(5)\qquad \begin{aligned} 0 &= |C'| \cdot |B - \lambda I| \cdot |C| = |C'(B - \lambda I)C| \\ &= |C'BC - \lambda I| = |D - \lambda I| \\ &= \begin{vmatrix} d_1 - \lambda & 0 & \cdots & 0 \\ 0 & d_2 - \lambda & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_p - \lambda \end{vmatrix} = \prod_{i=1}^{p} (d_i - \lambda). \end{aligned}$$

Thus the characteristic roots of B are the diagonal elements of the transformed matrix D.

If λ_i is a characteristic root of B, then a vector x_i not identically $\mathbf{0}$ satisfying

$$(6)\qquad (B = \lambda_i I)x_i = \mathbf{0}$$

is called a *characteristic vector* (or eigenvector) of the matrix B corresponding to the characteristic root λ_i. Any scalar multiple of x_i is also a characteristic vector. When B is symmetric, $x_i'(B - \lambda_i I) = \mathbf{0}$. If the roots are distinct, $x_j'Bx_i = 0$ and $x_j'x_i = 0$, $i \neq j$. Let $c_i = (1/\|x_i\|)x_i$ be the ith normalized characteristic vector, and let $C=(c_1,\ldots,c_p)$. Then $C'C=I$ and $BC=CD$. These lead to (4). If a characteristic root has multiplicity m, then a set of m corresponding characteristic vectors can be replaced by m linearly independent linear combinations of them. The vectors can be chosen to satisfy (6) and $x_j'x_i=0$ and $x_j'Bx_i=0$, $i\neq j$.

A characteristic vector lies in the direction of the principal axis (see Chapter 11). The characteristic roots of B are proportional to the squares of the reciprocals of the lengths of the principal axes of the ellipsoid

$$x'Bx=1 \tag{7}$$

since this becomes under the rotation $y=Cx$

$$1=y'Dy=\sum_{i=1}^{p} d_i y_i^2. \tag{8}$$

For a pair of matrices A (nonsingular) and B we shall also consider equations of the form

$$|B-\lambda A|=0. \tag{9}$$

The roots of such equations are of interest because of their invariance under certain transformations. In fac , for nonsingular C, the roots of

$$|C'BC-\lambda(C'AC)|=0 \tag{10}$$

are the same as those of (9) since

$$|C'BC-\lambda C'AC|=|C'(B-\lambda A)C|=|C'|\cdot|B-\lambda A|\cdot|C| \tag{11}$$

and $|C'|=|C|\neq 0$.

By Corollary A.1.6 we have that if A is positive definite there is a matrix E such that $E'AE=I$. Let $E'BE=B^*$. From Theorem A.2.1 we deduce that there exists an orthogonal matrix C such that $C'B^*C=D$, where D is diagonal. Defining EC as F, we have the following theorem:

Theorem A.2.2. *Given B positive semidefinite and A positive definite, there exists a nonsingular matrix F such that*

$$F'BF=\begin{pmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \lambda_p\end{pmatrix}, \tag{12}$$

$$F'AF=I, \tag{13}$$

where $\lambda_1\geq\cdots\geq\lambda_p$ (≥ 0) are the roots of (9). If B is positive definite, $\lambda_i>0$ $i=1,\ldots,p$.

Corresponding to each root λ_i there is a vector x_i satisfying

$$(14) \qquad (B - \lambda_i A)x_i = 0$$

and $x_i' A x_i = 1$. If the roots are distinct $x_j' B x_i = 0$ and $x_j' A x_i = 0$, $i \neq j$. Then $F = (x_1, \dots, x_p)$. If a root has multiplicity m, then a set of m linearly independent x_i's can be replaced by m linearly independent combinations of them. The vectors can be chosen to satisfy (14) and $x_j' B x_i = 0$ and $x_j' A x_i = 0$, $i \neq j$.

Theorem A.2.3 (The Singular Value Decomposition). *Given an $n \times p$ matrix X, $n \geq p$, there exists an $n \times n$ orthogonal matrix P, a $p \times p$ orthogonal matrix Q, and an $n \times p$ matrix D consisting of a $p \times p$ diagonal positive semidefinite matrix and an $(n-p) \times p$ zero matrix such that*

$$(15) \qquad X = PDQ.$$

Proof. From Theorem A.2.1, there exists a $p \times p$ orthogonal matrix Q and a diagonal matrix E such that

$$(16) \qquad QX'XQ' = \begin{pmatrix} E_1 & 0 \\ 0 & 0 \end{pmatrix},$$

where E_1 is diagonal and positive definite. Let $XQ' = Y = (Y_1 \ \ Y_2)$, where the number of columns of Y_1 is the order of E_1. Then $Y_2' Y_2 = 0$, and hence $Y_2 = 0$. Let $P_1 = Y_1 E_1^{-\frac{1}{2}}$. Then $P_1' P_1 = I$. An $n \times n$ orthogonal matrix $P = (P_1 \ \ P_2)$ satisfying the theorem is obtained by adjoining P_2 to make P orthogonal. Then the upper left-hand corner of D is $E_1^{\frac{1}{2}}$, and the rest of D consists of zeros. ■

Theorem A.2.4. *Let A be positive definite and B be positive semidefinite. Then*

$$(17) \qquad \lambda_p \leq \frac{x'Bx}{x'x} \leq \lambda_1,$$

where λ_1 and λ_p are the largest and smallest roots of (1), *and*

$$(18) \qquad \lambda_p \leq \frac{x'Bx}{x'Ax} \leq \lambda_1,$$

where λ_1 and λ_p are the largest and smallest roots of (9).

Proof. The inequalities (17) were essentially proved in Section 11.2, and can also be derived from (4). The inequalities (18) follow from Theorem A.2.2. ■

A square matrix A is *idempotent* if $A^2 = A$. If λ satisfies $|A - \lambda I| = 0$, there exists a vector $x \neq \mathbf{0}$ such that $\lambda x = Ax = A^2 x$. However, $A^2 x = A(Ax) = A\lambda x = \lambda^2 x$. Thus $\lambda^2 = \lambda$, and λ is either 0 or 1. The multiplicity of $\lambda = 1$ is the rank of A. If A is $p \times p$, then $I_p - A$ is idempotent of rank $p -$ (rank A), and A and $I_p - A$ are orthogonal. If A is symmetric, there is an orthogonal matrix O such that

$$(19) \qquad OAO' = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad O(I - A)O' = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}.$$

A.3. PARTITIONED VECTORS AND MATRICES

Consider the matrix A defined by (1) of Section A.1. Let

$$(1) \qquad \begin{aligned} A_{11} &= (a_{ij}), & i &= 1, \dots, p, & i &= 1, \dots, q, \\ A_{12} &= (a_{ij}), & i &= 1, \dots, p, & j &= q+1, \dots, n, \\ A_{21} &= (a_{ij}), & i &= p+1, \dots, m, & j &= 1, \dots, q, \\ A_{22} &= (a_{ij}), & i &= p+1, \dots, m, & j &= q+1, \dots, n. \end{aligned}$$

Then we can write

$$(2) \qquad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

We say that A has been *partitioned* into submatrices A_{ij}. Let B $(m \times n)$ be partitioned similarly into submatrices B_{ij}, $i, j = 1, 2$. Then

$$(3) \qquad A + B = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \end{pmatrix}.$$

Now partition C $(n \times r)$ as

$$(4) \qquad C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

where C_{11} and C_{12} have q rows and C_{11} and C_{21} have s columns. Then

$$(5) \qquad AC = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} A_{11}C_{11} + A_{12}C_{21} & A_{11}C_{12} + A_{12}C_{22} \\ A_{21}C_{11} + A_{22}C_{21} & A_{21}C_{12} + A_{22}C_{22} \end{pmatrix}.$$

To verify this, consider an element in the first p rows and first s columns of AC. The i, jth element is

$$\sum_{k=1}^{n} a_{ik}c_{kj}, \qquad i \le p, \quad j \le s. \tag{6}$$

This sum can be written

$$\sum_{k=1}^{q} a_{ik}c_{kj} + \sum_{k=q+1}^{n} a_{ik}c_{kj}. \tag{7}$$

The first sum is the i, jth element of $A_{11}C_{11}$, the second sum is the i, jth element of $A_{12}C_{21}$, and therefore the entire sum (6) is the i, jth element of $A_{11}C_{11} + A_{12}C_{21}$. In a similar fashion we can verify that the other submatrices of AC can be written as in (5).

We note in passing that if A is partitioned as in (2), then the transpose of A can be written

$$A' = \begin{pmatrix} A'_{11} & A'_{21} \\ A'_{12} & A'_{22} \end{pmatrix}. \tag{8}$$

If $A_{12} = \mathbf{0}$ and $A_{21} = \mathbf{0}$, then for A positive definite and A_{11} square,

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & A_{22}^{-1} \end{pmatrix}. \tag{9}$$

The matrix on the right exists because A_{11} and A_{22} are nonsingular. That the right-hand matrix is the inverse of A is verified by multiplication:

$$\begin{pmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & A_{22}^{-1} \end{pmatrix} = \begin{pmatrix} I & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix}, \tag{10}$$

which is a partitioned form of I_p.

We also note that

$$\begin{vmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{vmatrix} = \begin{vmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & I \end{vmatrix} \cdot \begin{vmatrix} I & \mathbf{0} \\ \mathbf{0} & A_{22} \end{vmatrix} = |A_{11}| \cdot |A_{22}|. \tag{11}$$

The evaluation of the first determinant in the middle is made by expanding according to minors of the last row; the only nonzero element in the sum is the last, which is 1 times a determinant of the same form with I of order one less. The procedure is repeated until $|A_{11}|$ is the minor. Similarly,

$$(12) \qquad \begin{vmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{vmatrix} = \begin{vmatrix} I & 0 \\ 0 & A_{22} \end{vmatrix} \cdot \begin{vmatrix} A_{11} & A_{12} \\ 0 & I \end{vmatrix} = |A_{11}| \cdot |A_{22}|.$$

A useful fact is that if A_1 of q rows and p columns is of rank q, there exists a matrix A_2 of $p-q$ rows and p columns such that

$$(13) \qquad A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

is nonsingular. This statement is verified by numbering the columns of A so that A_{11} consisting of the first q columns of A_1 is nonsingular (at least one $q \times q$ minor of A_1 is different from zero) and then taking A_2 as $(0 \ \ I)$; then

$$(14) \qquad |A| = \begin{vmatrix} A_{11} & A_{12} \\ 0 & I \end{vmatrix} = |A_{11}|,$$

which is not equal to zero.

Theorem A.3.1. *Let the square matrix A be partitioned as in (2) so that A_{22} is square. If A_{22} is nonsingular, let*

$$(15) \qquad B = \begin{pmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{pmatrix}, \qquad C = \begin{pmatrix} I & 0 \\ -A_{22}^{-1}A_{21} & I \end{pmatrix}.$$

Then

$$(16) \qquad BA = \begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \qquad AC = \begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

$$(17) \qquad BAC = \begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ 0 & A_{22} \end{pmatrix}.$$

If A is symmetric, $C = B'$.

Theorem A.3.2. *Let the square matrix A be partitioned as in (2) so that A_{22} is square. If A_{22} is nonsingular,*

$$(18) \qquad |A| = |A_{11} - A_{12}A_{22}^{-1}A_{21}| \cdot |A_{22}|.$$

Proof. Equation (18) follows from (16) because $|B| = 1$. ■

Corollary A.3.1. *For C nonsingular*

$$\begin{vmatrix} C & y \\ y' & 1 \end{vmatrix} = |C - yy'| = \begin{vmatrix} 1 & y' \\ y & C \end{vmatrix} = |C|\left(1 - y'C^{-1}y\right). \tag{19}$$

Theorem A.3.3. *Let the nonsingular matrix A be partitioned as in* (2) *so that A_{22} is square. If A_{22} is nonsingular, let $A_{11\cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$. Then*

$$A^{-1} = \begin{pmatrix} A_{11\cdot 2}^{-1} & -A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1} & A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1} + A_{22}^{-1} \end{pmatrix}. \tag{20}$$

Proof. From Theorem A.3.1,

$$A = B^{-1}\begin{pmatrix} A_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{pmatrix} C^{-1}. \tag{21}$$

Hence

$$\begin{aligned} A^{-1} &= C\begin{pmatrix} A_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{pmatrix}^{-1} B \\ &= \begin{pmatrix} I & \mathbf{0} \\ -A_{22}^{-1}A_{21} & I \end{pmatrix}\begin{pmatrix} A_{11\cdot 2}^{-1} & \mathbf{0} \\ \mathbf{0} & A_{22}^{-1} \end{pmatrix}\begin{pmatrix} I & -A_{12}A_{22} \\ \mathbf{0} & I \end{pmatrix}. \end{aligned} \tag{22}$$

Multiplication gives the desired result. ■

Corollary A.3.2. *If $x' = (x^{(1)\prime}\ \ x^{(2)\prime})$, then*

$$x'A^{-1}x = \left(x^{(1)} - A_{12}A_{22}^{-1}x^{(2)}\right)'A_{11\cdot 2}^{-1}\left(x^{(1)} - A_{12}A_{22}^{-1}x^{(2)}\right) + x^{(2)\prime}A_{22}^{-1}x^{(2)}. \tag{23}$$

Proof. From the theorem

$$\begin{aligned} x'A^{-1}x &= x^{(1)\prime}A_{11\cdot 2}^{-1}x^{(1)} - x^{(1)\prime}A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1}x^{(2)} \\ &\quad - x^{(2)\prime}A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1}x^{(1)} + x^{(2)\prime}\left(A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1} + A_{22}^{-1}\right)x^{(2)}, \end{aligned} \tag{24}$$

which is equal to the right-hand side of (23). ■

Theorem A.3.4. *Let the nonsingular matrix A be partitioned as in* (2) *so that A_{22} is square. If A_{22} is nonsingular,*

$$(25)\quad \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} = A_{22}^{-1}A_{21}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}A_{12}A_{22}^{-1} + A_{22}^{-1}.$$

Proof. The lower right-hand corner of A^{-1} is the right-hand side of (25) by Theorem A.3.3 and is also the left-hand side of (25) by interchange of 1 and 2. ■

Theorem A.3.5. *Let U be $p \times m$. The conditions for $I_p - UU'$, $I_m - U'U$, and*

$$(26)\quad \begin{pmatrix} I_p & U \\ U' & I_m \end{pmatrix}$$

to be positive definite are the same.

Proof. We have

$$(27)\quad \begin{pmatrix} v' & w' \end{pmatrix} \begin{pmatrix} I_p & U \\ U' & I_m \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} = v'v + v'Uw + w'U'v + w'w$$
$$= v'(I_m - UU')v + (U'v + w)'(U'v + w).$$

The second term on the right-hand side is nonnegative; the first term is positive for all $v \neq 0$ if and only if $I_m - U'U$ is positive definite. Reversing the roles of v and w shows that (26) is positive definite if and only if $I_p - UU'$ is positive definite. ■

A.4. SOME MISCELLANEOUS RESULTS

Theorem A.4.1. *Let C be $p \times p$, positive semidefinite, and of rank r $(\leq p)$. Then there is a nonsingular matrix A such that*

$$(1)\quad ACA' = \begin{pmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Proof. Since C is of rank r, there is a $(p - r) \times p$ matrix A_2 such that

$$(2)\quad A_2 C = \mathbf{0}.$$

Choose B $(r \times p)$ such that

$$(3)\quad \begin{pmatrix} B \\ A_2 \end{pmatrix}$$

is nonsingular. Then

$$\begin{pmatrix} B \\ A_2 \end{pmatrix} C (B' \quad A_2') = \begin{pmatrix} BC \\ 0 \end{pmatrix} (B' \quad A_2') = \begin{pmatrix} BCB' & 0 \\ 0 & 0 \end{pmatrix}. \tag{4}$$

This matrix is of rank r, and therefore BCB' is nonsingular. By Corollary A.1.6 there is a nonsingular matrix D such that $D(BCB')D' = I_r$. Then

$$A = \begin{pmatrix} DB \\ A_2 \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} B \\ A_2 \end{pmatrix} \tag{5}$$

is a nonsingular matrix such that (1) holds. ■

Lemma A.4.1. *If E is $p \times p$, symmetric, and nonsingular, there is a nonsingular matrix F such that*

$$FEF' = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \tag{6}$$

where the order of I is the number of positive characteristic roots of E and the order of $-I$ is the number of negative characteristic roots of E.

Proof. From Theorem A.2.1 we know there is an orthogonal matrix G such that

$$GEG' = \begin{pmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & h_p \end{pmatrix}, \tag{7}$$

where $h_1 \geq \cdots \geq h_q > 0 > h_{q+1} \geq \cdots \geq h_p$ are the characteristic roots of E. Let

$$K = \begin{pmatrix} 1/\sqrt{h_1} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1/\sqrt{h_q} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/\sqrt{-h_{q+1}} & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1/\sqrt{-h_p} \end{pmatrix}. \tag{8}$$

Then

$$KGEG'K' = (KG)E(KG)' = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}. \qquad \blacksquare \tag{9}$$

Corollary A.4.1. *Let C be $p \times p$, symmetric, and of rank r $(\leq p)$. Then there is a nonsingular matrix A such that*

$$ACA' = \begin{pmatrix} I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{10}$$

where the order of I is the number of positive characteristic roots of C and the order of $-I$ is the number of negative characteristic roots, the sum of the orders being r.

Proof. The proof is the same as that of Theorem A.4.1 except that Lemma A.4.1 is used instead of Corollary A.1.6. ■

Lemma A.4.2. *Let A be $n \times m$ $(n > m)$ such that*

$$A'A = I_m. \tag{11}$$

There exists an $n \times (n-m)$ matrix B such that $(A \;\; B)$ is orthogonal.

Proof. Since A is of rank m, there exists an $n \times (n-m)$ matrix C such that $(A \;\; C)$ is nonsingular. Take D as $C - AA'C$; then $D'A = 0$. Let E $[(n-m) \times (n-m)]$ be such that $E'D'DE = I$. Then B can be taken as DE. ■

Lemma A.4.3. *Let x be a vector of n components. Then there exists an orthogonal matrix O such that*

$$Ox = \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{12}$$

where $c = \sqrt{x'x}$.

Proof. Let the first row of O be $(1/c)x'$. The other rows may be chosen in any way to make the matrix orthogonal. ■

Lemma A.4.4. *Let $B = (b_{ij})$ be a $p \times p$ matrix. Then*

$$\frac{\partial |B|}{\partial b_{ij}} = B_{ij}, \qquad i, j = 1, \ldots, p. \tag{13}$$

Proof. The expansion of $|\boldsymbol{B}|$ by elements of the ith row is

$$(14)\qquad |\boldsymbol{B}| = \sum_{h=1}^{p} b_{ih}B_{ih}.$$

Since B_{ih} does not contain b_{ij}, the lemma follows. ∎

Lemma A.4.5. *Let $b_{ij} = \beta_{ij}(c_1, \dots, c_n)$ be the i, jth element of a $p \times p$ matrix $\boldsymbol{B}$. Then for $g = 1, \dots, n$,*

$$(15)\qquad \frac{\partial|\boldsymbol{B}|}{\partial c_g} = \sum_{i,h=1}^{p} \frac{\partial|\boldsymbol{B}|}{\partial b_{ih}} \cdot \frac{\partial \beta_{ih}(c_1, \dots, c_n)}{\partial c_g} = \sum_{i,h=1}^{p} B_{ih} \frac{\partial \beta_{ih}(c_1, \dots, c_n)}{\partial c_g}.$$

Theorem A.4.2. *If $\boldsymbol{A} = \boldsymbol{A}'$,*

$$(16)\qquad \frac{\partial|\boldsymbol{A}|}{\partial a_{ii}} = A_{ii},$$

$$(17)\qquad \frac{\partial|\boldsymbol{A}|}{\partial a_{ij}} = 2A_{ij}, \qquad i \neq j.$$

Proof. Equation (16) follows from the expansion of $|\boldsymbol{A}|$ according to elements of the ith row. To prove (17) let $b_{ij} = b_{ji} = a_{ij}$, $i, j = 1, \dots, p$, $i \leq j$. Then by Lemma A.4.5,

$$(18)\qquad \frac{\partial|\boldsymbol{B}|}{\partial a_{ij}} = B_{ij} + B_{ji},$$

Since $|\boldsymbol{A}| = |\boldsymbol{B}|$ and $B_{ij} = B_{ji} = A_{ij} = A_{ji}$, (17) follows. ∎

Theorem A.4.3.

$$(19)\qquad \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x},$$

where $\partial/\partial \boldsymbol{x}$ denotes taking partial derivatives with respect to each component of $\boldsymbol{x}$ and arranging the partial derivatives in a column.

Proof. Let $\boldsymbol{h}$ be a column vector of as many components as $\boldsymbol{x}$. Then

$$(20)\qquad \begin{aligned}(\boldsymbol{x}+\boldsymbol{h})'\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{h}) &= \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} + \boldsymbol{h}'\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{A}\boldsymbol{h} + \boldsymbol{h}'\boldsymbol{A}\boldsymbol{h} \\ &= \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{h}'\boldsymbol{A}\boldsymbol{x} + \boldsymbol{h}'\boldsymbol{A}\boldsymbol{h}.\end{aligned}$$

The partial derivative vector is the vector multiplying $\boldsymbol{h}'$ in the second term on the right. ∎

Definition A.4.1. *Let $\boldsymbol{A} = (a_{ij})$ be a $p \times m$ matrix and $\boldsymbol{B} = (b_{\alpha\beta})$ be a $q \times n$ matrix. The $pq \times mn$ matrix with $a_{ij}b_{\alpha\beta}$ as the element in the i, αth row and the*

j, βth column is called the Kronecker *or* direct product *of A and B and is denoted by $A \otimes B$; that is,*

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pm}B \end{pmatrix}. \tag{21}$$

Some properties are the following when the orders of matrices permit the indicated operations:

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD), \tag{22}$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}. \tag{23}$$

Theorem A.4.4. *Let the* ith *characteristic root of A ($p \times p$) be λ_i and the corresponding characteristic vector be $x_i = (x_{1i}, \ldots, x_{pi})'$, and let the* αth *root of B ($q \times q$) be ν_α and the corresponding characteristic vector be y_α, $\alpha = 1, \ldots, q$. Then the i, αth root of $A \otimes B$ is $\lambda_i \nu_\alpha$, and the corresponding characteristic vector is $x_i \otimes y_\alpha = (x_{1i}y'_\alpha, \ldots, x_{pi}y'_\alpha)'$, $i = 1, \ldots, p$, $\alpha = 1, \ldots, q$.*

Proof.

$$\begin{aligned}(A \otimes B)(x_i \otimes y_\alpha) &= \begin{pmatrix} a_{11}B & \cdots & a_{1p}B \\ \vdots & & \vdots \\ a_{p1}B & \cdots & a_{pp}B \end{pmatrix} \begin{pmatrix} x_{1i}y_\alpha \\ \vdots \\ x_{pi}y_\alpha \end{pmatrix} \\ &= \begin{pmatrix} \sum_j a_{1j}x_{ji}By_\alpha \\ \vdots \\ \sum_j a_{pj}x_{ji}By_\alpha \end{pmatrix} \\ &= \begin{pmatrix} \lambda_i x_{1i}By_\alpha \\ \vdots \\ \lambda_i x_{pi}By_\alpha \end{pmatrix} = \lambda_i \nu_\alpha \begin{pmatrix} x_{1i}y_\alpha \\ \vdots \\ x_{pi}y_\alpha \end{pmatrix}. \end{aligned} \tag{24}$$

■

Theorem A.4.5

$$|A \otimes B| = |A|^q |B|^p. \tag{25}$$

Proof. The determinant of any matrix is the product of its roots; therefore

$$|A \otimes B| = \prod_{i=1}^{p} \prod_{\alpha=1}^{q} \lambda_i \nu_\alpha = \left(\prod_{i=1}^{p} \lambda_i \right)^q \left(\prod_{\alpha=1}^{q} \nu_\alpha \right)^p. \tag{26}$$ ■

Definition A.4.2. *If the* $p \times m$ *matrix* $A = (a_1, \dots, a_m)$, *then* $\operatorname{vec} A = (a_1', \dots, a_m')'$.

Some properties of the vec operator [e.g., Magnus (1988)] are

$$\operatorname{vec} ABC = (C' \otimes A) \operatorname{vec} B, \tag{27}$$

$$\operatorname{vec} xy' = y \otimes x. \tag{28}$$

Theorem A.4.6. *The Jacobian of the transformation* $E = Y^{-1}$ *(from* E *to* Y*) is* $|Y|^{-2p}$, *where* p *is the order of* E *and* Y.

Proof. From $EY = I$, we have

$$\left(\frac{\partial}{\partial \theta} E \right) Y + E \left(\frac{\partial}{\partial \theta} Y \right) = 0, \tag{29}$$

where

$$\left(\frac{\partial}{\partial \theta} E \right) = \begin{pmatrix} \frac{\partial e_{11}}{\partial \theta} & \cdots & \frac{\partial e_{1p}}{\partial \theta} \\ \vdots & & \vdots \\ \frac{\partial e_{p1}}{\partial \theta} & \cdots & \frac{\partial e_{pp}}{\partial \theta} \end{pmatrix}. \tag{30}$$

Then

$$\left(\frac{\partial}{\partial \theta} E \right) = -E \left(\frac{\partial}{\partial \theta} Y \right) E = -Y^{-1} \left(\frac{\partial}{\partial \theta} Y \right) Y^{-1}. \tag{31}$$

If $\theta = y_{\alpha\beta}$, then

$$\left(\frac{\partial}{\partial y_{\alpha\beta}} E \right) = -E \varepsilon_{\alpha\beta} E = -\mathbf{e}_{\cdot\alpha} \mathbf{e}_{\beta\cdot}, \tag{32}$$

where $\varepsilon_{\alpha\beta}$ is a $p \times p$ matrix with all elements 0 except the element in the αth row and βth column, which is 1; and $e_{\cdot\alpha}$ is the αth column of E and $e_{\beta\cdot}$ is its βth row. Thus $\partial e_{ij}/\partial y_{\alpha\beta} = -e_{i\alpha}e_{\beta j}$. Then the Jacobian is the determinant of a $p^2 \times p^2$ matrix

$$\text{(33)} \quad \operatorname{mod}\left|\frac{\partial e_{ij}}{\partial y_{\alpha\beta}}\right| = |e_{i\alpha}e_{\beta j}| = |E \otimes E'| = |E|^p|E'|^p = |E|^{2p} = |Y|^{-2p}. \qquad \blacksquare$$

Theorem A.4.7. *Let A and B be symmetric matrices with characteristic roots $a_1 \geq a_2 \geq \cdots \geq a_p$ and $b_1 \geq b_2 \geq \cdots \geq b_p$, respectively, and let H be a $p \times p$ orthogonal matrix. Then*

$$\text{(34)} \qquad \max_H \operatorname{tr} HAH'B = \sum_{j=1}^p a_j b_j, \qquad \min_H HA'H'B = \sum_{j=1}^p a_j b_{p+1-j}.$$

Proof. Let $A = H_a D_a H_a'$ and $B = H_b D_b H_b'$, where H_a and H_b are orthogonal and D_a and D_b are diagonal with diagonal elements $a_1, \dots, a_p$ and $b_1, \dots, b_p$ respectively. Then

$$\text{(35)} \qquad \begin{aligned} \max_{H^*} \operatorname{tr} H^*AH^{*\prime}B &= \max_{H^*} \operatorname{tr} H^*H_aD_aH_a'H^{*\prime}H_bD_bH_b' \\ &= \max_{H^*} \operatorname{tr} H_b'H^*H_aD_a(H_b'H^*H_a)'D_b \\ &= \max_H \operatorname{tr} HD_aH'D_b, \end{aligned}$$

where $H = H_b'H^*H_a$. We have

$$\text{(36)} \quad \begin{aligned} \operatorname{tr} HD_aH'D_b &= \sum_{i=1}^p (HD_aH')_{ii}b_i \\ &= \sum_{i=1}^{p-1}\sum_{j=1}^{i} (HD_aH')_{jj}(b_i - b_{i+1}) + b_p\sum_{j=1}^p (HD_aH')_{jj} \\ &\leq \sum_{i=1}^{p-1}\sum_{j=1}^{i} a_j(b_i - b_{i+1}) + b_p\sum_{j=1}^p a_j \\ &= \sum_{i=1}^p a_i b_i \end{aligned}$$

by Lemma A.4.6 below. The minimum in (34) is treated as the negative of the maximum with B replaced by $-B$ [von Neumann (1937)]. $\blacksquare$

Lemma A.4.6. *Let $P=(p_{ij})$ be a doubly stochastic matrix $(p_{ij}\ge 0$, $\sum_{i=1}^{p} p_{ij}=1$, $\sum_{j=1}^{p} p_{ij}=1)$. Let $y_1 \ge y_2 \ge \cdots \ge y_p$. Then*

$$\sum_{i=1}^{k} y_i \ge \sum_{i=1}^{k}\sum_{j=1}^{n} p_{ij}y_j, \qquad k=1,\dots,p. \tag{37}$$

Proof.

$$\sum_{i=1}^{k}\sum_{j=1}^{p} p_{ij}y_j = \sum_{j=1}^{p} g_j y_j, \tag{38}$$

where $g_j=\sum_{i=1}^{k} p_{ij}$, $j=1,\dots,p$ $(0\le g_j\le 1$, $\sum_{j=1}^{p} g_j=k)$. Then

$$\begin{aligned}\sum_{j=1}^{p} g_j y_j-\sum_{j=1}^{k} y_j &= -\sum_{j=1}^{k} y_j + y_k\left(k-\sum_{j=1}^{p} g_j\right)+\sum_{j=1}^{p} g_j y_j \\ &= \sum_{j=1}^{k}(y_j-y_k)(g_j-1)+\sum_{j=k+1}^{p}(y_j-y_k)g_j \\ &\le 0. \qquad \blacksquare\end{aligned} \tag{39}$$

Corollary A.4.2. *Let A be a symmetric matrix with characteristic roots $a_1\ge a_2\ge \cdots a_p$. Then*

$$\max_{R'R=I_k} \operatorname{tr} R'AR = \sum_{i=1}^{k} a_i. \tag{40}$$

Proof. In Theorem A.4.7 let

$$B=\begin{pmatrix} I_k & 0 \\ 0 & 0\end{pmatrix}. \qquad \blacksquare \tag{41}$$

Theorem A.4.8.

$$|I+xC| = 1+x\operatorname{tr} C+O(x^2). \tag{42}$$

Proof. The determinant (42) is a polynomial in x of degree p; the coefficient of the linear term is the first derivative of the determinant evaluated at $x=0$. In Lemma A.4.5 let $n=1$, $c_1=x$, $\beta_{ih}(x)=\delta_{ih}+xc_{ih}$, where $\delta_{ii}=1$ and $\delta_{ih}=0$, $i\ne h$. Then $d\beta_{ih}(x)/dx=c_{ih}$, $B_{ii}=1$ for $x=0$, and $B_{ih}=0$ for $x=0$, $i\ne h$. Thus

$$\left.\frac{d|B(x)|}{dx}\right|_{x=0} = \sum_{i=1}^{p} c_{ii}. \qquad \blacksquare \tag{43}$$

A.5. GRAM–SCHMIDT ORTHOGONALIZATION AND THE SOLUTION OF LINEAR EQUATIONS

A.5.1. Gram–Schmidt Orthogonalization

The derivation of the Wishart density in Section 7.2 included the Gram–Schmidt orthogonalization of a set of vectors; we shall review that development here. Consider the p linearly independent n-dimensional vectors $v_1, \dots, v_p$ ($p \le n$). Define $w_1 = v_1$,

$$(1) \qquad w_i = v_i - \sum_{j=1}^{i-1} \frac{v_i' w_j}{\|w_j\|^2} w_j, \qquad i = 2, \dots, p.$$

Then $w_i \neq \mathbf{0}$, $i = 1, \dots, p$, because $v_1, \dots, v_p$ are linearly independent, and $w_i' w_j = 0$, $i \neq j$, as was proved by induction in Section 7.2. Let $u_i = (1/\|w_i\|) w_i$, $i = 1, \dots, p$. Then $u_1, \dots, u_p$ are *orthonormal*; that is, they are orthogonal and of unit length. Let $U = (u_1, \dots, u_p)$. Then $U'U = I$. Define $t_{ii} = \|w_i\|$ (> 0),

$$(2) \qquad t_{ij} = \frac{v_i' w_j}{\|w_j\|} = v_i' u_j, \qquad j = 1, \dots, i-1, \quad i = 2, \dots, p,$$

and $t_{ij} = 0$, $j = i+1, \dots, p$, $i = 1, \dots, p-1$. Then $T = (t_{ij})$ is a lower triangular matrix. We can write (1) as

$$(3) \qquad v_i = \|w_i\| u_i + \sum_{j=1}^{i-1} (v_i' u_j) u_j = \sum_{j=1}^{i} t_{ij} u_j, \qquad i = 1, \dots, p,$$

that is,

$$(4) \qquad V = (v_1, \dots, v_p) = UT'.$$

Then

$$(5) \qquad A = V'V = TU'UT' = TT'$$

as shown in Section 7.2. Note that if V is square, we have decomposed an arbitrary nonsingular matrix into the product of an orthogonal matrix and an upper triangular matrix with positive diagonal elements; this is sometimes known as the *QR decomposition*. The matrices U and T in (4) are unique.

These operations can be done in a different order. Let $V = (v_1^{(0)}, \dots, v_p^{(0)})$. For $k = 1, \dots, p-1$ define recursively

$$(6) \qquad t_{kk} = \|v_k^{(k-1)}\|, \qquad u_k = \frac{1}{\|v_k^{(k-1)}\|} v_k^{(k-1)} = \frac{1}{t_{kk}} v_k^{(k-1)},$$

$$(7) \qquad t_{jk} = v_j^{(k-1)\prime} u_k, \qquad j = k+1, \dots, p,$$

$$(8) \qquad v_j^{(k)} = v_j^{(k-1)} - t_{jk} u_k, \qquad j = k+1, \dots, p.$$

Finally $t_{pp} = \|v_p^{(p-1)}\|$ and $u_p = (1/t_{pp})v_p^{(p-1)}$. The same orthonormal vectors $u_1, \dots, u_p$ and the same triangular matrix (t_{ij}) are given by the two procedures.

The numbering of the columns of V is arbitrary. For numerical stability it is usually best at any given stage to select the largest of $\|v_j^{(k-1)}\|$ to call t_{kk}.

Instead of constructing w_i as orthogonal to $w_1, \dots, w_{i-1}$, we can equivalently construct it as orthogonal to $v_1, \dots, v_{i-1}$. Let $w_1 = v_1$, and define

(9) $$w_i = v_i + \sum_{j=1}^{i-1} f_{ij} v_j$$

such that

(10) $$\begin{aligned} 0 = v_h' w_i &= v_h' v_i + \sum_{j=1}^{i-1} f_{ij} v_h' v_j \\ &= a_{hi} + \sum_{j=1}^{i-1} a_{hj} f_{ij}, \qquad h = 1, \dots, i-1. \end{aligned}$$

Let $F = (f_{ij})$, where $f_{ii} = 1$ and $f_{ij} = 0$, $i < j$. Then

(11) $$W = (w_1, \dots, w_p) = VF'.$$

Let D_t be the diagonal matrix with $\|w_j\| = t_{jj}$ as the jth diagonal element. Then $U = WD_t^{-1} = VF'D_t^{-1}$. Comparison with $V = UT'$ shows that $F = DT^{-1}$. Since $A = TT'$, we see that $FA = DT'$ is upper triangular. Hence F is the matrix defined in Theorem A.1.2.

There are other methods of accomplishing the QR decomposition that may be computationally more efficient or more stable. A *Householder matrix* has the form $H = I_n - 2\alpha\alpha'$, where $\alpha'\alpha = 1$, and is orthogonal and symmetric. Such a matrix H_1 (i.e., a vector α) can be selected so that the first column of $H_1 V$ has 0's in all positions except the first, which is positive. The next matrix has the form

(12) $$H_2 = \begin{pmatrix} 1 & 0 \\ 0 & I_{n-1} \end{pmatrix} - 2\begin{pmatrix} 0 \\ \alpha \end{pmatrix}(0 \quad \alpha') = \begin{pmatrix} 1 & 0 \\ 0 & I_{n-1} - \alpha\alpha' \end{pmatrix}.$$

The $(n-1)$-component vector α is chosen so that the second column of $H_1 V$ has all 0's except the first two components, the second being positive. This process is continued until

(13) $$H_{p-1} \cdots H_2 H_1 V = \begin{bmatrix} T' \\ 0 \end{bmatrix},$$

where T' is upper triangular and $\mathbf{0}$ is $(n-p)\times p$. Let

$$H' = H_1 \cdots H_{p-1} = (H^{(1)} \quad H^{(2)}), \tag{14}$$

where $H^{(1)}$ has p columns. Then from (13) we obtain $V = H^{(1)}T'$. Since the decomposition is unique, $H^{(1)} = U$.

Another procedure uses *Givens matrices*. A Givens matrix G_{ij} is I except for the elements $g_{ii} = \cos\theta = g_{jj}$ and $g_{ij} = \sin\theta = -g_{ji}$, $i \neq j$. It is orthogonal. Multiplication of V on the left by such a matrix leaves all rows unchanged except the ith and jth; θ can be chosen so that the i, jth element of $G_{ij}V$ is 0. Givens matrices $G_{21}, \dots, G_{n1}$ can be chosen in turn so $G_{n1} \cdots G_{21}V$ has all 0's in the first column except the first element, which is positive. Next $G_{32}, \dots, G_{n2}$ can be selected in turn so that when they are applied the resulting matrix has 0's in the second column except for the first two elements. Let

$$G' = G'_{21} \cdots G'_{n1}G'_{32} \cdots G'_{n,p-1} = (G^{(1)} \quad G^{(2)}). \tag{15}$$

Then we obtain

$$V = G'\begin{bmatrix} T' \\ \mathbf{0} \end{bmatrix} = G^{(1)}T', \tag{16}$$

and $G^{(1)} = U$.

A.5.2. Solution of Linear Equations

In the computation of regression coefficients and other statistics, we need to solve linear equations

$$Ax = y, \tag{17}$$

where A is $p \times p$ and positive definite. One method of solution is Gaussian elimination of variables, or pivotal condensation. In the proof of Theorem A.1.2 we constructed a lower triangular matrix F with diagonal elements 1 such that $FA = A^*$ is upper triangular. If $Fy = y^*$, then the equation is

$$A^*x = y^*. \tag{18}$$

In coordinates this is

$$\sum_{j=1}^{p} a_{ij}^* x_j = y_i^*. \tag{19}$$

Let $a_{ij}^{**} = a_{ij}^{*}/a_{ii}^{*}$, $y_i^{**} = y_i^{*}/a_{ii}^{*}$, $j = i, i+1, \ldots, p$, $i = 1, \ldots, p$. Then

$$x_i = y_i^{**} - \sum_{j=i+1}^{p} a_{ij}^{**} x_j; \tag{20}$$

these equations are to be solved successively for $x_p, x_{p-1}, \ldots, x_1$. The calculation of $FA = A^*$ is known as the *forward* solution, and the solution of (18) as the *backward* solution.

Since $FAF' = A^*F' = D^2$ diagonal, (20) is $A^{**}x = y^{**}$, where $A^{**} = D^{-2}A^*$ and $y^{**} = D^{-2}y^*$. Solving this equation gives

$$x = A^{**-1}y^{**} = F'y^{**}. \tag{21}$$

The computation is

$$x = F_1' \cdots F_{p-1}' D^{-2} F_{p-1} \cdots F_1 y. \tag{22}$$

The multiplier of y in (22) indicates a sequence of row operations which yields A^{-1}.

The operations of the forward solution transform A to the upper triangular matrix A^*. As seen in Section A.5.1, the triangularization of a matrix can be done by a sequence of Householder transformations or by a sequence of Givens transformations.

From $FA = A^*$, we obtain

$$|A| = \prod_{i=1}^{p} a_{ii}^{(i)}, \tag{23}$$

which is the product of the diagonal elements of A^*, resulting from the forward solution. We also have

$$\begin{aligned} y'A^{-1}y &= (Fy)'D^{-2}(Fy) = y^{*\prime}D^{-2}y^* \\ &= y^{*\prime}y^{**}. \end{aligned} \tag{24}$$

The forward solution gives a computation for the quadratic form which occurs in T^2 and other statistics.

For more on matrix computations consult Golub and Von Loan (1989).

APPENDIX B

Tables

TABLE B.1
WILKS' LIKELIHOOD CRITERION: FACTORS $C(p, m, M)$
TO ADJUST TO $\chi^2_{p,m}$, WHERE $M = n - p + 1$

$M \backslash m$	5% Significance Level								
					$p = 3$				
	2	4	6	8	10	12	14	16	18
1	1.295	1.422	1.535	1.632	1.716	1.791	1.857	1.916	1.971
2	1.109	1.174	1.241	1.302	1.359	1.410	1.458	1.501	1.542
3	1.058	1.099	1.145	1.190	1.232	1.272	1.309	1.344	1.377
4	1.036	1.065	1.099	1.133	1.167	1.199	1.229	1.258	1.286
5	1.025	1.046	1.072	1.100	1.127	1.154	1.179	1.204	1.228
6	1.018	1.035	1.056	1.078	1.101	1.123	1.145	1.167	1.188
7	1.014	1.027	1.044	1.063	1.082	1.101	1.121	1.139	1.158
8	1.011	1.022	1.036	1.052	1.068	1.085	1.102	1.119	1.135
9	1.009	1.018	1.030	1.043	1.058	1.073	1.088	1.102	1.117
10	1.007	1.015	1.025	1.037	1.050	1.063	1.076	1.089	1.103
12	1.005	1.011	1.019	1.028	1.038	1.048	1.059	1.070	1.081
15	1.003	1.008	1.013	1.020	1.027	1.035	1.043	1.052	1.060
20	1.002	1.004	1.008	1.012	1.017	1.022	1.028	1.034	1.040
30	1.001	1.002	1.004	1.006	1.009	1.011	1.015	1.018	1.021
60	1.000	1.001	1.001	1.002	1.002	1.003	1.004	1.006	1.007
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	12.5916	21.0261	28.8693	36.4150	43.7730	50.9985	58.1240	65.1708	72.1532

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

TABLE B.1 (*Continued*)

	5% Significance Level								
	$p=3$		$p=4$						
$M \backslash m$	20	22	2	4	6	8	10	12	14
1	2.021	2.067	1.407	1.451	1.517	1.583	1.644	1.700	1.751
2	1.580	1.616	1.161	1.194	1.240	1.286	1.331	1.373	1.413
3	1.408	1.438	1.089	1.114	1.148	1.183	1.218	1.252	1.284
4	1.313	1.338	1.057	1.076	1.102	1.130	1.159	1.186	1.213
5	1.251	1.273	1.040	1.055	1.076	1.099	1.122	1.145	1.168
6	1.208	1.227	1.030	1.042	1.059	1.078	1.097	1.118	1.137
7	1.176	1.193	1.023	1.033	1.047	1.063	1.080	1.097	1.115
8	1.151	1.167	1.018	1.027	1.038	1.052	1.067	1.082	1.097
9	1.132	1.147	1.015	1.022	1.032	1.044	1.057	1.070	1.084
10	1.116	1.129	1.012	1.018	1.027	1.038	1.049	1.061	1.073
12	1.092	1.103	1.009	1.014	1.020	1.029	1.038	1.047	1.058
15	1.069	1.078	1.006	1.009	1.014	1.020	1.027	1.035	1.042
20	1.046	1.052	1.003	1.006	1.009	1.013	1.017	1.022	1.027
30	1.025	1.029	1.002	1.003	1.004	1.006	1.009	1.011	1.014
60	1.008	1.009	1.000	1.001	1.001	1.002	1.003	1.003	1.004
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	79.0819	85.9649	15.5073	26.2962	36.4150	46.1943	55.7585	65.1708	74.4683

TABLE B.1 (*Continued*)

	5% Significance Level								
	$p=4$			$p=5$					
$M \backslash m$	16	18	20	2	4	6	8	10	12
1	1.799	1.843	1.884	1.503	1.483	1.514	1.556	1.600	1.643
2	1.450	1.485	1.518	1.209	1.216	1.245	1.280	1.315	1.350
3	1.314	1.343	1.371	1.120	1.130	1.154	1.182	1.211	1.240
4	1.239	1.264	1.288	1.079	1.089	1.108	1.131	1.155	1.179
5	1.190	1.212	1.233	1.056	1.065	1.081	1.100	1.120	1.141
6	1.157	1.176	1.194	1.042	1.050	1.063	1.079	1.097	1.114
7	1.132	1.149	1.165	1.033	1.040	1.051	1.065	1.080	1.095
8	1.113	1.128	1.143	1.026	1.032	1.042	1.054	1.067	1.081
9	1.098	1.111	1.125	1.022	1.027	1.035	1.046	1.057	1.070
10	1.086	1.098	1.110	1.018	1.023	1.030	1.039	1.050	1.061
12	1.068	1.078	1.088	1.013	1.017	1.023	1.030	1.038	1.047
15	1.050	1.058	1.066	1.009	1.011	1.016	1.021	1.028	1.034
20	1.033	1.039	1.045	1.005	1.007	1.010	1.013	1.018	1.022
30	1.018	1.021	1.024	1.002	1.003	1.005	1.007	1.009	1.012
60	1.005	1.007	1.008	1.001	1.001	1.001	1.002	1.003	1.004
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	83.6753	92.8083	101.879	18.3070	31.4104	43.7730	55.7585	67.5048	79.0819

TABLE B.1 (*Continued*)

	5% Significance Level								
	$p = 5$		$p = 6$					$p = 7$	
$M \backslash m$	14	16	2	6	8	10	12	2	4
1	1.683	1.722	1.587	1.520	1.543	1.573	1.605	1.662	1.550
2	1.383	1.415	1.254	1.255	1.279	1.307	1.335	1.297	1.263
3	1.267	1.294	1.150	1.163	1.184	1.208	1.232	1.178	1.165
4	1.203	1.226	1.100	1.116	1.134	1.154	1.175	1.121	1.116
5	1.161	1.181	1.072	1.088	1.103	1.120	1.138	1.089	1.087
6	1.132	1.150	1.055	1.069	1.082	1.097	1.113	1.068	1.068
7	1.111	1.127	1.043	1.056	1.068	1.081	1.095	1.054	1.055
8	1.095	1.109	1.035	1.046	1.057	1.068	1.081	1.044	1.045
9	1.082	1.095	1.029	1.039	1.048	1.059	1.070	1.036	1.038
10	1.072	1.083	1.024	1.034	1.042	1.051	1.061	1.031	1.032
12	1.057	1.066	1.018	1.025	1.032	1.040	1.048	1.023	1.024
15	1.042	1.049	1.012	1.018	1.023	1.029	1.035	1.016	1.017
20	1.027	1.033	1.007	1.011	1.014	1.018	1.023	1.010	1.011
30	1.014	1.018	1.003	1.006	1.007	1.010	1.012	1.005	1.005
60	1.004	1.006	1.001	1.002	1.002	1.003	1.004	1.001	1.001
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	90.5312	101.879	21.0261	50.9985	65.1708	79.0819	92.8083	23.6848	41.3371

TABLE B.1 (*Continued*)

	5% Significance Level								
	$p = 7$			$p = 8$		$p = 9$			$p = 10$
$M \backslash m$	6	8	10	2	8	2	4	6	2
1	1.530	1.538	1.557	1.729	1.538	1.791	1.614	1.558	1.847
2	1.266	1.282	1.303	1.336	1.288	1.373	1.309	1.293	1.408
3	1.173	1.189	1.208	1.206	1.195	1.232	1.201	1.196	1.257
4	1.124	1.139	1.155	1.142	1.144	1.162	1.144	1.144	1.182
5	1.095	1.108	1.122	1.105	1.113	1.121	1.110	1.112	1.137
6	1.075	1.086	1.099	1.081	1.091	1.094	1.088	1.090	1.107
7	1.062	1.071	1.083	1.065	1.076	1.076	1.071	1.074	1.087
8	1.051	1.060	1.070	1.053	1.064	1.062	1.060	1.062	1.072
9	1.043	1.051	1.060	1.044	1.055	1.052	1.050	1.053	1.061
10	1.037	1.044	1.053	1.038	1.048	1.045	1.043	1.046	1.052
12	1.029	1.034	1.042	1.028	1.038	1.034	1.033	1.035	1.039
15	1.020	1.024	1.031	1.019	1.027	1.023	1.023	1.025	1.028
20	1.013	1.016	1.019	1.012	1.017	1.014	1.015	1.016	1.017
30	1.006	1.008	1.010	1.006	1.009	1.007	1.007	1.008	1.009
60	1.002	1.002	1.003	1.001	1.003	1.002	1.002	1.002	1.002
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	58.1240	74.4683	90.5312	26.2962	83.6753	28.8693	50.9985	72.1532	31.4104

TABLE B.1 (*Continued*)

	1% Significance Level							
	$p = 3$							
$M \backslash m$	2	4	6	8	10	12	14	16
1	1.356	1.514	1.649	1.763	1.862	1.949	2.026	2.095
2	1.131	1.207	1.282	1.350	1.413	1.470	1.523	1.571
3	1.070	1.116	1.167	1.216	1.262	1.306	1.346	1.384
4	1.043	1.076	1.113	1.150	1.187	1.221	1.254	1.285
5	1.030	1.054	1.082	1.112	1.141	1.170	1.198	1.224
6	1.022	1.040	1.063	1.087	1.112	1.136	1.159	1.182
7	1.016	1.031	1.050	1.070	1.091	1.111	1.132	1.152
8	1.013	1.025	1.041	1.058	1.075	1.093	1.111	1.129
9	1.010	1.021	1.034	1.048	1.064	1.080	1.095	1.111
10	1.009	1.017	1.028	1.041	1.055	1.069	1.082	1.097
12	1.006	1.012	1.021	1.031	1.042	1.053	1.064	1.076
15	1.004	1.009	1.014	1.021	1.030	1.038	1.047	1.056
20	1.002	1.005	1.009	1.013	1.019	1.024	1.030	1.036
30	1.001	1.002	1.004	1.007	1.009	1.012	1.016	1.019
60	1.000	1.001	1.001	1.002	1.003	1.004	1.005	1.006
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	16.8119	26.2170	34.8053	42.9798	50.8922	58.6192	66.2062	73.6826

TABLE B.1 (*Continued*)

	1% Significance Level							
	$p = 3$			$p = 4$				
$M \backslash m$	18	20	22	2	4	6	8	10
1	2.158	2.216	2.269	1.490	1.550	1.628	1.704	1.774
2	1.616	1.657	1.696	1.192	1.229	1.279	1.330	1.379
3	1.420	1.453	1.485	1.106	1.132	1.168	1.207	1.244
4	1.315	1.344	1.371	1.068	1.088	1.115	1.146	1.176
5	1.249	1.274	1.297	1.047	1.063	1.085	1.109	1.134
6	1.204	1.226	1.246	1.035	1.048	1.066	1.086	1.107
7	1.171	1.190	1.209	1.027	1.037	1.052	1.070	1.088
8	1.146	1.163	1.180	1.021	1.030	1.043	1.058	1.073
9	1.127	1.142	1.157	1.017	1.025	1.036	1.048	1.062
10	1.111	1.125	1.139	1.014	1.021	1.030	1.041	1.054
12	1.087	1.099	1.110	1.010	1.015	1.023	1.031	1.041
15	1.065	1.074	1.083	1.007	1.010	1.016	1.022	1.029
20	1.043	1.049	1.056	1.004	1.006	1.010	1.014	1.019
30	1.023	1.027	1.031	1.002	1.003	1.005	1.007	1.009
60	1.007	1.009	1.010	1.000	1.001	1.001	1.002	1.003
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	81.0688	88.3794	95.6257	20.0902	31.9999	42.9798	53.4858	63.6907

TABLE B.1 (*Continued*)

	1% Significance Level							
	$p = 4$					$p = 5$		
$M \backslash m$	12	14	16	18	20	2	4	6
1	1.838	1.896	1.949	1.999	2.045	1.606	1.589	1.625
2	1.424	1.467	1.507	1.545	1.580	1.248	1.253	1.284
3	1.280	1.314	1.347	1.378	1.408	1.141	1.150	1.175
4	1.205	1.234	1.261	1.287	1.313	1.092	1.101	1.121
5	1.159	1.183	1.207	1.230	1.252	1.065	1.074	1.090
6	1.128	1.149	1.169	1.189	1.208	1.049	1.056	1.070
7	1.106	1.124	1.142	1.160	1.177	1.038	1.044	1.056
8	1.089	1.105	1.121	1.137	1.153	1.031	1.036	1.046
9	1.076	1.091	1.105	1.119	1.133	1.025	1.030	1.039
10	1.066	1.079	1.092	1.105	1.118	1.021	1.025	1.033
12	1.051	1.062	1.073	1.083	1.094	1.015	1.019	1.025
15	1.037	1.045	1.053	1.062	1.071	1.010	1.013	1.017
20	1.024	1.029	1.035	1.041	1.047	1.006	1.008	1.011
30	1.012	1.015	1.019	1.022	1.026	1.003	1.004	1.005
60	1.004	1.005	1.006	1.007	1.008	1.001	1.001	1.001
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	73.6826	83.5134	93.2168	102.8168	112.3292	23.2093	37.5662	50.8922

TABLE B.1 (*Continued*)

	1% Significance Level							
	$p = 5$					$p = 6$		
$M \backslash m$	8	10	12	14	16	2	6	8
1	1.672	1.721	1.768	1.813	1.855	1.707	1.631	1.656
2	1.321	1.359	1.396	1.431	1.465	1.300	1.294	1.319
3	1.204	1.235	1.265	1.294	1.323	1.175	1.183	1.205
4	1.145	1.171	1.196	1.221	1.245	1.116	1.129	1.148
5	1.110	1.131	1.153	1.174	1.196	1.084	1.097	1.113
6	1.087	1.105	1.124	1.143	1.161	1.063	1.076	1.090
7	1.071	1.087	1.103	1.119	1.136	1.050	1.061	1.074
8	1.059	1.073	1.087	1.102	1.116	1.040	1.051	1.062
9	1.050	1.062	1.075	1.088	1.101	1.033	1.043	1.052
10	1.043	1.054	1.065	1.077	1.089	1.028	1.037	1.045
12	1.033	1.041	1.051	1.060	1.070	1.021	1.028	1.035
15	1.023	1.030	1.037	1.044	1.052	1.014	1.020	1.024
20	1.015	1.019	1.024	1.029	1.034	1.008	1.012	1.015
30	1.007	1.010	1.012	1.015	1.019	1.004	1.006	1.008
60	1.002	1.003	1.004	1.005	1.006	1.001	1.002	1.002
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	63.6907	76.1539	88.3794	100.4251	112.3291	26.2170	58.6192	73.6826

TABLE B.1 (*Continued*)

$M \backslash m$	1% Significance Level						
	$p = 6$		$p = 7$				
	10	12	2	4	6	8	10
1	1.687	1.722	1.797	1.667	1.642	1.648	1.666
2	1.348	1.378	1.348	1.305	1.306	1.321	1.342
3	1.230	1.255	1.207	1.188	1.194	1.210	1.229
4	1.169	1.191	1.140	1.130	1.138	1.152	1.169
5	1.131	1.150	1.102	1.097	1.105	1.117	1.132
6	1.106	1.122	1.078	1.076	1.083	1.094	1.107
7	1.087	1.102	1.062	1.061	1.067	1.077	1.089
8	1.074	1.086	1.050	1.050	1.056	1.065	1.075
9	1.063	1.075	1.042	1.042	1.047	1.055	1.065
10	1.055	1.065	1.035	1.036	1.041	1.048	1.056
12	1.042	1.051	1.026	1.027	1.031	1.037	1.044
15	1.030	1.037	1.018	1.019	1.022	1.025	1.032
20	1.019	1.024	1.011	1.012	1.014	1.017	1.020
30	1.010	1.013	1.005	1.006	1.007	1.007	1.011
60	1.003	1.004	1.001	1.002	1.002	1.003	1.003
∞	1.000	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	88.3794	102.816	29.1412	48.2782	66.2062	83.5134	100.425

TABLE B.1 (*Continued*)

$M \backslash m$	1% Significance Level					
	$p = 8$		$p = 9$			$p = 10$
	2	8	2	4	6	2
1	1.879	1.646	1.953	1.740	1.671	2.021
2	1.394	1.326	1.436	1.355	1.333	1.476
3	1.238	1.215	1.267	1.226	1.218	1.296
4	1.163	1.158	1.185	1.161	1.158	1.207
5	1.120	1.123	1.138	1.122	1.122	1.155
6	1.092	1.099	1.107	1.096	1.098	1.121
7	1.074	1.082	1.086	1.078	1.080	1.098
8	1.060	1.069	1.070	1.065	1.067	1.081
9	1.050	1.059	1.059	1.055	1.058	1.068
10	1.043	1.051	1.050	1.047	1.050	1.058
12	1.032	1.040	1.038	1.036	1.037	1.044
15	1.022	1.028	1.026	1.026	1.027	1.031
20	1.013	1.018	1.016	1.016	1.017	1.019
30	1.007	1.009	1.008	1.008	1.009	1.010
60	1.002	1.003	1.002	1.002	1.003	1.003
∞	1.000	1.000	1.000	1.000	1.000	1.000
χ^2_{pm}	31.9999	93.2168	34.8053	58.6192	81.0688	37.5662

TABLE B.2

TABLES OF SIGNIFICANCE POINTS FOR THE LAWLEY–HOTELLING TRACE TEST

$$\Pr\left\{\frac{n}{m} W \geq x_\alpha\right\} = \alpha$$

$n \backslash m$	5% Significance Level								
	$p = 2$								
	2	3	4	5	6	8	10	12	15
2	9.859*	10.659*	11.098*	11.373*	11.562*	11.804*	11.952*	12.052*	12.153*
3	58.428	58.915	59.161	59.308	59.407	59.531	59.606	59.655	59.705
4	23.999	23.312	22.918	22.663	22.484	22.250	22.104	22.003	21.901
5	15.639	14.864	14.422	14.135	13.934	13.670	13.504	13.391	13.275
6	12.175	11.411	10.975	10.691	10.491	10.228	10.063	9.949	9.832
7	10.334	9.594	9.169	8.893	8.697	8.440	8.277	8.164	8.048
8	9.207	8.488	8.075	7.805	7.614	7.361	7.201	7.090	6.975
10	7.909	7.224	6.829	6.570	6.386	6.141	5.984	5.875	5.761
12	7.190	6.528	6.146	5.894	5.715	5.474	5.320	5.212	5.100
14	6.735	6.090	5.717	5.470	5.294	5.057	4.905	4.798	4.686
18	6.193	5.571	5.209	4.970	4.798	4.566	4.416	4.309	4.198
20	6.019	5.405	5.047	4.810	4.640	4.410	4.260	4.154	4.042
25	5.724	5.124	4.774	4.542	4.374	4.147	3.998	3.892	3.780
30	5.540	4.949	4.604	4.374	4.209	3.983	3.835	3.729	3.617
35	5.414	4.829	4.488	4.260	4.096	3.872	3.724	3.618	3.505
40	5.322	4.742	4.404	4.178	4.014	3.791	3.643	3.538	3.425
50	5.198	4.625	4.290	4.066	3.904	3.682	3.535	3.429	3.315
60	5.118	4.549	4.217	3.994	3.833	3.611	3.465	3.359	3.245
70	5.062	4.496	4.165	3.944	3.783	3.562	3.416	3.310	3.196
80	5.020	4.457	4.127	3.907	3.747	3.526	3.380	3.274	3.159
100	4.963	4.403	4.075	3.856	3.696	3.476	3.330	3.224	3.109
200	4.851	4.298	3.974	3.757	3.598	3.380	3.234	3.127	3.012
∞	4.744	4.197	3.877	3.661	3.504	3.287	3.141	3.035	2.918

*Multiply by 10^2.

TABLE B.2 (*Continued*)

	1% Significance Level								
	$p = 2$								
$n \backslash m$	2	3	4	5	6	8	10	12	15
2	2.467†	2.667†	2.776†	2.844†	2.891†	2.952†	2.989†	3.014†	3.039†
3	2.985*	2.990*	2.992*	2.994*	2.995*	2.996*	2.997*	2.997*	2.998*
4	74.275	71.026	69.244	68.116	67.337	66.332	65.712	65.290	64.862
5	38.295	35.567	34.070	33.121	32.465	31.615	31.088	30.729	30.364
6	26.118	23.794	22.517	21.706	21.143	20.413	19.958	19.648	19.332
7	20.388	18.326	17.191	16.469	15.967	15.313	14.905	14.626	14.341
8	17.152	15.268	14.229	13.567	13.106	12.504	12.127	11.868	11.603
10	13.701	12.038	11.120	10.531	10.121	9.582	9.243	9.010	8.769
12	11.920	10.388	9.541	8.996	8.615	8.113	7.796	7.577	7.351
14	10.844	9.399	8.597	8.082	7.720	7.242	6.939	6.729	6.511
18	9.617	8.278	7.533	7.053	6.714	6.265	5.979	5.780	5.572
20	9.236	7.932	7.206	6.736	6.406	5.966	5.685	5.489	5.284
25	8.604	7.360	6.666	6.217	5.899	5.476	5.204	5.013	4.813
30	8.219	7.013	6.339	5.903	5.593	5.180	4.914	4.726	4.529
35	7.959	6.780	6.120	5.692	5.389	4.982	4.720	4.535	4.339
40	7.773	6.613	5.964	5.542	5.243	4.841	4.582	4.398	4.204
50	7.523	6.389	5.754	5.341	5.048	4.653	4.397	4.216	4.023
60	7.363	6.247	5.621	5.214	4.924	4.534	4.280	4.100	3.908
70	7.252	6.148	5.529	5.125	4.838	4.451	4.199	4.020	3.829
80	7.171	6.075	5.461	5.061	4.775	4.391	4.140	3.961	3.770
100	7.059	5.976	5.369	4.972	4.690	4.308	4.059	3.881	3.691
200	6.843	5.785	5.191	4.803	4.525	4.150	3.903	3.727	3.538
∞	6.638	5.604	5.023	4.642	4.369	4.000	3.757	3.582	3.393

†Multiply by 10^4
*Multiply by 10^2

TABLE B.2 (*Continued*)

	5% Significance Level								
					$p = 3$				
$n \backslash m$	3	4	5	6	8	10	12	15	20
3	25.930*	26.996*	27.665*	28.125*	28.712*	29.073*	29.316*	29.561*	29.809
4	1.188*	1.193*	1.196*	1.198*	1.200*	1.202*	1.203*	1.204*	1.205*
5	42.474	41.764	41.305	40.983	40.562	40.300	40.120	39.937	39.750
6	25.456	24.715	24.235	23.899	23.458	23.182	22.992	22.799	22.600
7	18.752	18.056	17.605	17.288	16.870	16.608	16.427	16.241	16.051
8	15.308	14.657	14.233	13.934	13.540	13.290	13.118	12.941	12.758
10	11.893	11.306	10.921	10.649	10.287	10.057	9.897	9.732	9.560
12	10.229	9.682	9.323	9.068	8.727	8.509	8.357	8.198	8.033
14	9.255	8.736	8.394	8.149	7.822	7.612	7.465	7.311	7.150
16	8.618	8.118	7.788	7.553	7.236	7.031	6.887	6.736	6.577
18	8.170	7.685	7.364	7.135	6.825	6.624	6.483	6.334	6.177
20	7.838	7.365	7.051	6.826	6.522	6.325	6.185	6.038	5.882
25	7.294	6.841	6.539	6.323	6.029	5.837	5.700	5.556	5.401
30	6.965	6.524	6.231	6.020	5.732	5.543	5.409	5.265	5.112
35	6.745	6.313	6.025	5.818	5.534	5.348	5.214	5.072	4.919
40	6.588	6.162	5.878	5.673	5.393	5.208	5.076	4.934	4.781
50	6.377	5.961	5.682	5.481	5.205	5.022	4.891	4.750	4.597
60	6.243	5.832	5.558	5.359	5.086	4.904	4.774	4.633	4.480
70	6.150	5.744	5.471	5.274	5.003	4.823	4.693	4.553	4.399
80	6.082	5.679	5.408	5.212	4.943	4.763	4.634	4.493	4.339
100	5.989	5.590	5.322	5.128	4.860	4.682	4.552	4.413	4.258
200	5.810	5.419	5.156	4.965	4.702	4.525	4.397	4.257	4.102
∞	5.640	5.256	4.999	4.812	4.552	4.377	4.250	4.110	3.954

*Multiply by 10^2.

TABLE B.2 (*Continued*)

	1% Significance Level								
	$p = 3$								
$n\backslash m$	3	4	5	6	8	10	12	15	20
3	6.484†	6.750†	6.917†	7.031†	7.178†	7.267†	7.328†	7.389†	7.451†
4	5.990*	5.995*	5.998*	6.000*	6.002*	6.003*	6.005*	6.006*	6.007*
5	1.274*	1.242*	1.222*	1.208*	1.190*	1.179*	1.172*	1.164*	1.156*
6	59.507	57.032	55.462	54.377	52.973	52.102	51.509	50.906	50.292
7	37.994	35.993	34.721	33.840	32.695	31.984	31.498	31.002	30.496
8	28.308	26.599	25.511	24.755	23.771	23.157	22.737	22.308	21.868
10	19.737	18.355	17.471	16.855	16.050	15.544	15.197	14.840	14.472
12	15.973	14.765	13.990	13.448	12.737	12.288	11.978	11.659	11.328
14	13.905	12.803	12.096	11.599	10.945	10.530	10.243	9.946	9.638
16	12.610	11.581	10.918	10.452	9.836	9.444	9.172	8.890	8.596
18	11.729	10.751	10.120	9.676	9.087	8.712	8.450	8.178	7.893
20	11.091	10.152	9.545	9.117	8.549	8.186	7.932	7.668	7.390
25	10.075	9.201	8.634	8.233	7.699	7.356	7.115	6.803	6.596
30	9.479	8.644	8.102	7.718	7.205	6.874	6.641	6.395	6.135
35	9.087	8.280	7.755	7.382	6.883	6.560	6.332	6.091	5.834
40	8.811	8.023	7.511	7.146	6.656	6.339	6.115	5.877	5.623
50	8.448	7.686	7.189	6.836	6.360	6.050	5.831	5.597	5.346
60	8.220	7.474	6.988	6.642	6.174	5.870	5.653	5.422	5.172
70	8.063	7.329	6.850	6.509	6.047	5.746	5.531	5.302	5.053
80	7.948	7.224	6.750	6.412	5.955	5.656	5.443	5.215	4.967
100	7.793	7.081	6.614	6.281	5.830	5.534	5.323	5.096	4.850
200	7.498	6.808	6.356	6.032	5.593	5.304	5.096	4.873	4.627
∞	7.222	6.554	6.116	5.801	5.373	5.089	4.885	4.664	4.419

†Multiply by 10^4.
*Multiply by 10^2.

TABLE B.2 (*Continued*)

	5% Significance Level								
	$p = 4$								
$n\backslash m$	4	5	6	8	10	12	15	20	25
4	49.964*	51.204*	52.054*	53.142*	53.808*	54.258*	54.71*	55.17*	55.46*
5	1.996*	2.001*	2.005*	2.009*	2.011*	2.013*	2.015*	2.016*	2.017*
6	65.715	64.999	64.497	63.841	63.432	63.151	62.866	62.573	62.396
7	37.343	36.629	36.129	35.474	35.064	34.782	34.495	34.200	34.019
8	26.516	25.868	25.413	24.814	24.437	24.178	23.912	23.639	23.471
10	17.875	17.326	16.938	16.424	16.098	15.872	15.640	15.399	15.250
12	14.338	13.848	13.500	13.037	12.741	12.535	12.321	12.099	11.961
14	12.455	12.002	11.680	11.248	10.972	10.778	10.577	10.366	10.234
16	11.295	10.868	10.563	10.154	9.890	9.705	9.512	9.309	9.181
18	10.512	10.104	9.812	9.419	9.165	8.986	8.798	8.600	8.475
20	9.950	9.556	9.274	8.893	8.645	8.471	8.287	8.093	7.970
25	9.059	8.688	8.422	8.062	7.826	7.659	7.482	7.293	7.173
30	8.538	8.182	7.927	7.578	7.350	7.188	7.015	6.829	6.710
35	8.197	7.852	7.603	7.263	7.040	6.880	6.710	6.526	6.408
40	7.957	7.619	7.375	7.041	6.821	6.664	6.495	6.313	6.195
50	7.640	7.313	7.075	6.750	6.535	6.380	6.214	6.033	5.916
60	7.442	7.120	6.887	6.568	6.356	6.203	6.038	5.858	5.740
70	7.305	6.988	6.758	6.443	6.232	6.081	5.917	5.738	5.620
80	7.206	6.892	6.665	6.351	6.143	5.992	5.829	5.650	5.532
100	7.071	6.762	6.537	6.228	6.021	5.872	5.710	5.531	5.413
200	6.814	6.514	6.295	5.993	5.791	5.644	5.484	5.305	5.186
∞	6.574	6.282	6.069	5.774	5.576	5.431	5.272	5.094	4.974

*Multiply by 10^2.

TABLE B.2 (*Continued*)

	1% Significance Level								
					$p = 4$				
$n \backslash m$	4	5	6	8	10	12	15	20	25
4	12.491†	12.800†	13.012†	13.283†	13.449†	13.561†	13.67†	13.79†	13.87†
5	9.999*	10.004*	10.008*	10.012*	10.014*	10.016*	10.018*	10.02*	10.02*
6	1.938*	1.906*	1.885*	1.857*	1.840*	1.828*	1.816*	1.804*	1.797*
7	85.053	82.731	81.125	79.047	77.759	76.882	75.989	75.082	74.522
8	51.991	50.178	48.921	47.290	46.276	45.583	44.877	44.156	43.715
10	29.789	28.478	27.566	26.376	25.632	25.121	24.597	24.060	23.731
12	21.965	20.889	20.138	19.154	18.534	18.108	17.668	17.215	16.936
14	18.142	17.199	16.539	15.670	15.121	14.742	14.349	13.943	13.691
16	15.916	15.059	14.457	13.662	13.157	12.807	12.444	12.066	11.831
18	14.473	13.674	13.112	12.368	11.894	11.564	11.221	10.863	10.639
20	13.466	12.710	12.177	11.470	11.018	10.703	10.374	10.030	9.814
25	11.924	11.237	10.751	10.103	9.687	9.395	9.089	8.766	8.562
30	11.055	10.409	9.951	9.338	8.943	8.665	8.372	8.060	7.863
35	10.499	9.880	9.440	8.851	8.470	8.200	7.915	7.611	7.418
40	10.114	9.514	9.087	8.514	8.142	7.879	7.600	7.301	7.110
50	9.614	9.040	8.631	8.079	7.720	7.465	7.194	6.902	6.713
60	9.305	8.747	8.319	7.311	7.460	7.210	6.943	6.655	6.468
70	9.095	8.549	8.158	7.630	7.284	7.037	6.774	6.488	6.301
80	8.944	8.405	8.020	7.498	7.157	6.912	6.651	6.367	6.181
100	8.739	8.211	7.833	7.321	6.985	6.744	6.486	6.204	6.019
200	8.354	7.848	7.484	6.990	6.664	6.429	6.176	5.898	5.714
∞	8.000	7.513	7.163	6.686	6.369	6.140	5.892	5.616	5.432

†Multiply by 10^4
*Multiply by 10^2

TABLE B.2 (*Continued*)

	5% Significance Level								
					$p = 5$				
$n \backslash m$	5	6	8	10	12	15	20	25	40
5	81.991*	83.352*	85.093*	86.160†	86.88†	—	—	—	—
6	3.009*	3.014*	3.020*	3.024†	3.027†	3.029†	3.032†	—	—
7	93.762	93.042	92.102	91.515	91.113	90.705	90.29	90.04	—
8	51.339	50.646	49.739	49.170	48.780	48.382	47.973	47.723	47.35
10	27.667	27.115	26.387	25.927	25.610	25.284	24.947	24.740	24.422
12	20.169	19.701	19.079	18.683	18.409	18.124	17.830	17.647	17.365
14	16.643	16.224	15.666	15.309	15.059	14.800	14.530	14.361	14.100
16	14.624	14.239	13.722	13.389	13.157	12.914	12.659	12.499	12.250
18	13.326	12.963	12.476	12.161	11.939	11.708	11.463	11.310	11.068
20	12.424	12.078	11.612	11.310	11.097	10.874	10.637	10.488	10.252
25	11.046	10.728	10.297	10.016	9.817	9.606	9.381	9.239	9.010
30	10.270	9.969	9.559	9.291	9.099	8.896	8.679	8.539	8.314
35	9.774	9.484	9.088	8.828	8.642	8.444	8.230	8.093	7.869
40	9.429	9.147	8.761	8.507	8.325	8.130	7.919	7.783	7.561
50	8.982	8.711	8.339	8.092	7.915	7.725	7.518	7.383	7.161
60	8.706	8.441	8.077	7.836	7.662	7.474	7.269	7.135	6.912
70	8.517	8.257	7.899	7.661	7.489	7.304	7.100	6.967	6.743
80	8.381	8.124	7.770	7.535	7.365	7.181	6.978	6.845	6.621
100	8.197	7.945	7.597	7.365	7.197	7.014	6.813	6.680	6.455
200	7.850	7.607	7.271	7.045	6.881	6.702	6.503	6.370	6.142
∞	7.531	7.295	6.970	6.750	6.590	6.414	6.217	6.084	5.850

†Multiply by 10^4.
*Multiply by 10^2.

TABLE B.2 (*Continued*)

	1% Significance Level								
	$p = 5$								
$n\backslash m$	5	6	8	10	12	15	20	25	40
5	20.495*	20.834*	21.267*	21.53*	—	—	—	—	—
6	15.014*	15.019*	15.025*	15.029*	15.033*	15.03*	15.06*	—	—
7	2.735*	2.704*	2.665*	2.640*	2.623*	2.606*	2.590*	2.579*	—
8	1.150*	1.128*	1.099*	1.081*	1.069*	1.057*	1.044*	1.036*	—
10	48.048	46.670	44.877	43.758	42.992	42.210	41.408	40.921	—
12	31.108	30.065	28.701	27.846	27.257	26.653	26.031	25.648	25.06
14	24.016	23.145	22.001	21.279	20.781	20.268	19.736	19.408	18.90
16	20.240	19.472	18.459	17.817	17.373	16.913	16.435	16.138	15.678
18	17.929	17.228	16.302	15.713	15.304	14.878	14.435	14.159	13.727
20	16.380	15.727	14.862	14.310	13.925	13.525	13.105	12.843	12.431
25	14.107	13.529	12.759	12.265	11.918	11.555	11.172	10.930	10.547
30	12.880	12.345	11.629	11.167	10.842	10.500	10.136	9.906	9.538
35	12.115	11.607	10.926	10.486	10.174	9.845	9.494	9.271	8.911
40	11.593	11.105	10.448	10.022	9.720	9.401	9.058	8.839	8.484
50	10.928	10.465	9.841	9.434	9.144	8.836	8.504	8.290	7.940
60	10.523	10.076	9.471	9.076	8.794	8.493	8.167	7.956	7.609
70	10.251	9.814	9.223	8.835	8.559	8.263	7.941	7.732	7.386
80	10.055	9.626	9.045	8.663	8.390	8.097	7.779	7.571	7.225
100	9.793	9.374	8.806	8.432	8.164	7.876	7.561	7.355	7.009
200	9.306	8.907	8.363	8.004	7.745	7.465	7.157	6.953	6.606
∞	8.863	8.482	7.961	7.615	7.365	7.093	6.790	6.588	6.236

*Multiply by 10^2.

TABLE B.2 (*Continued*)

	5% Significance Level								
	$p = 6$								
$n\backslash m$	6	8	10	12	15	20	25	30	35
10	45.722	44.677	44.019	43.567	43.103	42.626	42.334	42.136	41.993
12	28.959	28.121	27.590	27.223	26.843	26.451	26.209	26.044	25.925
14	22.321	21.600	21.141	20.821	20.489	20.144	19.929	19.783	19.677
16	18.858	18.210	17.795	17.505	17.202	16.886	16.688	16.553	16.455
18	16.755	16.157	15.772	15.501	15.218	14.921	14.735	14.607	14.513
20	15.351	14.788	14.424	14.168	13.899	13.615	13.436	13.313	13.223
25	13.293	12.786	12.456	12.222	11.975	11.711	11.544	11.428	11.343
30	12.180	11.705	11.395	11.173	10.939	10.687	10.526	10.414	10.331
35	11.484	11.031	10.733	10.520	10.293	10.049	9.892	9.782	9.700
40	11.009	10.571	10.282	10.075	9.853	9.614	9.460	9.351	9.270
50	10.402	9.983	9.706	9.507	9.293	9.060	8.908	8.801	8.721
60	10.031	9.625	9.355	9.160	8.951	8.721	8.572	8.465	8.385
70	9.781	9.383	9.118	8.927	8.720	8.494	8.345	8.239	8.159
80	9.601	9.209	8.948	8.759	8.555	8.330	8.182	8.076	7.996
100	9.360	8.976	8.720	8.534	8.333	8.110	7.963	7.857	7.777
200	8.910	8.542	8.295	8.115	7.919	7.701	7.555	7.449	7.369
500	8.659	8.300	8.059	7.882	7.689	7.473	7.328	7.222	7.140
1000	8.579	8.223	7.983	7.808	7.616	7.400	7.255	7.149	7.067
∞	8.500	8.146	7.908	7.734	7.543	7.328	7.183	7.077	6.994

TABLE B.2 (*Continued*)

	1% Significance Level								
	$p = 6$								
$n \backslash m$	6	8	10	12	15	20	25	30	35
10	86.397	83.565	81.804	80.602	79.376	78.124	77.360	76.845	76.474
12	46.027	44.103	42.899	42.073	41.227	40.359	39.826	39.466	39.206
14	32.433	30.918	29.966	29.309	28.634	27.936	27.507	27.215	27.004
16	25.977	24.689	23.875	23.311	22.729	22.126	21.753	21.498	21.314
18	22.292	21.146	20.418	19.913	19.389	18.844	18.505	18.273	18.105
20	19.935	18.886	18.217	17.752	17.267	16.761	16.445	16.229	16.071
25	16.642	15.737	15.156	14.749	14.324	13.875	13.592	13.397	13.254
30	14.944	14.118	13.586	13.211	12.816	12.398	12.133	11.949	11.814
35	13.913	13.138	12.635	12.281	11.906	11.506	11.252	11.074	10.943
40	13.223	12.482	12.000	11.659	11.298	10.911	10.663	10.490	10.361
50	12.358	11.661	11.206	10.882	10.538	10.167	9.927	9.759	9.633
60	11.839	11.169	10.730	10.417	10.083	9.721	9.486	9.320	9.196
70	11.493	10.841	10.413	10.107	9.779	9.424	9.192	9.028	8.905
80	11.246	10.607	10.187	9.886	9.563	9.212	8.983	8.819	8.697
100	10.917	10.295	9.886	9.592	9.276	8.930	8.703	8.541	8.419
200	10.312	9.723	9.333	9.052	8.748	8.412	8.190	8.030	7.908
500	9.980	9.409	9.030	8.755	8.458	8.128	7.907	7.747	7.625
1000	9.874	9.308	8.933	8.661	8.365	8.037	7.817	7.657	7.534
∞	9.770	9.210	8.838	8.568	8.274	7.948	7.728	7.568	7.446

TABLE B.2 (*Continued*)

	5% Significance Level							
	$p = 7$							
$n \backslash m$	8	10	12	15	20	25	30	35
10	85.040	84.082	83.426	82.755	82.068	81.648	81.364	81.159
12	42.850	42.126	41.627	41.113	40.583	40.257	40.037	39.877
14	29.968	29.373	28.961	28.534	28.091	27.817	27.631	27.495
16	24.038	23.519	23.158	22.781	22.389	22.145	21.978	21.857
18	20.692	20.222	19.893	19.549	19.189	18.964	18.809	18.696
20	18.561	18.125	17.819	17.498	17.159	16.947	16.800	16.694
25	15.587	15.202	14.930	14.642	14.337	14.143	14.009	13.911
30	14.049	13.693	13.440	13.172	12.884	12.701	12.573	12.478
35	13.113	12.776	12.535	12.278	12.002	11.825	11.700	11.608
40	12.485	12.160	11.927	11.679	11.411	11.237	11.115	11.025
50	11.695	11.386	11.165	10.927	10.668	10.500	10.381	10.292
60	11.219	10.921	10.706	10.475	10.221	10.056	9.938	9.850
70	10.901	10.610	10.400	10.173	9.923	9.760	9.643	9.555
80	10.674	10.388	10.181	9.957	9.710	9.548	9.432	9.344
100	10.371	10.091	9.889	9.669	9.426	9.265	9.150	9.062
200	9.812	9.545	9.350	9.138	8.902	8.744	8.629	8.542
500	9.504	9.244	9.054	8.846	8.613	8.456	8.342	8.254
1000	9.405	9.148	8.959	8.753	8.521	8.365	8.250	8.162
∞	9.308	9.053	8.866	8.661	8.431	8.275	8.160	8.072

TABLE B.2 (*Continued*)

$n\backslash m$	1% Significance Level							
	$p = 7$							
	8	10	12	15	20	25	30	35
10	185.93	182.94	180.90	178.83	176.73	175.44	174.57	173.92
12	71.731	69.978	68.779	67.552	66.296	65.528	65.010	64.636
14	44.255	42.978	42.099	41.197	40.269	39.698	39.311	39.032
16	33.097	32.057	31.339	30.599	29.834	29.361	29.039	28.806
18	27.273	26.374	25.750	25.105	24.435	24.019	23.735	23.529
20	23.757	22.949	22.388	21.804	21.195	20.816	20.556	20.367
25	19.117	18.440	17.965	17.469	16.947	16.619	16.392	16.227
30	16.848	16.239	15.810	15.360	14.882	14.580	14.370	14.216
35	15.512	14.945	14.544	14.121	13.670	13.383	13.183	13.036
40	14.634	14.095	13.713	13.309	12.876	12.599	12.405	12.262
50	13.553	13.049	12.691	12.310	11.899	11.634	11.448	11.309
60	12.914	12.432	12.088	11.720	11.323	11.065	10.882	10.746
70	12.492	12.024	11.690	11.332	10.942	10.689	10.509	10.374
80	12.193	11.736	11.408	11.056	10.673	10.422	10.244	10.110
100	11.797	11.353	11.034	10.691	10.316	10.070	9.894	9.761
200	11.077	10.658	10.356	10.028	9.667	9.427	9.254	9.123
500	10.685	10.230	9.987	9.668	9.314	9.078	8.906	8.774
1000	10.561	10.160	9.869	9.553	9.202	8.966	8.795	8.663
∞	10.439	10.043	9.755	9.441	9.092	8.857	8.686	8.555

TABLE B.2 (*Continued*)

	5% Significance Level							
	$p = 8$							
$n \backslash m$	8	10	12	15	20	25	30	35
14	42.516	41.737	41.198	40.641	40.066	39.711	39.470	39.296
16	31.894	31.242	30.788	30.318	29.829	29.525	29.318	29.167
18	26.421	25.847	25.446	25.028	24.591	24.319	24.132	23.996
20	23.127	22.605	22.239	21.856	21.454	21.201	21.028	20.902
25	18.770	18.324	18.009	17.677	17.325	17.102	16.947	16.834
30	16.626	16.221	15.934	15.629	15.303	15.095	14.950	14.843
35	15.356	14.977	14.707	14.418	14.109	13.910	13.771	13.668
40	14.518	14.156	13.898	13.621	13.322	13.129	12.994	12.893
50	13.482	13.142	12.898	12.636	12.351	12.165	12.034	11.936
60	12.866	12.540	12.305	12.051	11.774	11.593	11.465	11.368
70	12.459	12.142	11.912	11.665	11.393	11.215	11.088	10.992
80	12.169	11.858	11.634	11.390	11.122	10.946	10.820	10.725
100	11.785	11.483	11.264	11.026	10.763	10.590	10.465	10.370
200	11.084	10.798	10.589	10.362	10.108	9.939	9.816	9.722
500	10.701	10.423	10.221	9.999	9.751	9.584	9.461	9.367
1000	10.579	10.304	10.104	9.884	9.637	9.470	9.348	9.254
∞	10.459	10.188	9.989	9.771	9.526	9.360	9.238	9.144

TABLE B.2 (*Continued*)

	1% Significance Level							
	$p = 8$							
$n \backslash m$	8	10	12	15	20	25	30	35
14	65.793	64.035	62.828	61.592	60.323	59.545	59.019	58.639
16	44.977	43.633	42.707	41.754	40.771	40.164	39.753	39.456
18	35.265	34.146	33.373	32.573	31.745	31.232	30.882	30.629
20	29.786	28.808	28.129	27.425	26.691	26.235	25.924	25.697
25	23.001	22.212	21.661	21.085	20.480	20.100	19.838	19.647
30	19.867	19.173	18.686	18.173	17.631	17.288	17.051	16.876
35	18.077	17.440	16.991	16.516	16.011	15.690	15.466	15.301
40	16.924	16.324	15.900	15.451	14.970	14.662	14.447	14.288
50	15.528	14.975	14.582	14.163	13.711	13.420	13.216	13.063
60	14.715	14.190	13.815	13.414	12.980	12.698	12.499	12.351
70	14.184	13.677	13.313	12.925	12.502	12.226	12.031	11.885
80	13.810	13.315	12.960	12.580	12.165	11.894	11.701	11.556
100	13.317	12.839	12.496	12.127	11.722	11.457	11.267	11.124
200	12.429	11.983	11.660	11.311	10.925	10.669	10.484	10.343
500	11.951	11.521	11.210	10.871	10.495	10.244	10.061	9.921
1000	11.800	11.375	11.067	10.732	10.359	10.109	9.927	9.787
∞	11.652	11.233	10.928	10.597	10.227	9.978	9.796	9.656

TABLE B.2 (*Continued*)

	5% Significance Level						
	$p = 10$						
$n\backslash m$	10	12	15	20	25	30	35
14	98.999	98.013	97.002	95.963	95.326	94.9	94.6
16	58.554	57.814	57.050	56.260	55.772	55.44	55.20
18	43.061	42.454	41.824	41.169	40.762	40.485	40.284
20	35.146	34.620	34.071	33.497	33.140	32.895	32.716
25	26.080	25.660	25.219	24.753	24.458	24.255	24.107
30	22.140	21.773	21.384	20.970	20.706	20.523	20.388
35	19.955	19.618	19.260	18.876	18.630	18.458	18.331
40	18.569	18.252	17.914	17.550	17.316	17.151	17.029
50	16.913	16.622	16.309	15.969	15.748	15.592	15.476
60	15.960	15.684	15.385	15.059	14.847	14.695	14.582
70	15.341	15.074	14.786	14.469	14.261	14.113	14.002
80	14.907	14.647	14.365	14.055	13.851	13.705	13.595
100	14.338	14.087	13.814	13.513	13.313	13.170	13.061
200	13.319	13.085	12.828	12.542	12.351	12.212	12.106
500	12.774	12.548	12.301	12.023	11.836	11.699	11.594
1000	12.602	12.379	12.134	11.859	11.674	11.538	11.432
∞	12.434	12.214	11.972	11.700	11.515	11.380	11.275

TABLE B.2 (*Continued*)

$n\backslash m$	1% Significance Level						
	$p = 10$						
	10	12	15	20	25	30	35
14	180.90	178.28	175.62	172.91	171.24	170	—
16	89.068	87.414	85.270	83.980	82.91	82.2	81.7
18	59.564	58.328	57.055	55.742	54.933	54.384	53.990
20	45.963	44.951	43.905	42.821	42.150	41.693	41.362
25	31.774	31.029	30.253	29.440	28.932	28.583	28.328
30	26.115	25.489	24.832	24.139	23.701	23.399	23.177
35	23.116	22.556	21.966	21.338	20.939	20.663	20.459
40	21.267	20.749	20.201	19.615	19.241	18.980	18.787
50	19.114	18.646	18.148	17.611	17.266	17.023	16.842
60	17.901	17.462	16.992	16.484	16.154	15.922	15.748
70	17.124	16.703	16.252	15.762	15.443	15.216	15.046
80	16.583	16.175	15.738	15.260	14.948	14.726	14.559
100	15.881	15.490	15.069	14.608	14.305	14.088	13.925
200	14.641	14.280	13.889	13.457	13.169	12.962	12.803
500	13.986	13.641	13.266	12.848	12.569	12.366	12.210
1000	13.780	13.441	13.070	12.658	12.381	12.179	12.023
∞	13.581	13.246	12.881	12.472	12.198	11.997	11.842

TABLE B.3

Tables of Significance Points for the Bartlett–Nanda–Pillai Trace Test

$$\Pr\left\{\frac{n+m}{m} V \geq x_\alpha\right\} = \alpha$$

***p* = 2**

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	13	5.499	4.250	3.730	3.430	3.229	3.082	2.970	2.881	2.808	2.747	2.545	2.431
	15	5.567	4.310	3.782	3.476	3.271	3.122	3.008	2.917	2.842	2.779	2.572	2.455
	19	5.659	4.396	3.858	3.546	3.336	3.183	3.066	2.972	2.895	2.831	2.616	2.493
	23	5.718	4.453	3.911	3.595	3.383	3.228	3.109	3.013	2.935	2.869	2.650	2.524
	27	5.759	4.495	3.950	3.632	3.418	3.261	3.141	3.045	2.966	2.899	2.677	2.548
	33	5.801	4.539	3.992	3.672	3.456	3.299	3.178	3.081	3.001	2.934	2.709	2.578
	43	5.845	4.586	4.037	3.716	3.499	3.341	3.219	3.122	3.041	2.974	2.746	2.613
	63	5.891	4.635	4.086	3.764	3.547	3.389	3.266	3.169	3.088	3.020	2.791	2.657
	83	5.914	4.661	4.112	3.790	3.573	3.415	3.293	3.195	3.114	3.046	2.818	2.683
	123	5.938	4.688	4.139	3.818	3.601	3.443	3.321	3.223	3.143	3.075	2.847	2.713
	243	5.962	4.715	4.168	3.846	3.630	3.472	3.351	3.254	3.174	3.106	2.880	2.748
	∞	5.991	4.744	4.197	3.877	3.661	3.504	3.384	3.287	3.208	3.141	2.918	2.788
.01	13	7.499	5.409	4.570	4.094	3.780	3.555	3.383	3.248	3.138	3.047	2.751	2.587
	15	7.710	5.539	4.671	4.180	3.857	3.625	3.448	3.309	3.196	3.101	2.795	2.625
	19	8.007	5.732	4.824	4.312	3.976	3.734	3.550	3.405	3.287	3.188	2.867	2.686
	23	8.206	5.868	4.935	4.409	4.064	3.815	3.627	3.478	3.356	3.255	2.923	2.735
	27	8.349	5.970	5.019	4.483	4.131	3.878	3.686	3.534	3.410	3.307	2.968	2.775
	33	8.500	6.080	5.111	4.566	4.207	3.950	3.754	3.600	3.473	3.368	3.021	2.823
	43	8.660	6.201	5.214	4.659	4.294	4.032	3.833	3.675	3.547	3.439	3.085	2.881
	63	8.831	6.333	5.329	4.764	4.393	4.127	3.925	3.765	3.634	3.525	3.163	2.955
	83	8.920	6.404	5.392	4.823	4.449	4.181	3.977	3.815	3.684	3.574	3.210	3.000
	123	9.012	6.478	5.459	4.885	4.508	4.238	4.033	3.871	3.739	3.628	3.263	3.052
	243	9.108	6.556	5.529	4.951	4.572	4.301	4.095	3.932	3.800	3.689	3.323	3.113
	∞	9.210	6.638	5.604	5.023	4.642	4.369	4.163	4.000	3.867	3.757	3.393	3.185

***p* = 3**

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	14	6.989	5.595	5.019	4.684	4.458	4.293	4.165	4.063	3.979	3.908	3.672	3.537
	16	7.095	5.673	5.082	4.738	4.507	4.338	4.207	4.103	4.017	3.944	3.702	3.563
	20	7.243	5.787	5.177	4.822	4.583	4.409	4.274	4.166	4.077	4.002	3.751	3.606
	24	7.341	5.866	5.245	4.883	4.639	4.461	4.323	4.213	4.122	4.046	3.790	3.640
	28	7.410	5.925	5.295	4.929	4.682	4.501	4.362	4.250	4.158	4.081	3.821	3.668
	34	7.482	5.987	5.351	4.980	4.730	4.547	4.406	4.293	4.200	4.121	3.857	3.702
	44	7.559	6.055	5.412	5.037	4.784	4.599	4.457	4.342	4.248	4.169	3.901	3.743
	64	7.639	6.129	5.480	5.101	4.846	4.660	4.516	4.400	4.305	4.225	3.955	3.795
	84	7.681	6.168	5.517	5.137	4.880	4.693	4.549	4.433	4.338	4.257	3.986	3.826
	124	7.724	6.209	5.556	5.174	4.917	4.730	4.585	4.469	4.374	4.293	4.022	3.862
	244	7.768	6.251	5.597	5.214	4.957	4.769	4.624	4.508	4.413	4.333	4.063	3.904
	∞	7.815	6.296	5.640	5.257	4.999	4.812	4.667	4.552	4.457	4.377	4.110	3.954
.01	14	8.971	6.855	5.970	5.457	5.112	4.862	4.669	4.516	4.390	4.285	3.939	3.743
	16	9.245	7.006	6.083	5.551	5.195	4.937	4.738	4.581	4.451	4.343	3.986	3.783
	20	9.639	7.236	6.258	5.698	5.326	5.056	4.849	4.684	4.549	4.436	4.063	3.850
	24	9.910	7.403	6.387	5.808	5.424	5.146	4.933	4.764	4.625	4.509	4.124	3.903
	28	10.106	7.528	6.486	5.893	5.501	5.217	4.999	4.827	4.685	4.567	4.174	3.948
	34	10.317	7.667	6.598	5.990	5.588	5.298	5.076	4.900	4.756	4.635	4.233	4.001
	44	10.545	7.821	6.724	6.101	5.690	5.393	5.167	4.986	4.839	4.715	4.305	4.067
	64	10.790	7.994	6.867	6.230	5.809	5.505	5.274	5.090	4.939	4.813	4.394	4.151
	84	10.920	8.088	6.947	6.301	5.876	5.569	5.335	5.150	4.998	4.871	4.448	4.203
	124	11.056	8.188	7.032	6.379	5.948	5.639	5.403	5.216	5.063	4.935	4.510	4.263
	244	11.196	8.294	7.124	6.463	6.028	5.716	5.478	5.290	5.136	5.007	4.581	4.334
	∞	11.345	8.406	7.222	6.554	6.116	5.801	5.562	5.372	5.218	5.089	4.664	4.419

$p=4$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	15	8.331	6.859	6.245	5.885	5.642	5.462	5.323	5.212	5.119	5.041	4.779	4.627
	17	8.472	6.952	6.318	5.947	5.696	5.512	5.369	5.255	5.160	5.080	4.811	4.654
	21	8.671	7.091	6.429	6.043	5.782	5.591	5.443	5.324	5.225	5.143	4.864	4.701
	25	8.805	7.190	6.510	6.114	5.846	5.650	5.498	5.377	5.276	5.191	4.906	4.738
	29	8.901	7.263	6.571	6.168	5.896	5.696	5.542	5.418	5.316	5.230	4.939	4.768
	35	9.004	7.343	6.640	6.229	5.952	5.749	5.593	5.467	5.363	5.275	4.980	4.805
	45	9.113	7.431	6.716	6.298	6.017	5.811	5.652	5.524	5.418	5.329	5.029	4.851
	65	9.229	7.528	6.802	6.378	6.092	5.883	5.721	5.592	5.485	5.395	5.090	4.910
	85	9.291	7.580	6.849	6.421	6.134	5.923	5.761	5.631	5.523	5.432	5.127	4.945
	125	9.354	7.635	6.899	6.469	6.179	5.968	5.804	5.674	5.566	5.475	5.168	4.987
	245	9.419	7.693	6.952	6.519	6.228	6.016	5.852	5.721	5.613	5.522	5.216	5.035
	∞	9.488	7.754	7.009	6.574	6.282	6.069	5.905	5.774	5.667	5.576	5.272	5.094
.01	15	10.293	8.188	7.276	6.737	6.373	6.105	5.898	5.731	5.594	5.479	5.095	4.874
	17	10.619	8.360	7.401	6.840	6.462	6.184	5.971	5.799	5.658	5.539	5.144	4.916
	21	11.095	8.625	7.598	7.003	6.604	6.313	6.089	5.909	5.762	5.638	5.225	4.987
	25	11.428	8.818	7.744	7.126	6.712	6.411	6.180	5.995	5.844	5.716	5.290	5.044
	29	11.672	8.966	7.858	7.222	6.798	6.490	6.253	6.064	5.909	5.779	5.344	5.091
	35	11.938	9.131	7.987	7.332	6.897	6.581	6.338	6.145	5.986	5.853	5.408	5.149
	45	12.228	9.318	8.135	7.460	7.012	6.688	6.439	6.241	6.079	5.942	5.487	5.221
	65	12.545	9.529	8.306	7.610	7.149	6.816	6.561	6.358	6.192	6.052	5.586	5.314
	85	12.715	9.645	8.402	7.695	7.227	6.890	6.632	6.426	6.258	6.117	5.646	5.371
	125	12.893	9.769	8.505	7.787	7.313	6.971	6.710	6.503	6.333	6.190	5.715	5.439
	245	13.080	9.902	8.617	7.889	7.408	7.062	6.798	6.588	6.417	6.273	5.796	5.519
	∞	13.277	10.045	8.739	8.000	7.513	7.163	6.897	6.686	6.513	6.369	5.892	5.616

$p=5$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	16	9.589	8.071	7.430	7.052	6.795	6.605	6.457	6.338	6.239	6.155	5.873	5.706
	18	9.761	8.179	7.512	7.120	6.854	6.659	6.506	6.384	6.282	6.196	5.906	5.735
	22	10.007	8.340	7.639	7.228	6.949	6.745	6.586	6.458	6.352	6.263	5.961	5.784
	26	10.176	8.457	7.732	7.308	7.021	6.810	6.647	6.516	6.407	6.316	6.006	5.823
	30	10.298	8.544	7.803	7.370	7.077	6.862	6.696	6.562	6.451	6.358	6.042	5.856
	36	10.429	8.641	7.883	7.440	7.141	6.922	6.752	6.616	6.503	6.408	6.086	5.896
	46	10.571	8.748	7.974	7.521	7.216	6.992	6.819	6.680	6.565	6.468	6.140	5.945
	66	10.724	8.868	8.077	7.615	7.303	7.075	6.899	6.757	6.640	6.541	6.208	6.009
	86	10.805	8.933	8.134	7.667	7.353	7.122	6.944	6.802	6.684	6.584	6.248	6.049
	126	10.890	9.002	8.195	7.724	7.407	7.174	6.995	6.851	6.733	6.633	6.295	6.095
	246	10.978	9.076	8.261	7.786	7.466	7.232	7.052	6.907	6.788	6.688	6.350	6.150
	∞	11.071	9.154	8.332	7.853	7.531	7.296	7.115	6.970	6.851	6.750	6.414	6.217
.01	16	11.534	9.451	8.521	7.966	7.587	7.306	7.088	6.912	6.767	6.644	6.230	5.989
	18	11.902	9.642	8.658	8.077	7.682	7.391	7.165	6.983	6.833	6.707	6.281	6.033
	22	12.449	9.939	8.876	8.255	7.835	7.528	7.291	7.100	6.943	6.810	6.366	6.106
	26	12.837	10.159	9.040	8.390	7.954	7.635	7.389	7.192	7.030	6.893	6.434	6.166
	30	13.125	10.328	9.168	8.497	8.048	7.720	7.468	7.266	7.100	6.960	6.491	6.216
	36	13.442	10.518	9.314	8.621	8.158	7.820	7.561	7.354	7.183	7.040	6.559	6.277
	46	13.790	10.735	9.483	8.765	8.287	7.939	7.673	7.460	7.284	7.137	6.644	6.355
	66	14.176	10.984	9.681	8.936	8.442	8.083	7.808	7.589	7.409	7.258	6.752	6.455
	86	14.385	11.122	9.793	9.034	8.531	8.167	7.888	7.666	7.483	7.330	6.818	6.518
	126	14.606	11.270	9.914	9.142	8.630	8.260	7.977	7.752	7.567	7.412	6.895	6.592
	246	14.839	11.431	10.047	9.260	8.740	8.364	8.077	7.849	7.663	7.506	6.985	6.681
	∞	15.086	11.605	10.193	9.392	8.863	8.482	8.192	7.961	7.773	7.615	7.093	6.790

$p=6$

α	n \ m	1	2	3	4	5	6	7	8	9	10	15	20
.05	17	10.794	9.247	8.585	8.193	7.926	7.728	7.573	7.448	7.344	7.256	6.956	6.778
	19	10.993	9.367	8.676	8.268	7.990	7.785	7.625	7.496	7.389	7.299	6.990	6.808
	23	11.282	9.550	8.817	8.386	8.093	7.878	7.711	7.576	7.464	7.369	7.048	6.858
	27	11.483	9.684	8.922	8.475	8.172	7.950	7.777	7.638	7.523	7.425	7.095	6.899
	31	11.630	9.784	9.003	8.545	8.234	8.007	7.830	7.688	7.570	7.471	7.134	6.934
	37	11.790	9.897	9.094	8.624	8.306	8.073	7.892	7.747	7.627	7.525	7.181	6.976
	47	11.964	10.024	9.199	8.716	8.390	8.151	7.966	7.817	7.694	7.590	7.239	7.029
	67	12.154	10.166	9.319	8.824	8.490	8.245	8.056	7.903	7.778	7.672	7.312	7.099
	87	12.255	10.245	9.387	8.885	8.547	8.299	8.108	7.954	7.827	7.720	7.357	7.142
	127	12.362	10.328	9.459	8.951	8.609	8.359	8.165	8.010	7.882	7.774	7.409	7.193
	247	12.474	10.417	9.538	9.024	8.678	8.425	8.230	8.074	7.945	7.836	7.470	7.254
	∞	12.592	10.513	9.623	9.104	8.755	8.500	8.303	8.146	8.017	7.908	7.543	7.328
.01	17	12.722	10.664	9.721	9.157	8.767	8.478	8.252	8.069	7.917	7.788	7.351	7.093
	19	13.126	10.874	9.873	9.277	8.869	8.567	8.332	8.143	7.986	7.853	7.403	7.137
	23	13.736	11.202	10.111	9.469	9.034	8.714	8.465	8.266	8.100	7.961	7.490	7.213
	27	14.173	11.446	10.292	9.617	9.162	8.828	8.570	8.363	8.192	8.048	7.561	7.275
	31	14.501	11.635	10.433	9.734	9.264	8.921	8.655	8.442	8.267	8.119	7.621	7.328
	37	14.865	11.850	10.596	9.871	9.384	9.030	8.756	8.537	8.356	8.204	7.693	7.392
	47	15.270	12.097	10.787	10.032	9.527	9.160	8.878	8.652	8.466	8.309	7.783	7.474
	67	15.723	12.382	11.011	10.224	9.700	9.319	9.027	8.794	8.602	8.440	7.899	7.581
	87	15.970	12.542	11.138	10.335	9.800	9.413	9.115	8.878	8.683	8.520	7.971	7.649
	127	16.233	12.715	11.278	10.457	9.912	9.517	9.215	8.974	8.776	8.610	8.055	7.729
	247	16.513	12.903	11.432	10.593	10.037	9.635	9.328	9.084	8.883	8.715	8.154	7.827
	∞	16.812	13.108	11.602	10.745	10.178	9.770	9.458	9.210	9.008	8.838	8.274	7.948

$p=7$

α	n \ m	1	2	3	4	5	6	7	8	9	10	15	20
.05	18	11.961	10.396	9.719	9.316	9.040	8.835	8.675	8.545	8.437	8.345	8.031	7.843
	20	12.184	10.528	9.817	9.396	9.109	8.896	8.730	8.596	8.484	8.390	8.057	7.874
	24	12.513	10.731	9.972	9.525	9.220	8.996	8.821	8.680	8.563	8.464	8.127	7.926
	28	12.744	10.880	10.088	9.622	9.306	9.073	8.892	8.746	8.626	8.523	8.176	7.969
	32	12.915	10.994	10.178	9.699	9.374	9.135	8.950	8.800	8.676	8.572	8.216	8.005
	38	13.102	11.123	10.281	9.787	9.453	9.208	9.017	8.864	8.737	8.630	8.266	8.049
	48	13.308	11.267	10.399	9.890	9.547	9.294	9.098	8.941	8.811	8.701	8.328	8.106
	68	13.534	11.433	10.537	10.012	9.658	9.398	9.197	9.036	8.902	8.789	8.407	8.180
	88	13.657	11.524	10.614	10.082	9.722	9.459	9.255	9.092	8.956	8.842	8.456	8.226
	128	13.786	11.623	10.698	10.158	9.793	9.526	9.320	9.155	9.018	8.903	8.513	8.281
	248	13.923	11.728	10.790	10.242	9.872	9.602	9.394	9.226	9.089	8.972	8.580	8.348
	∞	14.067	11.842	10.890	10.334	9.960	9.687	9.477	9.309	9.170	9.053	8.661	8.431
.01	18	13.374	11.841	10.895	10.321	9.923	9.627	9.395	9.206	9.049	8.915	8.460	8.188
	20	14.310	12.069	11.056	10.448	10.031	9.721	9.479	9.283	9.121	8.982	8.512	8.233
	24	14.974	12.426	11.314	10.655	10.207	9.876	9.619	9.412	9.240	9.095	8.602	8.310
	28	15.456	12.694	11.510	10.815	10.344	9.999	9.731	9.515	9.337	9.186	8.676	8.374
	32	15.822	12.902	11.665	10.943	10.455	10.098	9.821	9.599	9.416	9.261	8.738	8.429
	38	16.230	13.141	11.845	11.092	10.586	10.215	9.930	9.700	9.511	9.351	8.814	8.496
	48	16.688	13.416	12.056	11.269	10.742	10.357	10.061	9.824	9.628	9.463	8.909	8.582
	68	17.206	13.737	12.306	11.482	10.932	10.532	10.224	9.978	9.776	9.605	9.033	8.696
	88	17.491	13.919	12.449	11.605	11.043	10.635	10.321	10.070	9.865	9.691	9.110	8.768
	128	17.796	14.116	12.607	11.743	11.168	10.751	10.431	10.176	9.967	9.790	9.201	8.854
	248	18.124	14.333	12.782	11.897	11.308	10.883	10.557	10.298	10.085	9.906	9.309	8.960
	∞	18.475	14.571	12.977	12.070	11.468	11.034	10.703	10.439	10.223	10.043	9.441	9.092

$p=8$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	19	13.101	11.524	10.835	10.423	10.141	9.930	9.766	9.632	9.521	9.426	9.100	8.904
	21	13.346	11.667	10.941	10.509	10.214	9.995	9.824	9.685	9.570	9.472	9.136	8.935
	25	13.710	11.889	11.109	10.647	10.333	10.101	9.920	9.774	9.652	9.550	9.198	8.988
	29	13.970	12.054	11.235	10.753	10.425	10.184	9.996	9.844	9.718	9.612	9.249	9.032
	33	14.163	12.180	11.334	10.837	10.499	10.250	10.057	9.902	9.772	9.663	9.292	9.070
	39	14.377	12.323	11.448	10.934	10.585	10.329	10.130	9.970	9.837	9.725	9.344	9.116
	49	14.614	12.487	11.580	11.048	10.688	10.423	10.218	10.053	9.917	9.801	9.409	9.176
	69	14.877	12.674	11.734	11.183	10.811	10.538	10.326	10.156	10.016	9.897	9.494	9.254
	89	15.021	12.779	11.822	11.261	10.883	10.605	10.390	10.218	10.075	9.955	9.547	9.304
	129	15.173	12.892	11.918	11.347	10.962	10.680	10.462	10.288	10.143	10.021	9.609	9.363
	249	15.335	13.015	12.023	11.442	11.051	10.765	10.544	10.367	10.221	10.098	9.682	9.436
	∞	15.507	13.148	12.138	11.549	11.152	10.862	10.638	10.459	10.312	10.188	9.771	9.526
.01	19	14.999	12.992	12.043	11.463	11.060	10.758	10.521	10.328	10.167	10.030	9.558	9.275
	21	15.463	13.235	12.215	11.598	11.174	10.857	10.610	10.409	10.241	10.099	9.612	9.321
	25	16.177	13.620	12.491	11.819	11.360	11.021	10.757	10.543	10.366	10.216	9.704	9.399
	29	16.700	13.910	12.703	11.991	11.507	11.151	10.874	10.651	10.467	10.310	9.780	9.465
	33	17.100	14.137	12.871	12.128	11.626	11.256	10.970	10.740	10.550	10.389	9.844	9.521
	39	17.549	14.398	13.067	12.290	11.766	11.383	11.086	10.848	10.651	10.485	9.924	9.591
	49	18.058	14.702	13.297	12.482	11.935	11.536	11.227	10.980	10.776	10.604	10.024	9.681
	69	18.640	15.058	13.573	12.716	12.143	11.725	11.403	11.146	10.934	10.756	10.155	9.801
	89	18.962	15.261	13.733	12.853	12.265	11.838	11.509	11.247	11.030	10.849	10.238	9.877
	129	19.310	15.484	13.909	13.005	12.403	11.965	11.630	11.362	11.142	10.956	10.335	9.970
	249	19.684	15.729	14.106	13.177	12.559	12.112	11.769	11.495	11.271	11.083	10.453	10.083
	∞	20.090	16.000	14.327	13.371	12.738	12.280	11.930	11.652	11.424	11.233	10.597	10.227

$p=10$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	21	15.322	13.733	13.027	12.603	12.311	12.093	11.922	11.782	11.666	11.566	11.222	11.013
	23	15.604	13.897	13.147	12.700	12.393	12.164	11.985	11.840	11.719	11.616	11.260	11.045
	27	16.033	14.154	13.340	12.857	12.526	12.282	12.091	11.937	11.808	11.699	11.325	11.100
	31	16.344	14.347	13.487	12.978	12.631	12.375	12.176	12.015	11.881	11.768	11.380	11.146
	35	16.580	14.497	13.603	13.075	12.716	12.451	12.245	12.079	11.941	11.824	11.426	11.186
	41	16.843	14.669	13.737	13.188	12.816	12.541	12.328	12.156	12.014	11.893	11.482	11.236
	51	17.140	14.868	13.895	13.323	12.936	12.651	12.430	12.251	12.104	11.979	11.555	11.301
	71	17.476	15.100	14.083	13.486	13.083	12.786	12.556	12.371	12.218	12.089	11.650	11.388
	91	17.662	15.231	14.191	13.581	13.169	12.866	12.632	12.443	12.288	12.156	11.710	11.443
	131	17.861	15.375	14.310	13.687	13.266	12.957	12.718	12.526	12.368	12.234	11.781	11.511
	251	18.076	15.532	14.443	13.806	13.376	13.061	12.818	12.622	12.461	12.325	11.867	11.594
	∞	18.307	15.705	14.591	13.940	13.501	13.180	12.933	12.735	12.572	12.434	11.972	11.700
.01	21	17.197	15.234	14.284	13.698	13.288	12.980	12.736	12.537	12.371	12.228	11.733	11.432
	23	17.707	15.507	14.476	13.849	13.413	13.088	12.832	12.624	12.449	12.301	11.789	11.478
	27	18.505	15.941	14.788	14.096	13.621	13.268	12.993	12.769	12.584	12.426	11.885	11.559
	31	19.101	16.273	15.029	14.290	13.786	13.413	13.123	12.888	12.693	12.528	11.965	11.628
	35	19.562	16.535	15.222	14.447	13.920	13.531	13.230	12.986	12.785	12.614	12.034	11.687
	41	20.088	16.839	15.448	14.632	14.080	13.674	13.359	13.106	12.897	12.720	12.119	11.761
	51	20.692	17.196	15.718	14.855	14.274	13.848	13.519	13.255	13.037	12.852	12.229	11.858
	71	21.394	17.623	16.045	15.130	14.516	14.067	13.722	13.445	13.216	13.024	12.374	11.989
	91	21.790	17.868	16.236	15.292	14.660	14.199	13.845	13.561	13.327	13.130	12.466	12.074
	131	22.221	18.141	16.450	15.475	14.824	14.350	13.986	13.695	13.456	13.254	12.577	12.177
	251	22.692	18.444	16.690	15.683	15.012	14.525	14.152	13.853	13.608	13.402	12.712	12.307
	∞	23.209	18.783	16.964	15.923	15.231	14.730	14.346	14.041	13.791	13.581	12.881	12.472

TABLE B.4

Tables of Significance Points for the Roy Maximum Root Test

$$\Pr\left\{\frac{m+n}{m} R \geq x_\alpha\right\} = \alpha$$

$p=2$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	13	5.499	3.736	3.011	2.605	2.342	2.157	2.018	1.910	1.823	1.752	1.527	1.407
	15	5.567	3.807	3.078	2.668	2.401	2.211	2.069	1.959	1.869	1.796	1.562	1.436
	19	5.659	3.905	3.173	2.759	2.487	2.293	2.148	2.033	1.940	1.864	1.618	1.484
	23	5.718	3.971	3.239	2.822	2.548	2.352	2.204	2.087	1.993	1.915	1.661	1.521
	27	5.759	4.018	3.286	2.868	2.593	2.396	2.247	2.129	2.033	1.954	1.696	1.552
	33	5.801	4.068	3.336	2.918	2.643	2.445	2.294	2.175	2.079	1.998	1.736	1.588
	43	5.845	4.120	3.391	2.973	2.697	2.498	2.347	2.228	2.131	2.049	1.783	1.631
	63	5.891	4.176	3.449	3.032	2.757	2.558	2.407	2.288	2.190	2.109	1.840	1.686
	83	5.914	4.205	3.480	3.064	2.789	2.591	2.440	2.321	2.223	2.142	1.873	1.718
	123	5.938	4.235	3.512	3.097	2.823	2.626	2.476	2.356	2.259	2.178	1.909	1.755
	243	5.962	4.265	3.545	3.132	2.859	2.663	2.513	2.395	2.298	2.217	1.951	1.797
	∞	5.991	4.297	3.580	3.169	2.897	2.702	2.554	2.436	2.340	2.261	1.998	1.847
.01	13	7.499	4.675	3.610	3.040	2.681	2.432	2.249	2.109	1.997	1.907	1.625	1.478
	15	7.710	4.834	3.742	3.154	2.782	2.523	2.333	2.186	2.069	1.973	1.676	1.519
	19	8.007	5.064	3.937	3.325	2.936	2.664	2.463	2.307	2.182	2.080	1.758	1.587
	23	8.206	5.223	4.074	3.448	3.048	2.768	2.559	2.397	2.268	2.161	1.823	1.641
	27	8.349	5.339	4.176	3.540	3.133	2.847	2.634	2.468	2.335	2.225	1.876	1.686
	33	8.500	5.465	4.287	3.642	3.228	2.936	2.718	2.548	2.412	2.299	1.938	1.740
	43	8.660	5.600	4.409	3.755	3.334	3.037	2.815	2.641	2.501	2.386	2.013	1.807
	63	8.831	5.747	4.543	3.881	3.454	3.153	2.926	2.749	2.607	2.488	2.105	1.891
	83	8.920	5.825	4.616	3.950	3.520	3.217	2.989	2.810	2.666	2.547	2.160	1.943
	123	9.012	5.906	4.692	4.022	3.591	3.285	3.056	2.877	2.732	2.612	2.222	2.002
	243	9.108	5.991	4.772	4.100	3.666	3.360	3.130	2.950	2.804	2.683	2.292	2.072
	∞	9.210	6.080	4.856	4.182	3.747	3.440	3.209	3.029	2.884	2.763	2.373	2.154

$p=3$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	14	6.989	4.517	3.544	3.010	2.669	2.430	2.254	2.117	2.008	1.919	1.639	1.491
	16	7.095	4.617	3.634	3.092	2.745	2.501	2.319	2.178	2.065	1.973	1.682	1.526
	20	7.243	4.760	3.767	3.215	2.859	2.608	2.420	2.274	2.156	2.059	1.751	1.585
	24	7.341	4.858	3.859	3.302	2.942	2.686	2.495	2.345	2.224	2.124	1.805	1.631
	28	7.410	4.929	3.927	3.367	3.004	2.746	2.552	2.400	2.277	2.176	1.849	1.669
	34	7.482	5.004	4.001	3.439	3.073	2.812	2.616	2.462	2.338	2.234	1.901	1.715
	44	7.559	5.086	4.081	3.517	3.149	2.887	2.689	2.534	2.408	2.303	1.962	1.771
	64	7.639	5.173	4.169	3.604	3.235	2.972	2.773	2.616	2.489	2.383	2.038	1.842
	84	7.681	5.220	4.216	3.651	3.282	3.019	2.820	2.663	2.535	2.429	2.082	1.885
	124	7.724	5.268	4.265	3.701	3.332	3.069	2.870	2.713	2.586	2.479	2.132	1.934
	244	7.768	5.318	4.317	3.754	3.386	3.123	2.924	2.768	2.641	2.535	2.189	1.991
	∞	7.815	5.370	4.371	3.810	3.443	3.181	2.983	2.828	2.701	2.596	2.253	2.059
.01	14	8.971	5.416	4.106	3.412	2.978	2.680	2.462	2.295	2.163	2.055	1.724	1.552
	16	9.245	5.613	4.265	3.548	3.098	2.787	2.559	2.384	2.245	2.132	1.782	1.598
	20	9.639	5.905	4.507	3.757	3.284	2.956	2.714	2.527	2.378	2.257	1.877	1.676
	24	9.910	6.111	4.681	3.910	3.422	3.082	2.831	2.636	2.481	2.354	1.954	1.740
	28	10.106	6.264	4.811	4.026	3.528	3.180	2.922	2.722	2.562	2.431	2.016	1.792
	34	10.317	6.431	4.955	4.156	3.647	3.292	3.027	2.821	2.657	2.521	2.091	1.857
	44	10.545	6.614	5.116	4.303	3.784	3.420	3.148	2.938	2.768	2.628	2.182	1.937
	64	10.790	6.815	5.296	4.469	3.940	3.568	3.291	3.075	2.901	2.757	2.295	2.040
	84	10.920	6.923	5.393	4.560	4.027	3.652	3.372	3.153	2.977	2.832	2.363	2.103
	124	11.056	7.037	5.49'	4.658	4.120	3.742	3.460	3.239	3.062	2.915	2.441	2.177
	244	11.196	7.157	5.603	4.763	4.221	3.841	3.556	3.334	3.155	3.008	2.530	2.264
	∞	11.345	7.284	5.725	4.875	4.331	3.948	3.663	3.440	3.260	3.112	2.634	2.369

$p=4$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	15	8.331	5.211	4.013	3.365	2.955	2.670	2.459	2.297	2.168	2.063	1.736	1.564
	17	8.472	5.336	4.123	3.464	3.045	2.752	2.536	2.368	2.235	2.126	1.784	1.603
	21	8.671	5.517	4.287	3.613	3.182	2.880	2.655	2.481	2.341	2.227	1.864	1.670
	25	8.805	5.644	4.403	3.721	3.283	2.975	2.745	2.566	2.422	2.304	1.928	1.724
	29	8.901	5.736	4.490	3.802	3.360	3.048	2.814	2.632	2.486	2.365	1.979	1.769
	35	9.004	5.837	4.585	3.893	3.446	3.130	2.893	2.708	2.559	2.436	2.040	1.822
	45	9.113	5.946	4.690	3.993	3.543	3.224	2.984	2.796	2.645	2.520	2.114	1.889
	65	9.229	6.065	4.806	4.106	3.653	3.332	3.090	2.900	2.746	2.619	2.206	1.974
	85	9.291	6.128	4.869	4.168	3.714	3.392	3.149	2.959	2.804	2.677	2.261	2.026
	125	9.354	6.195	4.935	4.234	3.779	3.457	3.214	3.023	2.868	2.740	2.322	2.087
	245	9.419	6.265	5.005	4.304	3.850	3.527	3.284	3.093	2.939	2.811	2.393	2.157
	∞	9.488	6.338	5.080	4.380	3.926	3.603	3.361	3.171	3.017	2.890	2.475	2.242
.01	15	10.293	6.080	4.549	3.744	3.244	2.901	2.651	2.461	2.310	2.188	1.812	1.618
	17	10.619	6.308	4.731	3.898	3.378	3.021	2.760	2.559	2.401	2.273	1.875	1.668
	21	11.095	6.650	5.010	4.137	3.589	3.211	2.933	2.720	2.550	2.412	1.981	1.754
	25	11.428	6.896	5.213	4.315	3.748	3.356	3.067	2.844	2.666	2.521	2.066	1.825
	29	11.672	7.080	5.368	4.451	3.872	3.469	3.172	2.942	2.759	2.609	2.137	1.884
	35	11.938	7.284	5.542	4.606	4.013	3.600	3.294	3.057	2.868	2.713	2.221	1.956
	45	12.228	7.510	5.737	4.782	4.175	3.752	3.437	3.193	2.998	2.837	2.326	2.048
	65	12.545	7.762	5.958	4.984	4.364	3.930	3.607	3.356	3.155	2.989	2.458	2.166
	85	12.715	7.899	6.080	5.096	4.470	4.031	3.704	3.450	3.246	3.078	2.538	2.240
	125	12.893	8.045	6.210	5.218	4.585	4.142	3.812	3.555	3.348	3.178	2.630	2.326
	245	13.080	8.199	6.350	5.349	4.710	4.263	3.930	3.671	3.462	3.290	2.737	2.430
	∞	13.277	8.363	6.500	5.491	4.848	4.397	4.062	3.801	3.591	3.418	2.863	2.555

$p=5$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	16	9.589	5.856	4.448	3.694	3.218	2.890	2.648	2.463	2.316	2.196	1.824	1.630
	18	9.761	6.002	4.575	3.806	3.320	2.982	2.734	2.542	2.390	2.265	1.878	1.674
	22	10.007	6.218	4.765	3.978	3.477	3.128	2.869	2.669	2.509	2.378	1.967	1.747
	26	10.176	6.370	4.903	4.104	3.593	3.237	2.971	2.765	2.601	2.465	2.037	1.807
	30	10.298	6.483	5.006	4.200	3.683	3.321	3.051	2.842	2.674	2.535	2.095	1.857
	36	10.429	6.606	5.121	4.307	3.785	3.418	3.144	2.930	2.759	2.617	2.165	1.918
	46	10.571	6.742	5.249	4.428	3.901	3.529	3.251	3.034	2.859	2.714	2.250	1.994
	66	10.724	6.892	5.392	4.566	4.034	3.658	3.377	3.156	2.979	2.832	2.357	2.092
	86	10.805	6.973	5.471	4.643	4.108	3.731	3.448	3.227	3.048	2.900	2.421	2.152
	126	10.890	7.058	5.554	4.724	4.189	3.810	3.527	3.304	3.125	2.976	2.494	2.223
	246	10.978	7.148	5.643	4.812	4.276	3.897	3.613	3.390	3.210	3.061	2.578	2.306
	∞	11.071	7.244	5.738	4.907	4.371	3.992	3.708	3.485	3.306	3.157	2.676	2.407
.01	16	11.534	6.701	4.963	4.055	3.492	3.108	2.829	2.616	2.448	2.312	1.895	1.680
	18	11.902	6.954	5.163	4.223	3.638	3.238	2.945	2.722	2.546	2.403	1.962	1.733
	22	12.449	7.340	5.473	4.487	3.870	3.446	3.135	2.896	2.707	2.554	2.076	1.825
	26	12.837	7.620	5.703	4.685	4.047	3.606	3.282	3.033	2.835	2.673	2.169	1.902
	30	13.125	7.832	5.880	4.840	4.186	3.733	3.400	3.142	2.938	2.770	2.246	1.966
	36	13.442	8.069	6.079	5.016	4.346	3.881	3.537	3.271	3.060	2.886	2.339	2.046
	46	13.790	8.335	6.306	5.219	4.532	4.053	3.699	3.425	3.206	3.026	2.456	2.147
	66	14.176	8.635	6.566	5.454	4.750	4.259	3.894	3.611	3.385	3.198	2.604	2.279
	86	14.385	8.800	6.711	5.587	4.874	4.377	4.007	3.720	3.490	3.301	2.695	2.362
	126	14.606	8.977	6.867	5.731	5.010	4.506	4.132	3.841	3.608	3.416	2.800	2.461
	246	14.839	9.165	7.036	5.888	5.159	4.650	4.272	3.978	3.741	3.547	2.924	2.579
	∞	15.086	9.367	7.218	6.060	5.324	4.810	4.428	4.132	3.893	3.697	3.070	2.724

$p=6$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
	17	10.794	6.470	4.861	4.005	3.468	3.098	2.827	2.620	2.455	2.322	1.908	1.693
	19	10.993	6.634	5.001	4.128	3.579	3.199	2.920	2.705	2.535	2.396	1.965	1.740
	23	11.282	6.880	5.216	4.320	3.753	3.359	3.068	2.844	2.665	2.519	2.062	1.819
	27	11.483	7.056	5.372	4.462	3.884	3.481	3.182	2.951	2.767	2.615	2.139	1.884
	31	11.630	7.188	5.491	4.571	3.985	3.576	3.272	3.036	2.848	2.693	2.203	1.939
.05	37	11.790	7.334	5.625	4.695	4.101	3.686	3.376	3.136	2.943	2.785	2.280	2.006
	47	11.964	7.495	5.774	4.836	4.235	3.813	3.498	3.253	3.057	2.894	2.375	2.090
	67	12.154	7.675	5.944	4.997	4.390	3.963	3.643	3.394	3.194	3.028	2.495	2.200
	87	12.255	7.774	6.038	5.088	4.477	4.048	3.727	3.476	3.274	3.107	2.568	2.268
	127	12.362	7.878	6.138	5.185	4.573	4.141	3.818	3.566	3.363	3.195	2.652	2.348
	247	12.474	7.989	6.246	5.291	4.676	4.244	3.920	3.667	3.463	3.294	2.749	2.444
	∞	12.592	8.107	6.362	5.405	4.790	4.357	4.033	3.780	3.576	3.408	2.864	2.561
	17	12.722	7.296	5.360	4.352	3.730	3.306	2.998	2.764	2.580	2.431	1.974	1.739
	19	13.126	7.570	5.574	4.531	3.885	3.444	3.122	2.877	2.683	2.526	2.044	1.795
	23	13.736	7.992	5.912	4.817	4.135	3.667	3.325	3.063	2.855	2.687	2.165	1.892
	27	14.173	8.303	6.164	5.034	4.328	3.841	3.484	3.210	2.993	2.816	2.264	1.974
	31	14.501	8.541	6.360	5.204	4.480	3.980	3.612	3.329	3.104	2.921	2.347	2.043
.01	37	14.865	8.808	6.583	5.400	4.657	4.142	3.763	3.470	3.237	3.047	2.448	2.128
	47	15.270	9.112	6.839	5.628	4.864	4.334	3.943	3.640	3.399	3.201	2.576	2.238
	67	15.723	9.458	7.136	5.895	5.111	4.565	4.161	3.848	3.598	3.393	2.739	2.383
	87	15.970	9.650	7.303	6.047	5.252	4.699	4.289	3.971	3.716	3.507	2.840	2.475
	127	16.233	9.856	7.484	6.213	5.408	4.847	4.431	4.108	3.850	3.637	2.958	2.584
	247	16.513	10.079	7.682	6.395	5.580	5.012	4.591	4.264	4.002	3.786	3.097	2.717
	∞	16.812	10.319	7.897	6.596	5.772	5.198	4.772	4.442	4.177	3.959	3.264	2.882

$p=7$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
	18	11.961	7.063	5.258	4.304	3.708	3.298	2.999	2.770	2.589	2.442	1.989	1.753
	20	12.184	7.243	5.411	4.437	3.827	3.406	3.098	2.861	2.674	2.522	2.049	1.802
	24	12.513	7.516	5.647	4.647	4.016	3.580	3.258	3.011	2.814	2.653	2.151	1.887
	28	12.744	7.714	5.821	4.803	4.160	3.713	3.382	3.127	2.924	2.757	2.235	1.956
	32	12.915	7.863	5.954	4.925	4.272	3.817	3.481	3.220	3.012	2.842	2.304	2.015
.05	38	13.102	8.030	6.104	5.063	4.401	3.939	3.596	3.330	3.117	2.942	2.388	2.088
	48	13.308	8.216	6.275	5.222	4.551	4.081	3.732	3.461	3.243	3.063	2.492	2.180
	68	13.534	8.426	6.471	5.407	4.727	4.251	3.895	3.619	3.396	3.213	2.625	2.300
	88	13.657	8.541	6.579	5.511	4.827	4.348	3.990	3.711	3.486	3.301	2.706	2.376
	128	13.786	8.665	6.697	5.624	4.937	4.455	4.095	3.814	3.588	3.401	2.800	2.465
	248	13.923	8.797	6.824	5.747	5.058	4.573	4.211	3.929	3.702	3.514	2.910	2.573
	∞	14.067	8.938	6.961	5.882	5.191	4.705	4.343	4.060	3.833	3.645	3.041	2.705
	18	13.874	7.872	5.744	4.640	3.960	3.498	3.163	2.908	2.707	2.545	2.050	1.797
	20	14.310	8.164	5.971	4.829	4.124	3.642	3.292	3.026	2.816	2.646	2.124	1.855
	24	14.974	8.619	6.332	5.133	4.389	3.879	3.507	3.222	2.997	2.814	2.250	1.956
	28	15.456	8.957	6.605	5.367	4.595	4.065	3.676	3.379	3.143	2.951	2.355	2.042
	32	15.822	9.218	6.818	5.551	4.759	4.214	3.814	3.506	3.262	3.063	2.443	2.115
.01	38	16.230	9.514	7.063	5.765	4.952	4.390	3.977	3.659	3.406	3.199	2.551	2.206
	48	16.688	9.852	7.346	6.015	5.179	4.600	4.173	3.843	3.581	3.366	2.688	2.324
	68	17.206	10.243	7.679	6.313	5.452	4.855	4.413	4.072	3.799	3.575	2.866	2.481
	88	17.491	10.461	7.867	6.483	5.610	5.003	4.554	4.207	3.929	3.701	2.976	2.581
	128	17.796	10.697	8.073	6.670	5.785	5.169	4.713	4.360	4.078	3.846	3.106	2.700
	248	18.124	10.954	8.298	6.878	5.980	5.356	4.894	4.535	4.248	4.012	3.260	2.847
	∞	18.475	11.233	8.546	7.108	6.200	5.567	5.099	4.736	4.446	4.207	3.447	3.030

$p=8$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	19	13.101	7.640	5.645	4.594	3.941	3.493	3.166	2.916	2.719	2.559	2.067	1.812
	21	13.346	7.834	5.808	4.737	4.067	3.607	3.270	3.012	2.808	2.643	2.130	1.863
	25	13.710	8.132	6.063	4.962	4.270	3.792	3.441	3.171	2.956	2.782	2.238	1.952
	29	13.970	8.350	6.253	5.131	4.425	3.935	3.574	3.295	3.074	2.893	2.326	2.025
	33	14.163	8.515	6.399	5.264	4.547	4.049	3.680	3.396	3.169	2.984	2.400	2.088
	39	14.377	8.701	6.566	5.416	4.688	4.181	3.806	3.515	3.283	3.092	2.490	2.165
	49	14.614	8.912	6.757	5.593	4.854	4.338	3.955	3.658	3.420	3.224	2.603	2.264
	69	14.877	9.151	6.977	5.800	5.050	4.526	4.136	3.832	3.589	3.388	2.747	2.395
	89	15.021	9.283	7.101	5.917	5.163	4.634	4.241	3.935	3.689	3.486	2.836	2.478
	129	15.173	9.426	7.235	6.046	5.287	4.755	4.358	4.050	3.802	3.597	2.940	2.576
	249	15.335	9.579	7.381	6.187	5.424	4.689	4.491	4.180	3.931	3.725	3.063	2.695
	∞	15.507	9.745	7.541	6.342	5.577	5.040	4.640	4.329	4.078	3.872	3.210	2.843
.01	19	14.999	8.435	6.119	4.921	4.185	3.685	3.323	3.048	2.832	2.658	2.125	1.853
	21	15.463	8.743	6.357	5.119	4.355	3.836	3.458	3.171	2.944	2.762	2.201	1.913
	25	16.177	9.226	6.739	5.439	4.634	4.084	3.682	3.376	3.134	2.938	2.332	2.018
	29	16.700	9.589	7.030	5.687	4.853	4.280	3.861	3.541	3.287	3.081	2.442	2.107
	33	17.100	9.871	7.259	5.885	5.028	4.439	4.007	3.676	3.414	3.200	2.535	2.184
	39	17.549	10.194	7.524	6.115	5.234	4.627	4.181	3.839	3.566	3.344	2.649	2.280
	49	18.058	10.565	7.833	6.387	5.480	4.854	4.392	4.037	3.754	3.523	2.795	2.405
	69	18.640	10.998	8.199	6.713	5.778	5.131	4.653	4.284	3.990	3.749	2.986	2.573
	89	18.962	11.242	8.408	6.901	5.952	5.294	4.808	4.432	4.132	3.886	3.105	2.680
	129	19.310	11.508	8.638	7.109	6.146	5.478	4.983	4.601	4.295	4.044	3.246	2.810
	249	19.684	11.798	8.891	7.341	6.364	5.685	5.183	4.794	4.483	4.228	3.415	2.970
	∞	20.090	12.117	9.173	7.601	6.610	5.922	5.413	5.019	4.703	4.445	3.622	3.171

$p=10$

α	$n \backslash m$	1	2	3	4	5	6	7	8	9	10	15	20
.05	21	15.322	8.761	6.395	5.158	4.392	3.869	3.489	3.199	2.970	2.785	2.217	1.925
	23	15.604	8.979	6.577	5.315	4.531	3.994	3.602	3.303	3.067	2.875	2.285	1.980
	27	16.033	9.320	6.864	5.566	4.756	4.199	3.790	3.477	3.229	3.028	2.403	2.075
	31	16.344	9.573	7.082	5.759	4.931	4.359	3.939	3.616	3.360	3.151	2.500	2.156
	35	16.580	9.769	7.252	5.912	5.071	4.488	4.060	3.730	3.467	3.253	2.581	2.225
	41	16.843	9.992	7.448	6.090	5.234	4.641	4.203	3.865	3.596	3.376	2.682	2.311
	51	17.140	10.247	7.676	6.299	5.429	4.824	4.376	4.030	3.754	3.527	2.810	2.423
	71	17.476	10.543	7.944	6.548	5.663	5.047	4.590	4.235	3.952	3.719	2.977	2.572
	91	17.662	10.709	8.097	6.691	5.800	5.178	4.716	4.358	4.070	3.834	3.081	2.668
	131	17.861	10.890	8.265	6.850	5.952	5.324	4.858	4.496	4.206	3.967	3.204	2.783
	251	18.076	11.087	8.450	7.027	6.122	5.490	5.021	4.656	4.363	4.122	3.350	2.925
	∞	18.307	11.303	8.654	7.224	6.315	5.679	5.207	4.840	4.546	4.304	3.529	3.102
.01	21	17.197	9.534	6.851	5.470	4.624	4.051	3.636	3.322	3.075	2.877	2.271	1.962
	23	17.707	9.867	7.107	5.682	4.806	4.211	3.779	3.452	3.194	2.987	2.351	2.025
	27	18.505	10.399	7.523	6.029	5.107	4.478	4.021	3.672	3.397	3.175	2.491	2.137
	31	19.101	10.805	7.846	6.302	5.346	4.693	4.216	3.851	3.564	3.330	2.608	2.233
	35	19.562	11.125	8.103	6.522	5.541	4.868	4.376	4.000	3.702	3.460	2.709	2.315
	41	20.088	11.495	8.405	6.782	5.772	5.078	4.570	4.180	3.871	3.619	2.835	2.420
	51	20.692	11.928	8.761	7.093	6.052	5.335	4.808	4.403	4.081	3.819	2.996	2.558
	71	21.394	12.441	9.190	7.473	6.397	5.654	5.107	4.686	4.350	4.076	3.211	2.745
	91	21.790	12.735	9.439	7.695	6.601	5.845	5.287	4.857	4.515	4.234	3.347	2.867
	131	22.221	13.059	9.716	7.944	6.832	6.062	5.494	5.055	4.705	4.418	3.510	3.016
	251	22.692	13.417	10.025	8.225	7.094	6.310	5.732	5.285	4.928	4.636	3.707	3.201
	∞	23.209	13.816	10.373	8.545	7.395	6.598	6.010	5.556	5.193	4.895	3.952	3.438

TABLE B.5

SIGNIFICANCE POINTS FOR THE MODIFIED LIKELIHOOD RATIO TEST OF EQUALITY OF COVARIANCE MATRICES BASED ON EQUAL SAMPLE SIZES

$\Pr\{-2\log\lambda^* \geq x\} = 0.05$

$n_g \backslash q$	2	3	4	5	6	7	8	9	10
					$p = 2$				
3	12.18	18.70	24.55	30.09	35.45	40.68	45.81	50.87	55.87
4	10.70	16.65	22.00	27.07	31.97	36.76	41.45	46.07	50.64
5	9.97	15.63	20.73	25.56	30.23	34.79	39.26	43.67	48.02
6	9.53	15.02	19.97	24.66	29.19	33.61	37.95	42.22	46.45
7	9.24	14.62	19.46	24.05	28.49	32.82	37.07	41.26	45.40
8	9.04	14.33	19.10	23.62	27.99	32.26	36.45	40.57	44.65
9	8.88	14.11	18.83	23.30	27.62	31.84	35.98	40.06	44.08
10	8.76	13.94	18.61	23.05	27.33	31.51	35.61	36.65	43.64
					$p = 3$				
5	19.2	30.5	41.0	51.0	60.7	70.3	79.7	89.0	98.3
6	17.57	28.24	38.06	47.49	56.68	65.69	74.58	83.37	92.09
7	16.59	26.84	36.29	45.37	54.21	62.89	71.45	79.91	88.29
8	15.93	25.90	35.10	43.93	52.54	60.99	69.33	77.56	85.72
9	15.46	25.22	34.24	42.90	51.34	59.62	67.79	75.86	83.86
10	15.11	24.71	33.59	42.11	50.42	58.58	66.62	74.57	82.45
11	14.83	24.31	33.08	41.50	49.71	57.76	65.71	73.56	81.35
12	14.61	23.99	32.67	41.01	49.13	57.11	64.97	72.75	80.46
13	14.43	23.73	32.33	40.60	48.66	56.57	64.37	72.08	79.72
					$p = 4$				
6	30.07	48.63	65.91	82.6	98.9	115.0	131.0	—	—
7	27.31	44.69	60.90	76.56	91.89	107.0	121.9	137.0	152.0
8	25.61	42.24	57.77	72.78	87.46	101.9	116.2	130.4	144.6
9	24.46	40.56	55.62	70.17	84.42	98.45	112.3	126.1	139.8
10	23.62	39.34	54.05	68.27	82.19	95.91	109.5	122.9	136.3
11	22.98	38.41	52.85	66.81	80.49	93.95	107.3	120.5	133.6
12	22.48	37.67	51.90	65.66	79.14	92.41	105.5	118.5	131.5
13	22.08	37.08	51.13	64.73	78.04	91.16	104.1	117.0	129.7
14	21.75	36.59	50.50	63.96	77.14	90.12	103.0	115.7	128.3
15	21.47	36.17	49.97	63.31	76.38	89.25	102.0	114.6	127.1

TABLE B.5 (*Continued*)

$n_g \backslash q$	2	3	4	5	6	7	$n_g \backslash q$	2	3	4	5
			$p = 5$						$p = 6$		
8	39.29	65.15	89.46	113.0	—	—	10	49.95	84.43	117.0	—
9	36.70	61.40	84.63	107.2	129.3	151.5					
10	34.92	58.79	81.25	103.1	124.5	145.7	11	47.43	80.69	112.2	142.9
							12	45.56	77.90	108.6	138.4
11	33.62	56.86	78.76	100.0	120.9	141.6	13	44.11	75.74	105.7	135.0
12	32.62	55.37	76.83	97.68	118.2	138.4	14	42.96	74.01	103.5	132.2
13	31.83	54.19	75.30	95.81	116.0	135.9	15	42.03	72.59	101.6	129.9
14	31.19	53.24	74.06	94.29	114.2	133.8					
15	30.66	52.44	73.02	93.03	112.7	132.1	16	41.25	71.41	100.1	128.0
							17	40.59	70.41	98.75	126.4
16	30.21	51.77	72.14	91.95	111.4	130.6	18	40.02	69.55	97.63	125.0
							19	39.53	68.80	96.64	123.8
							20	39.11	68.14	95.78	122.7

TABLE B.6

CORRECTION FACTORS FOR SIGNIFICANCE POINTS FOR THE SPHERICITY TEST

$n \backslash p$	5% Significance Level					
	3	4	5	6	7	8
4	1.217					
5	1.074	1.322				
6	1.038	1.122	1.383			
7	1.023	1.066	1.155	1.420		
8	1.015	1.041	1.088	1.180	1.442	
9	1.011	1.029	1.057	1.098	1.199	1.455
10	1.008	1.021	1.040	1.071	1.121	1.214
12	1.005	1.013	1.023	1.039	1.060	1.093
14	1.004	1.008	1.015	1.024	1.037	1.054
16	1.003	1.006	1.011	1.017	1.025	1.035
18	1.002	1.005	1.008	1.012	1.018	1.025
20	1.002	1.004	1.006	1.010	1.014	1.019
24	1.001	1.002	1.004	1.006	1.009	1.012
28	1.001	1.002	1.003	1.004	1.006	1.008
34	1.000	1.001	1.002	1.003	1.004	1.005
42	1.000	1.001	1.001	1.002	1.002	1.003
50	1.000	1.000	1.001	1.001	1.002	1.002
100	1.000	1.000	1.000	1.000	1.000	1.000
χ^2	11.0705	16.9190	23.6848	31.4104	40.1133	49.8018

TABLE B.6 (*Continued*)

$n \backslash p$	1% Significance Level					
	3	4	5	6	7	8
4	1.266					
5	1.091	1.396				
6	1.046	1.148	1.471			
7	1.028	1.079	1.186	1.511		
8	1.019	1.049	1.103	1.213	1.542	
9	1.013	1.034	1.067	1.123	1.234	1.556
10	1.010	1.025	1.047	1.081	1.138	1.250
12	1.006	1.015	1.027	1.044	1.068	1.104
14	1.004	1.010	1.018	1.028	1.041	1.060
16	1.003	1.007	1.012	1.019	1.028	1.039
18	1.002	1.005	1.009	1.014	1.020	1.028
20	1.002	1.004	1.007	1.011	1.015	1.021
24	1.001	1.003	1.005	1.007	1.010	1.013
28	1.001	1.002	1.003	1.005	1.007	1.009
34	1.001	1.001	1.002	1.003	1.004	1.006
42	1.000	1.001	1.001	1.002	1.003	1.003
50	1.000	1.001	1.001	1.001	1.002	1.002
100	1.000	1.000	1.000	1.000	1.000	1.001
χ^2	15.0863	21.6660	29.1412	37.5662	46.9629	57.3421

TABLE B.7[†]

SIGNIFICANCE POINTS FOR THE MODIFIED LIKELIHOOD RATIO TEST $\Sigma = \Sigma_0$

$\Pr\{-2 \log \lambda_1^* \geq x\} = 0.05$

n	5%	1%	n	5%	1%	n	5%	1%	n	5%	1%
	$p = 2$			$p = 3$			$p = 5$			$p = 6$	
2	13.50	19.95	4	18.8	25.6	9	32.5	40.0	12	40.9	49.0
3	10.64	15.56	5	16.82	22.68	10	31.4	38.6	13	40.0	47.8
4	9.69	14.13							14	39.3	47.0
5	9.22	13.42	6	15.81	21.23	11	30.55	37.51	15	38.7	46.2
			7	15.19	20.36	12	29.92	36.72			
6	8.94	13.00	8	14.77	19.78	13	29.42	36.09	16	38.22	45.65
7	8.75	12.73	9	14.47	19.36	14	29.02	35.57	17	37.81	45.13
8	8.62	12.53	10	14.24	19.04	15	28.68	35.15	18	37.45	44.70
9	8.52	12.38							19	37.14	44.32
10	8.44	12.26	11	14.06	18.80	16	28.40	34.79	20	36.87	43.99
			12	13.92	18.61	17	28.15	34.49	21	36.63	43.69
	$p = 4$		13	13.80	18.45	18	27.94	34.23			
7	25.8	30.8	14	13.70	18.31	19	27.76	34.00	22	36.41	43.43
8	24.06	29.33	15	13.62	18.20	20	27.60	33.79	24	36.05	42.99
9	23.00	28.36							26	35.75	42.63
10	22.28	27.66							28	35.49	42.32
									30	35.28	42.07
11	21.75	27.13									
12	21.35	26.71									
13	21.03	26.38									
14	20.77	26.10									
15	20.56	25.87									
	$p = 7$			$p = 8$			$p = 9$			$p = 10$	
18	48.6	56.9	24	58.4	67.1	28	70.1	79.6	34	(82.3)	(92.4)
19	48.2	56.3	26	57.7	66.3	30	69.4	78.8	36	81.7	91.8
20	47.7	55.8	28	57.09	65.68				38	81.2	91.2
21	47.34	55.36	30	56.61	65.12	32	68.8	78.17	40	80.7	90.7
22	47.00	54.96				34	68.34	77.60			
			32	56.20	64.64	36	(67.91)	(77.08)	45	79.83	89.63
24	46.43	54.28	34	55.84	64.23	38	(67.53)	(76.65)	50	79.13	88.83
26	45.97	53.73	36	55.54	63.87	40	67.21	76.29	55	78.57	88.20
28	45.58	53.27	38	55.26	63.55				60	78.13	87.68
30	45.25	52.88	40	55.03	63.28	45	66.54	75.51	65	77.75	87.26
32	44.97	52.55				50	66.02	74.92			
34	44.73	52.27				55	65.61	74.44	70	77.44	86.89
						60	65.28	74.06	75	77.18	86.59

[†]Entries in parentheses have been interpolated or extrapolated into Korin's table. p = number of variates; N = number of observations; $n = N - 1$. $\lambda_1^* = n \log|\Sigma_0| - np - n \log|S| + n \operatorname{tr}(S\Sigma_0^{-1})$, where S is the sample covariance matrix.

References

At the end of each reference in brackets is a list of sections in which that reference is used.

Abramowitz, Milton, and Irene Stegun (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards. U.S. Government Printing Office, Washington, D.C. [8.5]

Abruzzi, Adam (1950), *Experimental Procedures and Criteria for Estimating and Evaluating Industrial Productivity*, doctoral dissertation, Columbia University Library. [9.7, 9.P]

Adrian, Robert (1808), Research concerning the probabilities of the errors which happen in making observations, etc., *The Analyst or Mathematical Museum*, **1**, 93–109. [1.2]

Ahn, S. K., and G. C. Reinsel (1988), Nested reduced-rank autoregressive models for multiple time series, *Journal of American Statistical Association*, **83**, 849–856. [12.7]

Aitken, A. C. (1937), Studies in practical mathematics, II. The evaluation of the latent roots and latent vectors of a matrix, *Proceedings of the Royal Society of Edinburgh*, **57**, 269–305. [11.4]

Amemiya, Yasuo, and T. W. Anderson (1990), Asymptotic chi-square tests for a large class of factor analysis models, *Annals of Statistics*, **18**, 1453–1463. [14.6]

Anderson, R. L., and T. A. Bancroft (1952), *Statistical Theory in Research*, McGraw-Hill, New York. [8.P]

Anderson, T. W. (1946a), The non-central Wishart distribution and certain problems of multivariate statistics, *Annals of Mathematical Statistics*, **17**, 409–431. (Correction, **35** (1964), 923–924.) [14.4]

Anderson, T. W. (1946b), Analysis of multivariate variance, unpublished. [10.6]

An Introduction to Multivariate Statistical Analysis, Third Edition. By T. W. Anderson
ISBN 0-471-36091-0 Copyright © 2003 John Wiley & Sons, Inc.

Anderson, T. W. (1950), Estimation of the parameters of a single equation by the limited-information maximum-likelihood method, *Statistical Inference in Dynamic Economic Models* (Tjalling C. Koopmans, ed.) John Wiley & Sons, Inc., New York. [12.8]

Anderson, T. W. (1951a), Classification by multivariate analysis, *Pyschometrika*, **16**, 31–50. [6.5, 6.9]

Anderson, T. W. (1951b), Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statistics* **22**, 327–351. (Correction, *Annals of Statistics*, **8** (1980), 1400.) [12.6, 12.7]

Anderson, T. W. (1951c), The asymptotic distribution of certain characteristic roots and vectors, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Jerzy Neyman, ed.), University of California, Berkeley, 105–130. [12.7]

Anderson, T. W. (1955a), Some statistical problems in relating experimental data to predicting performance of a production process, *Journal of the American Statistical Association*, **50**, 163–177. [8.P]

Anderson, T. W. (1955b), The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities, *Proceedings of the American Mathematical Society*, **6**, 170–176. [8.10]

Anderson, T. W. (1957), Maximum likelihood estimates for a multivariate normal distribution when some obervations are missing, *Journal of the American Statistical Association*, **52**, 676–687. [4.P]

Anderson, T. W. (1963a), Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics*, **34**, 122–148. [11.3, 11.6, 11.7, 13.5]

Anderson, T. W. (1963b), A test for equality of means when covariance matrices are unequal, *Annals of Mathematical Statistics*, **34**, 671–672. [5.P]

Anderson, T. W. (1965a), Some optimum confidence bounds for roots of determinantal equations, *Annals of Mathematical Statistics*, **36**, 468–488. [11.6]

Anderson, T. W. (1965b), Some properties of confidence regions and tests of parameters in multivariate distributions (with discussion), *Proceedings of the IBM Scientific Computing Symposium in Statistics, October 21–23, 1963*, IBM Data Processing Division, White Plains, New York, 15–28. [11.6]

Anderson, T. W. (1969), Statistical inference for covariance matrices with linear structure, *Multivariate Analysis II* (P. R. Krishnaiah, ed.), Academic, New York, 55–66. [3.P]

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, John Wiley & Sons, Inc., New York. [8.11]

Anderson, T. W. (1973a), Asymptotically efficient estimation of covariance matrices with linear structure, *Annals of Statistics*, **1**, 135–141. [14.3]

Anderson, T. W. (1973b), An asymptotic expansion of the distribution of the Studentized classification statistic W, *Annals of Statistics*, 1, 964–972. [6.6]

Anderson, T. W. (1973c), Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions, *Discriminant Analysis and Applications* (T. Cacoullos, ed.), Academic, New York, 17–35. [6.6]

Anderson, T. W. (1974), An asymptotic expansion of the distribution of the limited information maximum likelihood estimate of a coefficient in a simultaneous equation system, *Journal of the American Statistical Association*, **69**, 565–573. (Correction, **71** (1976), 1010.) [12.7]

Anderson, T. W. (1976), Estimation of linear functional relationships: approximate distributions and connections with simultaneous equations in econometrics (with discussion), *Journal of the Royal Statistical Society B*, **38**, 1–36. [12.7]

Anderson, T. W. (1977), Asymptotic expansions of the distributions of estimates in simultaneous equations for alternative parameter sequences, *Econometrica*, **45**, 509–518. [12.7]

Anderson, T. W. (1984a), Estimating linear statistical relationships, *Annals of Statistics*, **12**, 1–45. [10.6, 12.6, 12.7, 14.1]

Anderson, T. W. (1984b), Asymptotic distribution of an estimator of linear functional relationships, unpublished. [12.7]

Anderson, T. W. (1987), Multivariate linear relations, *Proceedings of the Second International Tampere Conference in Statistics* (Tarmo Pukkila and Simo Puntanen, eds.), Tampere, Finland, 9–36. [12.7]

Anderson, T. W. (1989a), The asymptotic distribution of the likelihood ratio criterion for testing rank in multivariate components of variance, *Journal of Multivariate Analysis*, **30**, 72–79. [10.6, 12.7]

Anderson, T. W. (1989b), The asymptotic distribution of characteristic roots and vectors in multivariate components of variance, *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin* (Leon Jay Gleser, Michael D. Perlman, S. James Press, and Allan R. Sampson, eds.), Springer Verlag, New York, 177–196. [13.6]

Anderson, T. W. (1989c), Linear latent variable models and covariance structures, *Journal of Econometrics*, **41**, 91–119. [Correction **43** (1990), 395.] [12.8]

Anderson, T. W. (1993), Nonnormal multivariate distributions: Inference based on elliptically contoured distributions, *Multivariate Analysis: Future Directions* (C. R. Rao, ed.), North-Holland, Amsterdam, 1–25. [3.6]

Anderson, T. W. (1994), Inference in linear models, *Proceedings of the International Symposium on Multivariate Analysis and its Applications* (T. W. Anderson, K. T. Fang, and I. Olkin, eds.), Institute of Mathematical Statistics, 1–20. [12.7, 12.8, 14.3]

Anderson, T. W. (1999a), Asymptotic theory for canonical correlation analysis, *Journal of Multivariate Analysis*, **70**, 1–29. [13.7]

Anderson, T. W. (1999b), Asymptotic distribution of the reduced rank regression estimator under general conditions, *Annals of Statistics*, **27**, 1141–1154. [13.7]

Anderson, T. W. (2002), Specification and misspecification in reduced rank regression, *Sankhyā*, **64A**, 1–13. [12.7]

Anderson, T. W., and Yasuo Amemiya (1988a), The asymptotic normal distribution of estimators in factor analysis under general conditions, *Annals of Statistics*, **16**, 759–771. [14.6]

Anderson, T. W., and Yasuo Amemiya (1988b), Asymptotic distributions in factor analysis and linear structural relations, *Proceedings of the International Conference on Advances in Multivariate Statistical Analysis* (S. Das Gupta and J. K. Ghosh, eds.), Indian Statistical Institute. Calcutta, 1–22. [14.6]

Anderson, T. W., and R. R. Bahadur (1962), Classification into two multivariate normal distributions with different covariance matrices, *Annals of Mathematical Statistics*. **33**, 420–431. [6.10]

Anderson. T. W., and Kai-Tai Fang (1990a), On the theory of multivariate elliptically contoured distributions and their applications, *Statistical Inference in Elliptically Contoured and Related Distributions* (Kai-Tai Fang and T. W. Anderson, eds.), Allerton Press, Inc., New York, 1–23. [4.5]

Anderson. T. W., and Kai-Tai Fang (1990b), Inference in multivariate elliptically contoured distributions based on maximum likelihood, *Statistical Inference in Elliptically Contoured and Related Distributions* (Kai-Tai Fang and T. W. Anderson, eds.). Allerton Press, Inc., New York, 201–216. [3.6]

Anderson. T. W., and Kai-Tai Fang (1992), Theory and applications of elliptically contoured and related distributions, *The Development of Statistics: Recent Contributions from China* (X. R. Chen, K. T. Fang, and C. C. Yang, eds.), Longman Scientific and Technical, Harlow, Essex, 41–62. [10.11]

Anderson, T. W., Kai-Tai Fang, and Huang Hsu (1986), Maximum likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions, *Canadian Journal of Statistics*, **14**, 55–59. [Reprinted in *Statistical Inference in Elliptically Contoured and Related Distributions* (Kai-Tai Fang and T. W. Anderson, eds.), Allerton Press, Inc., New York, 1990, 217–223. [3.6]

Anderson, T. W., Somesh Das Gupta, and George P. H. Styan (1972), *A Bibliography of Multivariate Statistical Analysis*, Oliver & Boyd, Edinburgh. (Reprinted by Robert E. Krieger, Malabar, Florida, 1977.) [Preface]

Anderson, T. W., Naoto Kunitomo, and Takamitsu Sawa (1983a), Evaluation of the distribution function of the limited information maximum likelihood estimator, *Econometrica*, **50**, 1009–1027. [12.7]

Anderson, T. W., Naoto Kunitomo, and Takamitsu Sawa (1983b), Comparison of the densities of the TSLS and LIMLK estimators, *Global Econometrics, Essays in Honor of Lawrence R. Klein* (F. Gerard Adams and Bert Hickman, eds.), MIT, Cambridge, MA, 103–124. [12.7]

Anderson, T. W., and I. Olkin (1985), Maximum likelihood estimation of the parameters of a multivariate normal distribution, *Linear Algebra and Its Applications*, **70**, 147–171. [3.2]

Anderson, T. W., and Michael D. Perlman (1987), Consistency of invariant tests for the multivariate analysis of variance, *Proceedings of the Second International Tampere Conference in Statistics* (Tarmo Pukkila and Simo Puntaner, eds.), Tampere, Finland, 225–243. [8.10]

Anderson, T. W., and Michael D. Perlman (1993), Parameter consistency of invariant tests for MANOVA and related multivariate hypotheses, *Statistics and Probability: A Raghu Raj Bahadur Festschift* (J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy, and B. L. S. Prakasa Rao, eds.), Wiley Eastern Limited, 37–62. [8.10]

Anderson, T. W., and Herman Rubin (1949), Estimation of the parameters of a single equation in a complete system of stochastic equations, *Annals of Mathematical Statistics*, **20**, 46–63. [Reprinted in *Readings in Econometric Theory* (J. Malcolm Dowling and Fred R. Glahe, eds.), Colorado Associated University, 1970, 358–375.] [12.8]

Anderson, T. W., and Herman Rubin (1950), The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations, *Annals of Mathematical Statistics*, **21**, 570–582. [Reprinted in *Readings in Econometric Theory* (J. Malcolm Dowling and Fred R. Glahe, eds.), Colorado Associated University, 1970, 376–388.] [12.8]

Anderson, T. W., and Herman Rubin (1956), Statistical inference in factor analysis, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Jerzy Neyman, ed.), Vol. V, University of California, Berkeley and Los Angeles, 111–150. [14.2, 14.3, 14.4, 14.6]

Anderson, T. W., and Takamitsu Sawa (1973), Distributions of estimates of coefficients of a single equation in a simultaneous system and their asymptotic expansions, *Econometrica*, **41**, 683–714. [12.7]

Anderson, T. W., and Takamitsu Sawa (1977), Tables of the distribution of the maximum likelihood estimate of the slope coefficient and approximations, Technical Report No. 234, Economics Series, Institute for Mathematical Studies in the Social Sciences, Stanford University, April. [12.7]

Anderson, T. W., and Takamitsu Sawa (1979), Evaluation of the distribution function of the two-stage least squares estimate, *Econometrica*, **47**, 163–182. [12.7]

Anderson, T. W., and Takamitsu Sawa (1982), Exact and approximate distributions of the maximum likelihood estimator of a slope coefficient, *Journal of the Royal Statistical Society B*, **44**, 52–62. [12.7]

Anderson, T. W., and George P. H. Styan (1982), Cochran's theorem, rank additivity and tripotent matrices, *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, eds.), North-Holland, Amsterdam, 1–23. [7.4]

Anderson, T. W., and Akimichi Takemura (1982), A new proof of admissibility of tests in multivariate analysis, *Journal of Multivariate Analysis*, **12**, 457–468. [8.10]

Andersson, Steen A., David Madigan, and Michael D. Perlman (2001), Alternative Markov properties for chain graphs, *Scandinavian Journal of Statistics*, **28**, 33–85. [15.2]

Barnard, M. M. (1935), The secular variations of skull characters in four series of Egyptian skulls, *Annals of Eugenics*, **6**, 352–371. [8.8]

Barndorff-Nielson, O. E. (1978), *Information and Exponential Families in Statistical Theory*, John Wiley & Sons, New York. [15.5]

Barnes, E. W. (1899), The theory of the gamma function, *Messenger of Mathematics*, **29**, 64–129. [8.5]

Bartlett, M. S. (1934), The vector representation of a sample, *Proceedings of the Cambridge Philosophical Society*, **30**, 327–340. [8.3]

Bartlett, M. S. (1937a), Properties of sufficiency and statistical tests, *Proceedings of the Royal Society of London A*, **160**, 268–282. [10.2]

Bartlett, M. S. (1937b), The statistical conception of mental factors, *British Journal of Psychology*, **28**, 97–104. [14.7]

Bartlett, M. S. (1938), Further aspects of the theory of multiple regression, *Proceedings of the Cambridge Philosophical Society*, **34**, 33–40. [14.7]

Bartlett, M. S. (1939), A note on tests of significance in multivariate analysis, *Proceedings of the Cambridge Philosophical Society*, **35**, 180–185. [7.2, 8.6]

Bartlett, M. S. (1947), Multivariate analysis, *Journal of the Royal Statistical Society, Supplement*, **9**, 176–197. [8.8]

Bartlett, M. S. (1950), Tests of significance in factor analysis, *British Journal of Psychology (Statistics Section)*, **3**, 77–85. [14.3]

Basmann, R. L. (1957), A generalized classical method of linear estimation of coefficients in a structural equation, *Econometrica*, **25**, 77–83. [12.8]

Basmann, R. L. (1961), A note on the exact finite sample frequency functions of generalized classical linear estimators in two leading overidentified cases, *Journal of the American Statistical Association*, **56**, 619–636. [12.8]

Basmann, R. L. (1963), A note on the exact finite sample frequency functions of generalized classical linear estimators in a leading three-equation case, *Journal of the American Statistical Association*, **58**, 161–171. [12.8]

Bennett, B. M. (1951), Note on a solution of the generalized Behrens–Fisher problem, *Annals of the Institute of Statistical Mathematics*, **2**, 87–90. [5.5]

Berger, J. O. (1975), Minimax estimation of location vectors for a wide class of densities, *Annals of Statistics*, **3**, 1318–1328. [3.5]

Berger, J. O. (1976), Admissibility results for generalized Bayes estimators of coordinates of a location vector, *Annals of Statistics*, **4**, 334–356. [3.5]

Berger, J. O. (1980a), A robust generalized Bayes estimator and confidence region for a multivariate normal mean, *Annals of Statistics*, **8**, 716–761. [3.5]

Berger, J. O. (1980b), *Statistical Decision Theory, Foundations, Concepts and Methods*, Springer-Verlag, New York. [3.4, 6.2, 6.7]

Berndt, Ernst R., and Eugene Savin (1977), Conflict among criteria for testing hypotheses in the multivariate linear regression model, *Econometrica*, **45**, 1263–1277. [8.6]

Bhattacharya, P. K. (1966), Estimating the mean of a multivariate normal population with general quadratic loss function, *Annals of Mathematical Statistics*, **37**, 1819–1824. [3.5]

Birnbaum, Allan (1955), Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests, *Annals of Mathematical Statistics*, **26**, 21–36. [5.6, 8.10]

Björck, A., and G. Golub (1973), Numerical methods for computing angles between linear subspaces, *Mathematics of Computation*, **27**, 579–594. [12.3]

Blackwell, David, and M. A. Girshick (1954), *Theory of Games and Statistical Decisions*, John Wiley & Sons, New York. [6.2, 6.7]

Blalock, H. M., Jr (ed.) (1971), *Causal Models in the Social Sciences*, Aldine-Atheston, Chicago. [15.1]

Bonnesen, T., and W. Fenchel (1948), *Theorie der Konvexen Körper*, Chelsea, New York. [8.10]

Bose, R. C. (1936a), On the exact distribution and moment-coefficients of the D^2-statistic, *Sankyā*, **2**, 143–154. [3.3]

Bose, R. C. (1936b), A note on the distribution of differences in mean values of two samples drawn from two multivariate normally distributed populations, and the definition of the D^2-statistic, *Sankhyā*, **2**, 379–384. [3.3]

Bose, R. C., and S. N. Roy (1938), The distribution of the studentised D^2-statistic, *Sankhyā*, **4**, 19–38. [5.4]

Bowker, A. H. (1960), A representation of Hotelling's T^2 and Anderson's classification statistic W in terms of simple statistics, *Contributions to Probability and Statistics* (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.), Stanford University, Stanford, California, 142–149. [5.2]

Bowker, A. H., and R. Stigreaves (1961), An asymptotic expansion for the distribution function of the W-classification statistic, *Studies in Item Analysis and Prediction* (Herbert Solomon, ed.), Stanford University, Stanford, California, 293–310. [6.6]

Box, G. E. P. (1949), A general distribution theory for a class of likelihood criteria, *Biometrika*, **36**, 317–346. [8.5, 9.4, 10.4, 10.5]

Bravais, Auguste (1846), Analyse mathématique sur les probabilités des erreurs de situation d'un point, *Mémoires Présentés par Divers Savants à l'Académie Royale des Sciences de l'Institut de France*, **9**, 255–332. [1.2]

Brown, G. W. (1939), On the power of the L_1 test for equality of several variances, *Annals of Mathematical Statistics*, **10**, 119–128. [10.2]

Browne, M. W. (1974), Generalized least squares estimates in the analysis of covariance structures, *South African Statistical Journal*, **8**, 1–24. [14.3]

Chambers, John M. (1977), *Computational Methods for Data Analysis*, John Wiley & Sons, New York. [12.3]

Chan, Tony F., Gene H. Golub, and Randall J. LeVeque (1981), Algorithms for computing the sample variance: analysis and recommendations, unpublished. [3.2]

Chernoff, Herman (1956), Large sample theory: parametric case, *Annals of Mathematical Statistics*, **27**, 1–22. [8.6]

Chernoff, Herman, and N. Divinsky (1953), The computation of maximum likelihood estimates of linear structural equations, *Studies in Econometric Method* (W. C. Hood and T. C. Koopmans, eds.), John Wiley & Sons, Inc., New York. [12.8]

Chu, S. Sylvia, and K. C. S. Pillai (1979), Power comparisons of two-sided tests of equality of two covariance matrices based on six criteria, *Annals of the Institute of Statistical Mathematics*, **31**, 185–205. [10.6]

Chung, Kai Lai (1974), *A Course in Probability Theory*, 2nd ed., Academic, New York. [2.2]

Clemm, D. S., P. R. Krishnaiah, and V. B. Waikar (1973), Tables for the extreme roots of the Wishart matrix, *Journal of Statistical Computation and Simulation*, **2**, 65–92. [10.8]

Clunies-Ross, C. W., and R. H. Riffenburgh (1960), Geometry and linear discrimination, *Biometrika*, **47**, 185–189. [6.10]

Cochran, W. G. (1934), The distribution of quadratic forms in a normal system, with applications to the analysis of covariance, *Proceedings of the Cambridge Philosophical Society*, **30**, 178–191. [7.4]

Constantine, A. G. (1963), Some non-central distribution problems in multivariate analysis, *Annals of Mathematical Statistics*, **34**, 1270–1285. [12.4]

Constantine, A. G. (1966), The distribution of Hotelling's generalised T_0^2, *Annals of Mathematical Statistics*, **37**, 215–225. [8.6]

Consul, P. C. (1966), On the exact distributions of the likelihood ratio criteria for testing linear hypotheses about regression coefficients, *Annals of Mathematical Statistics*, **37**, 1319–1330. [8.4]

Consul, P. C. (1967a), On the exact distributions of likelihood ratio criteria for testing independence of sets of variates under the null hypothesis, *Annals of Mathematical Statistics*, **38**, 1160–1169. [9.3]

Consul, P. C. (1967b), On the exact distributions of the criterion W for testing sphericity and in a p-variate normal distribution, *Annals of Mathematical Statistics*, **38**, 1170–1174. [10.7]

Courant, R., and D. Hilbert (1953), *Methods of Mathematical Physics*, Interscience, New York. [9.10]

Cox, D. R., and N. Wermuth (1996), *Multivariate Dependencies*, Chapman and Hall, London. [15.1]

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University, Princeton. [2.6, 3.2, 3.4, 6.5, 7.2]

Dahm, P. Fred, and Wayne A. Fuller (1981), Generalized least squares estimation of the functional multivariate linear errors in variables model, unpublished. [14.3]

Daly, J. F. (1940), On the unbiased character of likelihood-ratio tests for independence in normal systems, *Annals of Mathematical Statistics*, **11**, 1–32. [9.2]

Darlington, R. B., S. L. Weinberg, and H. J. Walberg (1973), Canonical variate analysis and related techniques, *Review of Educational Research*, **43**, 433–454. [12.2]

Das Gupta, Somesh (1965), Optimum classification rules for classification into two multivariate normal populations, *Annals of Mathematical Statistics*, **36**, 1174–1184. [6.6]

Das Gupta, S., T. W. Anderson, and G. S. Mudholkar (1964), Monotonicity of the power functions of some tests of the multivariate linear hypothesis, *Annals of Mathematical Statistics*, **35**, 200–205. [8.10]

David, F. N. (1937), A note on unbiased limits for the correlation coefficient, *Biometrika*, **29**, 157–160. [4.2]

David, F. N. (1938), *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, Cambridge University, Cambridge. [4.2]

Davis, A. W. (1968), A system of linear differential equations for the distribution of Hotelling's generalized T_0^2, *Annals of Mathematical Statistics*, **39**, 815–832. [8.6]

Davis, A. W. (1970a), Exact distributions of Hotelling's generalized T_0^2, *Biometrika*, **57**, 187–191. [Preface, 8.6]

Davis, A. W. (1970b), Further applications of a differential equation for Hotelling's generalized T_0^2, *Annals of the Institute of Statistical Mathematics*, **22**, 77–87. [Preface, 8.6]

Davis, A. W. (1971), Percentile approximations for a class of likelihood ratio criteria, *Biometrika*, **58**, 349–356. [10.8, 10.9]

Davis, A. W. (1972a), On the marginal distributions of the latent roots of the multivariate beta matrix, *Annals of Mathematical Statistics*, **43**, 1664–1670. [8.6]

Davis, A. W. (1972b), On the distributions of the latent roots and traces of certain random matrices, *Journal of Multivariate Analysis*, **2**, 189–200. [8.6]

Davis, A. W. (1980), Further tabulation of Hotelling's generalized T_0^2, *Communications in Statistics*, **B9**, 321–336. [Preface, 8.6]

Davis, A. W., and J. B. F. Field (1971), Tables of some multivariate test criteria, Technical Report No. 32, Division of Mathematical Statistics, C.S.I.R.O., Canberra, Australia. [10.8]

Davis, Harold T. (1933), *Tables of the Higher Mathematical Functions*, Vol. I, Principia Press, Bloomington, Indiana. [8.5]

Davis, Harold T. (1935), *Tables of the Higher Mathematical Functions*, Vol. II, Principia Press, Bloomington, Indiana. [8.5]

Deemer, Walter L., and Ingram Olkin (1951), The Jacobians of certain matrix transformations useful in multivariate analysis. Based on lectures of P. L. Hsu at the University of North Carolina, 1947, *Biometrika*, **38**, 345–367. [13.2]

De Groot, Morris H. (1970), *Optimal Statistical Decisions*, McGraw-Hill, New York. [3.4, 6.2, 6.7]

Dempster, A. P. (1972), Covariance selection, *Biometrics*, **28**, 157–175. [15.5]

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38. [14.3]

Diaconis, Persi, and Bradley Efron (1983), Computer-intensive methods in statistics, *Scientific American*, **248**, 116–130. [4.3]

Eaton, M. L., and M. D. Perlman (1974), A monotonicity property of the power functions of some invariant tests for MANOVA, *Annals of Statistics*, **2**, 1022–1028. [8.10]

Edwards, D. (1995), *Introduction to Graphical Modelling*, Springer-Verlag, New York. [15.1]

Efron, Bradley (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia. [4.2]

Efron, Bradley, and Carl Morris (1977), Stein's paradox in statistics, *Scientific American*, **236**, 119–127. [3.5]

Elfving, G. (1947), A simple method of deducing certain distributions connected with multivariate sampling, *Skandinavisk Aktuarietidskrift*, **30**, 56–74. [7.2]

Fang, Kai-Tai, Samuel Kotz, and Kai-Wang Ng (1990), *Symmetric Multivariate and Related Distributions*, Chapman and Hall, New York. [2.7]

Fang, Kai-Tai, and Yaó-Ting Zhang (1990), *Generalized Multivariate Analysis*, Springer-Verlag, New York. [2.7, 3.6, 10.11, 7.9]

Ferguson, Thomas Shelburne (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Academic, New York. [3.4, 6.2, 6.7]

Fisher, R. A. (1915), Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika*, **10**, 507–521. [4.2, 7.2]

Fisher, R. A. (1921), On the "probable error" of a coefficient of correlation deduced from a small sample, *Metron*, **1**, Part 4, 3–32. [4.2]

Fisher, R. A. (1924), The distribution of the partial correlation coefficient, *Metron*, **3**, 329–332. [4.3]

Fisher, R. A. (1928), The general sampling distribution of the multiple correlation coefficient, *Proceedings of the Royal Society of London*, *A*, **121**, 654–673. [4.4]

Fisher, R. A. (1936), The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7, 179–188. [5.3, 6.5, 11.5]

Fisher, R. A. (1939), The sampling distribution of some statistics obtained from non-linear equations, *Annals of Eugenics*, **9**, 238–249. [13.2]

Fisher, R. A. (1947a), *The Design of Experiments* (4th ed.), Oliver and Boyd, Edinburgh. [8.9]

Fisher, R. A. (1947b), The analysis of covariance method for the relation between a part and the whole, *Biometrics*, **3**, 65–68. [3.P]

Fisher, R. A., and F. Yates (1942), *Statistical Tables for Biological, Agricultural and Medical Research* (2nd ed.), Oliver and Boyd, Edinburgh. [4.2]

Fog, David (1948), The geometrical method in the theory of sampling, *Biometrika*, **35**, 46–54. [7.2]

Foster, F. G. (1957), Upper percentage points of the generalized beta distribution, II, *Biometrika*, **44**, 441–453. [8.6]

Foster, F. G. (1958), Upper percentage points of the generalized beta distribution, III, *Biometrika*, **45**, 492–503. [8.6]

Foster, F. G., and D. H. Rees (1957), Upper percentage points of the generalized beta distribution, I, *Biometrika*, **44**, 237–247. [8.6]

Frets, G. P. (1921), Heredity of head form in man, *Genetica*, **3**, 193–384. [3.P]

Frisch, R. (1929), Correlation and scatter in statistical variables, *Nordic Statistical Journal*, **8**, 36–102. [7.5]

Frydenberg, M. (1990), The chain graph Markov property, *Scandinavian Journal of Statistics*, **17**, 333–353. [15.2]

Fujikoshi, Y. (1973), Monotonicity of the power functions of some tests in general MANOVA models, *Annals of Statistics*, **1**, 388–391. [8.6]

Fujikoshi, Y. (1974), The likelihood ratio tests for the dimensionality of regression coefficients, *Journal of Multivariate Analysis*, **4**, 327–340. [12.4]

Fujikoshi, Y., and M. Kanazawa (1976), The ML classification statistic in covariate discriminant analysis and its asymptotic expansions, *Essays in Probability and Statistics*, 305–320. [6.6]

Fuller, Wayne A., Sastry, G. Pantula, and Yasuo Amemiya (1982), The covariance matrix of estimators for the factor model, unpublished. [14.4]

Gabriel, K. R. (1969), Simultaneous test procedures—some theory of multiple comparisons, *Annals of Mathematical Statistics*, **40**, 224–250. [8.7]

Gajjar, A. V. (1967), Limiting distributions of certain transformations of multiple correlation coefficients, *Metron*, **26**, 189–193. [4.4]

Galton, Francis (1889), *Natural Inheritance*, MacMillan, London. [1.2, 2.5]

Gauss, K. F. (1823), *Theory of the Combination of Observations*, Göttingen. [1.2]

Giri, N. (1977), *Multivariate Statistical Inference*, Academic, New York. [7.2, 7.7]

Giri, N., and J. Kiefer (1964), Local and asymptotic minimax properties of multivariate tests, *Annals of Mathematical Statistics*, **35**, 21–35. [5.6]

Giri, N., J. Kiefer, and C. Stein (1963), Minimax character of Hotelling's T^2 test in the simplest case, *Annals of Mathematical Statistics*, **34**, 1524–1535. [5.6]

Girshick, M. A. (1939), On the sampling theory of roots of determinantal equations. *Annals of Mathematical Statistics*, **10**, 203–224. [13.2]

Gleser, Leon Jay (1981), Estimation in a multivariate "errors in variables" regression model: large sample results, *Annals of Statistics*, **9**, 24–44. [12.7]

Glynn, W. J., and R. J. Muirhead (1978), Inference in canonical correlation analysis. *Journal of Multivariate Analysis*, **8**, 468–478. [12.4]

Golub, Gene H., and Franklin T. Luk (1976), Singular value decomposition: applications and computations, unpublished. [12.3]

Golub, Gene H., and Charles F. Van Loan (1989), *Matrix Computations* (2nd ed.). Johns Hopkins University Press, Baltimore. [11.4, 12.3, A.5]

Grubbs, F. E. (1954), Tables of 1% and 5% probability levels of Hotelling's generalized T^2 statistics, Technical Note No. 926, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland. [8.6]

Gupta, Shanti S. (1963), Bibliography on the multivariate normal integrals and related topics, *Annals of Mathematical Statistics*, **34**, 829–838. [2.3]

Gurland, John (1968), A relatively simple form of the distribution of the multiple correlation coefficient, *Journal of the Royal Statistical Society B*, **30**, 276–283. [4.4]

Gurland, J., and R. Milton (1970), Further consideration of the distribution of the multiple correlation coefficient, *Journal of the Royal Statistical Society B*, **32**, 381–394. [4.4]

Haavelmo, T. (1944), The probability approach in econometrics, *Econometrica*, **12**, Supplement, 1–118. [12.7]

Haff, L. R. (1980), Empirical Bayes estimation of the multivariate normal covariance matrix, *Annals of Statistics*, **8**, 586–597. [7.8]

Halmos, P. R. (1950), *Measure Theory*, D. van Nostrand, New York. [4.5, 13.3]

Harris, Bernard, and Andrew P. Soms (1980). The use of the tetrachoric series for evaluating multivariate normal probabilities, *Journal of Multivariate Analysis*, **10**, 252–267. [2.3]

Hayakawa, Takesi (1967), On the distribution of the maximum latent root of a positive definite symmetric random matrix, *Annals of the Institute of Statistical Mathematics*, **19**, 1–17. [8.6]

Heck, D. L. (1960), Charts of some upper percentage points of the distribution of the largest characteristic root, *Annals of Mathematical Statistics*, **31**, 625–642. [8.6]

Hickman, W. Braddock (1953), *The Volume of Corporate Bond Financing Since 1900*, Princeton University, Princeton, 82–90. [10.7]

Hoel, Paul G. (1937), A significance test for component analysis, *Annals of Mathematical Statistics*, **8**, 149–158. [7.5]

Hooker, R. H. (1907), The correlation of the weather and crops, *Journal of the Royal Statistical Society*, **70**, 1–42. [4.2]

Hotelling, Harold (1931), The generalization of Student's ratio. *Annals of Mathematical Statistics*, **2**, 360–378. [5.1, 5.P]

Hotelling, Harold (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**, 417–441, 498–520. [11.1, 14.3]

Hotelling, Harold (1935), The most predictable criterion, *Journal of Educational Psychology*, **26**, 139–142. [12.2]

Hotelling, Harold (1936), Relations between two sets of variates, *Biometrika*, **28**, 321–377. [12.1]

Hotelling, Harold (1947), Multivariate quality control, illustrated by the air testing of sample bombsights. *Techniques of Statistical Analysis* (C. Eisenhart, M. Hastay, and W. A. Wallis, eds.). McGraw-Hill, New York, 111–184. [8.6]

Hotelling, Harold (1951), A generalized T test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Jerzy Neyman, ed.), University of California, Los Angeles and Berkeley, 23–41. [8.6, 10.7]

Hotelling, Harold (1953), New light on the correlation coefficient and its transforms (with discussion), *Journal of the Royal Statistical Society B*, **15**, 193–232. [4.2]

Howe, W. G. (1955). *Some Contributions to Factor Analysis*, U.S. Atomic Energy Commission Report. Oak Ridge National Laboratory, Oak Ridge, Tennessee. [14.2, 14.6]

Hsu, P. L. (1938), Notes on Hotelling's generalized T, *Annals of Mathematical Statistics*, **9**, 231–243. [5.4]

Hsu, P. L. (1939a), On the distribution of the roots of certain determinantal equations, *Annals of Eugenics*, **9**, 250–258. [13.2]

Hsu, P. L. (1939b), A new proof of the joint product moment distribution, *Proceedings of the Cambridge Philosophical Society*, **35**, 336–338. [7.2]

Hsu, P. L. (1945), On the power functions for the E^2-test and the T^2-test, *Annals of Mathematical Statistics*, **16**, 278–286. [5.6]

Hudson, M. (1974), Empirical Bayes estimation, Technical Report No. 58, NSF contract GP 30711X-2, Department of Statistics, Stanford University. [3.5]

Immer, F. R., H. D. Hayes, and LeRoy Powers (1934), Statistical determination of barley varietal adaptation, *Journal of the American Society of Agronomy*, **26**, 403–407. [8.9]

Ingham, A. E. (1933), An integral which occurs in statistics, *Proceedings of the Cambridge Philosophical Society*, **29**, 271–276. [7.2]

Ito, K. (1956), Asymptotic formulae for the distribution of Hotelling's generalized T_0^2 statistic, *Annals of Mathematical Statistics*, **27**, 1091–1105. [8.6]

Ito, K. (1960), Asymptotic formulae for the distribution of Hotelling's generalized T_0^2 statistic, II, *Annals of Mathematical Statistics*, **31**, 1148–1153. [8.6]

Izenman, A. J. (1975), Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis*, **5**, 248–264. [12.7]

Izenman, Alan Julian (1980), Assessing dimensionality in multivariate regression, *Analysis of Variance, Handbook of Statistics*, Vol. 1 (P. R. Krishnaiah, ed.), North-Holland, Amsterdam, 571–591. [3.P]

James, A. T. (1954), Normal multivariate analysis and the orthogonal group, *Annals of Mathematical Statistics*, **25**, 40–75. [7.2]

James, A. T. (1964), Distributions of matrix variates and latent roots derived from normal samples, *Annals of Mathematical Statistics*, **35**, 475–501. [8.6]

James, W., and C. Stein (1961), Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Jerzy Neyman, ed.), Vol. **I**, 361–379. University of California, Berkeley. [3.5, 7.8]

Japanese Standards Association (1972). *Statistical Tables and Formulas with Computer Applications.* [Preface]

Jennrich, Robert I., and Dorothy T. Thayer (1973), A note on Lawley's formulas for standard errors in maximum likelihood factor analysis, *Psychometrika*, **38**, 571–580. [14.3]

Johansen, S. (1995), *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press. [12.7]

Jolicoeur, Pierre, and J. E. Mosimann (1960), Size and shape variation in the painted turtle, a principal component analysis, *Growth*, **24**, 339–354. Also in *Benchmark Papers in Systematic and Evolutionary Biology* (E. H. Bryant and W. R. Atchley, eds.), **2** (1975), 86–101. [11.P]

Jöreskog, K. G. (1969) A general approach to confirmatory maximum likelihood factor analysis, *Psychometrika*, **34**, 183–202. [14.2]

Jöreskog, K. G., and Arthur S. Goldberger (1972), Factor analysis by generalized least squares, *Psychometrika*, **37**, 243–260. [14.3]

Kaiser, Henry F. (1958), The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, **23**, 187–200. [14.5]

Kanazawa, M. (1979), The asymptotic cut-off point and comparison of error probabilities in covariate discriminant analysis, *Journal of the Japan Statistical Society*, **9**, 7–17. [6.6]

Kelley, T. L. (1928), *Crossroads in the Mind of Man*, Stanford University, Stanford. [4.P, 9.P]

Kendall, M. G., and Alan Stuart (1973), *The Advanced Theory of Statistics* (3rd ed.), Vol. 2, Charles Griffin, London. [12.6]

Kennedy, William J., Jr., and James E. Gentle (1980), *Statistical Computing*, Marcel Dekker, New York. [12.3]

Khatri, C. G. (1963), Joint estimation of the parameters of multivariate normal populations, *Journal of Indian Statistical Association*, **1**, 125–133. [7.2]

Khatri, C. G. (1966), A note on a large sample distribution of a transformed multiple correlation coefficient, *Annals of the Institute of Statistical Mathematics*, **18**, 375–380. [4.4]

Khatri, C. G. (1972), On the exact finite series distribution of the smallest or the largest root of matrices in three situations, *Journal of Multivariate Analysis*, **2**, 201–207. [8.6]

Khatri, C. G., and K. C. Sreedharan Pillai (1966), On the moments of the trace of a matrix and approximations to its non-central distribution, *Annals of Mathematical Statistics*, **37**, 1312–1318. [8.6]

Khatri, C. G., and K. C. S. Pillai (1968), On the non-central distributions of two test criteria in multivariate analysis of variance, *Annals of Mathematical Statistics*, **39**, 215–226. [8.6]

Khatri, C. G., and K. V. Ramachandran (1958), Certain multivariate distribution problems, I (Wishart's distribution), *Journal of the Maharaja Sayajairo, University of Baroda*, **7**, 79–82. [7.2]

Kiefer, J. (1957), Invariance, minimax sequential estimation, and continuous time processes, *Annals of Mathematical Statistics*, **28**, 573–601. [7.8]

Kiefer, J. (1966), Multivariate optimality results, *Multivariate Analysis* (Parachuyi **R**. Krishnaiah, ed.), Academic, New York, 255–274. [7.8]

Kiefer, J., and R. Schwartz (1965), Admissible Bayes character of T^2-, R^2-, and other fully invariant tests for classical multivariate normal problems, *Annals of Mathematical Statistics*, **36**, 747–770. [5.P, 9.9, 10.10]

Klotz, Jerome, and Joseph Putter (1969), Maximum likelihood estimation of the multivariate covariance components for the balanced one-way layout. *Annals of Mathematical Statistics*, **40**, 1100–1105. [10.6]

Kolmogorov, A. (1950), *Foundations of the Theory of Probability*, Chelsea, New York. [2.2]

Konishi, Sadanori (1978a), An approximation to the distribution of the sample correlation coefficient, *Biometrika*, **65**, 654–656. [4.2]

Konishi, Sadanori (1978b), Asymptotic expansions for the distributions of statistics based on a correlation matrix, *Canadian Journal of Statistics*, **6**, 49–56. [4.2]

Konishi, Sadanori (1979), Asymptotic expansions for the distributions of functions of a correlation matrix, *Journal of Multivariate Analysis*, 9, 259–266. [4.2]

Koopmans, T. C., and Olav Reiersøl (1950), The identification of structural characteristics, *Annals of Mathematical Statistics*, **21**, 165–181. [14.2]

Korin, B. P. (1968), On the distribution of a statistic used for testing a covariance matrix, *Biometrika*, **55**, 171–178. [10.8]

Korin, B. P. (1969), On testing the equality of k covariance matrices, *Biometrika*, **56**, 216–218. [10.5]

Kramer, K. H. (1963), Tables for constructing confidence limits on the multiple correlation coefficient, *Journal of the American Statistical Association*, **58**, 1082–1085. [4.4]

Krishnaiah, P. R. (1978), Some recent developments on real multivariate distributions, *Development in Statistics* (P. R. Krishnaiah, ed.), Vol. 1, Academic, New York, 135–169. [8.6]

Krishnaiah, P. R. (1980), Computations of some multivariate distributions, *Analysis of Variance, Handbook of Statistics*, Vol. 1 (P. R. Krishnaiah, ed.), North-Holland, Amsterdam, 745–971.

Krishnaiah, P. R., and T. C. Chang (1972), On the exact distributions of the traces of $S_1(S_1+S_2)^{-1}$ and $S_1S_2^{-1}$, *Sankhyā*, *A*, **34**, 153–160. [8.6]

Krishnaiah, P. R., and F. J. Schuurmann (1974), On the evaluation of some distributions that arise in simultaneous tests for the equality of the latent roots of the covariance matrix, *Journal of Multivariate Analysis*, **4**, 265–282. [10.7]

Kshirsagar, A. M. (1959), Bartlett decomposition and Wishart distribution, *Annals of Mathematical Statistics*, **30**, 239–241. [7.2]

Kudo, H. (1955), On minimax invariant estimates of the transformation parameter, *Natural Science Report*, **6**, 31–73, Ochanomizu University, Tokyo, Japan. [7.8]

Kunitomo, Naoto (1980), Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations, *Journal of the American Statistical Association*, 75, 693–700. [12.7]

Lachenbruch, P. A., and M. R. Mickey (1968), Estimation of error rates in discriminant analysis, *Technometrics*, **10**, 1–11. [6.6]

Laplace, P. S. (1811), Mémoire sur les intégrales définies et leur application aux probabilités, *Mémoires de l'Institut Impérial de France, Année 1810*, 279–347. [1.2]

Lauritzen, Steffen L. (1996), *Graphical Models*, Clarendon Press, Oxford. [15.5]

Lauritzen, Steffen L., and N. Wermuth (1989), Graphical models for associations between variables some of which are qualitative and some quantitative, *Annals of Statistics*, **17**, 31–57. [15.2]

Läuter, Jürgen, Ekkehard Glimm, and Siegfried Kropf (1996a), New multivariate tests for data with an inherent structure, *Biometrics Journal*, **38**, 5–23. [Correction: **40**. (1998), 1015.] [5.7]

Läuter, Jürgen, Ekkehard Glimm, and Siegfried Kropf (1996b), Multivariate tests based on left-spherically distributed linear scores, *Annals of Statistics*. **26**. 1972–1988. [5.7]

Läuter, Jürgen, Ekkehard Glimm, and Siegfried Kropf (1996c). Exact stable multivariate tests for applications in clinical research. *ASA Proceedings of the Biopharmaceutical Section*, 46–55. [5.7]

Lawley, D. N. (1938), A generalization of Fisher's z test, *Biometrika*, **30**, 180–187. [8.6]

Lawley, D. N. (1940), The estimation of factor loadings by the method of maximum likelihood, *Proceedings of the Royal Society of Edinburgh, Sec. A*, **60**, 64–82. [14.3]

Lawley, D. N. (1941), Further investigations in factor estimation, *Proceedings of the Royal Society of Edinburgh, Sec. A*, **61**, 176–185. [14.4]

Lawley, D. N. (1953), A modified method of estimation in factor analysis and some large sample results, *Uppsala Symposium on Psychological Factor Analysis. 17–19 March 1953*, Uppsala, Almqvist and Wiksell, 35–42. [14.3]

Lawley, D. N. (1958), Estimation in factor analysis under various initial assumptions. *British Journal of Statistical Psychology*, **11**, 1–12. [14.2, 14.6]

Lawley, D. N. (1959), Tests of significance in canonical analysis, *Biometrika*. **46**. 59–66. [12.4]

Lawley, D. N. (1967), Some new results in maximum likelihood factor analysis. *Proceedings of the Royal Society of Edinburgh, Sec. A*, **87**. 256–264. [14.3]

Lawley, D. N., and A. E. Maxwell (1971), *Factor Analysis as a Statistical Method* (2nd ed.), American Elsevier, New York. [14.3, 14.5]

Lee, Y. S. (1971a), Asymptotic formulae for the distribution of a multivariate test statistic: power comparisons of certain multivariate tests. *Biometrika*. **58**, 647–651. [8.6]

Lee, Y. S. (1971b), Some results on the sampling distribution of the multiple correlation coefficient, *Journal of the Royal Statistical Society B*, **33**. 117–130. [4.4]

Lee, Y. S. (1972), Tables of upper percentage points of the multiple correlation coefficient, *Biometrika*, **59**, 175–189. [4.4]

Lehmann, E. L. (1959), *Testing Statistical Hypotheses*, John Wiley & Sons, New York. [4.2, 5.6]

Lehmer, Emma (1944), Inverse tables of probabilities of errors of the second kind. *Annals of Mathematical Statistics*, **15**, 388–398. [5.4]

Loève, M. (1977), *Probability Theory I* (4th ed.), Springer-Verlag, New York. [2.2]

Loève, M. (1978), *Probability Theory II* (4th ed.), Springer-Verlag, New York. [2.2]

Madow, W. G. (1938), Contributions to the theory of multivariate statistical analysis, *Transactions of the American Mathematical Society*, **44**, 454–495. [7.2]

Magnus, Jan R. (1988), *Linear Structures*, Charles Griffin and Co., London. [A.4]

Magnus, J. R., and H. Neudecker (1979), The commutation matrix: some properties and applications, *The Annals of Statistics*, 7, 381–394. [3.6]

Mahalanobis, P. C. (1930), On tests and measures of group divergence, *Journal and Proceedings of the Asiatic Society of Bengal*, **26**, 541–588. [3.3]

Mahalanobis, P. C., R. C. Bose, and S. N. Roy (1937), Normalisation of statistical variates and the use of rectangular co-ordinates in the theory of sampling distributions, *Sankhyā*, **3**, 1–40. [7.2]

Mallows, C. L. (1961), Latent vectors of random symmetric matrices, *Biometrika*, **48**, 133–149. [11.6]

Mardia, K. V. (1970), Measures of multivariate skewness and kurtosis with applications, *Biometrika*, 57, 519–530. [3.6]

Mariano, Roberto S., and Takamitsu Sawa (1972), The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables, *Journal of the American Statistical Association*, **67**, 159–163. [12.7]

Maronna, R. A. (1976), Robust M-estimators of multivariate location and scatter, *Annals of Statistics*, 4, 51–67. [3.6]

Marshall, A. W., and I. Olkin (1979), *Inequalities: Theory of Majorization and Its Applications*, Academic, New York. [8.10]

Mathai, A. M. (1971), On the distribution of the likelihood ratio criterion for testing linear hypotheses on regression coefficients, *Annals of the Institute of Statistical Mathematics*, **23**, 181–197. [8.4]

Mathai, A. M., and R. S. Katiyar (1979), Exact percentage points for testing independence, *Biometrika*, **66**, 353–356. [9.3]

Mathai, A. M., and P. N. Rathie (1980), The exact non-null distribution for testing equality of covariance matrices, *Sankhyā A*, **42**, 78–87. [10.4]

Mathai, A. M., and R. K. Saxena (1973), *Generalized Hypergeometric Functions with Applications in Statistics and Physical Sciences*, Lecture Notes No. 348, Springer-Verlag, New York. [9.3]

Mauchly, J. W. (1940), Significance test for sphericity of a normal n-variate distribution, *Annals of Mathematical Statistics*, **11**, 204–209. [10.7]

Mauldon, J. G. (1955), Pivotal quantities for Wishart's and related distributions, and a paradox in fiducial theory, *Journal of the Royal Statistical Society B*, **17**, 79–85. [7.2]

McDonald, Roderick P. (2002), What can we learn from path equations: identifiability, constraints, equivalence, *Psychometrika*, **67**, 225–249. [15.1]

McLachlan, G. J. (1973), An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis, *Australian Journal of Statistics*, **15**, 210–214. [6.6]

McLachlan, G. J. (1974a), An asymptotic unbiased technique for estimating the error rates in discriminant analysis, *Biometrics*, **30**, 239–249. [6.6]

McLachlan, G. J. (1974b), Estimation of the errors of misclassification on the criterion of asymptotic mean square error, *Technometrics*, **16**, 255–260. [6.6]

McLachlan, G. J. (1974c), The asymptotic distributions of the conditional error rate and risk in discriminant analysis, *Biometrika*, **61**, 131–135. [6.6]

McLachlan, G. J. (1977), Constrained sample discrimination with the studentized classification statistic W, *Communications in Statistics—Theory and Methods*, **A6**, 575–583 [6.6]

Memon, A. Z., and M. Okamoto (1971), Asymptotic expansion of the distribution of The Z statistic in discriminant analysis, *Journal of Multivariate Analysis*, **1**, 294–307. [6.6]

Mijares, T. A. (1964), *Percentage Points of the Sum* $V_1^{(s)}$ *of s Roots* ($s = 1 - 50$), The Statistical Center, University of the Philippines, Manila, [8.6]

Mikhail, N. N. (1965), A comparison of tests of the Wilks–Lawley hypothesis in multivariate analysis, *Biometrika*, **52**, 149–156. [8.6]

Mood, A. M. (1951), On the distribution of the characteristic roots of normal second-moment matrices, *Annals of Mathematical Statistics*, **22**, 266–273. [13.2]

Morris, Blair, and Ingram Olkin (1964), Some estimation and testing problems for factor analysis models, unpublished. [10.6]

Mudholkar, G. S. (1966), On confidence bounds associated with multivariate analysis of variance and non-independence between two sets of variates, *Annals of Mathematical Statistics*, **37**, 1736–1746. [8.7]

Mudholkar, Govind S., and Madhusudan C. Trivedi (1980), A normal approximation for the distribution of the likelihood ratio statistic in multivariate analysis of variance, *Biometrika*, **67**, 485–488. [8.5]

Mudholkar, Govind S., and Madhusudan C. Trivedi (1981), A normal approximation for the multivariate likelihood ratio statistics, *Statistical Distributions in Scientific Work* (C. Taillie et al., eds.), Vol. 5, 219–230, D. Reidel Publishing. [8.5]

Muirhead, R. J. (1970), Asymptotic distributions of some multivariate tests, *Annals of Mathematical Statistics*, **41**, 1002–1010. [8.6]

Muirhead, R. J. (1980), The effects of elliptical distributions on some standard procedures involving correlation coefficients: a review, *Multivariate Statistical Analysis* (R. P. Gupta, ed.), 143–159. [4.5]

Muirhead, Robb J. (1982), *Aspects of Multivariate Statistical Theory*, John Wiley and Sons, New York. [2.7, 3.6, 7.7]

Muirhead, R. J., and C. M. Waternaux (1980), Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations, *Biometrika*, **67**, 31–43. [4.5]

Nagao, Hisao (1973a), On some test criteria for covariance matrix, *Annals of Statistics*, **1**, 700–709. [9.5, 10.2, 10.7, 10.8]

Nagao, Hisao (1973b), Asymptotic expansions of the distributions of Bartlett's test and sphericity test under the local alternatives, *Annals of the Institute of Statistical Mathematics*, **25**, 407–422. [10.5, 10.6]

Nagao, Hisao (1973c), Nonnull distributions of two test criteria for independence under local alternatives, *Journal of Multivariate Analysis*, **3**, 435–444. [9.4]

Nagarsenker, B. N., and K. C. S. Pillai (1972), *The Distribution of the Sphericity Test Criterion*, ARL 72-0154, Aerospace Research Laboratories. [Preface]

Nagarsenker, B. N., and K. C. S. Pillai (1973a), The distribution of the sphericity test criterion, *Journal of Multivariate Analysis*, **3**, 226–235. [10.7]

Nagarsenker, B. N., and K. C. S. Pillai (1973b), Distribution of the likelihood ratio criterion for testing a hypothesis specifying a covariance matrix, *Biometrika*, **60**, 359–364. [10.8]

Nagarsenker, B. N., and K. C. S. Pillai (1974), Distribution of the likelihood ratio criterion for testing $\Sigma = \Sigma_0, \mu = \mu_0$, *Journal of Multivariate Analysis*, **4**, 114–122. [10.9]

Nanda, D. N. (1948), Distribution of a root of a determinantal equation, *Annals of Mathematical Statistics*, **19**, 47–57. [8.6]

Nanda, D. N. (1950), Distribution of the sum of roots of a determinantal equation under a certain condition, *Annals of Mathematical Statistics*, **21**, 432–439. [8.6]

Nanda, D. N. (1951), Probability distribution tables of the larger root of a determinantal equation with two roots, *Journal of the Indian Society of Agricultural Statistics*, **3**, 175–177. [8.6]

Narain, R. D. (1948), A new approach to sampling distributions of the multivariate normal theory, I, *Journal of the Indian Society of Agricultural Statistics*, **1**, 59–69. [7.2]

Narain, R. D. (1950), On the completely unbiased character of tests of independence in multivariate normal systems, *Annals of Mathematical Statistics*, **21**, 293–298. [9.2]

National Bureau of Standards, United States (1959), *Tables of the Bivariate Normal Distribution Function and Related Functions*, U.S. Government Printing Office, Washington, D.C. [2.3]

Neveu, Jacques (1965), *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco. [2.2]

Ogawa, J. (1953), On the sampling distributions of classical statistics in multivariate analysis, *Osaka Mathematics Journal*, **5**, 13–52. [7.2]

Okamoto, Masashi (1963), An asymptotic expansion for the distribution of the linear discriminant function, *Annals of Mathematical Statistics*, **34**, 1286–1301. (Correction, **39** (1968), 1358–1359.) [6.6]

Okamoto, Masashi (1973), Distinctness of the eigenvalues of a quadratic form in a multivariate sample, *Annals of Statistics*, **1**, 763–765. [13.2]

Olkin, Ingram, and S. N. Roy (1954), On multivariate distribution theory, *Annals of Mathematical Statistics*, **25**, 329–339. [7.2]

Olson, C. L. (1974), Comparative robustness of six tests in multivariate analysis of variance, *Journal of the American Statistical Association*, **69**, 894–908. [8.6]

Pearl, Judea (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge. [15.1]

Pearson, E. S., and H. O. Hartley (1972), *Biometrika Tables for Statisticians*, Vol. **II**, Cambridge (England), Published for the Biometrika Trustees at the University Press. [Preface, 8.4]

Pearson, E. S., and S. S. Wilks (1933), Methods of statistical analysis appropriate for k samples of two variables, *Biometrika*, **25**, 353–378. [10.5]

Pearson, K. (1896), Mathematical contributions to the theory of evolution—III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society of London, Series A*, **187**, 253–318. [2.5, 3.2]

Pearson, K. (1900), On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine*, **50** (fifth series), 157–175. [3.3]

Pearson, K. (1901), On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, **2** (sixth series), 559–572. [11.2]

Pearson, K. (1930), *Tables for Statisticians and Biometricians*, Part I (3rd ed.), Cambridge University, Cambridge. [2.3]

Pearson, K. (1931), *Tables for Statisticians and Biometricians*, Part II, Cambridge University, Cambridge. [2.3, 6.8]

Perlman, M. D. (1980), Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations, *Annals of Statistics*, **8**, 247–263. [10.2, 10.3]

Perlman, M. D., and I. Olkin (1980), Unbiasedness of invariant tests for MANOVA and other multivariate problems, *Annals of Statistics*, **8**, 1326–1341. [8.10]

Phillips, P. C. B. (1980), The exact distribution of instrumental variable estimators in an equation containing $n+1$ endogenous variables, *Econometrica*, **48**, 861–878. [12.7]

Phillips, P. C. B. (1982), A new approach to small sample theory, unpublished, Cowles Foundation for Research in Economics, Yale University. [12.7]

Pillai, K. C. S. (1954), On some distribution problems in multivariate analysis, Mimeo Series No. 54, Institute of Statistics, University of North Carolina, Chapel Hill, North Carolina. [8.6]

Pillai, K. C. S. (1955), Some new test criteria in multivariate analysis, *Annals of Mathematical Statistics*, **26**, 117–121. [8.6]

Pillai, K. C. S. (1956), On the distribution of the largest or the smallest root of a matrix in multivariate analysis, *Biometrika*, **43**, 122–127. [8.6]

Pillai, K. C. S. (1960), *Statistical Tables for Tests of Multivariate Hypotheses*, Statistical Center, University of the Philippines, Manila. [8.6]

Pillai, K. C. S. (1964), On the moments of elementary symmetric functions of the roots of two matrices, *Annals of Mathematical Statistics*, **35**, 1704–1712. [8.6]

Pillai, K. C. S. (1965), On the distribution of the largest characteristic root of a matrix in multivariate analysis, *Biometrika*, **52**, 405–412. [8.6]

Pillai, K. C. S. (1967), Upper percentage points of the largest root of a matrix in multivariate analysis, *Biometroka*, **54**, 189–194. [8.6]

Pillai, K. C. S., and A. K. Gupta (1969), On the exact distribution of Wilks' criterion, *Biometrika*, **56**, 109–118. [8.4]

Pillai, K. C. S., and Y. S. Hsu (1979), Exact robustness studies of the test of independence based on four multivariate criteria and their distribution problems under violations, *Annals of the Institute of Statistical Mathematics*, **31**, Part A, 85–101. [8.6]

Pillai, K. C. S., and K. Jayachandran (1967), Power comparisons of tests of two multivariate hypotheses based on four criteria, *Biometrika*, **54**, 195–210. [8.6]

Pillai, K. C. S., and K. Jayachandran (1970), On the exact distribution of Pillai's $V^{(s)}$ criterion, *Journal of the American Statistical Association*, **65**, 447–454. [8.6]

Pillai, K. C. S., and T. A. Mijares (1959), On the moments of the trace of a matrix and approximations to its distribution, *Annals of Mathematical Statistics*, **30**, 1135–1140. [8.6]

Pillai, K. C. S., and B. N. Nagarsenker (1971), On the distribution of the sphericity test criterion in classical and complex normal populations having unknown covariance matrices, *Annals of Mathematical Statistics*, **42**, 764–767. [10.7]

Pillai, K. C. S., and P. Samson, Jr. (1959), On Hotelling's generalization of T^2, *Biometrika*, **46**, 160–168. [8.6]

Pillai, K. C. S., and T. Sugiyama (1969), Non-central distributions of the largest latent roots of three matrices in multivariate analysis, *Annals of the Institute of Statistical Mathematics*, **21**, 321–327. [8.6]

Pillai, K. C. S., and D. L. Young (1971), On the exact distribution of Hotelling's generalized T_0^2, *Journal of Multivariate Analysis*, **1**, 90–107. [8.6]

Plana, G. A. A. (1813), Mémoire sur divers problèmes de probabilité, *Mémoires de l'Académie Impériale de Turin, pour les Années 1811–1812*, **20**, 355–408. [1.2]

Pólya, G. (1949), Remarks on computing the probability integral in one and two dimensions, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), 63–78. [2.P]

Rao, C. R. (1948a), The utilization of multiple measurements in problems of biological classification, *Journal of the Royal Statistical Society B*, **10**, 159–193. [6.9]

Rao, C. R. (1948b), Tests of significance in multivariate analysis, *Biometrika*, **35**, 58–79. [5.3]

Rao, C. Radhakrishna (1951), An asymptotic expansion of the distribution of Wilks's criterion, *Bulletin of the International Statistical Institute*, **33**, Part 2, 177–180. [8.5]

Rao, C. R. (1952), *Advanced Statistical Methods in Biometric Research*, John Wiley & Sons, New York. [12.5]

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), John Wiley & Sons, New York. [4.2]

Rasch, G. (1948), A functional equation for Wishart's distribution, *Annals of Mathematical Statistics*, **19**, 262–266. [7.2]

Reiersøl, Olav (1950), On the identifiability of parameters in Thurstone's multiple factor analysis, *Psychometrika*, **15**, 121–149. [14.2]

Reinsel, G. C., and R. P. Velu (1998), *Multivariate Reduced-rank Regression*, Springer, New York. [12.7]

Richardson, D. H. (1968), The exact distribution of a structural coefficient estimator, *Journal of the American Statistical Association*, **63**, 1214–1226. [12.7]

Rothenberg, Thomas J. (1977), Edgeworth expansions for multivariate test statistics, IP-255, Center for Research in Management Science, University of California, Berkeley. [8.6]

Roy, S. N. (1939), p-statistics or some generalisations in analysis of variance appropriate to multivariate problems, *Sankhyā*, **4**, 381–396. [13.2]

Roy, S. N. (1945), The individual sampling distribution of the maximum, the minimum and any intermediate of the p-statistics on the null-hypothesis, *Sankhyā*, **7**, 133–158. [8.6]

Roy, S. N. (1953), On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics*, **24**, 220–238. [8.6, 10.6]

Roy, S. N. (1957), *Some Aspects of Multivariate Analysis*, John Wiley & Sons, New York. [8.6, 10.6, 10.8]

Ruben, Harold (1966), Some new results on the distribution of the sample correlation coefficient, *Journal of the Royal Statistical Society B*, **28**, 513–525. [4.2]

Rubin, Donald B., and Dorothy T. Thayer (1982), EM algorithms for ML factor analysis, *Psychometrika*, **47**, 69–76. [14.3]

Ryan, D. A. J., J. J. Hubert, E. M. Carter, J. B. Sprague, and J. Parrot (1992), A reduced-rank multivariate regression approach to joint toxicity experiments, *Biometrics*, **48**, 155–162. [12.7]

Šalaevskiĭ, O. V. (1968), The minimax character of Hotelling's T^2 test (Russian), *Doklady Akademii Nauk SSSR*, **180**, 1048–1050. [5.6]

Šalaevskiĭ, (Shalaevskii), O. V. (1971), Minimax character of Hotelling's T^2 test. I. *Investigations in Classical Problems of Probability Theory and Mathematical Statistics*, V. M. Kalinin and O. V. Shalaevskii (Seminar in Mathematics, V. I. Steklov Institute, Leningrad, Vol. 13), Consultants Bureau, New York. [5.6]

Sargan, J. D., and W. M. Mikhail (1971), A general approximation to the distribution of instrumental variables estimates, *Econometrica*, **39**, 131–169. [12.7]

Sawa, Takamitsu (1969), The exact sampling distribution of ordinary least squares and two-stage least squares estimators, *Jorunal of the American Statistical Association*, **64**, 923–937. [12.7]

Schatzoff, M. (1966a), Exact distributions of Wilks's likelihood ratio criterion, *Biometrika*, **53**, 347–358. [8.4]

Schatzoff, M. (1966b), Sensitivity comparisons among tests of the general linear hypothesis, *Journal of the American Statistical Association*, **61**, 415–435. [8.6]

Scheffé, Henry (1942), On the ratio of the variances of two normal populations, *Annals of Mathematical Statistics*, **13**, 371–388. [10.2]

Scheffé, Henry (1943), On solutions of the Behrens–Fisher problem, based on the *t*-distribution, *Annals of Mathematical Statistics*, **14**, 35–44. [5.5]

Schmidli, H. (1996), *Reduced-rank Regression*, Physica, Berlin. [12.7]

Schuurmann, F. J., P. R. Krishnaiah, and A. K. Chattopadhyay (1975), Exact percentage points of the distribution of the trace of a multivariate beta matrix, *Journal of Statistical Computation and Simulation*, **3**, 331–343. [8.6]

Schuurmann, F. J., and V. B. Waikar (1973), Tables for the power function of Roy's two-sided test for testing hypothesis $\Sigma = I$ in the bivariate case, *Communications in Statistics*, **1**, 271–280. [10.8]

Schuurmann, F. J., V. B. Waikar, and P. R. Krishnaiah (1975), Percentage points of the joint distribution of the extreme roots of the random matrix $S_1(S_1 + S_2)^{-1}$, *Journal of Statistical Computation and Simulation*, **2**, 17–38. [10.6]

Schwartz, R. (1967), Admissible tests in multivariate analysis of variance, *Annals of Mathematical Statistics*, **38**, 698–710. [8.10]

Serfling, Robert J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York. [4.2]

Simaika, J. B. (1941), On an optimum property of two important statistical tests, *Biometrika*, **32**, 70–80. [5.6]

Siotani, Minoru (1980), Asymptotic approximations to the conditional distributions of the classification statistic Z and its studentized form Z^*, *Tamkang Journal of Mathematics*, **11**, 19–32. [6.6]

Siotani, M., and R. H. Wang (1975), Further expansion formulae for error rates and comparison of the W- and the Z-procedures in discriminant analysis, Technical Report No. 33, Department of Statistics, Kansas State University, Manhattan, Kansas. [6.6]

Siotani, M., and R. H. Wang (1977), Asymptotic Expansions for Error Rates and Comparison of the W-Procedure and the Z-Procedure in Discriminant Analysis, *Multivariate Analysis IV*, North-Holland, Amsterdam, 523–545. [6.6]

Sitgreaves, Rosedith (1952), On the distribution of two random matrices used in classification procedures, *Annals of Mathematical Statistics*, **23**, 263–270. [6.5]

Solari, M. E. (1969), The "maximum likelihood solution" of the problem of estimating a linear functional relationship, *Journal of the Royal Statistical Society B*, **31**, 372–375. [14.4]

Spearman, Charles (1904), "General-intelligence," objectively determined and measured, *American Journal of Psychology*, **15**, 201–293. [14.2]

Speed, T. P., and H. Kiiveri (1986), Gaussian Markov distributions over finite graphs, *Annals of Statistics*, **14**, 138–150. [15.5]

Srivastava, M. S., and C. G. Khatri (1979), *An Introduction to Multivariate Statistics*, North-Holland, New York. [10.9, 13.P]

Stein, C. (1956a), The admissibility of Hotelling's T^2-test, *Annals of Mathematical Statistics*, **27**, 616–623. [5.6]

Stein, C. (1956b), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the Third Berkeley Symposium on Mathematical and Statistical Probability* (Jerzy Neyman, ed.), Vol. I, 197–206, University of California, Berkeley. [3.5]

Stein, C. (1974), Estimation of the parameters of a multivariate normal distribution. I. Estimation of the means, Technical Report No. 63, NSF Contract GP 30711X-2, Department of Statistics, Stanford University. [3.5]

Stoica, P., and M. Viberg (1996), Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions, *IEEE Transaction Signal Processing*, **44**, 3069–3078. [12.7]

Student (W. S. Gosset) (1908), The probable error of a mean, *Biometrika*, **6**, 1–25. [3.2]

Styan, George P. H. (1990), *The Collected Papers of T. W. Anderson: 1943–1985*, John Wiley & Sons, Inc., New York. [Preface]

Subrahmaniam, Kocherlota, and Kathleen Subrahmaniam (1973), *Multivariate Analysis: A Selected and Abstracted Bibliography, 1957–1972*, Marcel Dekker, New York. [Preface]

Sugiura, Nariaki, and Hisao Nagao (1968), Unbiasedness of some test criteria for the equality of one or two covariance matrices, *Annals of Mathematical Statistics*, **39**, 1686–1692. [10.8]

Sugiyama, T. (1967), Distribution of the largest latent root and the smallest latent root of the generalized B statistic and F statistic in multivariate analysis, *Annals of Mathematical Statistics*, **38**, 1152–1159. [8.6]

Sugiyama, T., and K. Fukutomi (1966), On the distribution of the extreme characteristic roots of the matrices in multivariate analysis, *Reports of Statistical Application Research, Union of Japanese Scientists and Engineers*, **13**. [8.6]

Sverdrup, Erling (1947), Derivation of the Wishart distribution of the second order sample moments by straightforward integration of a multiple integral, *Skandinavisk Aktuarietidskrift*, **30**, 151–166. [7.2]

Tang, P. C. (1938), The power function of the analysis of variance tests with tables and illustrations of their use, *Statistical Resecrch Memoirs*, **2**, 126–157. [5.4]

Theil, H. (assisted by J. S. Cramer, H. Moerman, and A. Russchen) (1961), *Economic Forecasts and Policy*, (2nd rev. ed.), North-Holland, Amsterdam. Contributions to Economic Analysis No. XV (first published 1958). [12.8]

Thomson, Godfrey H. (1934), Hotelling's method modified to give Spearman's "*g*," *Journal of Educational Psychology*, **25**, 366–374. [14.3]

Thomson, Godfrey H. (1951), *The Factorial Analysis of Human Ability* (5th ed.), University of London, London. [14.7]

Thurstone, L. L. (1947), *Multiple-Factor Analysis*, University of Chicago, Chicago. [14.2, 14.5]

Tsay, R. S., and G. C. Tiao (1985), Use of canonical analysis in time series model identification, *Biometrika*, **72**, 299–315. [12.7]

Tukey, J. W. (1949), Dyadic anova, an analysis of variance for vectors, *Human Biology*, **21**, 65–110. [8.9]

Tyler, David E. (1981), Asymptotic inference for eigenvectors, *Annals of Statistics*, **9**, 725–745. [11.7]

Tyler, David E. (1982), Radial estimates and the test for sphericity, *Biometrika*, **69**, 429–436. [3.6]

Tyler, David E. (1983a); Robustness and efficiency properties of scatter matrices, *Biometrika*, **70**, 411–420. [3.6]

Tyler, David E. (1983b), The asymptotic distribution of principal component roots under local alternatives to multiple roots, *Annals of Statistics*, **11**, 1232–1242. [11.7]

Tyler, David E. (1987), A distribution free *M*-estimator of multivariate scatter, *Annals of Statistics*, **15**, 234–251. [3.6]

Velu, R. P., G. C. Reinsel, and D. W. Wichern (1986), Reduced rank models for multiple time series, *Biometrika*, **73**, 105–118. [12.7]

von Mises, R. (1945), On the classification of observation data into distinct groups, *Annals of Mathematical Statistics*, **16**, 68–73. [6.8]

von Neumann, J. (1937), Some matrix-inequalities and metrization of matric-space, *Tomsk University Review*, **1** 286–300. Reprinted in *John von Neuman Collected Works* (A. H. Taub, ed.), **4** (1962), Pergamon, New York, 205–219. [A.4]

Wald, A. (1943), Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society*, **54**, 426–482. [4.2]

Wald, A. (1944), On a statistical problem arising in the classification of an individual into one of two groups, *Annals of Mathematical Statistics*, **15**, 145–162. [6.4, 6.5]

Wald, A. (1950), *Statistical Decision Functions*, John Wiley & Sons, New York. [6.2, 6.7, 8.10]

Wald, A., and R. Brookner (1941), On the distribution of Wilks' statistic for testing the independence of several groups of variates, *Annals of Mathematical Statistics*, **12**, 137–152. [8.4, 9.3]

Walker, Helen M. (1931), *Studies in the History of Statistical Method*, Williams and Wilkins, Baltimore. [1.1]

Welch, P. D., and R. S. Wimpress (1961), Two multivariate statistical computer programs and their application to the vowel recognition problem, *Journal of the Acoustical Society of America*, **33**, 426–434. [6.10]

Wermuth, N. (1980), Linear recursive equations, covariance selection and path analysis, *Journal of the American Statistical Association*, **75**, 963–972. [15.5]

Whittaker, E. T., and G. N. Watson (1943), *A Course of Modern Analysis*, Cambridge University, Cambridge. [8.5]

Whittaker, Joe (1990), *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Inc., Chichester. [15.1]

Wijsman, Robert A. (1979), Constructing all smallest simultaneous confidence sets in a given class, with applications to MANOVA, *Annals of Statistics*, **7**, 1003–1018. [8.7]

Wijsman, Robert A. (1980), Smallest simultaneous confidence sets with applications in multivariate analysis, *Multivariate Analysis*, V, 483–498. [8.7]

Wilkinson, James Hardy (1965), *The Algebraic Eigenvalue Problem*, Clarendon, Oxford. [11.4]

Wilkinson, J. H., and C. Reinsch (1971), *Linear Algebra*, Springer-Verlag, New York. [11.4]

Wilks, S. S. (1932), Certain generalizations in the analysis of variance, *Biometrika*, **24**, 471–494. [7.5, 8.3, 10.4]

Wilks, S. S. (1934), Moment-generating operators for determinants of product moments in samples from a normal system, *Annals of Mathematics*, **35**, 312–340. [8.3]

Wilks, S. S. (1935), On the independence of k sets of normally distributed statistical variables, *Econometrica*, **3**, 309–326. [8.4, 9.3, 9.P]

Wishart, John (1928), The generalised product moment distribution in samples from a normal multivariate population, *Biometrika*, **20A**, 32–52. [7.2]

Wishart, John (1948), Proofs of the distribution law of the second order moment statistics, *Biometrika*, **35**, 55–57. [7.2]

Wishart, John, and M. S. Bartlett (1933), The generalised product moment distribution in a normal system, *Proceedings of the Cambridge Philosophical Society*, **29**, 260–270. [7.2]

Wold, H. D. A. (1954), Causality and econometrics, *Econometrica* **22**, 162–177. [15.1]

Wold, H. D. A. (1960), A generalization of casual chain models, *Econometrica*, **28**, 443–463. [15.1]

Woltz, W. G., W. A. Reid, and W. E. Colwell (1948), Sugar and nicotine in cured bright tobacco as related to mineral element composition, *Proceedings of the Soil Sciences Society of America*, **13**, 385–387. [8.P]

Wright, Sewall (1921), Correlation and causation, *Journal of Agricultural Research*, **20**, 557–585. [15.1]

Wright, Sewall (1934), The method of path coefficients, *Annals of Mathematical Statistics*, **5**, 161–215. [15.1]

Yamauti, Ziro (1977), *Concise Statistical Tables*, Japanese Standards Association. [Preface]

Yates, F., and W. G. Cochran (1938), The analysis of groups of experiments, *Journal of Agricultural Science*, **28**, 556. [8.9]

Yule, G. U. (1897a), On the significance of Bravais' formulae for regression & c., in the case of skew correlation, *Proceedings of the Royal Society of London*, **60**, 477–489. [2.5]

Yule, G. U. (1897b), On the theory of correlation, *Journal of the Royal Statistical Society*, **60**, 812–854. [2.5]

Zehna, P. W. (1966), Invariance of maximum likelihood estimators, *Annals of Mathematical Statistics*, **37**, 744. [3.2]

Index