Theory of Multivariate Statistics

Martin Bilodeau
David Brenner

Springer

Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Times Series and Forecasting
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Christensen: Linear Models for Multivariate, Time Series, and Spatial Data
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Statistics in Scientific Investigation: Its Basis, Application, and Interpretation
Mueller: Basic Principles of Structural Equation Modeling
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume I: Probability for Statistics
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume II: Statistical Inference

Martin Bilodeau David Brenner

Theory of Multivariate Statistics

Springer

Martin Bilodeau
Université de Montréal
Faculté des Arts et des Sciences
Département de Mathématiques et de Statistique
C.P. 6128, Succursale Centre-ville
Montréal (Québec) H3C 3J7
Canada
bilodeau@dms.umontreal.ca

David Brenner
Department of Statistics
University of Toronto
Ontario M5S 3G4
Canada
brenner@utstat.toronto.edu

With 9 illustrations

Library of Congress Cataloging-in-Publication Data
Bilodeau, Martin, 1961–
Theory of multivariate statistics / Martin Bilodeau, David Brenner.
p. cm. — (Springer texts in statistics)
Includes bibliographical references and indexes.
ISBN 0-387-98739-8 (hardcover : alk. paper)
1. Multivariate analysis. I. Brenner, David. II. Title.
III. Series.
QA278.B55 1999
519.5'35—dc21 99-26378

ISBN 0-387-98739-8 Springer-Verlag New York Berlin Heidelberg SPIN 10709151

A la mémoire de mon père, Arthur, à ma mère, Annette, et à Kahina.
M. Bilodeau

To Rebecca and Deena.
D. Brenner

This page intentionally left blank

Preface

Our object in writing this book is to present the main results of the modern theory of multivariate statistics to an audience of advanced students who would appreciate a concise and mathematically rigorous treatment of that material. It is intended for use as a textbook by students taking a first graduate course in the subject, as well as for the general reference of interested research workers who will find, in a readable form, developments from recently published work on certain broad topics not otherwise easily accessible, as, for instance, robust inference (using adjusted likelihood ratio tests) and the use of the bootstrap in a multivariate setting. The references contains over 150 entries post-1982. The main development of the text is supplemented by over 135 problems, most of which are original with the authors.

A minimum background expected of the reader would include at least two courses in mathematical statistics, and certainly some exposure to the calculus of several variables together with the descriptive geometry of linear algebra. Our book is, nevertheless, in most respects entirely self-contained, although a definite need for genuine fluency in general mathematics should not be underestimated. The pace is brisk and demanding, requiring an intense level of active participation in every discussion. The emphasis is on rigorous proof and derivation. The interested reader would profit greatly, of course, from previous exposure to a wide variety of statistically motivating material as well, and a solid background in statistics at the undergraduate level would obviously contribute enormously to a general sense of familiarity and provide some extra degree of comfort in dealing with the kinds of challenges and difficulties to be faced in the relatively advanced work

of the sort with which our book deals. In this connection, a specific introduction offering comprehensive overviews of the fundamental multivariate structures and techniques would be well advised. The textbook *A First Course in Multivariate Statistics* by Flury (1997), published by Springer-Verlag, provides such background insight and general description without getting much involved in the "nasty" details of analysis and construction. This would constitute an excellent supplementary source. Our book is in most ways thoroughly orthodox, but in several ways novel and unique.

In Chapter 1 we offer a brief account of the prerequisite linear algebra as it will be applied in the subsequent development. Some of the treatment is peculiar to the usages of multivariate statistics and to this extent may seem unfamiliar.

Chapter 2 presents in review, the requisite concepts, structures, and devices from probability theory that will be used in the sequel. The approach taken in the following chapters rests heavily on the assumption that this basic material is well understood, particularly that which deals with equality-in-distribution and the Cramér-Wold theorem, to be used with unprecedented vigor in the derivation of the main distributional results in Chapters 4 through 8. In this way, our approach to multivariate theory is much more structural and directly algebraic than is perhaps traditional, tied in this fashion much more immediately to the way in which the various distributions arise either in nature or may be generated in simulation. We hope that readers will find the approach refreshing, and perhaps even a bit liberating, particularly those saturated in a lifetime of matrix derivatives and jacobians.

As a textbook, the first eight chapters should provide a more than adequate amount of material for coverage in one semester (13 weeks). These eight chapters, proceeding from a thorough discussion of the normal distribution and multivariate sampling in general, deal in random matrices, Wishart's distribution, and Hotelling's T^2, to culminate in the standard theory of estimation and the testing of means and variances.

The remaining six chapters treat of more specialized topics than it might perhaps be wise to attempt in a simple introduction, but would easily be accessible to those already versed in the basics. With such an audience in mind, we have included detailed chapters on multivariate regression, principal components, and canonical correlations, each of which should be of interest to anyone pursuing further study. The last three chapters, dealing, in turn, with asymptotic expansion, robustness, and the bootstrap, discuss concepts that are of current interest for active research and take the reader (gently) into territory not altogether perfectly charted. This should serve to draw one (gracefully) into the literature.

The authors would like to express their most heartfelt thanks to everyone who has helped with feedback, criticism, comment, and discussion in the preparation of this manuscript. The first author would like especially to convey his deepest respect and gratitude to his teachers, Muni Srivastava

of the University of Toronto and Takeaki Kariya of Hitotsubashi University, who gave their unstinting support and encouragement during and after his graduate studies. The second author is very grateful for many discussions with Philip McDunnough of the University of Toronto. We are indebted to Nariaki Sugiura for his kind help concerning the application of Sugiura's Lemma and to Rudy Beran for insightful comments, which helped to improve the presentation. Eric Marchand pointed out some errors in the literature about the asymptotic moments in Section 8.4.1. We would like to thank the graduate students at McGill University and Université de Montréal, Gulhan Alpargu, Diego Clonda, Isabelle Marchand, Philippe St-Jean, Gueye N'deye Rokhaya, Thomas Tolnai and Hassan Younes, who helped improve the presentation by their careful reading and problem solving. Special thanks go to Pierre Duchesne who, as part of his Master Memoir, wrote and tested the S-Plus function for the calculation of the robust S estimate in Appendix C.

M. Bilodeau
D. Brenner

This page intentionally left blank

Contents

List of Tables

This page intentionally left blank

List of Figures

This page intentionally left blank

1
Linear algebra

1.1 Introduction

Multivariate analysis deals with issues related to the observations of many, usually correlated, variables on units of a selected random sample. These units can be of any nature such as persons, cars, cities, etc. The observations are gathered as vectors; for each selected unit corresponds a vector of observed variables. An understanding of vectors, matrices, and, more generally, linear algebra is thus fundamental to the study of multivariate analysis. Chapter 1 represents our selection of several important results on linear algebra. They will facilitate a great many of the concepts in multivariate analysis. A useful reference for linear algebra is Strang (1980).

1.2 Vectors and matrices

To express the dependence of the $\mathbf{x} \in \mathbb{R}^n$ on its coordinates, we may write any of

$$\mathbf{x} = (x_i,\ i = 1, \ldots, n) = (x_i) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

In this manner, $\mathbf{x}$ is envisaged as a "column" vector. The transpose of $\mathbf{x}$ is the "row" vector $\mathbf{x}' \in \mathbb{R}_n$

$$\mathbf{x}' = (x_i)' = (x_1, \ldots, x_n).$$

An $m \times n$ matrix $\mathbf{A} \in \mathbb{R}_n^m$ may also be denoted in various ways:

$$\mathbf{A} = (a_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n) = (a_{ij}) = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}.$$

The transpose of $\mathbf{A}$ is the $n \times m$ matrix $\mathbf{A}' \in \mathbb{R}_m^n$:

$$\mathbf{A}' = (a_{ij})' = (a_{ji}) = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{mn} \end{pmatrix}.$$

A square matrix $\mathbf{S} \in \mathbb{R}_n^n$ satisfying $\mathbf{S} = \mathbf{S}'$ is termed symmetric. The product of the $m \times n$ matrix $\mathbf{A}$ by the $n \times p$ matrix $\mathbf{B}$ is the $m \times p$ matrix $\mathbf{C} = \mathbf{AB}$ for which

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

The trace of $\mathbf{A} \in \mathbb{R}_n^n$ is $\text{tr } \mathbf{A} = \sum_{i=1}^n a_{ii}$ and one verifies that for $\mathbf{A} \in \mathbb{R}_n^m$ and $\mathbf{B} \in \mathbb{R}_m^n$, $\text{tr } \mathbf{AB} = \text{tr } \mathbf{BA}$.

In particular, row vectors and column vectors are themselves matrices, so that for $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$, we have the scalar result

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i y_i = \mathbf{y}'\mathbf{x}.$$

This provides the standard inner product, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{y}$, in $\mathbb{R}^n$ with the associated "euclidian norm" (length or modulus)

$$|\mathbf{x}| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}.$$

The Cauchy-Schwarz inequality is now proved.

Proposition 1.1 $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq |\mathbf{x}|\ |\mathbf{y}|,\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, *with equality if and only if (iff)* $\mathbf{x} = \lambda \mathbf{y}$ *for some* $\lambda \in \mathbb{R}$.

Proof. If $\mathbf{x} = \lambda \mathbf{y}$, for some $\lambda \in \mathbb{R}$, the equality clearly holds. If not, $0 < |\mathbf{x} - \lambda \mathbf{y}|^2 = |\mathbf{x}|^2 - 2\lambda \langle \mathbf{x}, \mathbf{y} \rangle + \lambda^2 |\mathbf{y}|^2$, $\forall \lambda \in \mathbb{R}$; thus, the discriminant of the quadratic polynomial must satisfy $4\langle \mathbf{x}, \mathbf{y} \rangle^2 - 4|\mathbf{x}|^2 |\mathbf{y}|^2 < 0$. □

The cosine of the angle θ between the vectors $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{y} \neq \mathbf{0}$ is just

$$\cos(\theta) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{x}|\ |\mathbf{y}|}.$$

Orthogonality is another associated concept. Two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ will be said to be orthogonal iff $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. In contrast, the outer (or tensor) product of $\mathbf{x}$ and $\mathbf{y}$ is an $n \times n$ matrix

$$\mathbf{x}\mathbf{y}' = (x_i y_j)$$

and this product is not commutative.

The concept of orthonormal basis plays a major role in linear algebra. A set $\{\mathbf{v}_i\}$ of vectors in $\mathbb{R}^n$ is orthonormal if

$$\mathbf{v}_i'\mathbf{v}_j = \delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j. \end{cases}$$

The symbol δ_{ij} is referred to as the Kronecker delta. The Gram-Schmidt orthogonalization method gives a construction of an orthonormal basis from an arbitrary basis.

Proposition 1.2 *Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis of $\mathbb{R}^n$. Define*

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{v}_1/|\mathbf{v}_1|, \\ \mathbf{u}_i &= \mathbf{w}_i/|\mathbf{w}_i|, \end{aligned}$$

where $\mathbf{w}_i = \mathbf{v}_i - \sum_{j=1}^{i-1}(\mathbf{v}_i'\mathbf{u}_j)\mathbf{u}_j$, $i = 2, \ldots, n$. Then, $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is an orthonormal basis.

1.3 Image space and kernel

Now, a matrix may equally well be recognized as a function either of its column vectors or its row vectors:

$$\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n) = \begin{pmatrix} \mathbf{g}_1' \\ \vdots \\ \mathbf{g}_m' \end{pmatrix}$$

for $\mathbf{a}_j \in \mathbb{R}^m$, $j = 1, \ldots, n$ or $\mathbf{g}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$. If we then write $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)$ with $\mathbf{b}_j \in \mathbb{R}^n$, $j = 1, \ldots, p$, we find that

$$\mathbf{AB} = (\mathbf{Ab}_1, \ldots, \mathbf{Ab}_p) = (\mathbf{g}_i'\mathbf{b}_j)\,.$$

In particular, for $\mathbf{x} \in \mathbb{R}^n$, we have expressly that

$$\mathbf{Ax} = (\mathbf{a}_1, \ldots, \mathbf{a}_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i \mathbf{a}_i \tag{1.1}$$

or

$$\mathbf{Ax} = \begin{pmatrix} \mathbf{g}_1' \\ \vdots \\ \mathbf{g}_m' \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{g}_1'\mathbf{x} \\ \vdots \\ \mathbf{g}_m'\mathbf{x} \end{pmatrix}. \tag{1.2}$$

The orthogonal complement of a subspace $\mathcal{V} \subset \mathbb{R}^n$ is, by definition, the subspace

$$\mathcal{V}^\perp = \{\mathbf{y} \in \mathbb{R}^n : \ \mathbf{y} \perp \mathbf{x}, \ \forall \mathbf{x} \in \mathcal{V}\}.$$

Expression (1.1) identifies the image space of $\mathbf{A}$, Im $\mathbf{A} = \{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^n\}$, with the linear span of its column vectors and the expression (1.2) reveals the kernel, ker $\mathbf{A} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{Ax} = \mathbf{0}\}$, to be the orthogonal complement of the row space, equivalently ker $\mathbf{A} = (\text{Im } \mathbf{A}')^{\perp}$. The dimension of the subspace Im $\mathbf{A}$ is called the rank of $\mathbf{A}$ and satisfies rank $\mathbf{A}$ = rank $\mathbf{A}'$, whereas the dimension of ker $\mathbf{A}$ is called the nullity of $\mathbf{A}$. They are related through the following simple relation:

Proposition 1.3 *For any* $\mathbf{A} \in \mathbb{R}^m_n$, $n = \text{nullity } \mathbf{A} + \text{rank } \mathbf{A}$.

Proof. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_\nu\}$ be a basis of ker $\mathbf{A}$ and extend it to a basis

$$\{\mathbf{v}_1, \ldots, \mathbf{v}_\nu, \mathbf{v}_{\nu+1}, \ldots, \mathbf{v}_n\}$$

of $\mathbb{R}^n$. One can easily check $\{\mathbf{Av}_{\nu+1}, \ldots, \mathbf{Av}_n\}$ is a basis of Im $\mathbf{A}$. Thus, $n = \text{nullity } \mathbf{A} + \text{rank } \mathbf{A}$. □

1.4 Nonsingular matrices and determinants

We recall some basic facts about nonsingular (one-to-one) linear transformations and determinants.

By writing $\mathbf{A} \in \mathbb{R}^n_n$ in terms of its column vectors $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ with $\mathbf{a}_j \in \mathbb{R}^n$, $j = 1, \ldots, n$, it is clear that

$$\mathbf{A} \text{ is one-to-one} \iff \mathbf{a}_1, \ldots, \mathbf{a}_n \text{ is a basis} \iff \ker \mathbf{A} = \{\mathbf{0}\}$$

and also from the simple relation $n = \text{nullity } \mathbf{A} + \text{rank } \mathbf{A}$,

$$\mathbf{A} \text{ is one-to-one} \iff \mathbf{A} \text{ is one-to-one and onto.}$$

These are all equivalent ways of saying $\mathbf{A}$ has an inverse or that $\mathbf{A}$ is nonsingular. Denote by $\sigma(1), \ldots, \sigma(n)$ a permutation of $1, \ldots, n$ and by $n(\sigma)$ its parity. Let $\mathbf{S}_n$ be the group of all the $n!$ permutations. The determinant is, by definition, the unique function det : $\mathbb{R}^n_n \to \mathbb{R}$, denoted $|\mathbf{A}| = \det(\mathbf{A})$, that is,

(i) multilinear: linear in each of $\mathbf{a}_1, \ldots, \mathbf{a}_n$ separately

(ii) alternating: $\left|\left(\mathbf{a}_{\sigma(1)}, \ldots, \mathbf{a}_{\sigma(n)}\right)\right| = (-1)^{n(\sigma)} \left|(\mathbf{a}_1, \ldots, \mathbf{a}_n)\right|$

(iii) normed: $|\mathbf{I}| = 1$.

This produces the formula

$$|\mathbf{A}| = \sum_{\sigma \in \mathbf{S}_n} (-1)^{n(\sigma)} a_{1\sigma(1)} \cdots a_{n\sigma(n)}$$

by which one verifies

$$|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}| \text{ and } |\mathbf{A}'| = |\mathbf{A}|\,.$$

Determinants are usually calculated with a Laplace development along any given row or column. To this end, let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^n_n$. Now, define the minor $|m(i,j)|$ of a_{ij} as the determinant of the $(n-1)\times(n-1)$ "submatrix" obtained by deleting the ith row and the jth column of $\mathbf{A}$ and the cofactor of a_{ij} as $c(i,j) = (-1)^{i+j}|m(i,j)|$. Then, the Laplace development of $|\mathbf{A}|$ along the ith row is $|\mathbf{A}| = \sum_{j=1}^n a_{ij}\cdot c(i,j)$ and a similar development along the jth column is $|\mathbf{A}| = \sum_{i=1}^n a_{ij}\cdot c(i,j)$. By defining $\text{adj}(\mathbf{A}) = (c(j,i))$, the transpose of the matrix of cofactors, to be the adjoint of $\mathbf{A}$, it can be shown $\mathbf{A}^{-1} = |\mathbf{A}|^{-1}\ \text{adj}(\mathbf{A})$.

But then

Proposition 1.4 *$\mathbf{A}$ is one-to-one $\iff |\mathbf{A}| \neq 0$.*

Proof. $\mathbf{A}$ is one-to-one means it has an inverse $\mathbf{B}$, $|\mathbf{A}|\,|\mathbf{B}| = 1$ so $|\mathbf{A}| \neq 0$. But, conversely, if $|\mathbf{A}| \neq 0$, suppose $\mathbf{Ax} = \sum_{j=1}^n x_j\mathbf{a}_j = \mathbf{0}$, then substituting $\mathbf{Ax}$ for the ith column of $\mathbf{A}$

$$\left|\left(\mathbf{a}_1,\ldots,\sum_{j=1}^n x_j\mathbf{a}_j,\ldots,\mathbf{a}_n\right)\right| = x_i\,|\mathbf{A}| = 0,\ i = 1,\ldots,n$$

so that $\mathbf{x} = \mathbf{0}$, whereby $\mathbf{A}$ is one-to-one. □

In general, for $\mathbf{a}_j \in \mathbb{R}^n$, $j = 1,\ldots,k$, write $\mathbf{A} = (\mathbf{a}_1,\ldots,\mathbf{a}_k)$ and form the "inner product" matrix $\mathbf{A}'\mathbf{A} = (\mathbf{a}_i'\mathbf{a}_j) \in \mathbb{R}^k_k$. We find

Proposition 1.5 *For $\mathbf{A} \in \mathbb{R}^n_k$,*

1. ker $\mathbf{A}$ = ker $\mathbf{A}'\mathbf{A}$
2. rank $\mathbf{A}$ = rank $\mathbf{A}'\mathbf{A}$
3. *$\mathbf{a}_1,\ldots,\mathbf{a}_k$ are linearly independent in $\mathbb{R}^n \iff |\mathbf{A}'\mathbf{A}| \neq 0$.*

Proof. If $\mathbf{x} \in$ ker $\mathbf{A}$, then $\mathbf{Ax} = \mathbf{0} \Longrightarrow \mathbf{A}'\mathbf{Ax} = \mathbf{0}$, and, conversely, if $\mathbf{x} \in$ ker $\mathbf{A}'\mathbf{A}$, then

$$\mathbf{A}'\mathbf{Ax} = \mathbf{0} \Longrightarrow \mathbf{x}'\mathbf{A}'\mathbf{Ax} = \mathbf{0} = |\mathbf{Ax}|^2 \Longrightarrow \mathbf{Ax} = \mathbf{0}.$$

The second part follows from the relation $k =$ nullity $\mathbf{A}$ + rank $\mathbf{A}$ and the third part is immediate as ker $\mathbf{A} = \{\mathbf{0}\}$ iff ker $\mathbf{A}'\mathbf{A} = \{\mathbf{0}\}$. □

1.5 Eigenvalues and eigenvectors

We now briefly state some concepts related to eigenvalues and eigenvectors. Consider, first, the complex vector space $\mathbb{C}^n$. The conjuguate of $v = x+iy \in \mathbb{C}$, $x, y \in \mathbb{R}$, is $\overline{v} = x-iy$. The concepts defined earlier are anologous in this case. The Hermitian transpose of a column vector $\mathbf{v} = (v_i) \in \mathbb{C}^n$ is the row vector $\mathbf{v}^H = (\overline{v_i})'$. The inner product on $\mathbb{C}^n$ can then be written $\langle \mathbf{v}_1, \mathbf{v}_2\rangle =$

$\mathbf{v}_1^H\mathbf{v}_2$ for any $\mathbf{v}_1$, $\mathbf{v}_2 \in \mathbb{C}^n$. The Hermitian transpose of $\mathbf{A} = (a_{ij}) \in \mathbb{C}_n^m$ is $\mathbf{A}^H = (\overline{a_{ji}}) \in \mathbb{C}_m^n$ and satisfies for $\mathbf{B} \in \mathbb{C}_p^n$, $(\mathbf{AB})^H = \mathbf{B}^H\mathbf{A}^H$. The matrix $\mathbf{A} \in \mathbb{C}_n^n$ is termed Hermitian iff $\mathbf{A} = \mathbf{A}^H$. We now define what is meant by an eigenvalue. A scalar $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{A} \in \mathbb{C}_n^n$ if there exists a vector $\mathbf{v} \neq \mathbf{0}$ in $\mathbb{C}^n$ such that $\mathbf{Av} = \lambda\mathbf{v}$. Equivalently, $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{A}$ iff $|\mathbf{A} - \lambda\mathbf{I}| = 0$, which is a polynomial equation of degree n. Hence, there are n complex eigenvalues, some of which may be real, with possibly some repetitions (multiplicity). The vector $\mathbf{v}$ is then termed the eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ. Note that if $\mathbf{v}$ is an eigenvector, so is $\alpha\mathbf{v}$, $\forall\alpha \neq 0$ in $\mathbb{C}$, and, in particular, $\mathbf{v}/|\mathbf{v}|$ is a normalized eigenvector.

Now, before defining what is meant by $\mathbf{A}$ is "diagonalizable" we define a matrix $\mathbf{U} \in \mathbb{C}_n^n$ to be unitary iff $\mathbf{U}^H\mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^H$. This means that the columns (or rows) of $\mathbf{U}$ comprise an orthonormal basis of $\mathbb{C}^n$. We note immediately that if $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is an orthonormal basis of eigenvectors corresponding to eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, then $\mathbf{A}$ can be diagonalized by the unitary matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$; i.e., we can write

$$\mathbf{U}^H\mathbf{AU} = \mathbf{U}^H(\mathbf{Au}_1, \ldots, \mathbf{Au}_n) = \mathbf{U}^H(\lambda_1\mathbf{u}_1, \ldots, \lambda_n\mathbf{u}_n) = \operatorname{diag}(\boldsymbol{\lambda}),$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$. Another simple related property: If there exists a unitary matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ such that $\mathbf{U}^H\mathbf{AU} = \operatorname{diag}(\boldsymbol{\lambda})$, then $\mathbf{u}_i$ is an eigenvector corresponding to λ_i. To verify this, note that

$$\mathbf{Au}_i = \mathbf{U}\ \operatorname{diag}(\boldsymbol{\lambda})\mathbf{U}^H\mathbf{u}_i = \mathbf{U}\ \operatorname{diag}(\boldsymbol{\lambda})\mathbf{e}_i = \mathbf{U}\lambda_i\mathbf{e}_i = \lambda_i\mathbf{u}_i.$$

Two fundamental propositions concerning Hermitian matrices are the following.

Proposition 1.6 *If $\mathbf{A} \in \mathbb{C}_n^n$ is Hermitian, then all its eigenvalues are real.*

Proof.

$$\overline{\mathbf{v}^H\mathbf{Av}} = (\mathbf{v}^H\mathbf{Av})^H = \mathbf{v}^H\mathbf{A}^H\mathbf{v} = \mathbf{v}^H\mathbf{Av},$$

which means that $\mathbf{v}^H\mathbf{Av}$ is real for any $\mathbf{v} \in \mathbb{C}^n$. Now, if $\mathbf{Av} = \lambda\mathbf{v}$ for some $\mathbf{v} \neq \mathbf{0}$ in $\mathbb{C}^n$, then $\mathbf{v}^H\mathbf{Av} = \lambda\mathbf{v}^H\mathbf{v} = \lambda|\mathbf{v}|^2$. But since $\mathbf{v}^H\mathbf{Av}$ and $|\mathbf{v}|^2$ are real, so is λ. □

Proposition 1.7 *If $\mathbf{A} \in \mathbb{C}_n^n$ is Hermitian and $\mathbf{v}_1$ and $\mathbf{v}_2$ are eigenvectors corresponding to eigenvalues λ_1 and λ_2, respectively, where $\lambda_1 \neq \lambda_2$, then $\mathbf{v}_1 \perp \mathbf{v}_2$.*

Proof. Since $\mathbf{A}$ is Hermitian, $\mathbf{A} = \mathbf{A}^H$ and λ_i, $i = 1, 2$, are real. Then,

$$\begin{aligned}
\mathbf{Av}_1 = \lambda_1\mathbf{v}_1 \quad &\Longrightarrow \quad \mathbf{v}_1^H\mathbf{A}^H = \mathbf{v}_1^H\mathbf{A} = \lambda_1\mathbf{v}_1^H \Longrightarrow \mathbf{v}_1^H\mathbf{Av}_2 = \lambda_1\mathbf{v}_1^H\mathbf{v}_2, \\
\mathbf{Av}_2 = \lambda_2\mathbf{v}_2 \quad &\Longrightarrow \quad \mathbf{v}_1^H\mathbf{Av}_2 = \lambda_2\mathbf{v}_1^H\mathbf{v}_2.
\end{aligned}$$

Subtracting the last two expressions, $(\lambda_1 - \lambda_2)\mathbf{v}_1^H\mathbf{v}_2 = 0$ and, thus, $\mathbf{v}_1^H\mathbf{v}_2 = 0$. □

Proposition 1.7 immediately shows that if all the eigenvalues of $\mathbf{A}$, Hermitian, are distinct, then there exists an orthonormal basis of eigenvectors whereby $\mathbf{A}$ is diagonalizable. Toward proving this is true even when the eigenvalues may be of a multiple nature, we need the following proposition. However, before stating it, define $\mathbf{T} = (t_{ij}) \in \mathbb{R}_n^n$ to be a lower triangular matrix iff $t_{ij} = 0$, $i < j$. Similarly, $\mathbf{T} \in \mathbb{R}_n^n$ is termed upper triangular iff $t_{ij} = 0$, $i > j$.

Proposition 1.8 *Let $\mathbf{A} \in \mathbb{C}_n^n$ be any matrix. There exists a unitary matrix $\mathbf{U} \in \mathbb{C}_n^n$ such that $\mathbf{U}^H\mathbf{A}\mathbf{U}$ is upper triangular.*

Proof. The proof is by induction on n. The result is obvious for $n = 1$. Next, assume the proposition holds for n and prove it is true for $n + 1$. Let λ_1 be an eigenvalue of $\mathbf{A}$ and $\mathbf{u}_1$, $|\mathbf{u}_1| = 1$, be an eigenvector. Let $\mathbf{U}_1 = (\mathbf{u}_1, \mathbf{\Gamma})$ for some $\mathbf{\Gamma}$ such that $\mathbf{U}_1$ is unitary (such a $\mathbf{\Gamma}$ exists from the Gram-Schmidt method). Then,

$$\mathbf{U}_1^H\mathbf{A}\mathbf{U}_1 = \mathbf{U}_1^H(\lambda_1\mathbf{u}_1, \mathbf{A}\mathbf{\Gamma}) = \begin{pmatrix} \lambda_1 & \mathbf{u}_1^H\mathbf{A}\mathbf{\Gamma} \\ \mathbf{0} & \mathbf{B} \end{pmatrix},$$

where $\mathbf{B} = \mathbf{\Gamma}^H\mathbf{A}\mathbf{\Gamma} \in \mathbb{C}_n^n$. From the induction hypothesis, there exists $\mathbf{V}$ unitary such that $\mathbf{V}^H\mathbf{B}\mathbf{V} = \mathbf{T}$ is triangular. Define

$$\mathbf{U}_2 = \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{V} \end{pmatrix}$$

and it is clear that $\mathbf{U}_2$ is also unitary. Finally,

$$\begin{aligned} (\mathbf{U}_1\mathbf{U}_2)^H\mathbf{A}(\mathbf{U}_1\mathbf{U}_2) &= \mathbf{U}_2^H \begin{pmatrix} \lambda_1 & \mathbf{u}_1^H\mathbf{A}\mathbf{\Gamma} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \mathbf{U}_2 \\ &= \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{V}^H \end{pmatrix} \begin{pmatrix} \lambda_1 & \mathbf{u}_1^H\mathbf{A}\mathbf{\Gamma} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{V} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & \mathbf{u}_1^H\mathbf{A}\mathbf{\Gamma}\mathbf{V} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}, \end{aligned}$$

which is of the desired form. The proof is complete because $\mathbf{U} \equiv \mathbf{U}_1\mathbf{U}_2$ is unitary. □

As a corollary we obtain that Hermitian matrices are always diagonalizable.

Corollary 1.1 *Let $\mathbf{A} \in \mathbb{C}_n^n$ be Hermitian. There exists a unitary matrix $\mathbf{U}$ such that $\mathbf{U}^H\mathbf{A}\mathbf{U} = \text{diag}(\boldsymbol{\lambda})$.*

Proof. Proposition 1.8 showed there exists $\mathbf{U}$, unitary, such that $\mathbf{U}^H\mathbf{A}\mathbf{U}$ is triangular. However, if $\mathbf{A}$ is Hermitian, so is $\mathbf{U}^H\mathbf{A}\mathbf{U}$. The only matrices that are both Hermitian and triangular are the diagonal matrices. □

In the sequel, we will always use Corollary 1.1 for $\mathbf{S} \in \mathbb{R}_n^n$ symmetric. However, first note that when $\mathbf{S}$ is symmetric all its eigenvalues are real, whereby the eigenvectors can also be chosen to be real, they are the solutions of $(\mathbf{S} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$. When $\mathbf{U} \in \mathbb{R}_n^n$ is unitary, it is called an orthogonal

matrix instead. A matrix $\mathbf{H} \in \mathbb{R}_n^n$ is said to be orthogonal iff the columns (or rows) of $\mathbf{H}$ form an orthonormal basis of $\mathbb{R}^n$, i.e., $\mathbf{H}'\mathbf{H} = \mathbf{I} = \mathbf{H}\mathbf{H}'$. The group of orthogonal matrices in $\mathbb{R}_n^n$ will be denoted by

$$\mathbf{O}_n = \{\mathbf{H} \in \mathbb{R}_n^n : \ \mathbf{H}\mathbf{H}' = \mathbf{I}\}.$$

We have proven the "spectral decomposition:"

Proposition 1.9 *If $\mathbf{S} \in \mathbb{R}_n^n$ is symmetric, then there exists $\mathbf{H} \in \mathbf{O}_n$ such that $\mathbf{H}'\mathbf{S}\mathbf{H} = \mathrm{diag}(\boldsymbol{\lambda})$.*

The columns of $\mathbf{H}$ form an orthonormal basis of eigenvectors and $\boldsymbol{\lambda}$ is the vector of corresponding eigenvalues.

Now, a symmetric matrix $\mathbf{S} \in \mathbb{R}_n^n$ is said to be positive semidefinite, denoted $\mathbf{S} \geq \mathbf{0}$ or $\mathbf{S} \in \mathcal{PS}_n$, iff $\mathbf{v}'\mathbf{S}\mathbf{v} \geq 0$, $\forall \mathbf{v} \in \mathbb{R}^n$, and it is positive definite, denoted $\mathbf{S} > \mathbf{0}$ or $\mathbf{S} \in \mathcal{P}_n$, iff $\mathbf{v}'\mathbf{S}\mathbf{v} > 0, \forall \mathbf{v} \neq \mathbf{0}$. Finally, the positive semidefinite and positive definite matrices can be characterized in terms of eigenvalues.

Proposition 1.10 *Let $\mathbf{S} \in \mathbb{R}_n^n$ symmetric with eigenvalues $\lambda_1, \ldots, \lambda_n$.*

1. *$\mathbf{S} \geq \mathbf{0}$ iff $\lambda_i \geq 0$, $i = 1, \ldots, n$.*

2. *$\mathbf{S} > \mathbf{0}$ iff $\lambda_i > 0$, $i = 1, \ldots, n$.*

Note that if $\mathbf{S}$ is positive semidefinite, then from Proposition 1.9, we can write

$$\mathbf{S} = \mathbf{H}\mathbf{D}\mathbf{H}' = (\mathbf{H}\mathbf{D}^{1/2})(\mathbf{H}\mathbf{D}^{1/2})' = (\mathbf{H}\mathbf{D}^{1/2}\mathbf{H}')^2,$$

where $\mathbf{D} = \mathrm{diag}(\lambda_i)$ and $\mathbf{D}^{1/2} = \mathrm{diag}(\lambda_i^{1/2})$, so that for $\mathbf{A} = \mathbf{H}\mathbf{D}^{1/2}$, $\mathbf{S} = \mathbf{A}\mathbf{A}'$, or for $\mathbf{B} = \mathbf{H}\mathbf{D}^{1/2}\mathbf{H}'$, $\mathbf{S} = \mathbf{B}^2$. The positive semidefinite matrix $\mathbf{B}$ is often denoted $\mathbf{S}^{1/2}$ and is the square root of $\mathbf{S}$. If $\mathbf{S}$ is positive definite, we can also define $\mathbf{S}^{-1/2} = \mathbf{H}\mathbf{D}^{-1/2}\mathbf{H}'$, which satisfies $\left(\mathbf{S}^{-1/2}\right)^2 = \mathbf{S}^{-1}$. Finally, inequalities between matrices must be understood in terms of positive definiteness; i.e., for matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \geq \mathbf{B}$ (respectively $\mathbf{A} > \mathbf{B}$) means $\mathbf{A} - \mathbf{B} \geq \mathbf{0}$ (respectively $\mathbf{A} - \mathbf{B} > \mathbf{0}$).

A related decomposition which will prove useful for canonical correlations is the singular value decomposition (SVD).

Proposition 1.11 *Let $\mathbf{A} \in \mathbb{R}_n^m$ of* rank $\mathbf{A} = r$. *There exists $\mathbf{G} \in \mathbf{O}_m$, $\mathbf{H} \in \mathbf{O}_n$ such that*

$$\mathbf{A} = \mathbf{G} \begin{pmatrix} \mathbf{D}_{\boldsymbol{\rho}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{H}'$$

where $\mathbf{D}_{\boldsymbol{\rho}} = \mathrm{diag}(\rho_1, \ldots, \rho_r)$, $\rho_i > 0$, $i = 1, \ldots, r$.

Proof. Since $\mathbf{A}'\mathbf{A} \geq \mathbf{0}$, there exists $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_n) \in \mathbf{O}_n$ such that

$$\mathbf{A}'\mathbf{A} = \mathbf{H}\ \mathrm{diag}(\lambda_1, \ldots, \lambda_r, \mathbf{0})\ \mathbf{H}',$$

where $\lambda_i > 0$, $i = 1, \ldots, r$. For $j > r$, $|\mathbf{A}\mathbf{h}_j|^2 = \mathbf{h}_j'\mathbf{A}'\mathbf{A}\mathbf{h}_j = 0$ which means $\mathbf{A}\mathbf{h}_j = \mathbf{0}$. For $j \leq r$, define $\rho_j = \sqrt{\lambda_j}$ and $\mathbf{g}_j = \mathbf{A}\mathbf{h}_j/\rho_j$. Then, $\mathbf{g}_i'\mathbf{g}_j = \mathbf{h}_i'\mathbf{A}'\mathbf{A}\mathbf{h}_j/\rho_i\rho_j = \delta_{ij}$; i.e., $\mathbf{g}_1, \ldots, \mathbf{g}_r$ are orthonormal. By completing to an orthonormal basis of $\mathbb{R}^m$, we can find

$$\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_r, \mathbf{g}_{r+1}, \ldots, \mathbf{g}_m) \in \mathbf{O}_m.$$

Now,

$$\mathbf{g}_i'\mathbf{A}\mathbf{h}_j = \begin{cases} 0, & j > r \\ \rho_j\delta_{ij}, & j \leq r, \end{cases}$$

or in matrix notation,

$$\mathbf{G}'\mathbf{A}\mathbf{H} = \begin{pmatrix} \mathbf{D}_{\boldsymbol{\rho}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

□

In the SVD ρ_j^2, $j = 1, \ldots, r$, are the nonzero eigenvalues of $\mathbf{A}'\mathbf{A}$ and the columns of $\mathbf{H}$ are the eigenvectors.

1.6 Orthogonal projections

Now recall some basic facts about orthogonal projections. By definition, an *orthogonal projection*, $\mathbf{P}$, is simply a linear transformation for which $\mathbf{x} - \mathbf{P}\mathbf{x} \perp \mathbf{P}\mathbf{y}$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, but then, equivalently,

$$\begin{aligned} (\mathbf{x} - \mathbf{P}\mathbf{x})'(\mathbf{P}\mathbf{y}) = 0, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad &\iff \quad \mathbf{x}'\mathbf{P}\mathbf{y} = \mathbf{x}'\mathbf{P}'\mathbf{P}\mathbf{y}, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \\ &\iff \quad \mathbf{P}'\mathbf{P} = \mathbf{P} \\ &\iff \quad \mathbf{P} = \mathbf{P}' = \mathbf{P}^2. \end{aligned}$$

A matrix $\mathbf{P}$ such that $\mathbf{P} = \mathbf{P}' = \mathbf{P}^2$ is also called an idempotent matrix. Not surprisingly, an orthogonal projection is completely determined by its image.

Proposition 1.12 *If $\mathbf{P}_1$ and $\mathbf{P}_2$ are two orthogonal projections, then*

$$\text{Im } \mathbf{P}_1 = \text{Im } \mathbf{P}_2 \iff \mathbf{P}_1 = \mathbf{P}_2.$$

Proof. It holds since

$$\mathbf{x} - \mathbf{P}_1\mathbf{x} \perp \mathbf{P}_2\mathbf{y}, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \Longrightarrow \mathbf{P}_2 = \mathbf{P}_1'\mathbf{P}_2,$$

and, similarly, $\mathbf{P}_1 = \mathbf{P}_2'\mathbf{P}_1$, whence $\mathbf{P}_1 = \mathbf{P}_1' = \mathbf{P}_2$. □

If $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ is any basis for Im $\mathbf{P}$, we have explicitly

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{1.3}$$

To see this, simply write $\mathbf{P}\mathbf{x} = \mathbf{X}\mathbf{b}$, and orthogonality, $\mathbf{X}'(\mathbf{x} - \mathbf{X}\mathbf{b}) = \mathbf{0}$, determines the (unique) coefficients $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}$. In particular, for

any orthonormal basis $\mathbf{H}$, $\mathbf{P} = \mathbf{H}\mathbf{H}'$, where $\mathbf{H}'\mathbf{H} = \mathbf{I}_k$. Thus, incidentally, tr $\mathbf{P} = k$ and the dimension of the image space is expressed in the trace.

However, by this representation we see that for any two orthogonal projections, $\mathbf{P}_1 = \mathbf{H}\mathbf{H}'$ and $\mathbf{P}_2 = \mathbf{G}\mathbf{G}'$,

$$\mathbf{P}_1\mathbf{P}_2 = \mathbf{0} \Longleftrightarrow \mathbf{H}'\mathbf{G} = \mathbf{0} \Longleftrightarrow \mathbf{G}'\mathbf{H} = \mathbf{0} \Longleftrightarrow \mathbf{P}_2\mathbf{P}_1 = \mathbf{0}.$$

Definition 1.1 *$\mathbf{P}_1$ and $\mathbf{P}_2$ are said to be* mutually orthogonal projections *iff $\mathbf{P}_1$ and $\mathbf{P}_2$ are orthogonal projections such that $\mathbf{P}_1\mathbf{P}_2 = \mathbf{0}$. We write $\mathbf{P}_1 \perp \mathbf{P}_2$ when this is the case.*

Although orthogonal projection and orthogonal transformation are far from synonymous, there is, nevertheless, finally a very close connection between the two concepts. If we partition any orthogonal transformation $\mathbf{H} = (\mathbf{H}_1, \ldots, \mathbf{H}_k)$, then the brute algebraic fact

$$\mathbf{H}\mathbf{H}' = \mathbf{I} = \mathbf{H}_1\mathbf{H}_1' + \cdots + \mathbf{H}_k\mathbf{H}_k'$$

represents a precisely corresponding partition of the identity into mutually orthogonal projections.

As a last comment on othogonal projection, if $\mathbf{P}$ is the orthogonal projection on the subspace $\mathcal{V} \subset \mathbb{R}^n$, then $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, which satisfies $\mathbf{Q} = \mathbf{Q}' = \mathbf{Q}^2$ is also an othogonal projection. In fact, since $\mathbf{P}\mathbf{Q} = \mathbf{0}$, then Im $\mathbf{Q}$ and Im $\mathbf{P}$ are orthogonal subspaces and, thus, $\mathbf{Q}$ is the orthogonal projection on $\mathcal{V}^\perp$.

1.7 Matrix decompositions

Denote the groups of triangular matrices with positive diagonal elements as

$$\begin{aligned} \mathbf{L}_n^+ &= \{\mathbf{T} \in \mathbb{R}_n^n : \ \mathbf{T} \text{ is lower triangular}, t_{ii} > 0, \ i = 1, \ldots, n\}, \\ \mathbf{U}_n^+ &= \{\mathbf{T} \in \mathbb{R}_n^n : \ \mathbf{T} \text{ is upper triangular}, t_{ii} > 0, \ i = 1, \ldots, n\}. \end{aligned}$$

An important implication of Proposition 1.2 for matrices is the following matrix decomposition.

Proposition 1.13 *If $\mathbf{A} \in \mathbb{R}_n^n$ is nonsingular, then $\mathbf{A} = \mathbf{T}\mathbf{H}$ for some $\mathbf{H} \in \mathbf{O}_n$ and $\mathbf{T} \in \mathbf{L}_n^+$. Moreover, this decomposition is unique.*

Proof. The existence follows from the Gram-Schmidt method applied to the basis formed by the rows of $\mathbf{A}$. The rows of $\mathbf{H}$ form the orthonormal basis obtained at the end of that procedure and the elements of $\mathbf{T} = (t_{ij})$ are the coefficients needed to go from one basis to the other. By the Gram-Schmidt construction itself, it is clear that $\mathbf{T} \in \mathbf{L}_n^+$. For unicity, suppose $\mathbf{T}\mathbf{H} = \mathbf{T}_1\mathbf{H}_1$, where $\mathbf{T}_1 \in \mathbf{L}_n^+$ and $\mathbf{H}_1 \in \mathbf{O}_n$. Then, $\mathbf{T}_1^{-1}\mathbf{T} = \mathbf{H}_1\mathbf{H}'$ is a matrix in $\mathbf{L}_n^+ \cap \mathbf{O}_n$. But, $\mathbf{I}_n$ is the only such matrix (why?). Hence, $\mathbf{T} = \mathbf{T}_1$ and $\mathbf{H} = \mathbf{H}_1$. □

A slight generalization of Proposition 1.13 when $\mathbf{A} \in \mathbb{R}_n^p$ is of rank $\mathbf{A} = p$ is proposed in Problem 1.8.7. Another similar triangular decomposition, known in statistics as the Bartlett decomposition, for positive definite matrices can now be easily obtained.

Proposition 1.14 *If* $\mathbf{S} \in \mathcal{P}_n$, *then* $\mathbf{S} = \mathbf{T}\mathbf{T}'$ *for a unique* $\mathbf{T} \in \mathbf{L}_n^+$.

Proof. Since $\mathbf{S} > \mathbf{0}$, then $\mathbf{S} = \mathbf{H}\mathbf{D}\mathbf{H}'$, where $\mathbf{H} \in \mathbf{O}_n$ and $\mathbf{D} = \text{diag}(\lambda_i)$ with $\lambda_i > 0$. Let $\mathbf{D}^{1/2} = \text{diag}(\lambda_i^{1/2})$ and $\mathbf{A} = \mathbf{H}\mathbf{D}^{1/2}$. Then, we can write $\mathbf{S} = \mathbf{A}\mathbf{A}'$, where $\mathbf{A}$ is nonsingular. From Proposition 1.13, there exists $\mathbf{T} \in \mathbf{L}_n^+$ and $\mathbf{G} \in \mathbf{O}_n$ such that $\mathbf{A} = \mathbf{T}\mathbf{G}$. But, then, $\mathbf{S} = \mathbf{T}\mathbf{G}\mathbf{G}'\mathbf{T}' = \mathbf{T}\mathbf{T}'$. For unicity, suppose $\mathbf{T}\mathbf{T}' = \mathbf{T}_1\mathbf{T}_1'$, where $\mathbf{T}_1 \in \mathbf{L}_n^+$. Then, $\mathbf{T}_1^{-1}\mathbf{T}\mathbf{T}'\mathbf{T}_1'^{-1} = \mathbf{I}$, which implies that $\mathbf{T}_1^{-1}\mathbf{T} \in \mathbf{L}_n^+ \cap \mathbf{O}_n = \{\mathbf{I}\}$. Hence, $\mathbf{T} = \mathbf{T}_1$. □

Other notions of linear algebra such as Kronecker product and "vec" operator will be recalled when needed in the sequel.

1.8 Problems

1. Consider the partitioned matrix $\mathbf{S} = (s_{ij}) = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}$.

   (i) If $\mathbf{S}_{11}$ is nonsingular, prove that

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

   (ii) For $\mathbf{S} > \mathbf{0}$, prove Hadamard's inequality, $|\mathbf{S}| \leq \prod_i s_{ii}$.

   (iii) Let $\mathbf{S}$ and $\mathbf{S}_{11}$ be nonsingular. Prove that

   $$\mathbf{S}^{-1} = \begin{pmatrix} \mathbf{S}_{11}^{-1} + \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22.1}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & -\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22.1}^{-1} \\ -\mathbf{S}_{22.1}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & \mathbf{S}_{22.1}^{-1} \end{pmatrix},$$

   where $\mathbf{S}_{22.1} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$.

   (iv) Let $\mathbf{S}$ and $\mathbf{S}_{22}$ be nonsingular. Prove that

   $$\mathbf{S}^{-1} = \begin{pmatrix} \mathbf{S}_{11.2}^{-1} & -\mathbf{S}_{11.2}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1} \\ -\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11.2}^{-1} & \mathbf{S}_{22}^{-1} + \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11.2}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1} \end{pmatrix},$$

   where $\mathbf{S}_{11.2} = \mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$.

   Hint: Define

   $$\mathbf{A} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & \mathbf{I} \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} \mathbf{I} & -\mathbf{S}_{11}^{-1}\mathbf{S}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

   and consider the product $\mathbf{ASB}$.

2. Establish with the partitioning

   $$\begin{aligned} \mathbf{x} &= (\mathbf{x}_1', \mathbf{x}_2')', \\ \mathbf{S} &= \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \end{aligned}$$

that

$$\mathbf{x}'\mathbf{S}^{-1}\mathbf{x} = (\mathbf{x}_1 - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{x}_2)'\mathbf{S}_{11.2}^{-1}(\mathbf{x}_1 - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{x}_2) + \mathbf{x}_2'\mathbf{S}_{22}^{-1}\mathbf{x}_2.$$

3. For any $\mathbf{A} \in \mathbb{R}_q^p$, $\mathbf{B} \in \mathbb{R}_p^q$, prove the following:

   (i) $|\mathbf{I}_p + \mathbf{AB}| = |\mathbf{I}_q + \mathbf{BA}|$.
   Hint:

$$\begin{pmatrix} \mathbf{I}_p + \mathbf{AB} & \mathbf{A} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{A} \\ -\mathbf{B} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{B} & \mathbf{I}_q \end{pmatrix},$$

$$\begin{pmatrix} \mathbf{I}_p & \mathbf{A} \\ \mathbf{0} & \mathbf{I}_q + \mathbf{BA} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{B} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{A} \\ -\mathbf{B} & \mathbf{I}_q \end{pmatrix}.$$

   (ii) The nonzero eigenvalues of $\mathbf{AB}$ and $\mathbf{BA}$ are the same.

4. Prove Proposition 1.2.

5. Prove Proposition 1.10.

6. Show that if $\mathbf{P}$ defines an orthogonal projection, then the eigenvalues of $\mathbf{P}$ are either 0 or 1.

7. Demonstrate the slight generalizations of Proposition 1.13:

   (i) If $\mathbf{A} \in \mathbb{R}_p^n$ is of rank $\mathbf{A} = p$, then $\mathbf{A} = \mathbf{HT}$ for some $\mathbf{T} \in \mathbf{U}_p^+$ and $\mathbf{H}$ satisfying $\mathbf{H}'\mathbf{H} = \mathbf{I}_p$. Further, $\mathbf{T}$ and $\mathbf{H}$ are unique.
   Hint: For unicity, note that if $\mathbf{A} = \mathbf{HT} = \mathbf{H}_1\mathbf{T}_1$ with $\mathbf{T}_1 \in \mathbf{U}_p^+$ and $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{I}_p$, then Im $\mathbf{A}$ = Im $\mathbf{H}$ = Im $\mathbf{H}_1$ and $\mathbf{H}_1\mathbf{H}_1'$ is the orthogonal projection on Im $\mathbf{H}_1$.
   (ii) If $\mathbf{A} \in \mathbb{R}_p^n$ is of rank $\mathbf{A} = n$, then $\mathbf{A} = \mathbf{TH}$, where $\mathbf{T} \in \mathbf{L}_n^+$ and $\mathbf{HH}' = \mathbf{I}_n$. Further, $\mathbf{T}$ and $\mathbf{H}$ are unique.

8. Assuming $\mathbf{A}$ and $\mathbf{A} + \mathbf{uv}'$ are nonsingular, prove

$$(\mathbf{A} + \mathbf{uv}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{uv}'\mathbf{A}^{-1}}{(1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u})}.$$

9. **Vector differentiation.**
   Let $f(\mathbf{x})$ be a real valued function of $\mathbf{x} \in \mathbb{R}^n$. Define

$$\partial f(\mathbf{x})/\partial \mathbf{x} = (\partial f(\mathbf{x})/\partial x_i)\,.$$

   Verify

   (i) $\partial \mathbf{a}'\mathbf{x}/\partial \mathbf{x} = \mathbf{a}$,
   (ii) $\partial \mathbf{x}'\mathbf{Ax}/\partial \mathbf{x} = 2\mathbf{Ax}$, if $\mathbf{A}$ is symmetric.

10. **Matrix differentiation** [Srivastava and Khatri (1979), p. 37].
    Let $g(\mathbf{S})$ be a real-valued function of the symmetric matrix $\mathbf{S} \in \mathbb{R}_n^n$. Define $\partial f(\mathbf{S})/\partial \mathbf{S} = \left(\frac{1}{2}(1 + \delta_{ij})\partial f(\mathbf{S})/\partial s_{ij}\right)$. Verify

    (i) $\partial \mathrm{tr}(\mathbf{S}^{-1}\mathbf{A})/\partial \mathbf{S} = -\mathbf{S}^{-1}\mathbf{AS}^{-1}$, if $\mathbf{A}$ is symmetric,
    (ii) $\partial \ln|\mathbf{S}|/\partial \mathbf{S} = \mathbf{S}^{-1}$.

Hint for (ii): $\mathbf{S}^{-1} = |\mathbf{S}|^{-1}\text{adj}(\mathbf{S})$.

11. **Rayleigh's quotient.**
Assume $\mathbf{S} \geq \mathbf{0}$ in $\mathbb{R}_n^n$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Prove:
(i)

$$\lambda_n \leq \frac{\mathbf{x}'\mathbf{S}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \leq \lambda_1, \ \forall \mathbf{x} \neq \mathbf{0}.$$

(ii) For any fixed $j = 2, \ldots, n$,

$$\frac{\mathbf{x}'\mathbf{S}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \leq \lambda_j, \ \forall \mathbf{x} \neq \mathbf{0}$$

such that $\langle \mathbf{x}, \mathbf{x}_1 \rangle = \cdots = \langle \mathbf{x}, \mathbf{x}_{j-1} \rangle = 0$.

12. Demonstrate that if $\mathbf{A}$ is symmetric and $\mathbf{B} > \mathbf{0}$, then

$$\sup_{|\mathbf{h}|=1} \frac{\mathbf{h}'\mathbf{A}\mathbf{h}}{\mathbf{h}'\mathbf{B}\mathbf{h}} = \lambda_1(\mathbf{A}\mathbf{B}^{-1}),$$

where $\lambda_1(\mathbf{A}\mathbf{B}^{-1})$ denotes the largest eigenvalue of $\mathbf{A}\mathbf{B}^{-1}$.

13. Let $\mathbf{A}_m > \mathbf{0}$ in $\mathbb{R}_n^n$ $(m = 1, 2, \ldots)$ be a sequence. For any $\mathbf{A} \in \mathbb{R}_n^n$, define $||\mathbf{A}||^2 = \sum_{i,j} a_{ij}^2$ and let $\lambda_{1,m} \geq \cdots \geq \lambda_{n,m}$ be the ordered eigenvalues of $\mathbf{A}_m$. Prove that if $\lambda_{1,m} \to 1$ and $\lambda_{n,m} \to 1$, then $\lim_{m\to\infty} ||\mathbf{A}_m - \mathbf{I}|| = 0$.

14. In $\mathbb{R}^p$, prove that if $|\mathbf{x}_1| = |\mathbf{x}_2|$, then there exists $\mathbf{H} \in \mathbf{O}_p$ such that $\mathbf{H}\mathbf{x}_1 = \mathbf{x}_2$.
Hint: When $\mathbf{x}_1 \neq \mathbf{0}$, consider $\mathbf{H} \in \mathbf{O}_p$ with first row $\mathbf{x}_1'/|\mathbf{x}_1|$.

15. Show that for any $\mathbf{V} \in \mathbb{R}_n^n$ and any $m = 1, 2, \ldots$,
(i) if $(\mathbf{I} - t\mathbf{V})$ is nonsingular then [Srivastava and Khatri (1979), p. 33]

$$(\mathbf{I} - t\mathbf{V})^{-1} = \sum_{i=0}^{m} t^i \mathbf{V}^i + t^{m+1}\mathbf{V}^{m+1}(\mathbf{I} - t\mathbf{V})^{-1}.$$

(ii) If $\mathbf{V} > \mathbf{0}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $|t| < 1/\lambda_1$, then

$$(\mathbf{I} - t\mathbf{V})^{-1} = \sum_{i=0}^{\infty} t^i \mathbf{V}^i.$$

2
Random vectors

2.1 Introduction

A random vector is simply a vector whose components are random variables. The variables are the characteristics of interest that will be observed on each of the selected units in the sample. Questions related to probabilities of a variable to take on some values or probabilities of two or more variables to take on simultaneously values in a set are common in multivariate analysis. Chapter 2 gives a collection of important probability concepts on random vectors such as distribution functions, expected values, characteristic functions, discrete and absolutely continuous distributions, independence, etc.

2.2 Distribution functions

First, some basic notations concerning "rectangles" useful to describe the distribution function of a random vector are given. Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\} = [-\infty, \infty]$. It is convenient to define a partial order on $\bar{\mathbb{R}}^n$ by

$$\mathbf{x} \leq \mathbf{y} \text{ iff } x_i \leq y_i, \forall i = 1, \ldots, n,$$

and

$$\mathbf{x} < \mathbf{y} \text{ iff } x_i < y_i, \forall i = 1, \ldots, n.$$

This allows us to express "n-dimensional" rectangles in $\mathbb{R}^n$ succinctly:

$$I = (\mathbf{a}, \mathbf{b}] = \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{a} < \mathbf{x} \leq \mathbf{b}\} \text{ for any } \mathbf{a}, \ \mathbf{b} \in \bar{\mathbb{R}}^n.$$

The interior and closure of I are respectively

$$I^\circ = (\mathbf{a}, \mathbf{b}) = \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{a} < \mathbf{x} < \mathbf{b}\}$$

and

$$\bar{I} = [\mathbf{a}, \mathbf{b}] = \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$$

and the boundary of I is the "$(n-1)$-dimensional" relative complement

$$\partial I = \bar{I} - I^\circ.$$

Finally let the 2^n "corners" of I (a subset of $\bar{\mathbb{R}}^n$) be denoted by the cartesian product

$$\mathbf{a} \times \mathbf{b} = \times_{i=1}^n \{a_i, b_i\}.$$

Definition 2.1 *For* $\mathbf{x}$ *distributed on* $\mathbb{R}^n$*, the* distribution function *(d.f.) of* $\mathbf{x}$ *is the function* $F : \bar{\mathbb{R}}^n \to [0,1]$*, where* $F(\mathbf{t}) = P(\mathbf{x} \leq \mathbf{t})$*,* $\forall \mathbf{t} \in \bar{\mathbb{R}}^n$*. This is denoted* $\mathbf{x} \sim F$ *or* $\mathbf{x} \sim F_{\mathbf{x}}$*.*

A d.f. is automatically right-continuous; thus, if it is known on any dense subset $D \subset \mathbb{R}^n$, it is determined everywhere. This is because for any $\mathbf{t} \in \bar{\mathbb{R}}^n$, a sequence $\mathbf{d}_n$ may be chosen in D descending to $\mathbf{t}$: $\mathbf{d}_n \downarrow \mathbf{t}$.

From the d.f. may be computed the probability of any rectangle

$$P(\mathbf{a} < \mathbf{x} \leq \mathbf{b}) = \sum_{\mathbf{t} \in \mathbf{a} \times \mathbf{b}} (-1)^{N_{\mathbf{a}}(\mathbf{t})} F(\mathbf{t}), \ \forall \mathbf{a} < \mathbf{b},$$

where $N_{\mathbf{a}}(\mathbf{t}) = \sum_{i=1}^n \delta(a_i, t_i)$ counts the number of t_i's that are a_i's.

The borel subsets of $\mathbb{R}^n$ comprise the smallest σ-algebra containing the rectangles

$$\mathcal{B}^n = \sigma\left((\mathbf{a}, \mathbf{b}] : \ \mathbf{a}, \mathbf{b} \in \mathbb{R}^n\right).$$

The class $\mathcal{G}^n$ of all countable disjoint unions of rectangles contains all the open subsets of $\mathbb{R}^n$, and if we let $G = \sum_{i=1}^\infty (\mathbf{a}_i, \mathbf{b}_i]$ denote a generic element in this class, it follows that

$$P(\mathbf{x} \in G) = \sum_{i=1}^\infty P(\mathbf{a}_i < \mathbf{x} \leq \mathbf{b}_i).$$

By the Caratheodory extension theorem (C.E.T.), the probability of a general borel set $A \in \mathcal{B}^n$ is then uniquely determined by the formula

$$P_{\mathbf{x}}(A) \equiv P(\mathbf{x} \in A) = \inf_{A \subset G} P(\mathbf{x} \in G).$$

2.3 Equals-in-distribution

Definition 2.2 $\mathbf{x}$ *and* $\mathbf{y}$ *are* equidistributed *(identically distributed), denoted* $\mathbf{x} \stackrel{d}{=} \mathbf{y}$*, iff* $P_{\mathbf{x}}(A) = P_{\mathbf{y}}(A)$, $\forall A \in \mathcal{B}^n$.

On the basis of the previous section, it should be clear that for any dense $D \subset \mathbb{R}^n$:

Proposition 2.1 (C.E.T) $\mathbf{x} \stackrel{d}{=} \mathbf{y} \iff F_{\mathbf{x}}(\mathbf{t}) = F_{\mathbf{y}}(\mathbf{t}),\ \forall \mathbf{t} \in D.$

Although at first glance, $\stackrel{d}{=}$ looks like nothing more than a convenient shorthand symbol, there is an immediate consequence of the definition, deceptively simple to state and prove, that has powerful application in the sequel.

Let $g : \mathbb{R}^n \to \Omega$ where Ω is a completely arbitrary space.

Proposition 2.2 (Invariance) $\mathbf{x} \stackrel{d}{=} \mathbf{y} \Longrightarrow g(\mathbf{x}) \stackrel{d}{=} g(\mathbf{y})$.

Proof.

$$P\left(g(\mathbf{x}) \in B\right) = P\left(\mathbf{x} \in g^{-1}(B)\right) = P\left(\mathbf{y} \in g^{-1}(B)\right) = P\left(g(\mathbf{y}) \in B\right).$$

□

Example 2.1

$$\begin{aligned}
\mathbf{x} \stackrel{d}{=} \mathbf{y} \quad &\Longrightarrow \quad x_i \stackrel{d}{=} y_i,\ i = 1, \ldots, n \\
&\Longrightarrow \quad x_i x_j \stackrel{d}{=} y_i y_j,\ i, j = 1, \ldots, n \\
&\Longrightarrow \quad \prod_{i=1}^{n} x_i^{r_i} \stackrel{d}{=} \prod_{i=1}^{n} y_i^{r_i},\ \textit{for any } r_i,\ i = 1, \ldots, n \\
&\Longrightarrow \quad \textit{etc.}
\end{aligned}$$

2.4 Discrete distributions

Definition 2.3 *The* probability function *(p.f.) of* $\mathbf{x}$ *is the function*

$$p : \bar{\mathbb{R}}^n \to [0, 1] \textit{ where } p(\mathbf{t}) = P(\mathbf{x} = \mathbf{t}),\ \forall \mathbf{t} \in \bar{\mathbb{R}}^n.$$

The p.f. may be evaluated directly from the d.f.:

$$p(\mathbf{t}) = \lim_{\mathbf{s}_m \uparrow \mathbf{t}} P(\mathbf{s}_m < \mathbf{x} \leq \mathbf{t}),$$

where $\mathbf{s}_m \uparrow \mathbf{t}$ means $\mathbf{s}_1 < \mathbf{s}_2 < \cdots$ and $\mathbf{s}_m \to \mathbf{t}$ as $m \to \infty$. The subset $D = p^{-1}(0)^c$ where the p.f. is nonzero may contain at most a countable number of points. D is known as the discrete part of $\mathbf{x}$, and $\mathbf{x}$ is said to be discrete if it is "concentrated" on D:

Definition 2.4 $\mathbf{x}$ *is* discrete *iff* $P(\mathbf{x} \in D) = 1$.

One may verify that

$$\mathbf{x} \text{ is discrete } \iff P(\mathbf{x} \in A) = \sum_{\mathbf{t} \in A \cap D} p(\mathbf{t}),\ \forall A \in \mathcal{B}^n.$$

Thus, the distribution of $\mathbf{x}$ is entirely determined by its p.f. if and only if it is discrete, and in this case, we may simply write $\mathbf{x} \sim p$ or $\mathbf{x} \sim p_{\mathbf{x}}$.

2.5 Expected values

For any event A, we may consider the indicator function

$$I_A(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in A \\ 0, & \mathbf{x} \notin A. \end{cases}$$

It is clear that $I_A(\mathbf{x})$ is itself a discrete random variable, referred to as a *Bernoulli trial*, for which

$$P\left(I_A(\mathbf{x}) = 1\right) = P_{\mathbf{x}}(A) \text{ and } P\left(I_A(\mathbf{x}) = 0\right) = 1 - P_{\mathbf{x}}(A).$$

This is denoted $I_A(\mathbf{x}) \sim \text{Bernoulli}\left(P_{\mathbf{x}}(A)\right)$ and we define $E\ I_A(\mathbf{x}) = P_{\mathbf{x}}(A)$.

For any k mutually disjoint and exhaustive events $A_1, \ldots, A_k$ and k real numbers $a_1, \ldots, a_k$, we may form the simple function

$$s(\mathbf{x}) = a_1 I_{A_1}(\mathbf{x}) + \cdots + a_k I_{A_k}(\mathbf{x}).$$

Obviously, $s(\mathbf{x})$ is also discrete with

$$P\left(s(\mathbf{x}) = a_i\right) = P_{\mathbf{x}}(A_i),\ i = 1, \ldots, k.$$

By requiring that E be linear, we (are forced to) define

$$E\ s(\mathbf{x}) = a_1 P_{\mathbf{x}}(A_1) + \cdots + a_k P_{\mathbf{x}}(A_k).$$

The most general function for which we need ever compute an expected value may be directly expressed as a limit of a sequence of simple functions. Such a function $g(\mathbf{x})$ is said to be measurable and we may explicitly write

$$g(\mathbf{x}) = \lim_{N \to \infty} s_N(\mathbf{x}),$$

where convergence holds pointwise, i.e., for every fixed $\mathbf{x}$. If $g(\mathbf{x})$ is non-negative, it can be proven that we may always choose the sequence of simple functions to be themselves non-negative and nondecreasing as a sequence whereupon we define

$$E\ g(\mathbf{x}) = \lim_{N \to \infty} E\ s_N(\mathbf{x}) = \sup_N E\ s_N(\mathbf{x}).$$

Then, in general, we write $g(\mathbf{x})$ as the difference of its positive and negative parts

$$g(\mathbf{x}) = g^+(\mathbf{x}) - g^-(\mathbf{x}),$$

defined by

$$g^+(\mathbf{x}) = \begin{cases} g(\mathbf{x}), & g(\mathbf{x}) \geq 0 \\ 0, & g(\mathbf{x}) < 0, \end{cases}$$

$$g^-(\mathbf{x}) = \begin{cases} -g(\mathbf{x}), & g(\mathbf{x}) \leq 0 \\ 0, & g(\mathbf{x}) < 0, \end{cases}$$

and finish by defining

$$E\ g(\mathbf{x}) = \begin{cases} E\ g^+(\mathbf{x}) - E\ g^-(\mathbf{x}), & \text{if } E\ g^+(\mathbf{x}) < \infty \text{ or } E\ g^-(\mathbf{x}) < \infty \\ \text{"undefined,"} & \text{otherwise.} \end{cases}$$

We may sometimes use the Leibniz notation

$$E\ g(\mathbf{x}) = \int g(\mathbf{t}) dP_{\mathbf{x}}(\mathbf{t}) = \int g(\mathbf{t}) dF(\mathbf{t}).$$

One should verify the fundamental inequality $|E\ g(\mathbf{x})| \leq E\ |g(\mathbf{x})|$.

Let $\uparrow$ denote convergence of a monotonically nondecreasing sequence. Something is said to happen for almost all $\mathbf{x}$ if it fails to happen on a set A such that $P_{\mathbf{x}}(A) = 0$. The two main theorems concerning "continuity" of E are the following:

Proposition 2.3 (Monotone convergence theorem (M.C.T.)) *Suppose* $0 \leq g_1(\mathbf{x}) \leq g_2(\mathbf{x}) \leq \cdots$. *If* $g_N(\mathbf{x}) \uparrow g(\mathbf{x})$, *for almost all* $\mathbf{x}$, *then* $E\ g_N(\mathbf{x}) \uparrow E\ g(\mathbf{x})$.

Proposition 2.4 (Dominated convergence theorem (D.C.T.)) *If* $g_N(\mathbf{x}) \to g(\mathbf{x})$, *for almost all* $\mathbf{x}$, *and* $|g_N(\mathbf{x})| \leq h(\mathbf{x})$ *with* $E\ h(\mathbf{x}) < \infty$, *then* $E\ |g_N(\mathbf{x}) - g(\mathbf{x})| \to 0$ *and, thus, also* $E\ g_N(\mathbf{x}) \to E\ g(\mathbf{x})$.

It should be clear by the process whereby expectation is defined (in stages) that we have

Proposition 2.5 $\mathbf{x} \stackrel{d}{=} \mathbf{y} \iff E\ g(\mathbf{x}) = E\ g(\mathbf{y}),\ \forall g$ *measurable.*

2.6 Mean and variance

Consider the "linear functional" $\mathbf{t}'\mathbf{x} = \sum_{i=1}^n t_i x_i$ for each (fixed) $\mathbf{t} \in \mathbb{R}^n$, and the "euclidean norm" (length) $|\mathbf{x}| = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$. By any of three equivalent ways, for $p > 0$ one may say that the pth moment of $\mathbf{x}$ is finite:

$$\begin{aligned} E\ |\mathbf{t}'\mathbf{x}|^p < \infty,\ \forall \mathbf{t} \in \mathbb{R}^n &\iff E\ |x_i|^p < \infty,\ i = 1, \ldots, n \\ &\iff E\ |\mathbf{x}|^p < \infty. \end{aligned}$$

To show this, one must realize that $|x_i| \leq |\mathbf{x}| \leq \sum_{i=1}^n |x_i|$ and $\mathcal{L}_p = \{\mathbf{x} \in \mathbb{R}^n :\ E\ |\mathbf{x}|^p < \infty\}$ is a linear space (v. Problem 2.14.3).

From the simple inequality $a^r \leq 1 + a^p$, $\forall a \geq 0$ and $0 < r \leq p$, if we let $a = |\mathbf{x}|$ and take expectations, we get $E\ |\mathbf{x}|^r \leq 1 + E\ |\mathbf{x}|^p$. Hence, if for

$p > 0$, the pth moment of $\mathbf{x}$ is finite, then also the rth moment is finite, for any $0 < r \leq p$.

A product-moment of order p for $\mathbf{x} = (x_1, \ldots, x_n)'$ is defined by

$$E \prod_{i=1}^{n} x_i^{p_i},\ p_i \geq 0,\ i = 1, \ldots, n,\ \sum_{i=1}^{n} p_i = p.$$

A useful inequality to determine that a product-moment is finite is Hölder's inequality:

Proposition 2.6 (Hölder's inequality) *For any univariate random variables x and y,*

$$E\ |xy| \leq (E\ |x|^r)^{1/r} \cdot (E\ |y|^s)^{1/s},\ r > 1,\ \frac{1}{r} + \frac{1}{s} = 1.$$

From this inequality, if the pth moment of $\mathbf{x} \in \mathbb{R}^n$ is finite, then all product-moments of order p are also finite. This can be verified for $n = 2$, as Hölder's inequality gives

$$E\ |x_1^{p_1} x_2^{p_2}| \leq (E\ |x_1|^p)^{p_1/p} \cdot (E\ |x_2|^p)^{p_2/p},\ p_i \geq 0,\ i = 1, 2,\ p_1 + p_2 = p.$$

The conclusion for general n follows by induction.

If the first moment of $\mathbf{x}$ is finite we define the *mean* of $\mathbf{x}$ by

$$\boldsymbol{\mu} = E\ \mathbf{x} \stackrel{\text{def}}{=} (E\ x_i) = (\mu_i).$$

If the second moment of $\mathbf{x}$ is finite, we define the *variance* of $\mathbf{x}$ by

$$\boldsymbol{\Sigma} = \text{var}\ \mathbf{x} \stackrel{\text{def}}{=} (\text{cov}(x_i, x_j)) = (\sigma_{ij})\,.$$

In general, we define the expected value of any multiply indexed array of univariate random variables, $\boldsymbol{\xi} = (x_{ijk\ldots})$, componentwise by $E\ \boldsymbol{\xi} = (E\ x_{ijk\ldots})$. Vectors and matrices are thus only special cases and it is obvious that

$$\boldsymbol{\Sigma} = E\ (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' = E\ \mathbf{x}\mathbf{x}' - \boldsymbol{\mu}\boldsymbol{\mu}'.$$

It is also obvious that for any $\mathbf{A} \in \mathbb{R}_n^m$,

$$E\ \mathbf{A}\mathbf{x} = \mathbf{A}\boldsymbol{\mu} \text{ and var } \mathbf{A}\mathbf{x} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.$$

In particular, $E\ \mathbf{t}'\mathbf{x} = \mathbf{t}'\boldsymbol{\mu}$ and var $\mathbf{t}'\mathbf{x} = \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} \geq 0$, $\forall \mathbf{t} \in \mathbb{R}^n$. Now, the reader should verify that more generally

$$\text{cov}(\mathbf{s}'\mathbf{x}, \mathbf{t}'\mathbf{x}) = \mathbf{s}'\boldsymbol{\Sigma}\mathbf{t}$$

and that considered as a function of $\mathbf{s}$ and $\mathbf{t}$, the left-hand side defines a (pseudo) inner product. Thus, $\boldsymbol{\Sigma}$ is automatically positive semidefinite, $\boldsymbol{\Sigma} \geq \mathbf{0}$. But by this, we may immediately write $\boldsymbol{\Sigma} = \mathbf{H}\mathbf{D}\mathbf{H}'$ with $\mathbf{H}$ orthogonal and $\mathbf{D} = \text{diag}(\boldsymbol{\lambda})$, where the columns of $\mathbf{H}$ comprise an orthonormal basis of "eigenvectors" and the components of $\boldsymbol{\lambda} \geq \mathbf{0}$ list the corresponding

"eigenvalues." Accordingly, we may always "normalize" any $\mathbf{x}$ with $\mathbf{\Sigma} > \mathbf{0}$ by letting

$$\mathbf{z} = \mathbf{D}^{-1/2}\mathbf{H}'(\mathbf{x} - \boldsymbol{\mu}),$$

which represents a three-stage transformation of $\mathbf{x}$ in which we first relocate by $\boldsymbol{\mu}$, then rotate by $\mathbf{H}'$, and, finally, rescale by $\lambda_i^{-1/2}$ independently along each axis. We find, of course, that

$$E\ \mathbf{z} = \mathbf{0} \text{ and } \text{var}\ \mathbf{z} = \mathbf{I}.$$

The linear transformation $\mathbf{z} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ also satisfies $E\ \mathbf{z} = \mathbf{0}$ and $\text{var}\ \mathbf{z} = \mathbf{I}$.

When the vector $\mathbf{x} \in \mathbb{R}^n$ is partitioned as $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$, where $\mathbf{y} \in \mathbb{R}^r$, $\mathbf{z} \in \mathbb{R}^s$, and $n = r + s$, it is useful to define the covariance between two vectors. The covariance matrix between $\mathbf{y}$ and $\mathbf{z}$ is, by definition,

$$\text{cov}(\mathbf{y}, \mathbf{z}) = (\text{cov}(y_i, z_j)) \in \mathbb{R}_s^r.$$

Then, we may write

$$\text{var}(\mathbf{x}) = \begin{pmatrix} \text{var}(\mathbf{y}) & \text{cov}(\mathbf{y}, \mathbf{z}) \\ \text{cov}(\mathbf{z}, \mathbf{y}) & \text{var}(\mathbf{z}) \end{pmatrix}.$$

Sometimes, expected value of $\mathbf{y}$ is easier to calculate by conditioning on another random vector $\mathbf{z}$. In this regard, the conditional mean theorem and conditional variance theorem are stated. A general proof of the conditional mean theorem can be found in Billingsley (1995, Section 34).

Proposition 2.7 (Conditional mean formula) $E[E(\mathbf{y}|\mathbf{z})] = E\ \mathbf{y}$.

An immediate consequence is the conditional variance formula.

Proposition 2.8 (Conditional variance formula)

$$\text{var}\ \mathbf{y} = E[\text{var}(\mathbf{y}|\mathbf{z})] + \text{var}[E(\mathbf{y}|\mathbf{z})].$$

Example 2.2 *Define a group variable I such that*

$$\begin{aligned} P(I = 1) &= 1 - \epsilon, \\ P(I = 2) &= \epsilon. \end{aligned}$$

Conditionally on I, assume

$$\begin{aligned} x|I = 1 &\sim N(\mu_1, \sigma_1^2), \\ x|I = 2 &\sim N(\mu_2, \sigma_2^2). \end{aligned}$$

Then

$$\begin{aligned} f_x(x) &= (1 - \epsilon)(2\pi)^{-1/2} \frac{1}{\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x - \mu_1)^2\right) \\ &\quad + \epsilon(2\pi)^{-1/2} \frac{1}{\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x - \mu_2)^2\right) \end{aligned}$$

is a mixture or ϵ-contaminated normal density. It follows from the construction of x that

$$\begin{aligned} E\ x &= E[E(x|I)] = (1-\epsilon)\mu_1 + \epsilon\mu_2 \equiv \mu, \\ \text{var}\ x &= E[\text{var}(x|I)] + \text{var}[E(x|I)] \\ &= (1-\epsilon)\sigma_1^2 + \epsilon\sigma_2^2 + (1-\epsilon)(\mu_1-\mu)^2 + \epsilon(\mu_2-\mu)^2. \end{aligned}$$

2.7 Characteristic functions

We require only the most basic facts about characteristic functions.

Definition 2.5 *The* characteristic function *of* $\mathbf{x}$ *is the function* $c : \mathbb{R}^n \to \mathbb{C}$ *defined by*

$$c(\mathbf{t}) = c_{\mathbf{x}}(\mathbf{t}) = E\ e^{i\mathbf{t}'\mathbf{x}}.$$

Note:

1. $c(\mathbf{0}) = 1$, $|c(\mathbf{t})| \leq 1$ and $c(-\mathbf{t}) = \overline{c(\mathbf{t})}$.
2. $c(\mathbf{t})$ is uniformly continuous:

$$\begin{aligned} |c(\mathbf{t}) - c(\mathbf{s})| &= \left| E\ \left(e^{i(\mathbf{t}-\mathbf{s})'\mathbf{x}} - 1 \right) e^{i\mathbf{s}'\mathbf{x}} \right| \\ &\leq E\ \left| e^{i(\mathbf{t}-\mathbf{s})'\mathbf{x}} - 1 \right|. \end{aligned}$$

Since $\left| e^{i(\mathbf{t}-\mathbf{s})'\mathbf{x}} - 1 \right| \leq 2$, continuity follows by the D.C.T. Uniformity holds since $\left| e^{i(\mathbf{t}-\mathbf{s})'\mathbf{x}} - 1 \right|$ depends only on $\mathbf{t} - \mathbf{s}$.

The main result is perhaps the "inversion formula" proven in Appendix A:

$$P_{\mathbf{x}}(\mathbf{a}, \mathbf{b}] = \lim_{N\to\infty} \frac{1}{(2\pi)^n} \int_{(\mathbf{a},\mathbf{b}]} \int_{\mathbb{R}^n} e^{-i\mathbf{t}'\mathbf{x}} c(\mathbf{t}) e^{-\mathbf{t}'\mathbf{t}/2N^2} d\mathbf{t} d\mathbf{x},$$

$\forall \mathbf{a}, \mathbf{b}$ such that $P_{\mathbf{x}}(\partial(\mathbf{a}, \mathbf{b}]) = 0$. Thus, the C.E.T. may be applied immediately to produce the technically equivalent:

Proposition 2.9 (Uniqueness) $\mathbf{x} \stackrel{d}{=} \mathbf{y} \iff c_{\mathbf{x}}(\mathbf{t}) = c_{\mathbf{y}}(\mathbf{t}),\ \forall \mathbf{t} \in \mathbb{R}^n$.

Now if we consider the linear functionals of $\mathbf{x}$: $\mathbf{t}'\mathbf{x}$ with $\mathbf{t} \in \mathbb{R}^n$, it is clear that $c_{\mathbf{t}'\mathbf{x}}(s) = c_{\mathbf{x}}(s\mathbf{t})$, $\forall s \in \mathbb{R}$, $\mathbf{t} \in \mathbb{R}^n$, so that the characteristic function of $\mathbf{x}$ determines all those of $\mathbf{t}'\mathbf{x}$, $\mathbf{t} \in \mathbb{R}^n$ and vice versa.

Let $S^{n-1} = \{\mathbf{s} \in \mathbb{R}^n : \ |\mathbf{s}| = 1\}$ be the "unit sphere" in $\mathbb{R}^n$, and we have

Proposition 2.10 (Cramér-Wold) $\mathbf{x} \stackrel{d}{=} \mathbf{y} \iff \mathbf{t}'\mathbf{x} \stackrel{d}{=} \mathbf{t}'\mathbf{y},\ \forall \mathbf{t} \in S^{n-1}$.

Proof. Since $c_{\mathbf{t}'\mathbf{x}}(s) = c_{\mathbf{x}}(s\mathbf{t})$, $\forall s \in \mathbb{R}$, $\mathbf{t} \in \mathbb{R}^n$, it is clear that

$$\mathbf{x} \stackrel{d}{=} \mathbf{y} \Longleftrightarrow \mathbf{t}'\mathbf{x} \stackrel{d}{=} \mathbf{t}'\mathbf{y},\ \forall \mathbf{t} \in \mathbb{R}^n.$$

Since $\mathbf{t}'\mathbf{x} = |\mathbf{t}| \left(\frac{\mathbf{t}}{|\mathbf{t}|}\right)' \mathbf{x}$, $\forall \mathbf{t} \neq \mathbf{0}$, it is also clear that

$$\mathbf{t}'\mathbf{x} \stackrel{d}{=} \mathbf{t}'\mathbf{y},\ \forall \mathbf{t} \in \mathbb{R}^n \Longleftrightarrow \mathbf{t}'\mathbf{x} \stackrel{d}{=} \mathbf{t}'\mathbf{y},\ \forall \mathbf{t} \in S^{n-1}.$$

□

By this result, it is clear that one may reduce a good many issues concerning random vectors to the univariate level.

In the specific matter of computation, the reader should know that in the special case of a univariate random variable X:

> If $E\ e^{\pm\delta X} < \infty$ for any $\delta > 0$, the Laplace transform of X is determined in the strip $|\mathrm{Re}(z)| \leq \delta$ as the (absolutely convergent) power series
>
> $$L_X(z) = \sum_{n=0}^{\infty} E\ X^n z^n / n!,$$
>
> and since such a power series is completely determined by its coefficients, we find that one may legitimately obtain the characteristic function
>
> $$c_X(t) = L_X(it),\ \forall t \in \mathbb{R},$$
>
> by merely observing the coefficients in an expansion of the moment-generating function since they are necessarily the same as those of the Laplace transform:
>
> $$m_X(t) = L_X(t),\ \forall |t| \leq \delta.$$

Example 2.3 *Suppose* $f_z(s) = (2\pi)^{-1/2} e^{-s^2/2}$. *One easily computes the moment-generating function (m.g.f.), finding*

$$m_z(t) = e^{t^2/2},$$

which has the obvious expansion for every t, *whereupon*

$$c_z(t) = e^{-t^2/2}. \tag{2.1}$$

2.8 Absolutely continuous distributions

Lebesgue measure, λ, is the extension to all borel sets of our natural sense of volume measure in $\mathbb{R}^n$. Thus, we define

$$\lambda(\mathbf{a}, \mathbf{b}] \quad = \quad \prod_{i=1}^{n} (b_i - a_i),\ \forall \mathbf{a} < \mathbf{b} \text{ in } \mathbb{R}^n,$$

$$\lambda(G) \quad = \quad \sum_{i=1}^{\infty} \lambda(\mathbf{a}_i, \mathbf{b}_i], \ \forall G = \sum_{i=1}^{\infty} (\mathbf{a}_i, \mathbf{b}_i] \text{ in } \mathcal{G}^n,$$

and

$$\lambda(A) = \inf_{A \subset G} \lambda(G), \ \forall A \text{ in } \mathcal{B}^n.$$

As before, the C.E.T. guarantees that λ is a measure on $\mathcal{B}^n$. We will often denote Lebesgue measure explicitly as volume: $\lambda(A) = \text{vol}(A)$. Incidentally, something is said to happen "almost everywhere" (a.e.) if the set where it fails to happen has zero volume.

Now, the general conception of a random vector continuously distributed in space is that the probabilities of events will depend continuously on the volume of the events. Thus,

Definition 2.6 $\mathbf{x}$ *is* absolutely continuous*, denoted* $\mathbf{x} \ll \lambda$*, iff*

$$\forall \epsilon > 0, \ \exists \delta > 0 \text{ such that } \text{vol}(A) < \delta \Longrightarrow P(\mathbf{x} \in A) < \epsilon.$$

But, in that case,

Proposition 2.11 $\mathbf{x} \ll \lambda \Longleftrightarrow \text{vol}(A) = 0 \Longrightarrow P(\mathbf{x} \in A) = 0.$

Proof. Assume $\mathbf{x} \ll \lambda$. If $\text{vol}(A) = 0$ but $P(\mathbf{x} \in A) \neq 0$ we may take $\epsilon = P(\mathbf{x} \in A)/2$ to find the contradiction that $P(\mathbf{x} \in A) < \epsilon$. Conversely, suppose $\text{vol}(A) = 0 \Longrightarrow P(\mathbf{x} \in A) = 0$ but that $\mathbf{x} \not\ll \lambda$. Then, $\exists \epsilon_0 > 0$ such that $\forall n$, $\exists A_n$ with $\text{vol}(A_n) < 1/2^n$ but $P(\mathbf{x} \in A_n) \geq \epsilon_0$. Letting $A = \overline{\lim} A_n = \cap_{n=1}^{\infty} \cup_{k \geq n} A_k$, since $\cup_{k \geq n} A_k$ is a monotone sequence we find the contradiction that $\text{vol}(A) = 0$ but $P(\mathbf{x} \in A) \geq \epsilon_0$. □

Thus, a distribution which depends continuously on volume satisfies the relatively simple criterion

$$\mathbf{x} \ll \lambda \Longleftrightarrow \text{vol}(A) = 0 \Longrightarrow P(\mathbf{x} \in A) = 0.$$

However, it is on this particular criterion, by the theorem of Radon-Nikodym, that absolute continuity is characterized finally in terms of densities:

Proposition 2.12 (Radon-Nikodym) $\mathbf{x}$ *is absolutely continuous* $\Longleftrightarrow$ *there is a (a.e.-unique) probability density function (p.d.f.)* $f : \mathbb{R}^n \to [0, \infty)$ *such that*

$$P(\mathbf{x} \in A) = \int_A f(\mathbf{t}) d\mathbf{t}, \ \forall A \in \mathcal{B}^n.$$

But since the p.d.f. then determines such a distribution completely, we may simply write $\mathbf{x} \sim f$ or $\mathbf{x} \sim f_{\mathbf{x}}$.

It is, of course, by the extension process that defines expectation (in stages) that automatically

$$E \ g(\mathbf{x}) = \int g(\mathbf{t}) f(\mathbf{t}) d\mathbf{t}, \ \forall g \text{ measurable},$$

such that $E\, g(\mathbf{x})$ is defined.

Now in particular, the distribution function may itself be expressed as

$$F(\mathbf{t}) = P(\mathbf{x} \leq \mathbf{t}) = \int_{-\infty}^{\mathbf{t}} f(\mathbf{s}) d\mathbf{s},\ \forall \mathbf{t} \in \mathbb{R}^n.$$

In practice, we will often be able to invoke the fundamental theorems of calculus to obtain an explicit representation of the p.d.f. by simply differentiating the d.f.:

1. By the first fundamental theorem of calculus,
$$f(\mathbf{t}) = \partial^n F(\mathbf{t}) / \partial t_1 \cdots \partial t_n$$
at every $\mathbf{t}$ where $f(\mathbf{t})$ is continuous.
2. Also, by the second fundamental theorem, if
$$f(\mathbf{t}) = \partial^n F(\mathbf{t}) / \partial t_1 \cdots \partial t_n$$
exists and is continuous (a.e.) on some rectangle I, then
$$P(\mathbf{x} \in A) = \int_A f(\mathbf{t}) d\mathbf{t},\ \forall A \subset I.$$

Finally, in relation to the inversion formula, when the characteristic function $c(\mathbf{t})$ is absolutely integrable, i.e., $\int_{\mathbb{R}^n} |c(\mathbf{t})| d\mathbf{t} < \infty$, the corresponding distribution function is absolutely continuous with p.d.f. (v. Appendix A):

$$f(\mathbf{s}) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\mathbf{t}'\mathbf{s}} c(\mathbf{t}) d\mathbf{t}. \tag{2.2}$$

2.9 Uniform distributions

The most fundamental absolutely continuous distribution would, of course, be conveyed by volume measure itself. Consider any event C for which $0 < \mathrm{vol}(C) < \infty$.

Definition 2.7 $\mathbf{x}$ *is* uniformly distributed *on* C, *denoted* $\mathbf{x} \sim \mathrm{unif}(C)$, *iff*

$$P(\mathbf{x} \in A) = \mathrm{vol}(AC)/\mathrm{vol}(C),\ \forall A \in \mathcal{B}^n.$$

If $\mathrm{vol}(\partial C) = 0$, as is often the case, we may just as well include as exclude it, so if $\mathbf{x} \sim \mathrm{unif}(C)$, $\mathbf{y} \sim \mathrm{unif}(C^\circ)$ and $\mathbf{z} \sim \mathrm{unif}(\bar{C})$, then $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ are equidistributed:

$$\mathbf{x} \overset{d}{=} \mathbf{y} \overset{d}{=} \mathbf{z}.$$

Now, for $\mathbf{x} \sim \mathrm{unif}(C)$ we may immediately reexpress each probability as an integral:

$$P(\mathbf{x} \in A) = \int_A k \cdot I_C(\mathbf{t}) d\mathbf{t},\ \forall A \in \mathcal{B}^n,$$

where

$$I_C(\mathbf{t}) = \begin{cases} 1, & \mathbf{t} \in C \\ 0, & \mathbf{t} \notin C, \end{cases}$$

is the indicator function for C and $k = \text{vol}(C)^{-1}$. We thus have an explicit determination of "the" density for $\mathbf{x}$:

$$f(\mathbf{t}) = k\ I_C(\mathbf{t}),\ \forall \mathbf{t} \in \mathbb{R}^n.$$

Example 2.4 *For* $\mathbf{x} \sim \text{unif}([\mathbf{0}, \mathbf{1}])$ *on* $\mathbb{R}^n$*, the p.d.f. may be expressed as a simple product*

$$f(\mathbf{t}) = I_{[\mathbf{0},\mathbf{1}]}(\mathbf{t}) = \prod_{i=1}^{n} I_{[0,1]}(t_i),\ \forall \mathbf{t} \in \mathbb{R}^n,$$

from which

$$F(\mathbf{t}) = \prod_{i=1}^{n} \left(t_i I_{[0,1]}(t_i) + I_{(1,\infty)}(t_i) \right),\ \forall \mathbf{t} \in \mathbb{R}^n.$$

2.10 Joints and marginals

Consider $\mathbf{x}_i \sim F_i$ on $\mathbb{R}^{n_i}$, $i = 1, \ldots, k$, with

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_k \end{pmatrix} \sim F \text{ on } \mathbb{R}^n \text{ where } n = \sum_{i=1}^{k} n_i.$$

$\mathbf{x}$ is called the *joint* of $\mathbf{x}_1, \ldots, \mathbf{x}_k$ which are, in turn, called *marginals* of $\mathbf{x}$.

Since it is clear that

$$P(\mathbf{x} \leq \mathbf{t}) = P(\mathbf{x}_1 \leq \mathbf{t}_1, \ldots, \mathbf{x}_k \leq \mathbf{t}_k),\ \forall \mathbf{t} = \begin{pmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_k \end{pmatrix},$$

we will, by a slight abuse of our notation, write

$$F(\mathbf{t}) = F(\mathbf{t}_1, \ldots, \mathbf{t}_k),\ \forall \mathbf{t} = \begin{pmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_k \end{pmatrix}$$

to reflect this "partitioning." In this way, the distribution function is said to express the joint distribution of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, and the marginals may be recovered on the simple substitution of $\boldsymbol{\infty}$ in all but the ith place:

$$F_i(\mathbf{s}) = F(\boldsymbol{\infty}, \ldots, \mathbf{s}, \ldots, \boldsymbol{\infty}),\ \forall \mathbf{s} \in \mathbb{R}^{n_i}.$$

In the special case where $\mathbf{x}$ is absolutely continuous with p.d.f. $f(\mathbf{t}) = f(\mathbf{t}_1, \ldots, \mathbf{t}_k)$, it follows that each $\mathbf{x}_i$ is also absolutely continuous with p.d.f. $f_i(\mathbf{s})$ that is obtained by "integrating out" the other variables:

$$f_i(\mathbf{s}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{t}_1, \ldots, \mathbf{s}, \ldots, \mathbf{t}_k) \prod_{\substack{1 \le j \le k \\ j \ne i}} d\mathbf{t}_j, \ \forall \mathbf{s} \in \mathbb{R}^{n_i}.$$

This is by direct application of Fubini's theorem whereby we may interchange the order of integration in a product integral to verify that

$$P(\mathbf{x}_i \in A) = \int_A f_i(\mathbf{s}) d\mathbf{s}, \ \forall A \in \mathcal{B}^{n_i}$$

and, of course, in particular,

$$F_i(\mathbf{s}) = \int_{-\infty}^{\mathbf{s}} f_i(\mathbf{u}) d\mathbf{u}, \ \forall \mathbf{s} \in \mathbb{R}^{n_i}.$$

Koehler and Symanowski (1995) presented a method for constructing multivariate distributions with any specific set of univariate marginals. It provides a rich class of distributions for modeling multivariate data as well as a basis for easily simulating correlated observations. The inclusion of different association parameters for differents subsets of variables allows for many different patterns of associations. Their work follows those of Genest and MacKay (1986) and Marshall and Olkin (1988), among others. A tool called linkage [Li et al. (1996)] can be used for the construction of multivariate distributions with given multivariate marginals; Cuadras (1992) found related results.

Example 2.5 *The bivariate parametric family of d.f.'s on* $[0,1]^2$ *of Cook and Johnson* (1981) *is defined by*

$$F(t_1, t_2; \alpha) = \left[\frac{1}{t_1^\alpha} + \frac{1}{t_2^\alpha} - 1 \right]^{-1/\alpha}, \ \alpha > 0. \tag{2.3}$$

The case $\alpha = 0$ *can be defined by continuity. It has marginals*

$$\begin{aligned} F_1(t_1) &= F(t_1, 1; \alpha) = t_1, \\ F_2(t_2) &= F(1, t_2; \alpha) = t_2, \end{aligned}$$

which are identically distributed as unif$([0,1])$. *Multivariate distributions on* $[\mathbf{0}, \mathbf{1}]$ *with uniform marginals are often referred to as* copulas. *The slight modification*

$$F(t_1, t_2; \alpha) = \left[\frac{1}{F_1(t_1)^\alpha} + \frac{1}{F_2(t_2)^\alpha} - 1 \right]^{-1/\alpha}, \ \alpha > 0,$$

is a bivariate distribution with arbitrary marginals F_1 *and* F_2. *The bivariate parametric family of d.f.'s on* $[0,1]^2$ *of Frank* (1979) *[v. also Genest* (1987)*]*

$$F(t_1, t_2; \alpha) = \log_\alpha \left[1 + \frac{(\alpha^{t_1} - 1)(\alpha^{t_2} - 1)}{(\alpha - 1)} \right], \ \alpha > 0, \tag{2.4}$$

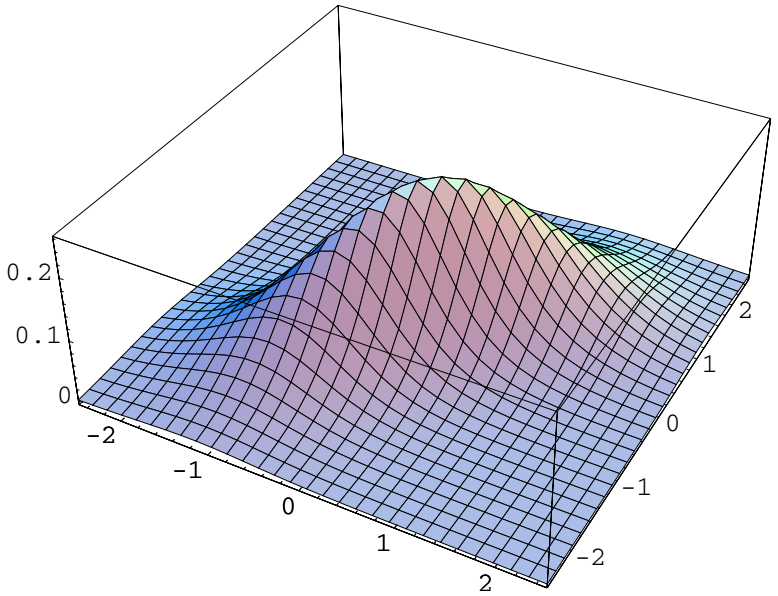

Figure 2.1. Bivariate Frank density with standard normal marginals and a correlation of 0.7.

(the case $\alpha = 1$ can be defined by continuity), where $\log_\alpha(\cdot)$ denotes logarithm in base α, is also a copula. Such distributions have found applications in modeling survival data [Oakes (1982)*, Carrière* (1994)*]. Figure 2.1 is a graph of a bivariate Frank density with standard normal marginals. The association parameter $\alpha = 0.00296$ using Nelsen* (1986) *corresponds to a correlation of* 0.7*.*

2.11 Independence

Definition 2.8 $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *are mutually statistically* independent *iff*

$$P(\mathbf{x}_1 \in A_1, \ldots, \mathbf{x}_k \in A_k) = \prod_{i=1}^{k} P(\mathbf{x}_i \in A_i), \ \forall A_i \in \mathcal{B}^{n_i}, \ i = 1, \ldots, k.$$

Denote pairwise independence ($k = 2$) simply $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$. By the extension process that defines expectation, we ultimately find:

Proposition 2.13

$$\mathbf{x}_1, \ldots, \mathbf{x}_k \text{ are independent } \iff E \prod_{i=1}^{k} g_i(\mathbf{x}_i) = \prod_{i=1}^{k} E\ g_i(\mathbf{x}_i),$$

$\forall g_1, \ldots, g_k$ *such that* $E\ |g_i(\mathbf{x}_i)| < \infty$.

and also (chiefly by the C.E.T.)

Proposition 2.14

$$\mathbf{x}_1, \ldots, \mathbf{x}_k \text{ are independent } \iff F(\mathbf{t}) = \prod_{i=1}^{k} F_i(\mathbf{t}_i), \ \forall \mathbf{t} \in \mathbb{R}^n.$$

In the special case where each $\mathbf{x}_i$ is absolutely continuous with p.d.f. $f_i(\mathbf{t}_i)$, we may conclude that $\mathbf{x}$ is as well, and we have:

Proposition 2.15

$$\mathbf{x}_1, \ldots, \mathbf{x}_k \text{ are independent } \iff f(\mathbf{t}) = \prod_{i=1}^{k} f_i(\mathbf{t}_i), \ \forall \mathbf{t} \in \mathbb{R}^n.$$

Finally, independence may also be characterized:

Proposition 2.16

$$\mathbf{x}_1, \ldots, \mathbf{x}_k \text{ are independent} \iff c_{\mathbf{x}}(\mathbf{t}) = \prod_{i=1}^{k} c_{\mathbf{x}_i}(\mathbf{t}_i), \ \forall \mathbf{t} \in \mathbb{R}^n.$$

Example 2.6 *For* $\mathbf{x} \sim \text{unif}([\mathbf{0}, \mathbf{1}])$ *on* $\mathbb{R}^n$*, it is clear that* $x_1, \ldots, x_n$ *are independently and identically distributed (i.i.d.) as* $\text{unif}([0, 1])$.

Example 2.7 *Let* $\mathbf{x} = (x_1, x_2)'$ *have Frank's d.f.* (2.4)*. Proposition* 2.14 *yields, after elementary calculus,* $x_1 \perp\!\!\!\perp x_2$ *iff* $\alpha = 1$.

2.12 Change of variables

We recall some basic calculus [Spivak (1965), p. 16]. Let $A \subset \mathbb{R}^n$ be open.

Definition 2.9 *The derivative of* $\boldsymbol{\phi} : A \to \mathbb{R}^m$*, at* $\mathbf{x} \in A$*, is the unique linear transformation* $\boldsymbol{\phi}'(\mathbf{x}) \in \mathbb{R}_n^m$ *such that*

$$\boldsymbol{\phi}(\mathbf{x} + \mathbf{h}) - \boldsymbol{\phi}(\mathbf{x}) = \boldsymbol{\phi}'(\mathbf{x})\mathbf{h} + o(\mathbf{h})$$

or, equivalently,

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\left| \boldsymbol{\phi}(\mathbf{x} + \mathbf{h}) - \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}'(\mathbf{x})\mathbf{h} \right|}{|\mathbf{h}|} = 0.$$

When $\boldsymbol{\phi}'(\mathbf{x})$ exists, all partial derivatives $\partial \phi_i(\mathbf{x}) / \partial x_j$ exist. This determines the derivative componentwise as

$$\boldsymbol{\phi}'(\mathbf{x}) = \left(\partial \phi_i(\mathbf{x}) / \partial x_j \right).$$

A condition for $\boldsymbol{\phi}'(\mathbf{x})$ to exist is that all partial derivatives $\partial \phi_i(\mathbf{x}) / \partial x_j$ exist in an open neighborhood of $\mathbf{x}$ and are continuous at $\mathbf{x}$. There are, of course, various notations for derivatives, all acceptable:

$$\boldsymbol{\phi}'(\mathbf{x}) = D\boldsymbol{\phi}(\mathbf{x}) = \partial \boldsymbol{\phi}(\mathbf{x}) / \partial \mathbf{x}.$$

The derivative satisfies the "chain rule"

$$(\boldsymbol{\phi} \circ \boldsymbol{\psi})'(\mathbf{x}) = \boldsymbol{\phi}'(\boldsymbol{\psi}(\mathbf{x}))\ \boldsymbol{\psi}'(\mathbf{x}).$$

In the very special case $m = n$, the *jacobian* of $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^n$ is, by definition, the absolute value of the determinant of $\boldsymbol{\phi}'(\mathbf{x})$ and is denoted by $|\boldsymbol{\phi}'(\mathbf{x})|_+$. Another common notation for the jacobian of the transformation $\mathbf{y} = \boldsymbol{\phi}(\mathbf{x})$ is

$$J(\mathbf{y} \to \mathbf{x}) = |\boldsymbol{\phi}'(\mathbf{x})|_+.$$

From the chain rule, it is made clear that if $\mathbf{z} = \boldsymbol{\phi}(\mathbf{y})$ and $\mathbf{y} = \boldsymbol{\psi}(\mathbf{x})$, then

$$\begin{aligned} J(\mathbf{z} \to \mathbf{x}) &= J(\mathbf{z} \to \mathbf{y}) \cdot J(\mathbf{y} \to \mathbf{x}), \\ J(\mathbf{y} \to \mathbf{x}) &= [J(\mathbf{x} \to \mathbf{y})]^{-1}. \end{aligned}$$

At any rate, we have an important and general result, easy to state, but the proof of which is by no means trivial [Spivak (1965), p. 67].

Proposition 2.17 *Let $\boldsymbol{\phi} : A \to \mathbb{R}^n$ be one-to-one and continuously differentiable on A. If $f : \boldsymbol{\phi}(A) \to \mathbb{R}$ is integrable, then*

$$\int_{\boldsymbol{\phi}(A)} f(\mathbf{x}) d\mathbf{x} = \int_A f\left(\boldsymbol{\phi}(\mathbf{y})\right)\ |\boldsymbol{\phi}'(\mathbf{y})|_+ d\mathbf{y}.$$

It is this result that is applied directly to obtain the standard "change of variables" formula for absolutely continuous random vectors.

Proposition 2.18 *If $\mathbf{x} \sim f$ on $\mathbb{R}^n$ and $C = \{\mathbf{x} : f(\mathbf{x}) > 0\}$ is open, for any $\boldsymbol{\phi} : C \to \mathbb{R}^n$ one-to-one and bi-differentiable with inverse $\boldsymbol{\psi} : \boldsymbol{\phi}(C) \to C$, let $\mathbf{y} = \boldsymbol{\phi}(\mathbf{x})$. Then, $\mathbf{y} \sim g$ with*

$$g(\mathbf{y}) = f\left(\boldsymbol{\psi}(\mathbf{y})\right)\ |\boldsymbol{\psi}'(\mathbf{y})|_+.$$

Proof.

$$\begin{aligned} P(\mathbf{y} \in B) &= P\left(\boldsymbol{\phi}(\mathbf{x}) \in B\right) = P\left(\mathbf{x} \in \boldsymbol{\psi}(B)\right) \\ &= \int_{\boldsymbol{\psi}(B)} f(\mathbf{x}) d\mathbf{x} = \int_B f\left(\boldsymbol{\psi}(\mathbf{y})\right) |\boldsymbol{\psi}'(\mathbf{y})|_+ d\mathbf{y}. \end{aligned}$$

□

By an abuse of notation, $\mathbf{y}$ (and $\mathbf{x}$) have two different meanings in Proposition 2.18: $\mathbf{y}$ is a random vector in $\mathbf{y} \sim g$, whereas it is any given point of $\mathbb{R}^n$ in the density $g(\mathbf{y})$.

Now, if the function $\boldsymbol{\phi}$ in question is simply a linear transformation, $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, it is already its own derivative everywhere on $\mathbb{R}^n$, $\boldsymbol{\phi}'(\mathbf{x}) = \mathbf{A}$, and the formula for change of variables greatly simplifies.

Suppose that $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$ is a nonsingular transformation. The group of all such nonsingular transformations is known as the general linear group and denoted by

$$\mathbf{G}_n = \{\mathbf{A} \in \mathbb{R}_n^n :\ \mathbf{A} \text{ is nonsingular}\} = \{\mathbf{A} \in \mathbb{R}_n^n :\ |\mathbf{A}| \neq 0\}.$$

Two examples are as follows:

Example 2.8 *If* $\mathbf{x} \sim f$ *and* $\mathbf{y} = \mathbf{A}\mathbf{x}$*, we find* $\mathbf{y} \sim g$*, where*

$$g(\mathbf{y}) = f(\mathbf{A}^{-1}\mathbf{y})\ |\mathbf{A}|_+^{-1}.$$

Example 2.9 $\mathbf{x} \sim \text{unif}(C),\ C \subset \mathbb{R}^n \Longrightarrow \mathbf{A}\mathbf{x} + \mathbf{b} \sim \text{unif}(\mathbf{A}C + \mathbf{b})$ *where* $\mathbf{A}C + \mathbf{b} = \{\mathbf{A}\mathbf{x} + \mathbf{b} :\ \mathbf{x} \in C\}$.

2.13 Jacobians

The derivation of jacobians is the difficult part in making transformations. It can be a daunting task. This section is directed to the derivation of more complicated jacobians. It can be skipped on a first reading and consulted when needed in the sequel. Although jacobians are useful for densities, our approach is to derive distributions without appealing, whenever possible, to densities. Derivations of densities appear mainly in the form of problems.

Proposition 2.19 *The jacobian of the transformation* $\mathbf{V} = \mathbf{A}\mathbf{W}\mathbf{A}'$, $\mathbf{W} \in \mathbb{R}_n^n$ *symmetric and* $\mathbf{A} \in \mathbb{R}_n^n$ *constant, is* $J(\mathbf{V} \to \mathbf{W}) = |\mathbf{A}|_+^{n+1}$.

Proof. The transformation is linear and, thus, the jacobian is necessarily a polynomial in the elements of $\mathbf{A}$, $p(\mathbf{A})$ say. If $\mathbf{W} = \mathbf{B}\mathbf{U}\mathbf{B}'$, then from the chain rule, we have

$$J(\mathbf{V} \to \mathbf{U}) = J(\mathbf{V} \to \mathbf{W}) \cdot J(\mathbf{W} \to \mathbf{U}),$$

i.e., $p(\mathbf{A}\mathbf{B}) = p(\mathbf{A})p(\mathbf{B})$. The only polynomials in the elements of a matrix satisfying this multiplicative rule are the integer powers of the determinant [MacDuffy (1943, p. 50)]. Hence, $p(\mathbf{A}) = |\mathbf{A}|^k$, for some integer k. We can find k by choosing $\mathbf{A} = a\mathbf{I}$. Since $\mathbf{V} = a^2\mathbf{W}$ and there are $\frac{1}{2}n(n+1)$ distinct elements, then $J(\mathbf{V} \to \mathbf{W}) = a^{n(n+1)} = |a\mathbf{I}|^{n+1}$. We found $k = n + 1$. □

James (1954) also used MacDuffy's characterization of the determinant for skew-symmetric matrices. At this point, we make some comments concerning the differential of a function of several variables. Our development here closely resembles that of Srivastava and Khatri (1979, p. 26). For a real-valued function $y = f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, the differential is defined as $dy = df = \sum_{i=1}^n \partial f(\mathbf{x})/\partial x_i \cdot dx_i$. For a vector-valued function $\mathbf{y} = \mathbf{f}(\mathbf{x})$, $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$, the differential is defined componentwise, i.e.,

$$d\mathbf{y} = d\mathbf{f} = \begin{pmatrix} df_1 \\ \vdots \\ df_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \partial f_1(\mathbf{x})/\partial x_i \cdot dx_i \\ \vdots \\ \sum_{i=1}^n \partial f_n(\mathbf{x})/\partial x_i \cdot dx_i \end{pmatrix} = \partial \mathbf{f}(\mathbf{x})/\partial \mathbf{x} \cdot d\mathbf{x},$$

where $\partial \mathbf{f}(\mathbf{x})/\partial \mathbf{x} = (\partial f_i(\mathbf{x})/\partial x_j) \in \mathbb{R}_n^n$ is the usual derivative of $\mathbf{f}(x)$. Hence, $d\mathbf{y}$ is a linear function of $d\mathbf{x}$ with jacobian

$$J(d\mathbf{y} \to d\mathbf{x}) = |\partial \mathbf{f}(\mathbf{x})/\partial \mathbf{x}|_+ = J(\mathbf{y} \to \mathbf{x}).$$

Note that $\mathbf{x}$ and $\mathbf{y}$ could be replaced by any "vectorized" array or matrix. For example, for $\mathbf{F}(\mathbf{X}) = (f_{ij}(\mathbf{X})) \in \mathbb{R}_n^m$, we can define the differential componentwise, i.e., $d\mathbf{F} = (df_{ij})$. The reader can then check (v. Problem 2.14.15)

$$\mathbf{F} = \mathbf{GH} \Longrightarrow d\mathbf{F} = \mathbf{G} \cdot d\mathbf{H} + d\mathbf{G} \cdot \mathbf{H}.$$

As an example, consider the inverse transformation.

Proposition 2.20 *The jacobian of the transformation* $\mathbf{V} = \mathbf{W}^{-1}$, $\mathbf{W} \in \mathbb{R}_n^n$ *nonsingular and symmetric, is* $J(\mathbf{V} \to \mathbf{W}) = |\mathbf{W}|_+^{-(n+1)}$.

Proof. Since $\mathbf{VW} = \mathbf{I}$, then $\mathbf{V} \cdot d\mathbf{W} + d\mathbf{V} \cdot \mathbf{W} = \mathbf{0}$, which implies $d\mathbf{V} = -\mathbf{W}^{-1} \cdot d\mathbf{W} \cdot \mathbf{W}^{-1}$. Hence, from Proposition 2.19, $J(\mathbf{V} \to \mathbf{W}) = J(d\mathbf{V} \to d\mathbf{W}) = |\mathbf{W}|_+^{-(n+1)}$. □

The jacobian of "conditional transformations" [Srivastava and Khatri (1979), p. 29], used to prove Propositions 2.22 and 2.23, may provide simplifications in some cases.

Proposition 2.21 *Let* $\mathbf{x}_i$ *and* $\mathbf{y}_i$ *in* $\mathbb{R}^{p_i}$, $i = 1, \ldots, r$, *be related through the system of "conditional transformations"*

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{f}_1(\mathbf{x}_1), \\ \mathbf{y}_2 &= \mathbf{f}_2(\mathbf{y}_1, \mathbf{x}_2), \\ &\vdots \\ \mathbf{y}_r &= \mathbf{f}_r(\mathbf{y}_1, \ldots, \mathbf{y}_{r-1}, \mathbf{x}_r), \end{aligned}$$

where each $\mathbf{f}_i$ *is differentiable. Then,*

$$J(\mathbf{y}_1, \ldots, \mathbf{y}_r \to \mathbf{x}_1, \ldots, \mathbf{x}_r) = \prod_{i=1}^{r} J(\mathbf{y}_i \to \mathbf{x}_i).$$

Proof. The jacobian has the triangular form

$$J = \begin{vmatrix} \partial \mathbf{y}_1/\partial \mathbf{x}_1 & 0 & \cdots & 0 \\ * & \partial \mathbf{y}_2/\partial \mathbf{x}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & \partial \mathbf{y}_r/\partial \mathbf{x}_r \end{vmatrix}_+,$$

and, thus, we get $J = \prod_{i=1}^{r} |\partial \mathbf{y}_i/\partial \mathbf{x}_i|_+$ immediately. □

As an example of jacobian via conditional transformations, consider the Bartlett decomposition of $\mathbf{W} > \mathbf{0}$ in $\mathbb{R}_n^n$ as $\mathbf{W} = \mathbf{TT}'$ for a unique $\mathbf{T} \in \mathbf{L}_n^+$ (v. Proposition 1.14). Due to symmetry, $\mathbf{W}$ has effectively $n(n+1)/2$ elements and, thus, the decomposition gives a transformation $\mathbf{f} : \mathbb{R}^{n(n+1)/2} \to \mathbb{R}^{n(n+1)/2}$ defined by $\mathbf{f}(\mathbf{W}) = \mathbf{T}$.

Proposition 2.22 *The jacobian of the transformation* $\mathbf{f}(\mathbf{W}) = \mathbf{T}$ *is*

$$J(\mathbf{W} \to \mathbf{T}) = 2^n \prod_{i=1}^{n} t_{ii}^{n-i+1}.$$

Proof. Partition $\mathbf{W}$ and $\mathbf{T}$ in conformity so that

$$\begin{pmatrix} w_{11} & \mathbf{w}_{21}' \\ \mathbf{w}_{21} & \mathbf{W}_{22} \end{pmatrix} = \begin{pmatrix} t_{11} & \mathbf{0}' \\ \mathbf{t}_{21} & \mathbf{T}_{22} \end{pmatrix} \begin{pmatrix} t_{11} & \mathbf{t}_{21}' \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix}.$$

Observe the system of conditional transformations

$$\begin{aligned} w_{11} &= t_{11}^2, \\ \mathbf{w}_{21} &= w_{11}^{1/2} \mathbf{t}_{21}, \\ \mathbf{W}_{22} &= \mathbf{w}_{21}\mathbf{w}_{21}'/w_{11} + \mathbf{T}_{22}\mathbf{T}_{22}' \end{aligned}$$

from which

$$J(\mathbf{W} \to \mathbf{T}) = (2t_{11})(w_{11}^{1/2})^{n-1} J(\mathbf{W}_{22} \to \mathbf{T}_{22}) = 2t_{11}^n J(\mathbf{W}_{22} \to \mathbf{T}_{22}).$$

The conclusion follows by induction. □

As another example, consider the transformation to polar coordinates on $\mathbb{R}^n$, $\mathbf{x} \mapsto (r, \theta_1, \ldots, \theta_{n-1})$ given by

$$\begin{aligned} x_1 &= r\sin(\theta_1)\sin(\theta_2)\cdots\sin(\theta_{n-2})\sin(\theta_{n-1}), \\ x_2 &= r\sin(\theta_1)\sin(\theta_2)\cdots\sin(\theta_{n-2})\cos(\theta_{n-1}), \\ x_3 &= r\sin(\theta_1)\sin(\theta_2)\cdots\cos(\theta_{n-1}), \\ &\vdots \\ x_{n-1} &= r\sin(\theta_1)\cos(\theta_2), \\ x_n &= r\cos(\theta_1), \end{aligned}$$

where $r > 0$ is the "radius" and $0 < \theta_i \leq \pi$, $i = 1, \ldots, n-2$, $0 < \theta_{n-1} \leq 2\pi$ are the "angles". The jacobian $J(\mathbf{x} \to r, \boldsymbol{\theta})$ is facilitated with the system of conditional transformations

$$\begin{aligned} y_1 &= x_1^2 + \cdots + x_{n-2}^2 + x_{n-1}^2 + x_n^2 &&= r^2, \\ y_2 &= x_1^2 + \cdots + x_{n-2}^2 + x_{n-1}^2 &&= y_1 \sin^2(\theta_1), \\ y_3 &= x_1^2 + \cdots + x_{n-2}^2 &&= y_2 \sin^2(\theta_2), \\ &\vdots \\ y_{n-1} &= x_1^2 + x_2^2 &&= y_{n-2} \sin^2(\theta_{n-2}), \\ y_n &= x_1^2 &&= y_{n-1} \sin^2(\theta_{n-1}). \end{aligned}$$

Proposition 2.23 *The jacobian of the transformation to polar coordinates in* $\mathbb{R}^n$ *is*

$$J(\mathbf{x} \to r, \boldsymbol{\theta}) = r^{n-1} \sin^{n-2}(\theta_1) \sin^{n-3}(\theta_2) \cdots \sin(\theta_{n-2}).$$

Proof. We give the main idea and the reader is asked in Problem 2.14.11 to complete the details. We have $J(\mathbf{y} \to r, \boldsymbol{\theta}) = J(\mathbf{y} \to \mathbf{x}) \cdot J(\mathbf{x} \to r, \boldsymbol{\theta})$. The jacobian $J(\mathbf{y} \to \mathbf{x})$ is trivial and $J(\mathbf{y} \to r, \boldsymbol{\theta})$ is evaluated using Proposition 2.21 on conditional transformations. □

Let $S^{n-1} = \{\mathbf{s} \in \mathbb{R}^n : |\mathbf{s}| = 1\}$ be the "unit sphere" in $\mathbb{R}^n$. The superscript $n-1$ refers to the dimension of this surface. At times, we would like to bypass the angles and consider directly the transformation

$$\begin{aligned} \mathbf{f} : \mathbb{R}^n \backslash \{\mathbf{0}\} &\to (0, \infty) \times S^{n-1}, \\ \mathbf{x} &\mapsto (r, \mathbf{u}) \end{aligned}$$

defined by $r = |\mathbf{x}|$ and $\mathbf{u} = \mathbf{x}/|\mathbf{x}| \in S^{n-1}$. Since [Courant (1936), p. 302]

$$\int_{|\mathbf{x}| \le R} g(|\mathbf{x}|) d\mathbf{x} = \int_0^R \int_{S^{n-1}} g(r) r^{n-1} dr d\mathbf{u},$$

where $d\mathbf{u}$ is the "area element" of S^{n-1}, then r^{n-1} is the jacobian.

Proposition 2.24 *The jacobian of the transformation* $\mathbf{x} \mapsto (r, \mathbf{u})$ *is*

$$J(\mathbf{x} \to r, \mathbf{u}) = r^{n-1}.$$

The jacobians of other transformations on k-surfaces (manifolds) in $\mathbb{R}^n$ are useful for sampling distributions of eigenvalues, for example, but their full understanding requires a knowledge of differential forms and integration on manifolds [Spivak (1965), James (1954)]. This will not be pursued here.

2.14 Problems

1. Show that $|E\ g(\mathbf{x})| \le E\ |g(\mathbf{x})|$ for any $g : \mathbb{R}^n \to \mathbb{R}$ such that $E\ |g(\mathbf{x})| < \infty$.

2. Prove the C_r inequality: For $\mathbf{x}$ and $\mathbf{y}$ distributed on $\mathbb{R}^k$,

   $$E\ |\mathbf{x} + \mathbf{y}|^r \le C_r \left[E\ |\mathbf{x}|^r + E\ |\mathbf{y}|^r\right],\ r > 0,$$

   where

   $$C_r = \begin{cases} 1, & 0 < r \le 1 \\ 2^{r-1}, & r \ge 1. \end{cases}$$

   Hint: Show the simple inequality $(a+b)^r \le C_r(a^r + b^r)$, $r > 0$, $a \ge 0$, $b \ge 0$.

3. For each $p > 0$ let $\mathcal{L}_p$ denote the collection of all random vectors on $\mathbb{R}^k$ for which the pth moment exists: $E\ |\mathbf{x}|^p < \infty$. Prove the following basic facts:

   (i) $\mathcal{L}_p$ is a vector space.
   (ii) For any $0 < r \le p$, $\mathcal{L}_p \subseteq \mathcal{L}_r$.

(iii) $E\ |\mathbf{x}|^p < \infty \iff E\ |x_i|^p < \infty,\ i = 1, \ldots, k \iff E\ |\mathbf{t}'\mathbf{x}|^p < \infty,\ \forall \mathbf{t} \in \mathbb{R}^k$.
(iv) $E\ |\mathbf{a}'\mathbf{x}|^p = 0$, for some $\mathbf{a} \in \mathbb{R}^k \Longrightarrow P(\mathbf{x} \in \mathbf{a}^\perp) = 1$.
(v) For any $\mathbf{x} \in \mathcal{L}_1$, $|E\ \mathbf{x}| \leq E\ |\mathbf{x}|$. Indicate also the precise circumstances under which equality occurs.

4. Prove that if the pth moment ($p > 0$) of $\mathbf{x} \in \mathbb{R}^n$ is finite, then all product-moments of $\mathbf{x}$ of order p are finite.

5. For $\mathbf{x}$ distributed on $\mathbb{R}^n$, consider the p.d.f. $f_{\mathbf{x}}(\mathbf{x}) = c|\mathbf{x}|^2 \cdot I_{[\mathbf{0},\mathbf{1}]}(\mathbf{x})$.
(i) Determine c.
(ii) Determine $E\ \mathbf{x}$ and var $\mathbf{x}$.
(iii) Determine $E\ \prod_{i=1}^n x_i^i$.
Hint: $E\ g(\mathbf{x}) = cE\ |\mathbf{u}|^2 g(\mathbf{u})$, where $\mathbf{u} \sim \text{unif}([\mathbf{0}, \mathbf{1}])$.

6. Let $\mathbf{A} \in \mathbb{R}_n^m$, $\mathbf{B} \in \mathbb{R}_q^p$, and $\mathbf{C} \in \mathbb{R}_q^m$ be constant and $\mathbf{X} \in \mathbb{R}_p^n$, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{y}, \mathbf{z} \in \mathbb{R}^q$ be random. Check the following:
(i) $E(\mathbf{AXB} + \mathbf{C}) = \mathbf{A}(E\ \mathbf{X})\mathbf{B} + \mathbf{C}$
(ii) $\text{cov}(\mathbf{Ax}, \mathbf{By}) = \mathbf{A}\ \text{cov}(\mathbf{x}, \mathbf{y})\mathbf{B}'$
(iii) $\text{cov}(\mathbf{x}, \mathbf{y} + \mathbf{z}) = \text{cov}(\mathbf{x}, \mathbf{y}) + \text{cov}(\mathbf{x}, \mathbf{z})$.

7. Prove the conditional variance formula.

8. **Pairwise versus mutual independence** [Bhat (1981)].
Let x and y be i.i.d. random variables taking the values $+1$ and -1 with probability $1/2$ each. Define $z = xy$.
i) Establish x, y, z are pairwise independent, but not mutually independent.
ii) Does $x \perp\!\!\!\perp z$ and $y \perp\!\!\!\perp z$ imply $\frac{x}{y} \perp\!\!\!\perp z$?

9. Let $\mathbf{x} = (x_1, x_2)'$ have the d.f. of Cook and Johnson (1981) as in expression (2.3). Demontrate $x_1 \perp\!\!\!\perp x_2$ iff $\alpha = 0$.

10. Given a bivariate copula d.f. $C(t_1, t_2)$, two measures of association are Spearman's ρ and Kendall's τ,

$$\rho = 12 \int_{[0,1]^2} t_1 t_2 dC(t_1, t_2) - 3,$$

$$\tau = 4 \int_{[0,1]^2} C(t_1, t_2) dC(t_1, t_2) - 1,$$

$|\rho| \leq 1$ and $|\tau| \leq 1$. Now, let $|\alpha| < 1/3$ in the bivariate Morgenstern copula

$$C(t_1, t_2) = t_1 t_2 [1 + 3\alpha(1 - t_1)(1 - t_2)].$$

Verify this copula is parameterized by Spearman's measure, or

$$\alpha = 12 \int_{[0,1]^2} t_1 t_2 dC(t_1, t_2) - 3.$$

11. Complete the proof of Proposition 2.23.

12. Demonstrate the jacobian of the transformation $\mathbf{T}_1 = \mathbf{A}\mathbf{T}_2$ for $\mathbf{T}_2 \in \mathbf{L}_n^+$ and $\mathbf{A} = (a_{ij})$ constant also in $\mathbf{L}_n^+$ is $J(\mathbf{T}_1 \to \mathbf{T}_2) = \prod_{i=1}^n a_{ii}^i$.

13. Demonstrate the jacobian of the transformation $\mathbf{U}_1 = \mathbf{A}\mathbf{U}_2$ for $\mathbf{U}_2 \in \mathbf{U}_n^+$ and $\mathbf{A} = (a_{ij})$ constant also in $\mathbf{U}_n^+$ is

$$J(\mathbf{U}_1 \to \mathbf{U}_2) = \prod_{i=1}^n a_{ii}^{n-i+1}.$$

14. Demonstrate the jacobian of the transformation $\mathbf{V} = \mathbf{A}\mathbf{W}\mathbf{A}'$, where $\mathbf{W} \in \mathbb{R}_n^n$ is skew-symmetric, i.e., $\mathbf{W} = -\mathbf{W}'$, is $J(\mathbf{V} \to \mathbf{W}) = |\mathbf{A}|_+^{n-1}$.

15. Suppose $\mathbf{F}(\mathbf{X}) = (f_{ij}(\mathbf{X})) \in \mathbb{R}_n^m$ and define the differential componentwise, i.e., $d\mathbf{F} = (df_{ij})$. Demonstrate that

$$\mathbf{F} = \mathbf{G}\mathbf{H} \Longrightarrow d\mathbf{F} = \mathbf{G} \cdot d\mathbf{H} + d\mathbf{G} \cdot \mathbf{H}.$$

16. Let

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} > \mathbf{0}$$

and define the transformation

$$\mathbf{f} : (\mathbf{V}_{11}, \mathbf{V}_{12}, \mathbf{V}_{22}) \mapsto (\mathbf{V}_{11.2}, \mathbf{V}_{12}, \mathbf{V}_{22}),$$

where $\mathbf{V}_{11.2} = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$.

(i) Prove $\mathbf{f}$ defines a one-to-one mapping.
(ii) Obtain $J(\mathbf{V}_{11}, \mathbf{V}_{12}, \mathbf{V}_{22} \to \mathbf{V}_{11.2}, \mathbf{V}_{12}, \mathbf{V}_{22}) = 1$.

3

Gamma, Dirichlet, and F distributions

3.1 Introduction

This chapter introduces some basic probability distributions useful in statistics. The gamma distribution, in particular, is the building block of many other distributions such as chi-square, F, and Dirichlet. The Dirichlet distribution, as defined in Section 3.3, has the important physical interpretation of proportion of time waited in a Poisson process. However, it has other applications such as the distribution of spacing variables (v. Problem 3.5.3) and the distribution theory (v. Section 4.5) related to spherical distributions, which play an important role in robustness.

3.2 Gamma distributions

Definition 3.1 Standard gamma*: $z \sim gamma(p)$ or $z \sim G(p)$ on $p > 0$ "degrees of freedom" iff*

$$f_z(z) = \begin{cases} \Gamma(p)^{-1}\ z^{p-1}e^{-z}, & z > 0 \\ 0, & z \leq 0. \end{cases}$$

The integrating constant, as it depends on $p > 0$, is known as the gamma function and is, in fact, defined by

$$\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}dt,\ p > 0.$$

One may verify some basic properties:

$$\Gamma(p+1) = p\Gamma(p),\ \Gamma(2) = \Gamma(1) = 1,\ \Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

and, in particular, $\Gamma(n) = (n-1)!$. Obviously,

$$E\ z^r = \Gamma(p+r)/\Gamma(p),\ \forall r > -p,$$

so that $E\ z = \text{var}\ z = p$. A more general gamma distribution is obtained by simply rescaling the standard gamma.

Definition 3.2 Scaled gamma*: $x \sim gamma(p, \theta)$ or $x \sim G(p, \theta)$ on $p > 0$ "degrees of freedom" and "scale" $\theta > 0$ iff*

$$x = \theta z,\ z \sim G(p).$$

Obviously,

$$f_x(x) = \begin{cases} \Gamma(p)^{-1}\ \theta^{-p}\ x^{p-1}e^{-x/\theta}, & x > 0 \\ 0, & x \le 0, \end{cases}$$

and, thus, the characteristic function is $c_x(t) = (1 - i\theta t)^{-p}$ with the "convolution" of gamma distributions as a corollary.

Corollary 3.1 *If x_i, $i = 1, \ldots, n$, are independent $G(p_i, \theta)$, then*

$$\sum_{i=1}^{n} x_i \sim G\left(\sum_{i=1}^{n} p_i, \theta\right).$$

In the special case where $p = 1$ we have the exponential distributions.

Definition 3.3

$$\begin{aligned} \text{Standard exponential} &: \quad z \sim \exp(1)\ \textit{iff}\ z \sim G(1). \\ \text{Scaled exponential} &: \quad x \sim \exp(\theta)\ \textit{iff}\ x = \theta z,\ z \sim \exp(1). \end{aligned}$$

The chi-square distribution is another special case.

Definition 3.4 Chi-square*: $y \sim \chi^2_m$ or $y \stackrel{d}{=} \chi^2_m$ iff*

$$y = 2z,\ z \sim G(\tfrac{1}{2}m).$$

Equivalently, $y \sim \chi^2_m$ iff $y \sim G(\frac{1}{2}m, 2)$, and the chi-square is a special case of the scaled gamma above. Thus, the gamma distribution occurs in common statistical practice as the chi-square (2×gamma≡chi-square). The characteristic function of $y \sim \chi^2_m$ is immediate: $c_y(t) = (1 - i2t)^{-m/2}$. One should, however, also recall how it describes "waiting time" in a Poisson process.

Recall that the Poisson process N_t arises first on purely physical considerations as a description of the number of "successes" in what is effectively an infinite number of independent bernoulli trials over the fixed time period t where the average number is known to be proportional to t. On these

assumptions,

$$X_n \sim \text{binomial}(n, p_n) \text{ and } E\ X_n \to \lambda t,$$

whereby

$$X_n \xrightarrow{d} N_t, \text{ where } N_t \sim \text{Poisson}(\lambda t).$$

One then has the (conjugate) waiting time process T_n to describe the amount of time to wait until at least n "successes." Since

$$T_n > t \Longleftrightarrow N_t < n,$$

we find

$$P(T_n > t) = P(N_t < n) = \sum_{i=0}^{n-1} e^{-\lambda t}(\lambda t)^i / i!$$

and differentiating produces the p.d.f.

$$f_n(t) = \lambda e^{-\lambda t}(\lambda t)^{n-1}/(n-1)!,\ t > 0,$$

whereby we discover

$$T_n \sim G(n, \lambda^{-1}) \text{ or, equivalently, } z_n = \lambda T_n \sim G(n).$$

The exponential itself is just as well predicated on a different intuition in that one may show that it is the unique distribution that has "no memory" in the explicit sense that

$$x \sim \exp(\theta) \Longleftrightarrow P(x > s + t | x > t) = P(x > s) > 0,\ \forall s, t > 0.$$

3.3 Dirichlet distributions

If the gamma is intuitively a waiting time, the Dirichlet, otherwise known as the multivariate beta, is simply the proportion of time waited.

Definition 3.5 Dirichlet*:* $\mathbf{x} \sim D_n(\mathbf{p}; p_{n+1})$ *or* $\mathbf{x} \sim \text{beta}_n(\mathbf{p}; p_{n+1})$, $\mathbf{p} = (p_1, \ldots, p_n)'$, $p_i > 0$, $i = 1, \ldots, n+1$ *iff*

$$\mathbf{x} \overset{d}{=} \frac{1}{T}\mathbf{z}$$

with $z_i \overset{indep}{\sim} G(p_i)$, $i = 1, \ldots, n+1$, $\mathbf{z} = (z_1, \ldots, z_n)'$, *and* $T = \sum_{i=1}^{n+1} z_i$.

The notation $\overset{\text{indep}}{\sim}$ means "independently distributed as."

Proposition 3.1 *The joint p.d.f. of* $\mathbf{x}$ *and* T *can be described as*

$$f_T(t) = \frac{1}{\Gamma(\sum_{i=1}^{n+1} p_i)} t^{\sum_{i=1}^{n+1} p_i - 1} e^{-t},\ t > 0 \left(i.e.,\ T \sim G\left(\sum_{i=1}^{n+1} p_i\right)\right),$$

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{\Gamma(\sum_{i=1}^{n+1} p_i)}{\prod_{i=1}^{n+1}\Gamma(p_i)} \prod_{i=1}^{n} x_i^{p_i-1} \left(1 - \sum_{i=1}^{n} x_i\right)^{p_{n+1}-1}, \ \mathbf{x} \in T^n,$$

where $T^n = \{\mathbf{x} \in \mathbb{R}^n : \ x_i > 0, \ \sum_{i=1}^n x_i < 1\}$. *Moreover,* $\mathbf{x} \perp\!\!\!\perp T$.

Proof. Using independence, the joint p.d.f. of the z_i's is

$$f_{\mathbf{z},z_{n+1}}(\mathbf{z}, z_{n+1}) = \frac{1}{\prod_{i=1}^{n+1}\Gamma(p_i)} \cdot \prod_{i=1}^{n+1} z_i^{p_i-1} \cdot \exp\left(-\sum_{i=1}^{n+1} z_i\right), \ z_i > 0, \ \forall i.$$

We simply transform from $(z_1, \ldots, z_{n+1})$ to $(x_1, \ldots, x_n, t)$, where

$$z_i = t x_i, \ i = 1, \ldots, n$$

and

$$z_{n+1} = t\left(1 - \sum_{i=1}^{n} x_i\right).$$

The jacobian is given by

$$\begin{aligned}
\left|\frac{\partial z_1, \ldots, z_{n+1}}{\partial x_1, \ldots, x_n, t}\right|_+ &= \left|\frac{\partial t x_1, \ldots, t x_n, t(1 - \sum_{i=1}^n x_i)}{\partial x_1, \ldots, x_n, t}\right|_+ \\
&= \begin{vmatrix} t & & 0 & x_1 \\ & \ddots & & \vdots \\ 0 & & t & x_n \\ -t & \cdots & -t & 1 - \sum_{i=1}^n x_i \end{vmatrix} \\
&= \begin{vmatrix} t & & 0 & x_1 \\ & \ddots & & \vdots \\ 0 & & t & x_n \\ 0 & \cdots & 0 & 1 \end{vmatrix} = t^n.
\end{aligned}$$

Thus, the joint p.d.f. of $(\mathbf{x}, T)$ is

$$\frac{1}{\prod_{i=1}^{n+1}\Gamma(p_i)} \prod_{i=1}^{n} x_i^{p_i-1} (1 - \sum_{i=1}^{n} x_i)^{p_{n+1}-1} t^{\sum_{i=1}^{n+1} p_i - 1} e^{-t}, \ \mathbf{x} \in T^n, \ t > 0,$$

and the conclusions are reached. □

Note that the Dirichlet where all the parameters are 1 is simply the uniform distribution on the triangular region T^n, $D_n(\mathbf{1}; 1) \equiv \text{unif}(T^n)$. Also, the Dirichlet distribution generalizes the *beta distribution*, $D_1(p_1; p_2) \stackrel{d}{=} \text{beta}(p_1; p_2)$, with p.d.f.

$$f_x(x) = B(p_1, p_2)^{-1} x^{p_1-1}(1-x)^{p_2-1}, \ 0 < x < 1,$$

where $B(p_1, p_2) = \Gamma(p_1)\Gamma(p_2)/\Gamma(p_1 + p_2)$, is the *beta function*.

The converse of Proposition 3.1 is almost obvious (by inverse change of variables); it need only be stated.

Proposition 3.2 *If* $z_i = Tx_i$, $i = 1, \ldots, n$ *and* $z_{n+1} = T(1 - \sum_{i=1}^{n} x_i)$ *with* $T \sim G(\sum_{i=1}^{n+1} p_i)$, $\mathbf{x} \sim D_n(\mathbf{p}; p_{n+1})$, *and* $\mathbf{x} \perp\!\!\!\perp T$, *then*

$$z_i \overset{indep}{\sim} G(p_i),\ i = 1, \ldots, n+1.$$

Four useful corollaries are also stated and the reader is asked to prove them. For $\mathbf{x} \sim D_n(\mathbf{p}; p_{n+1})$, let $p = \sum_{i=1}^{n+1} p_i$ denote the "grand total," noting that, by definition,

$$\mathbf{x} \overset{d}{=} \frac{1}{T}\mathbf{z}$$

with $z_i \overset{\text{indep}}{\sim} G(p_i)$, $i = 1, \ldots, n+1$, $\mathbf{z} = (z_1, \ldots, z_n)'$, and $T = \sum_{i=1}^{n+1} z_i$.

We find the following:

Corollary 3.2 (Marginal Dirichlet) *If* $\mathbf{x}_1 = (x_{i_1}, \ldots, x_{i_k})'$ *denotes any subset of the coordinates, then* $\mathbf{x}_1 \sim D_k(\mathbf{p}_1; q)$ *with* $\mathbf{p}_1 = (p_{i_1}, \ldots, p_{i_k})'$ *and* $p = q + \sum_{j=1}^{k} p_{i_j}$.

Corollary 3.3 *If* $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_m')'$ *is "partitioned" in any manner whatever so that we may write* $\mathbf{x}_i \sim D_{k_i}(\mathbf{p}_i; q_i)$, $i = 1, \ldots, m$, *define* $\mathbf{y}$ *by letting* $y_i = \mathbf{x}_i'\mathbf{1}$, *i.e., the total of the components of* $\mathbf{x}_i$, *with corresponding* $r_i = \mathbf{p}_i'\mathbf{1}$. *We find* $\mathbf{y} \sim D_m(\mathbf{r}; p_{n+1})$ *with* $\mathbf{r} = (r_1, \ldots, r_m)'$.

Corollary 3.4 *If* $S = \mathbf{x}'\mathbf{1} = \sum_{i=1}^{n} x_i$ *and again* $\mathbf{x}_1 = (x_{i_1}, \ldots, x_{i_k})'$, $k < n$, *is any subset, let* $\mathbf{w}_1 \overset{d}{=} \frac{1}{S}\mathbf{x}_1$. *We find* $\mathbf{w}_1 \sim D_k(\mathbf{p}_1; r)$ *with* $\mathbf{p}_1 = (p_{i_1}, \ldots, p_{i_k})'$ *as before but this time,* $p - p_{n+1} = r + \sum_{j=1}^{k} p_{i_j}$.

Corollary 3.5 (Conditional Dirichlet) *If*

$$\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')' \sim D_n\left((\mathbf{p}_1', \mathbf{p}_2')'; p_{n+1}\right),$$

where $\mathbf{x}_1, \mathbf{p}_1 \in \mathbb{R}^r$ *and* $\mathbf{x}_2, \mathbf{p}_2 \in \mathbb{R}^s$, $n = r + s$, *then*

$$\frac{\mathbf{x}_1}{1 - \mathbf{x}_2'\mathbf{1}} \mid \mathbf{x}_2 \sim D_r(\mathbf{p}_1; p_{n+1}).$$

We easily compute the moments of a Dirichlet distribution. By the converse representation in Proposition 3.2, if $T \sim G(p)$, $\mathbf{x} \sim D_n(\mathbf{p}; p_{n+1})$, and $\mathbf{x} \perp\!\!\!\perp T$, we have

$$T\mathbf{x} \overset{d}{=} \mathbf{z} \text{ with } z_i \overset{\text{indep}}{\sim} G(p_i),\ i = 1, \ldots, n.$$

This gives

$$T^r \prod_{i=1}^{n} x_i^{r_i} \overset{d}{=} \prod_{i=1}^{n} z_i^{r_i} \text{ with } r = \sum_{i=1}^{n} r_i,$$

so that

$$E\ T^r\ E\ \prod_{i=1}^{n} x_i^{r_i} = \prod_{i=1}^{n} E\ z_i^{r_i} \text{ for } r_i > -p_i,\ i = 1, \ldots, n.$$

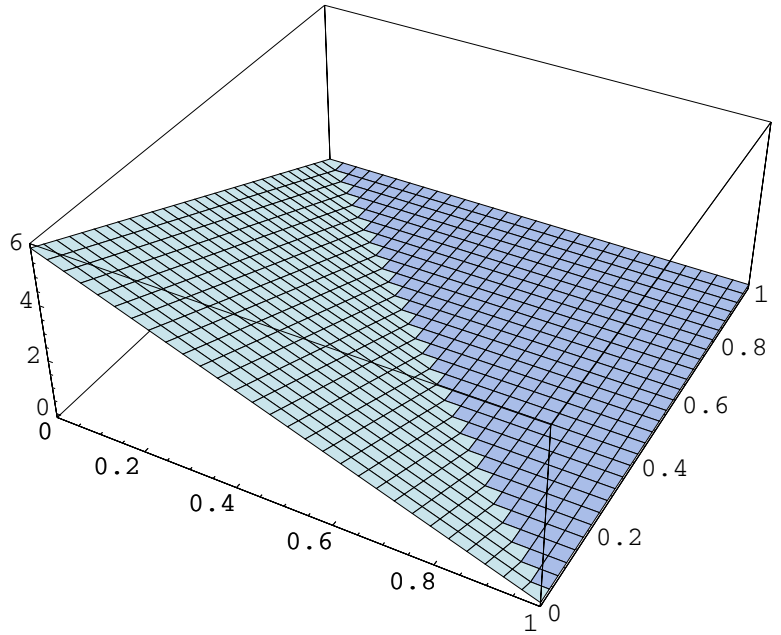

Figure 3.1. Bivariate Dirichlet density for values of the parameters $p_1 = p_2 = 1$ and $p_3 = 2$.

We find

$$E \prod_{i=1}^{n} x_i^{r_i} = \frac{\prod_{i=1}^{n} E\ z_i^{r_i}}{E\ T^r}.$$

In particular,

$$\begin{aligned} E\ x_i &= p_i/p,\ \forall i, \\ E\ x_i^2 &= (p_i+1)p_i/(p+1)p,\ \forall i, \\ \text{and } E\ x_i x_j &= p_i p_j/(p+1)p,\ \forall i \neq j, \end{aligned}$$

and letting $\boldsymbol{\theta} = \frac{1}{p}\mathbf{p}$ gives

$$E\ \mathbf{x} = \boldsymbol{\theta} \text{ and var } \mathbf{x} = \frac{1}{p+1}\left(\text{diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}'\right).$$

Figure 3.1 exhibits a bivariate Dirichlet density. Various characterizations of Dirichlet distributions can be found in the literature [Rao and Sinha (1988), Gupta and Richards (1990)].

3.4 F distributions

The ratio of two independent gammas is described by the F distribution, intuitively a relative waiting time.

Definition 3.6 F distribution*: $F \sim F(s_1, s_2)$ iff $F \stackrel{d}{=} \frac{y_1/s_1}{y_2/s_2}$, $y_i \stackrel{indep}{\sim} \chi^2_{s_i}$, $i = 1, 2$.*

One may easily obtain the moments of an F distribution and, in particular, its mean and variance. Sometimes, the distributions are more easily expressed in terms of the canonical F_c distribution:

Definition 3.7 Canonical F_c distribution*: $F \sim F_c(s_1, s_2)$ iff $F \stackrel{d}{=} y_1/y_2$, $y_i \stackrel{indep}{\sim} \chi^2_{s_i}$, $i = 1, 2$.*

One should also verify the simple relation

$$F \sim F_c(s_1, s_2) \Longleftrightarrow (1 + F)^{-1} \sim \text{beta}(\tfrac{1}{2}s_2; \tfrac{1}{2}s_1).$$

The noncentral chi-square and F distributions useful to describe the non-null distribution of some tests are defined in Section 4.3.

3.5 Problems

1. If $y \sim \chi^2_m$, then $E\, y^h = 2^h \Gamma\left(\frac{1}{2}m + h\right) / \Gamma\left(\frac{1}{2}m\right)$, $h > -\frac{1}{2}m$.
2. Prove Corollary 3.1.
3. Assume $\mathbf{x} \sim \text{unif}([\mathbf{0}, \mathbf{1}])$ in $\mathbb{R}^n$.
   (i) Define $\mathbf{y}$ by $y_1 = x_{(1)} = \min(\{x_1, \ldots, x_n\})$ and
   $$y_i = x_{(i)} = \min\left(\{x_1, \ldots, x_n\} - \{x_{(1)}, \ldots, x_{(i-1)}\}\right),$$
   $i = 2, \ldots, n$. Determine the distribution of $\mathbf{y}$.
   (ii) Define $\mathbf{z}$ by $z_1 = y_1$ and $z_i = y_i - y_{i-1}$, $i = 2, \ldots, n$, and determine the distribution of $\mathbf{z}$.
   (iii) Determine $E\, \mathbf{x}$, var $\mathbf{x}$, $E\, \mathbf{y}$, and var $\mathbf{y}$ as well as $E\, \mathbf{z}$ and var $\mathbf{z}$.
4. Prove Corollaries 3.2, 3.3, 3.4, and 3.5.
5. Show the simple equivalence
   $$F \sim F_c(s_1, s_2) \Longleftrightarrow (1 + F)^{-1} \sim \text{beta}(\tfrac{1}{2}s_2; \tfrac{1}{2}s_1).$$
6. Obtain the density of $F \sim F_c(s_1, s_2)$:
   $$f(F) = \frac{\Gamma\left(\frac{1}{2}(s_1 + s_2)\right)}{\Gamma\left(\frac{1}{2}s_1\right)\ \Gamma\left(\frac{1}{2}s_2\right)} \frac{F^{s_1/2-1}}{(1 + F)^{(s_1+s_2)/2}},\ F > 0.$$

4
Invariance

4.1 Introduction

Invariance is a distributional property of a random vector acted upon by a group of transformations. The simplest group of transformations $\{+1, -1\}$ leads to symmetric distributions by defining a random variable to be symmetric iff $x \stackrel{d}{=} -x$. Groups of transformations acting on random vectors commonly encountered are the permutations and orthogonal transformations. The permutation invariance gives the "exchangeable" random vectors and the invariance by orthogonal transformations defines the spherical distributions. Of great importance is the orthogonal group, since it specifies the physical basis for normality in the Maxwell-Hershell theorem. Spherical distributions will play a central role later in Chapter 13 to build the elliptical models useful in the study of robustness.

4.2 Reflection symmetry

Definition 4.1 *x is (reflection)* symmetric *iff* $x \stackrel{d}{=} -x$.

One immediately notes that if $x \stackrel{d}{=} -x$ and $E\ |x| < \infty$, then $E\ x = 0$ (why?). The distribution of a symmetric random variable x is completely determined by the distribution of its modulus $|x|$, as the next proposition shows.

Proposition 4.1 $x \stackrel{d}{=} -x \Longleftrightarrow x \stackrel{d}{=} s|x|$ *with* $s \perp\!\!\!\perp |x|$, $s \sim \text{unif}\{\pm 1\}$.

Proof. ($\Longrightarrow$) : Let F be the d.f. of x. Then,

$$\begin{aligned} P(s|x| \le t) &= \tfrac{1}{2}\left\{P(|x| \le t) + P(|x| \ge -t)\right\} \\ &= \tfrac{1}{2} \cdot \begin{cases} P(-t \le x \le t) + 1, & t \ge 0 \\ P(|x| \ge -t), & t < 0 \end{cases} \\ &= \tfrac{1}{2} \cdot \begin{cases} (2F(t) - 1) + 1, & t \ge 0 \\ 2F(t), & t < 0 \end{cases} \\ &= F(t). \end{aligned}$$

($\Longleftarrow$) : Since $s \sim \text{unif}\{\pm 1\}$ is symmetric, then

$$\begin{aligned} s \stackrel{d}{=} -s,\ s \perp\!\!\!\perp |x| &\Longrightarrow (s, |x|) \stackrel{d}{=} (-s, |x|) \\ &\Longrightarrow x \stackrel{d}{=} s|x| \stackrel{d}{=} -s|x| \stackrel{d}{=} -x. \end{aligned}$$

□

If $p(0) = P(x = 0) = 0$, we may specifically let $s = x/|x|$ be the sign of x and show that when x is symmetric, the sign of x is $+1$ or -1 with probability $\frac{1}{2}$ and is distributed independently of the modulus $|x|$.

Proposition 4.2 *If* $x \stackrel{d}{=} -x$ *and* $p(0) = 0$ *then* $x/|x| \perp\!\!\!\perp |x|$ *and* $x/|x| \sim \text{unif}\{\pm 1\}$.

Proof. Uniform:

$$x/|x| \stackrel{d}{=} -x/|x| \Longrightarrow P\left(x/|x| = 1\right) = P\left(x/|x| = -1\right) = \frac{1}{2}.$$

Independence:

$$\begin{aligned} P\left(|x| \le t, x/|x| = 1\right) &= P\left(|x| \le t, x > 0\right) \\ &= P(0 < x \le t) = \tfrac{1}{2}P(|x| \le t). \end{aligned}$$

□

Obviously, by these propositions, one may generate symmetric distributions at will.

Example 4.1 *Let* $x \stackrel{d}{=} -x$ *and* $|x| \sim \exp(1)$ *to obtain the "double exponential" (Laplace distribution) with p.d.f.* $f(x) = \frac{1}{2}\exp(-|x|)$.

4.3 Univariate normal and related distributions

Definition 4.2 Standard normal*:* $z \sim N(0,1)$ *iff* $f_z(z) = (2\pi)^{-1/2} e^{-\frac{1}{2}z^2}$.

The univariate normal is so intimately connected to the gamma(1/2) that these two may safely be thought of as synonymous. We easily verify:

Proposition 4.3 $z \sim N(0,1) \Longleftrightarrow z \stackrel{d}{=} -z$ *and* $\frac{1}{2}z^2 \sim G(\frac{1}{2})$.

From the convolution of gamma variables in Corollary 3.1, we obtain immediately

Proposition 4.4 $y \sim \chi^2_m \Longleftrightarrow y \stackrel{d}{=} \sum_{i=1}^m z_i^2$, *with* $z_1, \ldots, z_m$ *i.i.d.* $N(0,1)$.

By the representation in Proposition 4.3, we quickly produce the (integral) moments:

$$\begin{array}{lll} \text{n odd:} & E\ z^n & = 0 \text{ (i.e. } z^n \stackrel{d}{=} -z^n) \\ \text{n even:} & E\ z^n & = E\ z^{2k} \\ & & = 2^k\ E\ w^k,\ w \sim G(\frac{1}{2}) \\ & & = 2^k \Gamma(k+\frac{1}{2})/\Gamma(\frac{1}{2}) \\ & & = (n-1)(n-3)\cdots 3 \cdot 1. \end{array}$$

In particular, $E\ z = 0$ and $\text{var}\ z = 1$.

The more general normal is obtained by simply relocating and rescaling:

Definition 4.3 General normal*:* $x \sim N(\mu, \sigma^2)$ *iff* $x \stackrel{d}{=} \sigma z + \mu$, $z \sim N(0,1)$.

Clearly, the integral moments of x are simple polynomials in μ and σ^2

$$E\ x^n = \sum_{i=0}^n \binom{n}{i} \sigma^i E(z^i) \mu^{n-i},$$

and one may write these out explicitly. In particular,

$$E\ x = \mu,\ \text{var}\ x = \sigma^2,\ \text{and}\ E\ x^2 = \mu^2 + \sigma^2.$$

Also, immediate from (2.1) is the characteristic function

$$c_x(t) = e^{it\mu} c_z(\sigma t) = e^{it\mu - \frac{1}{2}\sigma^2 t^2}. \tag{4.1}$$

We digress somewhat from invariance considerations to introduce two important noncentral distributions. Motivated by the chi-square representation in Proposition 4.4, we now define the noncentral chi-square and F distributions and show some characterizations.

Definition 4.4 Noncentral chi-square*:* $y \sim \chi^2_m(\delta)$ *iff* $y \stackrel{d}{=} \sum_{i=1}^m x_i^2$, *with* $x_i \stackrel{indep}{\sim} N(\mu_i, 1)$, $i = 1, \ldots, m$, *and* $\delta = \sum_{i=1}^m \mu_i^2/2$.

Definition 4.5 Noncentral F*:* $F \sim F(s_1, s_2; \delta)$ *iff* $F \stackrel{d}{=} \frac{y_1/s_1}{y_2/s_2}$ *with* $y_1 \sim \chi^2_{s_1}(\delta)$, $y_2 \sim \chi^2_{s_2}$, *and* $y_1 \perp\!\!\!\perp y_2$.

Definition 4.6 Noncentral canonical F_c*:* $F \sim F_c(s_1, s_2; \delta)$ *iff* $F \stackrel{d}{=} y_1/y_2$, $y_1 \sim \chi^2_{s_1}(\delta)$, $y_2 \sim \chi^2_{s_2}$, *and* $y_1 \perp\!\!\!\perp y_2$.

Proposition 4.5 $y \sim \chi^2_m(\delta) \Longrightarrow c_y(t) = (1-2it)^{-m/2} \exp[\delta 2it/(1-2it)]$.

Proof. By definition, if $z_1, \ldots, z_{m-1}, x$ are independent with $z_i \sim N(0,1)$, $i = 1, \ldots, m-1$, and $x \sim N(\mu, 1)$, then $y = \sum_{i=1}^{m-1} z_i^2 + x^2 \sim \chi^2_m(\delta)$, where $\delta = \mu^2/2$. Using independence and the characteristic function of $\sum_{i=1}^{m-1} z_i^2 \sim \chi^2_{m-1}$, $c_y(t) = (1-2it)^{-(m-1)/2} \cdot c_{x^2}(t)$. By direct computation,

$$\begin{aligned} c_{x^2}(t) &= \int_{-\infty}^{\infty} e^{itx^2} (2\pi)^{-1/2} e^{-\frac{1}{2}(x-\mu)^2} dx \\ &= \exp\left[\frac{\mu^2 it}{(1-2it)}\right] \\ &\quad \cdot \int_{-\infty}^{\infty} (2\pi)^{-1/2} \exp\left\{-\frac{(1-2it)}{2}\left[x - \frac{\mu}{(1-2it)}\right]^2\right\} dx \\ &= (1-2it)^{-1/2} \exp[\delta 2it/(1-2it)]. \end{aligned}$$

Thus, $c_y(t) = (1-2it)^{-m/2} \exp[\delta 2it/(1-2it)]$. □

Proposition 4.6 *If $y \sim \chi^2_m(\delta)$, then*

$$P(y \le t) = \sum_{k=0}^{\infty} e^{-\delta} \frac{\delta^k}{k!} P\left(\chi^2_{m+2k} \le t\right);$$

i.e., y is a Poisson mixture of central chi-square distributions.

Proof. A Taylor series gives

$$\exp\left[\frac{\delta 2it}{(1-2it)}\right] = e^{-\delta} \exp\left[\frac{\delta}{(1-2it)}\right] = e^{-\delta} \sum_{k=0}^{\infty} \frac{\delta^k}{k!} (1-2it)^{-k}.$$

Hence,

$$c_y(t) = \sum_{k=0}^{\infty} e^{-\delta} \frac{\delta^k}{k!} (1-2it)^{-\frac{1}{2}(m+2k)}.$$

This means that if we define $K, u_0, u_1, \ldots$ mutually independent where $u_i \sim \chi^2_{m+2i}$ and $K \sim \text{Poisson}(\delta)$, then $y \stackrel{d}{=} \sum_{k=0}^{\infty} u_k \cdot I(K = k)$. Finally, since $y \mid K \stackrel{d}{=} u_K$, it comes

$$P(y \le t) = E\ P(y \le t \mid K) = \sum_{k=0}^{\infty} e^{-\delta} \frac{\delta^k}{k!} P\left(\chi^2_{m+2k} \le t\right).$$

□

A similar expansion for noncentral F_c and, of course, F distributions exist.

Proposition 4.7 *If $F \sim F_c(s_1, s_2; \delta)$, then*

$$P(F \le t) = \sum_{k=0}^{\infty} e^{-\delta} \frac{\delta^k}{k!} P\left(F_c(s_1 + 2k, s_2) \le t\right);$$

i.e., F is a Poisson mixture of central F_c distributions.

Proof. As in the proof of Proposition 4.6, if we define $K, y_2, u_0, u_1, \ldots$ mutually independent where $K \sim \text{Poisson}(\delta)$, $u_i \sim \chi^2_{s_1+2i}$ and $y_2 \sim \chi^2_{s_2}$, then

$$F \stackrel{d}{=} \sum_{k=0}^{\infty} (u_k/y_2) I(K = k).$$

Now, since $F \mid K \stackrel{d}{=} F_c(s_1 + 2K, s_2)$,

$$P(F \le t) = E\ P(F \le t \mid K) = \sum_{k=0}^{\infty} e^{-\delta} \frac{\delta^k}{k!} P\left(F_c(s_1 + 2k, s_2) \le t\right),$$

which concludes the proof. □

4.4 Permutation invariance

We represent any permutation σ of $1, \ldots, n$ by the linear transformation obtained by the corresponding permutation of the columns of the identity

$$\mathbf{J}_\sigma = \left(\mathbf{e}_{\sigma(1)}, \ldots, \mathbf{e}_{\sigma(n)}\right).$$

In the transpose,

$$\mathbf{J}'_\sigma \mathbf{x} = (x_{\sigma(1)}, \ldots, x_{\sigma(n)})'$$

permutes the elements of $\mathbf{x}$. We denote the group of all such permutations:

$$\mathbf{S}_n = \{\mathbf{J}_\sigma :\ \sigma = \text{permutation of } 1, \ldots, n\}.$$

Definition 4.7 $\mathbf{x}$ *is* permutationally invariant *(exchangeable) iff* $\mathbf{x} \stackrel{d}{=} \mathbf{J}\mathbf{x}$, $\forall \mathbf{J} \in \mathbf{S}_n$.

It is obviously a very special case when the x_i's are i.i.d. x as the characteristic function shows

$$c_{\mathbf{x}}(\mathbf{t}) = \prod_{i=1}^{n} c_x(t_i) = c_{\mathbf{x}}(\mathbf{J}'\mathbf{t}) = c_{\mathbf{J}\mathbf{x}}(\mathbf{t}),\ \forall \mathbf{J} \in \mathbf{S}_n.$$

Any subvector of an exchangeable random vector will also be exchangeable and all subvectors of the same dimension will be identically distributed. In particular,

$$x_i \stackrel{d}{=} x_1,\ \forall i \quad \text{and} \quad x_i x_j \stackrel{d}{=} x_1 x_2,\ \forall i \ne j.$$

If $\mathbf{x}$ is permutationally invariant, this forces the mean and variance to have a certain structure. Let $E\ x_1 = \mu$, $\text{var}\ x_1 = \sigma^2$, and $\text{cov}(x_1, x_2) = \rho\sigma^2$ to find

$$E\ \mathbf{x} = \mu\mathbf{1} \quad \text{and} \quad \text{var}\ \mathbf{x} = \sigma^2 \left\{(1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}'\right\}.$$

The inequality $\rho \geq -1/(n-1)$ must hold, as the eigenvalues of var $\mathbf{x}$ are positive. Furthermore, in statistical applications, the physical assumption that $\mathbf{x}$ is exchangeable is usually made independent of the sample size n; but this (additional) assumption forces $\rho \geq 0$. We verify also this fact by normalizing to $z_i = (x_i - \mu)/\sigma$, expressing ρ in terms of var $\bar{z}$ and then passing to the limit in n (v. Problem 4.6.9).

Permutationally invariant vectors are used, in particular, to model familial data where $\mathbf{x} = (x_1, \ldots, x_n)'$ represents a variable observed on n siblings of a family. The parameter ρ in that context is often referred to as the *intraclass correlation* coefficient. Results of a minimum variance unbiased estimation in an exchangeable model are available under the normality assumption [Yamato (1990)].

4.5 Orthogonal invariance

Recall that an *orthogonal transformation*, $\mathbf{H}$, is simply one which preserves length:

$$|\mathbf{Hx}| = |\mathbf{x}|, \ \forall \mathbf{x} \in \mathbb{R}^n.$$

However, it is clearly equivalent that $\mathbf{H}$ preserves the inner product:

$$|\mathbf{H}(\mathbf{x}+\mathbf{y})|^2 = |\mathbf{x}+\mathbf{y}|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \iff (\mathbf{Hx})'(\mathbf{Hy}) = \mathbf{x}'\mathbf{y}, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

and from this, it is equivalent that $\mathbf{H}$ be nonsingular with the inverse equal to its own transpose:

$$\mathbf{x}'\mathbf{H}'\mathbf{Hy} = \mathbf{x}'\mathbf{y}, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \iff \mathbf{H}'\mathbf{H} = \mathbf{I} = \mathbf{HH}';$$

this last expression is identical with the fact that the columns (or rows) of $\mathbf{H}$ determine an orthonormal basis of $\mathbb{R}^n$.

We denote the group of all orthogonal transformations by

$$\mathbf{O}_n = \{\mathbf{H} \in \mathbb{R}^n_n : \ \mathbf{H}'\mathbf{H} = \mathbf{I}\}.$$

The subgroup of all orthogonal transformations with positive "orientation" is known as the *rotation group*: $\mathbf{O}_n^+ = \{\mathbf{H} \in \mathbf{O}_n : \ |\mathbf{H}| = 1\}$.

Definition 4.8 $\mathbf{x}$ *is* orthogonally invariant *(spherical) iff* $\mathbf{x} \stackrel{d}{=} \mathbf{Hx}$, $\forall \mathbf{H} \in \mathbf{O}_n$.

It is clear that this includes reflection symmetry as a special case when $n = 1$, but the considerations become altogether more interesting when $n \geq 2$, and in what follows, we will assume this.

Since $\mathbf{S}_n \subset \mathbf{O}_n$ (subgroup), it is clear that $\mathbf{x}$ is permutationally invariant, but, in addition, we will also have reflection symmetry in the coordinates. Thus, in particular,

$$x_i \stackrel{d}{=} -x_i, \ \forall i \text{ and } x_i x_j \stackrel{d}{=} -x_i x_j, \ \forall i \neq j,$$

and this forces $\mu = \rho = 0$ by which we must always have

$$E\,\mathbf{x} = \mathbf{0} \text{ and } \operatorname{var}\mathbf{x} = \sigma^2\mathbf{I}.$$

The following result reveals that the distribution of any orthogonally invariant $\mathbf{x}$ is completely determined by its first coordinate; it is important in the sequel.

Proposition 4.8 $\mathbf{x}$ *is orthogonally invariant* $\Longleftrightarrow \mathbf{t}'\mathbf{x} \stackrel{d}{=} x_1,\ \forall \mathbf{t} \in S^{n-1}$.

Proof. ($\Longrightarrow$): Let $\mathbf{t}'$ be the first row of $\mathbf{H} \in \mathbf{O}_n$ and project onto the first coordinate.
($\Longleftarrow$): Let $\mathbf{H} \in \mathbf{O}_n$. Since both $\mathbf{H}'\mathbf{t}$ and $\mathbf{t}$ are unit vectors if $\mathbf{t}$ is, we have immediately

$$x_1 \stackrel{d}{=} \mathbf{t}'\mathbf{x} \stackrel{d}{=} \mathbf{t}'\mathbf{H}\mathbf{x},\ \forall \mathbf{H} \in \mathbf{O}_n,\ \forall \mathbf{t} \in S^{n-1}.$$

The Cramér-Wold theorem (v. Proposition 2.10) gives the conclusion. □

The characterization in Proposition 4.8 was used by Fang et al. (1993) to get a Wilcoxon-type goodness-of-fit test for orthogonal invariance. Koltchinskii and Li (1998) provide other clues to this testing problem.

Corollary 4.1 $\mathbf{x}$ *is orthogonally invariant* $\Longleftrightarrow \mathbf{x} \stackrel{d}{=} \mathbf{H}\mathbf{x},\ \forall \mathbf{H} \in \mathbf{O}_n^+$.

Proof. For any $\mathbf{t} \in S^{n-1}$, one can always find $\mathbf{H} \in \mathbf{O}_n^+$, whose first row is $\mathbf{t}'$. Since $\mathbf{x} \stackrel{d}{=} \mathbf{H}\mathbf{x}$, project on the first coordinate to obtain $x_1 \stackrel{d}{=} \mathbf{t}'\mathbf{x}$. □

With Corollary 4.1 the terms "orthogonal invariance" and "rotational invariance" can be used interchangingly and we will use the latter in the sequel.

Now, perhaps, the very most obviously rotationally invariant distribution is a spherical uniform. Accordingly, we consider $\mathbf{x} \sim \text{unif}(B^n)$ with $B^n = \{\mathbf{s} \in \mathbb{R}^n : |\mathbf{s}| < 1\}$ the "unit ball" in $\mathbb{R}^n$. The reader can show that if we define $\mathbf{y}$ by $y_i = x_i^2$, $i = 1, \ldots, n$, then $\mathbf{y} \sim D_n(\frac{1}{2}\mathbf{1}; 1)$ and $R = |\mathbf{x}| \sim \text{beta}(n; 1)$ (v. Problem 4.6.8). We simply "project" this distribution onto its $(n-1)$-dimensional boundary to obtain:

Definition 4.9 $\mathbf{u} \sim \text{unif}(S^{n-1})$ *iff* $\mathbf{u} \stackrel{d}{=} \mathbf{x}/|\mathbf{x}|$ *with* $\mathbf{x} \sim \text{unif}(B^n)$.

It is clear that $\mathbf{u}$ is rotationally invariant, inheriting this property directly from $\mathbf{x}$. What, however, may be a bit surprising is that $\mathbf{u}$ is *the only* rotationally invariant distribution on S^{n-1}:

Proposition 4.9 $\mathbf{z}$ *is rotationally invariant on* S^{n-1} *iff* $\mathbf{z} \sim \text{unif}(S^{n-1})$.

Proof. Assume $\mathbf{z}$ is rotationally invariant on S^{n-1}. Take $\mathbf{u} \perp\!\!\!\perp \mathbf{z}$, where $\mathbf{u} \sim \text{unif}(S^{n-1})$. Then,

$$\mathbf{z}'\mathbf{u} \mid \mathbf{z} \stackrel{d}{=} u_1 \mid \mathbf{z} \stackrel{d}{=} u_1 \stackrel{d}{=} \mathbf{z}'\mathbf{u},$$

but this is completely symmetric in $\mathbf{z}$ and $\mathbf{u}$, so

$$\mathbf{u}'\mathbf{z} \mid \mathbf{u} \stackrel{d}{=} z_1 \mid \mathbf{u} \stackrel{d}{=} z_1 \stackrel{d}{=} \mathbf{u}'\mathbf{z}$$

and we conclude

$$\mathbf{t}'\mathbf{u} \stackrel{d}{=} u_1 \stackrel{d}{=} z_1 \stackrel{d}{=} \mathbf{t}'\mathbf{z}, \ \forall \mathbf{t} \in S^{n-1}.$$

Using the Cramér-Wold Proposition 2.10, $\mathbf{u} \stackrel{d}{=} \mathbf{z}$. □

As a bonus, we also get the distribution of the cosine of the angle between two vectors independently and uniformly distributed on S^{n-1}.

Corollary 4.2 $\mathbf{u}, \mathbf{z} \stackrel{indep}{\sim} \text{unif}(S^{n-1}) \Longrightarrow (\mathbf{u}'\mathbf{z})^2 \sim \text{beta}\left(\frac{1}{2}; \frac{1}{2}(n-1)\right)$.

Proof. This is just a by-product of the proof of Proposition 4.9, in which we found $\mathbf{z}'\mathbf{u} \stackrel{d}{=} u_1$, where $u_1 \stackrel{d}{=} x_1/|\mathbf{x}|$ with $\mathbf{x} \sim \text{unif}(B^n)$. But, then, $(\mathbf{z}'\mathbf{u})^2 \stackrel{d}{=} x_1^2/|\mathbf{x}|^2$, where $(x_1^2, \ldots, x_n^2) \sim D_n(\frac{1}{2}\mathbf{1}; 1)$ by which, from Corollary 3.4, $x_1^2/|\mathbf{x}|^2 \sim D_1\left(\frac{1}{2}; \frac{1}{2}(n-1)\right)$. □

The culmination of all this is a fundamental representation in "polar" coordinates of the general rotationally invariant distribution with respect to the uniform distribution on the sphere. In this sense, the unif(S^{n-1}) will be referred to as the "unit spherical" distribution. This representation was used recently by Gupta and Song (1997) and Szablowski (1998) to define and characterize l_p-norm spherical distributions.

Proposition 4.10 $\mathbf{x}$ *is rotationally invariant* $\Longleftrightarrow \mathbf{x} \stackrel{d}{=} R\mathbf{u}$ *with* $R \stackrel{d}{=} |\mathbf{x}|$, $\mathbf{u} \sim \text{unif}(S^{n-1})$, $R \perp\!\!\!\perp \mathbf{u}$.

Proof. ($\Longleftarrow$): Since $\mathbf{H}\mathbf{u} \stackrel{d}{=} \mathbf{u}$, $\forall \mathbf{H} \in \mathbf{O}_n$, and $\mathbf{u} \perp\!\!\!\perp R$,

$$\begin{aligned} &\Longrightarrow \quad (R, \mathbf{H}\mathbf{u}) \stackrel{d}{=} (R, \mathbf{u}), \ \forall \mathbf{H} \in \mathbf{O}_n \\ &\Longrightarrow \quad \mathbf{H}R\mathbf{u} \stackrel{d}{=} R\mathbf{u}, \ \forall \mathbf{H} \in \mathbf{O}_n \\ &\Longrightarrow \quad \mathbf{H}\mathbf{x} \stackrel{d}{=} \mathbf{H}R\mathbf{u} \stackrel{d}{=} R\mathbf{u} \stackrel{d}{=} \mathbf{x}, \ \forall \mathbf{H} \in \mathbf{O}_n. \end{aligned}$$

($\Longrightarrow$): Let $R \stackrel{d}{=} |\mathbf{x}|$, $\mathbf{u} \sim \text{unif}(S^{n-1})$, $R \perp\!\!\!\perp \mathbf{u}$, and take any $\mathbf{v} \sim \text{unif}(S^{n-1})$, $\mathbf{v} \perp\!\!\!\perp \mathbf{x}$. Then,

(1) $(R, \mathbf{u}) \stackrel{d}{=} (|\mathbf{x}|, \mathbf{v}) \Longrightarrow R\mathbf{u} \stackrel{d}{=} |\mathbf{x}|\mathbf{v} \Longrightarrow Ru_1 \stackrel{d}{=} |\mathbf{x}|v_1$,

(2) $\mathbf{x}'\mathbf{v} \mid \mathbf{x} \stackrel{d}{=} |\mathbf{x}|v_1 \mid \mathbf{x} \Longrightarrow \mathbf{v}'\mathbf{x} \stackrel{d}{=} |\mathbf{x}|v_1$,

(3) $\mathbf{v}'\mathbf{x} \mid \mathbf{v} \stackrel{d}{=} x_1 \mid \mathbf{v} \Longrightarrow x_1 \stackrel{d}{=} \mathbf{v}'\mathbf{x}$,

and as easy as (1), (2), (3), $x_1 \stackrel{d}{=} Ru_1$. However, since $\mathbf{x}$ and $R\mathbf{u}$ are both rotationally invariant, then for any $\mathbf{t} \in S^{n-1}$,

$$\mathbf{t}'\mathbf{x} \stackrel{d}{=} x_1 \stackrel{d}{=} Ru_1 \stackrel{d}{=} \mathbf{t}'R\mathbf{u}$$

and from the Cramér-Wold Proposition 2.10, $\mathbf{x} \stackrel{d}{=} R\mathbf{u}$. □

Thus, any rotationally invariant $\mathbf{x}$ is completely determined by its "modulus" R. Moreover, if there is no probability at the origin, $p_{\mathbf{x}}(\mathbf{0}) = 0$, we have that the "direction" $\mathbf{u}$ of $\mathbf{x}$, $\mathbf{u} = \mathbf{x}/|\mathbf{x}|$, is unif$(S^{n-1})$ and distributed independently of the modulus $R = |\mathbf{x}|$.

Corollary 4.3 $\mathbf{x}$ *is rotationally invariant* $\iff$ $|\mathbf{x}| \perp\!\!\!\perp \mathbf{x}/|\mathbf{x}|$ *and* $\mathbf{x}/|\mathbf{x}| \sim$ unif(S^{n-1}).

Proof. ($\Longleftarrow$): By expressing $\mathbf{x} = |\mathbf{x}| \cdot \mathbf{x}/|\mathbf{x}|$, Proposition 4.10 gives the result.
($\Longrightarrow$): Simply note that $\mathbf{x} \stackrel{d}{=} R\mathbf{u}$, where $R \stackrel{d}{=} |\mathbf{x}|$, $\mathbf{u} \sim$ unif(S^{n-1}), and $R \perp\!\!\!\perp \mathbf{u}$. Thus, $(|\mathbf{x}|, \mathbf{x}/|\mathbf{x}|) \stackrel{d}{=} (R, \mathbf{u})$. □

Another proof due to Kariya and Eaton (1977) does not assume a density and relies on the unicity of a rotationally invariant distribution on S^{n-1}. We finish this section by obtaining a fundamental result that precisely specifies the physical basis for normality with rotational invariance at the very heart ($n \geq 2$).

Proposition 4.11 (Maxwell-Hershell) $x_1, \ldots, x_n$ *i.i.d.* $N(0, \sigma^2)$ $\iff$ $\mathbf{x}$ *is rotationally invariant and* $x_1, \ldots, x_n$ *are independent.*

Proof. ($\Longrightarrow$): Independence is given and since

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2}\frac{1}{\sigma^2}|\mathbf{x}|^2\right)$$

depends only on $|\mathbf{x}|$, rotational invariance is obvious.
($\Longleftarrow$): Let $c(t)$ be the characteristic function of x_1. Since x_1 is symmetric, $c(t) = c(-t) = \overline{c(t)}$, and, thus, $c(t) \in \mathbb{R}$ and, of course, $-1 \leq c(t) \leq 1$. Both hypotheses may be expressed with respect to the characteristic function

$$\prod_{i=1}^{n} c(t_i) = c_{\mathbf{x}}(\mathbf{t}) = c_{\mathbf{Hx}}(\mathbf{t}),\ \forall \mathbf{H} \in \mathbf{O}_n,\ \mathbf{t} \in \mathbb{R}^n.$$

But, of course,

$$c_{\mathbf{Hx}}(\mathbf{t}) = c_{\mathbf{x}}(\mathbf{H}'\mathbf{t}),\ \forall \mathbf{H} \in \mathbf{O}_n,\ \mathbf{t} \in \mathbb{R}^n,$$

and we may specifically choose $\mathbf{H} = (\,\mathbf{t}/|\mathbf{t}|, \quad \mathbf{\Gamma}\,) \in \mathbf{O}_n$ for some $\mathbf{\Gamma}$ so that $\mathbf{H}'\mathbf{t} = |\mathbf{t}|\mathbf{e}_1$ whence, altogether, $\prod_{i=1}^{n} c(t_i) = c(|\mathbf{t}|)$, $\forall \mathbf{t} \in \mathbb{R}^n$. Then, letting $\mathbf{t} = (s, t, 0, \ldots, 0)$ and defining $h(x) = c(\sqrt{x})$, $\forall x \geq 0$, we find Hamel's equation

$$h(s^2 + t^2) = h(s^2)\ h(t^2),\ \forall s, t.$$

But then,

$$\begin{aligned} h(x) &= h(2 \cdot x/2) = [h(x/2)]^2 \geq 0,\ \forall x \geq 0, \\ h(1) &= h(p \cdot 1/p) = [h(1/p)]^p,\ p = 1, 2, \ldots, \end{aligned}$$

and

$$h(p/q) = h(p \cdot 1/q) = [h(1/q)]^p = [h(1)]^{p/q},\ p, q = 1, 2, \ldots.$$

Since the rational numbers are dense and h is continuous, then $h(x) = [h(1)]^x = \exp(-kx)$, $\forall x \geq 0$, where $k \equiv -\ln h(1) \geq 0$. Finally, $c(t) = h(t^2) = \exp(-kt^2)$, $\forall t$. The case $h(1) = 0$ was excluded, as it would imply that $c(0) = 0$, a contradiction, and $h(1) = 1$ corresponds to $x_1 = 0$ w.p.1 (with probability 1), a degenerate normal with $\sigma^2 = 0$. □

Thus, we have now a fair understanding of the basic physics of normality. In the next chapter, we give a more mathematical treatment of the multivariate normal "family" in general; based on the theorem of Cramér and Wold, multivariate normality will be (by definition) directly equated to univariate normality of the linear functionals.

4.6 Problems

1. If $y \sim \chi^2_m(\delta)$, show that

$$\begin{aligned} E\ y &= m + 2\delta, \\ \text{var}\ y &= 2m + 8\delta. \end{aligned}$$

2. If $F \sim F(s_1, s_2; \delta)$, show that

$$\begin{aligned} E\ F &= \frac{s_2(s_1 + 2\delta)}{s_1(s_2 - 2)},\ s_2 > 2, \\ \text{var}\ F &= 2\frac{s_2^2}{s_1^2}\left[\frac{(s_1 + 2\delta)^2 + (s_1 + 4\delta)(s_2 - 2)}{(s_2 - 2)^2(s_2 - 4)}\right],\ s_2 > 4. \end{aligned}$$

3. Obtain the density of $F \sim F_c(s_1, s_2; \delta)$ using Problem 3.5.6:

$$f(F) = \sum_{k=0}^{\infty} e^{-\delta}\frac{\delta^k}{k!}\frac{\Gamma\left(\frac{1}{2}(s_1 + s_2 + 2k)\right)}{\Gamma\left(\frac{1}{2}(s_1 + 2k)\right)\ \Gamma\left(\frac{1}{2}s_2\right)}\frac{F^{(s_1+2k)/2-1}}{(1+F)^{(s_1+s_2+2k)/2}},$$

   $F > 0$.

4. Assume $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ has a spherical distribution. Show that $\mathbf{x}_1$ also has a spherical distribution.

5. Let $\mathbf{x} \in \mathbb{R}^n$ have a spherical distribution with a finite rth moment. Demonstrate that all product-moments of $\mathbf{x}$, $E(x_1^{s_1} \cdots x_n^{s_n})$, of order $s = \sum_{i=1}^n s_i \leq r$ are null provided one of the s_i is odd.

6. Let $\mathbf{x} = (x_1, \ldots, x_n)'$ have a spherical distribution. Prove the following:

   (i) $c_{\mathbf{x}}(\mathbf{t}) = c_{x_1}(|\mathbf{t}|)$ is a function of $|\mathbf{t}|$.

   (ii) If $\mathbf{x}$ is absolutely continuous, $\mathbf{x} \sim f$, then $f(\mathbf{x}) = f(|\mathbf{x}|\mathbf{e}_1)$ depends on $\mathbf{x}$ only through $|\mathbf{x}|$.

7. Assume $\mathbf{x} \in \mathbb{R}^n$ is rotationally invariant. Prove the mixture characterization

$$c_{\mathbf{x}}(\mathbf{t}) = \int_0^\infty c_{u_1}(|\mathbf{t}|r)dF(r),$$

where $\mathbf{u} = (u_1, \ldots, u_n)' \sim \text{unif}(S^{n-1})$ and F is the distribution function of $|\mathbf{x}|$ on $[0, \infty)$. This means any rotationally invariant distribution is a mixture of uniform distributions on spheres of varying radius $r \geq 0$ [Schoenberg (1938)].

8. Let $\mathbf{x} \sim \text{unif}(B^n)$, where $B^n = \{\mathbf{s} : \mathbf{s}'\mathbf{s} \leq 1\}$ is the "unit ball" in $\mathbb{R}^n$.
   (i) Define $\mathbf{y}$ by $y_i = x_i^2$, $i = 1, \ldots, n$, and determine the distribution of $\mathbf{y}$, the marginal distribution of each y_i, $i = 1, \ldots, n$, and, finally, the distribution of $R^2 = |\mathbf{x}|^2 = \mathbf{x}'\mathbf{x}$.
   (ii) Obtain $\text{vol}(B^n)$ using (i) and indicate the special cases $n = 1, 2, 3$.
   (iii) Determine $E\ \mathbf{x}$ and $\text{var}\ \mathbf{x}$ as well as $E\ R^2$ and $\text{var}\ R^2$.

   Hint: Realize that $\mathbf{y}$ is "concentrated" on

$$T^n = \{\mathbf{y} : y_i \geq 0, \sum_{i=1}^n y_i \leq 1\}.$$

9. Assume $\mathbf{x} \in \mathbb{R}^n$ is permutationally invariant $\forall n$ and $E\ |\mathbf{x}|^2 < \infty$. Let $S = g(\mathbf{x})$, where $g : \mathbb{R}^n \to \mathbb{R}$ is any (permutation) symmetric function, i.e., $g(\mathbf{Jt}) = g(\mathbf{t}),\ \forall \mathbf{J} \in \mathbf{S}_n,\ \forall \mathbf{t} \in \mathbb{R}^n$.
   (i) Prove $\rho \geq 0$.
   (ii) $E\left(f(\mathbf{x}) \mid S\right) \overset{\text{w.p.1}}{=} E\left(f(\mathbf{Jx}) \mid S\right),\ \forall \mathbf{J} \in \mathbf{S}_n$.
   (iii) $\text{cov}(x_1, x_2 \mid \bar{x}) \overset{\text{w.p.1}}{\leq} 0$.

10. Assume $\mathbf{x} \in \mathbb{R}^n$ has a "rotationally invariant" distribution such that $P(\mathbf{x} = \mathbf{0}) = 0$, i.e., $\mathbf{Hx} \overset{d}{=} \mathbf{x}$, $\forall \mathbf{H} \in \mathbf{O}_n$. Let $R = |\mathbf{x}|$ and $\mathbf{z} = \mathbf{x}/R$.
    (i) Prove that $\mathbf{z}$ has the same distribution as if $\mathbf{x}$ had been $\text{unif}(B^n)$.
    (ii) Prove that $R \perp\!\!\!\perp \mathbf{z}$.
    (iii) Determine $E\ \mathbf{z}$ and $\text{var}\ \mathbf{z}$.
    (iv) Determine $E\ \mathbf{x}$ and $\text{var}\ \mathbf{x}$ in terms of $E\ R^2$.

    Partition $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$, $\mathbf{x}_1 \in \mathbb{R}^k$ and $\mathbf{x}_2 \in \mathbb{R}^{n-k}$ and let $R_i = |\mathbf{x}_i|$, $i = 1, 2$.
    (v) Determine the distribution of R_1^2/R^2 and R_1^2/R_2^2.

11. Assume $\mathbf{u} \sim \text{unif}(S^{n-1})$ and $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2')'$, $\mathbf{u}_1 \in \mathbb{R}^k$.
    (i) Prove that the density of $\mathbf{u}_1$ is

$$f(\mathbf{u}_1) = \frac{\Gamma(\frac{1}{2}n)}{\pi^{k/2}\Gamma[\frac{1}{2}(n-k)]}(1 - \mathbf{u}_1'\mathbf{u}_1)^{(n-k)/2-1},\ 0 < \mathbf{u}_1'\mathbf{u}_1 < 1.$$

Hint: Show $(u_1^2, \ldots, u_k^2) \sim D_k(\frac{1}{2}\mathbf{1}; \frac{1}{2}(n-k))$ and consider the one-to-many transformation $u_i^2 \mapsto \pm u_i$.

(ii) Prove $|\mathbf{u}_1|^2 \sim \text{beta}(\frac{1}{2}k; \frac{1}{2}(n-k))$.

12. Assume $\mathbf{u} = (u_1, u_2, u_3)' \sim \text{unif}(S^2)$. Show that $u_1 \sim \text{unif}(-1, 1)$. Does this hold in other dimensions?

13. Let $\mathbf{x} = (x_1, \ldots, x_n)'$ have a spherical density $f_{\mathbf{x}}(\mathbf{x}) = g(|\mathbf{x}|^2)$ for some function $g : [0, \infty) \to [0, \infty)$. Let $\mathbf{x} = r\mathbf{u}$, where $r \geq 0$ denotes "radius" and $\mathbf{u} \in S^{n-1}$ represents "direction." Prove the following using $J(\mathbf{x} \to r, \mathbf{u}) = r^{n-1}$:

(i) $r \perp\!\!\!\perp \mathbf{u}$.

(ii) r^2 has density

$$f_{r^2}(s) = \tfrac{1}{2}\omega_n s^{n/2-1} g(s), \ s > 0,$$

where ω_n is the "area" of the unit sphere S^{n-1}.

(iii) With the special case $x_1, \ldots, x_n$ i.i.d. $N(0, 1)$, find the "area" ω_n.

(iv) What is the density of $\mathbf{u}$?

14. Let $\mathbf{x} \in \mathbb{R}^n$ have a spherical density $f_{\mathbf{x}}(\mathbf{x}) = g(|\mathbf{x}|^2)$ and

$$\mathbf{x} \mapsto r, \theta_1, \ldots, \theta_{n-1}$$

be the transformation to polar coordinates as in Proposition 2.23. Prove $\theta_{n-1} \sim \text{unif}(0, 2\pi)$. What can be said about the other angles?

15. Prove the following concerning spherical distributions:

(i) If $g(|\mathbf{x}|^2)$ is a density on $\mathbb{R}^n$ for some $g : [0, \infty) \to [0, \infty)$, then $\int_0^\infty r^{n-1} g(r^2) dr = \Gamma\left(\frac{1}{2}n\right) / (2\pi^{n/2})$.

(ii) If the kth moment of $\mathbf{x}$ is finite, i.e., $E\,|\mathbf{x}|^k < \infty$, then

$$\int_0^\infty r^{n+k-1} g(r^2) dr < \infty.$$

(iii) If the second moment of $\mathbf{x}$ is finite, then $\text{var}\,\mathbf{x} = \alpha\mathbf{I}$, where $\alpha = E\,x_1^2$. From Problem 4.6.6, $c_{\mathbf{x}}(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t})$ for some function ϕ. Prove $\alpha = -2\phi'(0)$.

5
Multivariate normal

5.1 Introduction

This chapter is entirely devoted to the multivariate normal distribution. In Section 5.2, the basic properties are demonstrated. Then, Sections 5.3 and 5.4 make the distinction between the nonsingular and the singular cases. In the nonsingular case, the density is derived while we explain the geometry of the singular case. Section 5.5 contains the conditional distribution in all its generality. Finally, the last section reaps the first benefits by considering some applications in univariate sampling, regression, and elementary correlation.

5.2 Definition and elementary properties

Let $\mathbf{\Sigma} = (\sigma_{ij}) \in \mathbb{R}_n^n$ be symmetric, positive semidefinite, and $\boldsymbol{\mu} \in \mathbb{R}^n$.

Definition 5.1 Multivariate normal*:*

$$\mathbf{x} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma}) \textit{ iff } \mathbf{t}'\mathbf{x} \sim N(\mathbf{t}'\boldsymbol{\mu}, \mathbf{t}'\mathbf{\Sigma}\mathbf{t}),\ \forall \mathbf{t} \in \mathbb{R}^n.$$

Note that $\mathbf{x}$ has product-moments of any order by the fact that this is true of $\mathbf{t}'\mathbf{x}$, $\forall \mathbf{t} \in \mathbb{R}^n$.

Proposition 5.1 $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma}) \Longrightarrow E\ \mathbf{x} = \boldsymbol{\mu}$ *and* var $\mathbf{x} = \mathbf{\Sigma}$.

Proof. Setting $\mathbf{t} = \mathbf{e}_i = (0, \ldots, 1, \ldots, 0)'$, we find the individual component $x_i = \mathbf{e}_i'\mathbf{x}$:

$$x_i \sim N(\mu_i, \sigma_{ii}), \text{ where } \mu_i = E\, x_i,\ \sigma_{ii} = \text{var}\, x_i.$$

Similarly, setting $\mathbf{t}$ to be a vector with 1's in the ith and jth components, $i \neq j$, and 0's elsewhere, we find

$$x_i + x_j \sim N(\mu_i + \mu_j, \sigma_{ii} + \sigma_{jj} + 2\sigma_{ij}),\ i \neq j.$$

However, since on the other hand for $i \neq j$, var $(x_i + x_j) =$ var $x_i +$ var $x_j +$ $2\text{cov}(x_i, x_j)$, then $\text{cov}(x_i, x_j) = \sigma_{ij}$. □

Proposition 5.2 *Let* $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$*, linear, and* $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. Then,* $\mathbf{Ax} \sim N_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$*.*

Proof. Let $\mathbf{y} = \mathbf{Ax}$ and merely note that

$$\mathbf{s}'\mathbf{y} = (\mathbf{A}'\mathbf{s})'\mathbf{x} \sim N\left(\mathbf{s}'\mathbf{A}\boldsymbol{\mu}, \mathbf{s}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{s}\right),\ \forall \mathbf{s} \in \mathbb{R}^m.$$

□

By specializing $\mathbf{A}$ to be the projection onto any particular subset of coordinates, we deduce immediately that all the marginal distributions are normal.

As a simple corollary on rotational invariance, we have

$$\mathbf{z} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}) \Longrightarrow \mathbf{Hz} \stackrel{d}{=} \mathbf{z},\ \forall \mathbf{H} \in \mathbf{O}_n.$$

The characteristic function for $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ derives from the univariate level (4.1):

$$c_{\mathbf{x}}(\mathbf{t}) = c_{\mathbf{t}'\mathbf{x}}(1) = \exp\left(-\tfrac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} + i\mathbf{t}'\boldsymbol{\mu}\right).$$

Example 5.1 *Although all marginals of* $\mathbf{x}$ *have a univariate normal distribution, the vector* $\mathbf{x}$ *itself may not have a multivariate normal distribution. Consider a random vector* $\mathbf{x}$ *whose distribution is a mixture of two multivariate normal distributions,*

$$c_{\mathbf{x}}(\mathbf{t}) = \alpha c_{\mathbf{x}_1}(\mathbf{t}) + (1-\alpha)c_{\mathbf{x}_2}(\mathbf{t}),\ 0 < \alpha < 1,$$

where

$$\begin{aligned} \mathbf{x}_1 &\sim N_n\left(\mathbf{0}, (1-\rho_1)\mathbf{I} + \rho_1\mathbf{1}\mathbf{1}'\right), \\ \mathbf{x}_2 &\sim N_n\left(\mathbf{0}, (1-\rho_2)\mathbf{I} + \rho_2\mathbf{1}\mathbf{1}'\right). \end{aligned}$$

Then, $c_{x_i}(t_i) = \alpha c_z(t_i) + (1-\alpha)c_z(t_i) = c_z(t_i)$*, where* $z \sim N(0,1)$*, which shows* $x_i \stackrel{d}{=} z$*,* $i = 1, \ldots, n$*, but* $\mathbf{x}$ *does not have a multivariate normal distribution. Other counterexamples can be given using copulas [v. Example* 2.5*].*

As a special case, we have the characteristic function for $\mathbf{z} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$:

$$c_{\mathbf{z}}(\mathbf{t}) = \exp\left(-\tfrac{1}{2}\sigma^2\mathbf{t}'\mathbf{t}\right) = \prod_{i=1}^{n} \exp(-\tfrac{1}{2}t_i^2\sigma^2) = \prod_{i=1}^{n} c_{z_i}(t_i)$$

by which

$$\mathbf{z} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}) \Longleftrightarrow z_1, \ldots, z_n \text{ i.i.d. } N(0, \sigma^2).$$

An implication is that if $\mathbf{z} = (\mathbf{z}_1', \mathbf{z}_2')' \sim N_n(\mathbf{0}, \mathbf{I})$, then $\mathbf{z}_1 \perp\!\!\!\perp \mathbf{z}_2$, and the density for $\mathbf{z}$ becomes

$$f_{\mathbf{z}}(\mathbf{z}) = (2\pi)^{-n/2} \exp\left(-\tfrac{1}{2}\mathbf{z}'\mathbf{z}\right) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(-\tfrac{1}{2}z_i^2\right).$$

It is also clear from the characteristic function that the family of multivariate normal is closed under translation:

$$\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Longrightarrow \mathbf{x} + \mathbf{b} \sim N_n(\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}),\ \forall \mathbf{b} \in \mathbb{R}^n.$$

Now, suppose that $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and write $\boldsymbol{\Sigma} = \mathbf{H}\mathbf{D}\mathbf{H}'$ with $\mathbf{H}$ orthogonal and $\mathbf{D} = \text{diag}(\boldsymbol{\lambda})$. We find, of course, that $\mathbf{y} = \mathbf{H}'(\mathbf{x} - \boldsymbol{\mu}) \sim N_n(\mathbf{0}, \mathbf{D})$, and if we then let $\mathbf{A} = \mathbf{H}\mathbf{D}^{1/2}$, we deduce the representation:

Proposition 5.3 $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Longleftrightarrow \mathbf{x} \stackrel{d}{=} \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ *for any* $\mathbf{A}$ *such that* $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$, $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I})$.

Finally, partition $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$, where $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{n_2}$, $n = n_1 + n_2$, with corresponding

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Concerning independence, we have the following necessary and sufficient condition.

Proposition 5.4 *Let* $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')' \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then,*

$$\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2 \Longleftrightarrow \boldsymbol{\Sigma}_{12} = \mathbf{0}.$$

Proof. ($\Longrightarrow$):

$$\begin{aligned} \mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2 &\Longrightarrow E\ g_1(\mathbf{x}_1)g_2(\mathbf{x}_2) = E\ g_1(\mathbf{x}_1)E\ g_2(\mathbf{x}_2),\ \forall g_1, g_2 \\ &\Longrightarrow \boldsymbol{\Sigma}_{12} = \mathbf{0}. \end{aligned}$$

($\Longleftarrow$): Assume $\boldsymbol{\Sigma}_{12} = \mathbf{0}$. Write $\boldsymbol{\Sigma}_{ii} = \mathbf{A}_{ii}\mathbf{A}_{ii}'$, $i = 1, 2$. Then, $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$, where

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix}.$$

Using the representation

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \stackrel{d}{=} \mathbf{A}\mathbf{z} + \boldsymbol{\mu} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{z}_1 + \boldsymbol{\mu}_1 \\ \mathbf{A}_{22}\mathbf{z}_2 + \boldsymbol{\mu}_2 \end{pmatrix},$$

where $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I})$, it is clear that since $\mathbf{z}_1 \perp\!\!\!\perp \mathbf{z}_2$, then $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2$. □

Another simple proof based on characteristic functions is proposed in Problem 5.7.3.

5.3 Nonsingular normal

When $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $|\boldsymbol{\Sigma}| = |\mathbf{A}\mathbf{A}'| = |\mathbf{A}|^2 = |\mathbf{D}| = \prod_{i=1}^n \lambda_i > 0$, define $\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ whereby we have an explicit density for $\mathbf{x}$ by simple change of variables:

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= f_{\mathbf{z}}\left(\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \cdot J(\mathbf{z} \to \mathbf{x}) \\ &= (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \end{aligned}$$

and, of course, from Proposition 4.4, then also

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{z}'\mathbf{z} \sim \chi^2_n.$$

The quantity $[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]^{1/2}$ is often called the *Mahalanobis distance* of $\mathbf{x}$ to $\boldsymbol{\mu}$.

Example 5.2 *The bivariate density function is just a special case. For*

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

we find

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} \sigma_1^{-2} & -\rho/\sigma_1\sigma_2 \\ -\rho/\sigma_1\sigma_2 & \sigma_2^{-2} \end{pmatrix}.$$

Thus, the bivariate density takes the form

$$\begin{aligned} f_{\mathbf{x}}(x_1, x_2) = \frac{1}{2\pi} \frac{1}{\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 \right.\right. \\ \left.\left. -2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right\}. \end{aligned}$$

A plot of this density is given in Figure 5.1.

The *contours*, which consists of the set of points of equal probability density, of a multivariate normal are the points $\mathbf{x}$ of equal Mahalanobis distance to $\boldsymbol{\mu}$,

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2,$$

for any constant $c > 0$. Letting $\mathbf{y} = \mathbf{H}'\mathbf{x}$, $\boldsymbol{\nu} = \mathbf{H}'\boldsymbol{\mu}$, where $\mathbf{H}$ diagonalizes $\boldsymbol{\Sigma}$, $\mathbf{H}'\boldsymbol{\Sigma}\mathbf{H} = \mathbf{D}$, then the contours are the ellipsoids

$$\sum_{i=1}^p (y_i - \nu_i)^2/d_i = c^2$$

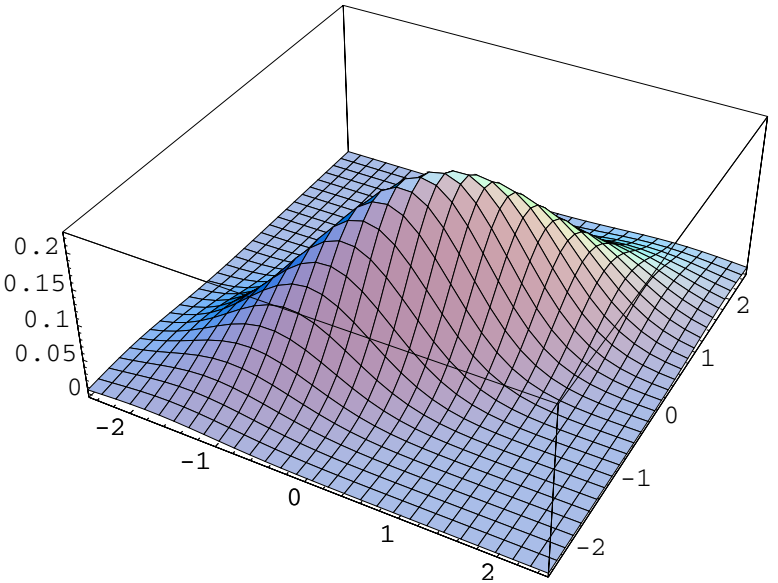

Figure 5.1. Bivariate normal density for values of the parameters $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.7$.

centered at $\boldsymbol{\nu}$ with principal axes of half length $cd_i^{1/2}$ supported by the eigenvectors in $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_p)$.

Example 5.3 *The contours of the bivariate normal density are in parametric form, and in the y coordinates,*

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} + c \begin{pmatrix} d_1^{1/2} \sin\theta \\ d_2^{1/2} \cos\theta \end{pmatrix}, \ 0 \le \theta \le 2\pi.$$

Thus, the contours in the original x coordinates are just

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + c \begin{pmatrix} h_{11} d_1^{1/2} \sin\theta + h_{12} d_2^{1/2} \cos\theta \\ h_{21} d_1^{1/2} \sin\theta + h_{22} d_2^{1/2} \cos\theta \end{pmatrix}, \ 0 \le \theta \le 2\pi.$$

A contour plot is given in Figure 5.2.

Example 5.4 *Using the transformation to polar coordinates on p.* 32*, the contours of the trivariate normal density are in parametric form, and in the y coordinates,*

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} + c \begin{pmatrix} d_1^{1/2} \sin\theta_1 \sin\theta_2 \\ d_2^{1/2} \sin\theta_1 \cos\theta_2 \\ d_3^{1/2} \cos\theta_1 \end{pmatrix}, \ 0 \le \theta_1 \le \pi, \ 0 \le \theta_2 \le 2\pi.$$

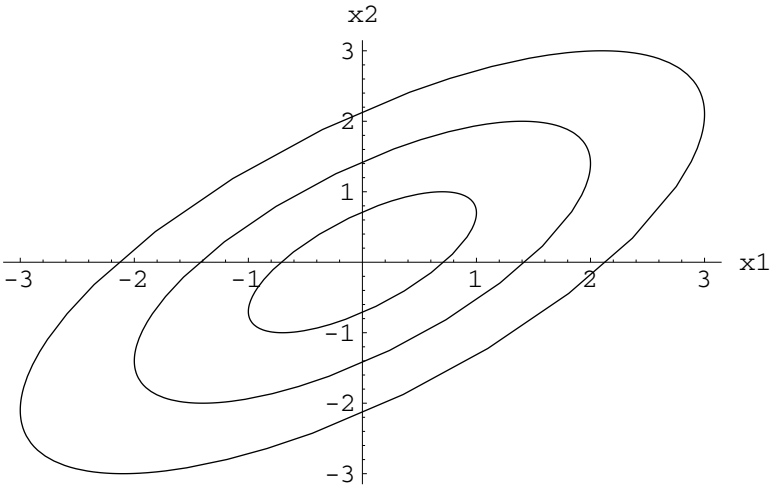

Figure 5.2. Contours of the bivariate normal density for values of the parameters $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.7$. Values of $c = 1,\ 2,\ 3$ were taken.

Thus, the contours in the original x coordinates are just

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + c \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} d_1^{1/2} \sin\theta_1 \sin\theta_2 \\ d_2^{1/2} \sin\theta_1 \cos\theta_2 \\ d_3^{1/2} \cos\theta_1 \end{pmatrix},$$

$0 \le \theta_1 \le \pi$, $0 \le \theta_2 \le 2\pi$. *The contour plot corresponding to $c = 1$ is given in Figure* 5.3 *when $\boldsymbol{\mu} = \mathbf{0}$ and*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{pmatrix}.$$

The corresponding eigenvalues of $d_1 = 18$, $d_2 = d_3 = 9$ give the typical ellipsoidal contours.

Still assuming $|\boldsymbol{\Sigma}| > 0$, we apply the Gram-Schmidt process to the basis formed by the row vectors of $\mathbf{A} = \mathbf{H}\mathbf{D}^{1/2}$, obtaining (uniquely) $\mathbf{A} = \mathbf{T}\mathbf{G}$ (v. Proposition 1.13) with $\mathbf{T} \in \mathbf{L}_n^+$, $\mathbf{G} \in \mathbf{O}_n$, where

$$\mathbf{L}_n^+ = \{\mathbf{T} \in \mathbf{G}_n :\ \mathbf{T} \text{ is lower triangular},\ t_{ii} > 0,\ i = 1, \ldots, n\}.$$

Then, $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}' = \mathbf{T}\mathbf{T}'$ for a unique $\mathbf{T} \in \mathbf{L}_n^+$ (v. Proposition 1.14). We have the "triangular" representation:

Proposition 5.5 (Triangular representation)

$$\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Longleftrightarrow \mathbf{x} \stackrel{d}{=} \mathbf{T}\mathbf{z} + \boldsymbol{\mu}$$

with $\mathbf{T} \in \mathbf{L}_n^+$ such that $\boldsymbol{\Sigma} = \mathbf{T}\mathbf{T}'$ and $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I})$.

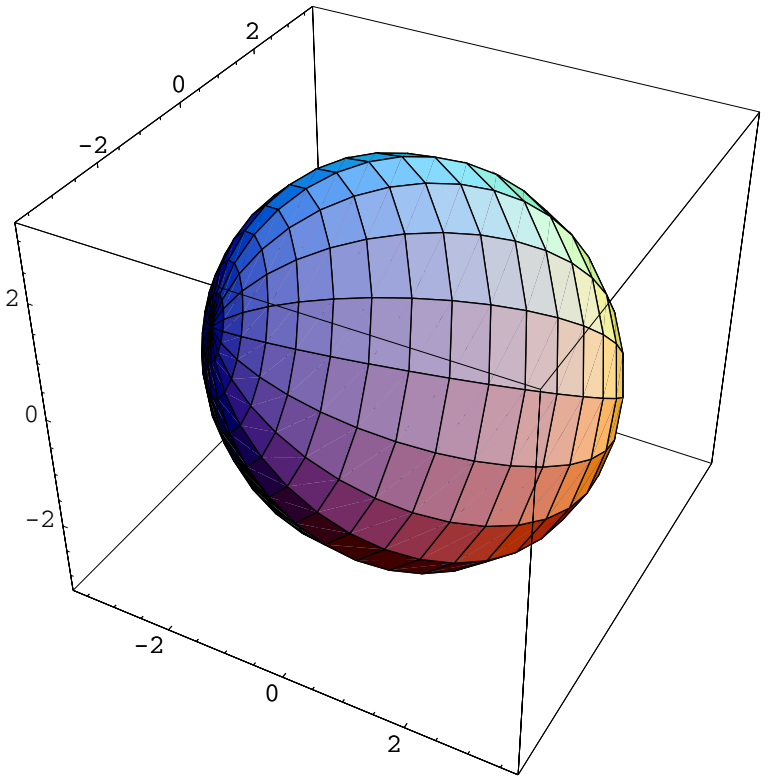

Figure 5.3. A contour of a trivariate normal density.

5.4 Singular normal

Now, for $x \sim N(\mu, \sigma^2)$, we know that $\sigma^2 = 0 \iff x = \mu$ w.p.1. This holds, since if $\sigma^2 = 0$, $P(x = \mu) = \lim_{n\to\infty} P(|x - \mu| < 1/n)$, but $P(|x - \mu| \geq 1/n) \leq \sigma^2 n^2 = 0$, $\forall n$. Thus, the normal family includes the "trivial" (constant) random variables as special cases. By Cramér-Wold Proposition 2.10, this also holds for random vectors $\mathbf{x} \in \mathbb{R}^n$ with $E\,\mathbf{x} = \boldsymbol{\mu}$ and $\operatorname{var}\mathbf{x} = \boldsymbol{\Sigma}$: $\boldsymbol{\Sigma} = \mathbf{0} \iff \mathbf{x} = \boldsymbol{\mu}$ w.p.1. However, if $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| = 0$, we may write

$$\boldsymbol{\Sigma} = \mathbf{H}\mathbf{D}\mathbf{H}' = (\mathbf{H}_1, \mathbf{H}_2) \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{H}_1' \\ \mathbf{H}_2' \end{pmatrix} = \mathbf{H}_1\mathbf{D}_1\mathbf{H}_1',$$

where $\mathbf{D}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ comprises the nonzero eigenvalues, $\mathbf{H}_1 = (\mathbf{h}_1, \ldots, \mathbf{h}_r)$ gives a basis for the column space of $\boldsymbol{\Sigma}$, $\operatorname{Im}\boldsymbol{\Sigma}$, and $\mathbf{H}_2 = (\mathbf{h}_{r+1}, \ldots, \mathbf{h}_n)$ gives a basis for the kernel, $\ker\boldsymbol{\Sigma}$. One should note that

$$\begin{aligned} \operatorname{Im}\mathbf{H}_2 &= (\operatorname{Im}\mathbf{H}_1)^\perp = \ker\boldsymbol{\Sigma}, \\ \operatorname{Im}\mathbf{H}_1 &= (\operatorname{Im}\mathbf{H}_2)^\perp = \operatorname{Im}\boldsymbol{\Sigma}. \end{aligned}$$

Then it is clear that $\mathbf{H}_2'\boldsymbol{\Sigma}\mathbf{H}_2 = \mathbf{0}$ and, thus, we find that $\mathbf{H}_2'(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$ w.p.1 or, equivalently, $\mathbf{x} - \boldsymbol{\mu} \in (\operatorname{Im}\mathbf{H}_2)^\perp$ w.p.1., whereas $\mathbf{H}_1'(\mathbf{x} - \boldsymbol{\mu}) \sim N_r(\mathbf{0}, \mathbf{D}_1)$ has a nonsingular normal distribution. Of course, this is yet equivalent to saying that $\mathbf{x} \in \boldsymbol{\mu} + \operatorname{Im}\boldsymbol{\Sigma}$ w.p.1 and one can then almost visualize $\mathbf{x}$ in this r-dimensional affine subspace of $\mathbb{R}^n$, $r < n$. A curious fact in this case is that $\operatorname{vol}(\boldsymbol{\mu} + \operatorname{Im}\boldsymbol{\Sigma}) = 0$ but $P_{\mathbf{x}}(\boldsymbol{\mu} + \operatorname{Im}\boldsymbol{\Sigma}) = 1$, therefore $\mathbf{x}$ cannot be absolutely continuous (v. Proposition 2.11).

It is worth recalling at this point that any constant random vector is automatically statistically independent of any other random vector, and so we might notice, in particular, the rather odd looking fact that

$$\mathbf{H}_2'(\mathbf{x} - \boldsymbol{\mu}) \perp\!\!\!\perp \mathbf{x}.$$

5.5 Conditional normal

By a suitable permutation, one may rearrange an arbitrary multivariate normal $\mathbf{x}$ so that any subset $\mathbf{x}_1$ of its coordinates are brought to the fore, and the overall distribution is expressed by

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_n\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right), \ \mathbf{x}_1 \in \mathbb{R}^{n_1}, \ \mathbf{x}_2 \in \mathbb{R}^{n_2}, \ n = n_1 + n_2.$$

We derive the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$.

First suppose $\boldsymbol{\Sigma}_{22}$ is nonsingular and note that for any $\mathbf{B}$,

$$\begin{pmatrix} \mathbf{I} & -\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B}' & \mathbf{I} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\mathbf{B}' - \mathbf{B}\boldsymbol{\Sigma}_{21} + \mathbf{B}\boldsymbol{\Sigma}_{22}\mathbf{B}' & \boldsymbol{\Sigma}_{12} - \mathbf{B}\boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{22}\mathbf{B}' & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

so by deliberately setting $\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, we find

$$\begin{pmatrix} \mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \\ \mathbf{x}_2 \end{pmatrix} \sim N_n \left(\begin{pmatrix} \boldsymbol{\mu}_1 - \mathbf{B}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11.2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right),$$

where

$$\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

However, independence means $\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \mid \mathbf{x}_2 \stackrel{d}{=} \mathbf{x}_1 - \mathbf{B}\mathbf{x}_2$, so that we have

$$\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \mid \mathbf{x}_2 \sim N_{n_1}(\boldsymbol{\mu}_1 - \mathbf{B}\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11.2}).$$

Since we may legitimately treat $\mathbf{x}_2$ as though constant (the full justification of this depending on the fact that we have a "regular" conditional distribution to which the Fubini theorem applies [Ash (1972)]) we may conclude that

$$\mathbf{x}_1 \mid \mathbf{x}_2 \sim N_{n_1}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11.2}).$$

In the singular case, if $|\boldsymbol{\Sigma}_{22}| = 0$, we may always write

$$\boldsymbol{\Sigma}_{22} = (\mathbf{H}_1, \mathbf{H}_2) \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{H}_1' \\ \mathbf{H}_2' \end{pmatrix} = \mathbf{H}_1 \mathbf{D} \mathbf{H}_1',$$

where $\mathbf{D} \in \mathbb{R}_k^k$ is nonsingular, and we may then take what is called a "pseudo-inverse" for $\boldsymbol{\Sigma}_{22}$:

$$\boldsymbol{\Sigma}_{22}^- = (\mathbf{H}_1, \mathbf{H}_2) \begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{H}_1' \\ \mathbf{H}_2' \end{pmatrix} = \mathbf{H}_1 \mathbf{D}^{-1} \mathbf{H}_1'.$$

We then have, of course, $\mathbf{H}_2'(\mathbf{x}_2 - \boldsymbol{\mu}_2) \stackrel{\text{w.p.1}}{=} \mathbf{0}$ and also

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{H}_1'\mathbf{x}_2 \end{pmatrix} \sim N_{n_1+k} \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \mathbf{H}_1'\boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\mathbf{H}_1 \\ \mathbf{H}_1'\boldsymbol{\Sigma}_{21} & \mathbf{D} \end{pmatrix} \right)$$

to which the results in the nonsingular case apply, immediately showing

$$\begin{pmatrix} \mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \\ \mathbf{H}_1'\mathbf{x}_2 \end{pmatrix} \sim N_{n_1+k} \left(\begin{pmatrix} \boldsymbol{\mu}_1 - \mathbf{B}\boldsymbol{\mu}_2 \\ \mathbf{H}_1'\boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11.2} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \right),$$

where $\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-$ and $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-\boldsymbol{\Sigma}_{21}$. But from this, $\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \perp\!\!\!\perp \mathbf{H}_1'\mathbf{x}_2$, and thus, overall, $\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2 \perp\!\!\!\perp \mathbf{x}_2$. We arrive at the completely general conclusion:

Proposition 5.6 $\mathbf{x}_1 \mid \mathbf{x}_2 \sim N_{n_1}(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2})$, *where*

$$\begin{aligned} \boldsymbol{\mu}_{1.2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-(\mathbf{x}_2 - \boldsymbol{\mu}_2), \\ \boldsymbol{\Sigma}_{11.2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-\boldsymbol{\Sigma}_{21}. \end{aligned}$$

5.6 Elementary applications

5.6.1 Sampling the univariate normal

Observe that

$$x_1, \ldots, x_n \text{ i.i.d. } N(\mu, \sigma^2) \Longleftrightarrow \mathbf{x} \sim N_n(\mu \mathbf{1}, \sigma^2 \mathbf{I}).$$

Letting $\mathbf{H} = (\, \mathbf{1}/\sqrt{n}, \quad \mathbf{\Gamma}\,) \in \mathbf{O}_n$ for some $\mathbf{\Gamma}$ and

$$\mathbf{w} = \mathbf{H}'\mathbf{x} \sim N_n(\sqrt{n}\mu \mathbf{e}_1, \sigma^2 \mathbf{I}),$$

obviously $w_1, \ldots, w_n$ are independent, and, of course,

$$w_1 = \sqrt{n}\bar{x}$$

and

$$w_2^2 + \cdots + w_n^2 = |\mathbf{w}|^2 - w_1^2 = |\mathbf{x}|^2 - n\bar{x}^2 = |\mathbf{x} - \bar{x}\mathbf{1}|^2 = (n-1)s_x^2,$$

where $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$ is the sample variance. Thus, we have the basic statistical result

$$\bar{x} \sim N(\mu, \sigma^2/n),\ (n-1)s_x^2 \stackrel{d}{=} \sigma^2 \chi_{n-1}^2, \text{ and } \bar{x} \perp\!\!\!\perp s_x^2$$

with its trivial algebraic corollary

$$\sqrt{n}(\bar{x} - \mu)/s_x \stackrel{d}{=} \sqrt{(n-1)}z/\chi_{n-1},\ z \sim N(0,1), \text{ and } z \perp\!\!\!\perp \chi_{n-1}^2.$$

We make the following definition (W.S. Gosset, "Student," 1908):

Definition 5.2 t-Distribution*: $t \stackrel{d}{=} t_p$ iff $t \stackrel{d}{=} \sqrt{p}z/\chi_p$, where $z \sim N(0,1)$ and $z \perp\!\!\!\perp \chi_p^2$.*

Thus, by definition, $\sqrt{n}(\bar{x}-\mu)/s_x \stackrel{d}{=} t_{n-1}$ is a pivotal quantity for μ. Clearly, $t \stackrel{d}{=} t_p \Longleftrightarrow t \stackrel{d}{=} -t$ and $t^2 \sim F(1,p)$. This provides a quick way of obtaining the integral moments of t_p. The Student's t-distribution sometimes plays a role in the dependent case. The intraclass correlation model is one such example.

Example 5.5 *Assume $\mathbf{x} \sim N_n\left(\mu\mathbf{1}, \sigma^2[(1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}']\right)$, where $-1/(n-1) \le \rho \le 1$. Let $\bar{x} = \sum_{i=1}^n x_i/n$, $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$, and $t = \sqrt{n}(\bar{x} - \mu)/s_x$. We determine a constant c such that $ct \sim t_{n-1}$. With the orthogonal transformation above, we still have*

$$\begin{aligned} \mathbf{w} &= \mathbf{H}'\mathbf{x}, \\ w_1 &= \sqrt{n}\bar{x}, \\ w_2^2 + \cdots + w_n^2 &= (n-1)s_x^2. \end{aligned}$$

Since $\mathbf{H}'\mathbf{1} = (\sqrt{n}, \mathbf{0}')'$, the distribution of $\mathbf{w}$ is

$$\mathbf{w} \sim N_n\left(\begin{pmatrix} \sqrt{n}\mu \\ \mathbf{0} \end{pmatrix}, \sigma^2\left[(1-\rho)\mathbf{I} + \rho\ \mathrm{diag}(n, 0, \ldots, 0)\right]\right).$$

Hence, $w_1 \perp\!\!\!\perp (w_2, \ldots, w_n)'$*, which implies* $\bar{x} \perp\!\!\!\perp s_x^2$*. The distribution of* $\bar{x}$ *and* s_x^2 *are given by*

$$\begin{aligned} \sqrt{n}\bar{x} &\sim N(\sqrt{n}\mu, \sigma^2[(1-\rho)+\rho n]), \\ (n-1)s_x^2 &\sim \sigma^2(1-\rho)\chi^2_{n-1}. \end{aligned}$$

Finally, we can conclude that $ct \sim t_{n-1}$ *by defining*

$$c = \left[\frac{1-\rho}{(1-\rho)+\rho n}\right]^{1/2}.$$

In fact, the Student's t-distribution has nothing to do with normal distributions. It is more related to the concept of spherical symmetry, as in the next example [Efron (1969)].

Example 5.6 *Assume* $\mathbf{x} \in \mathbb{R}^n$ *has a "rotationally invariant" distribution and* $P(\mathbf{x} = \mathbf{0}) = 0$*. We establish that* $\sqrt{n}\bar{x}/s_x \sim t_{n-1}$*, where, as usual,* $\bar{x} = \sum_{i=1}^n x_i/n$ *and* $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$*. Using Proposition* 4.10*, the representation* $\mathbf{x} \stackrel{d}{=} R\mathbf{u}$*, where* $\mathbf{u} \sim \text{unif}(S^{n-1})$ *and* $R \perp\!\!\!\perp \mathbf{u}$*, is valid. Hence,* $(\bar{x}, s_x) \stackrel{d}{=} (R\bar{u}, Rs_u)$ *and the distribution of*

$$\sqrt{n}\frac{\bar{x}}{s_x} \stackrel{d}{=} \sqrt{n}\frac{R\bar{u}}{Rs_u} = \sqrt{n}\frac{\bar{u}}{s_u}$$

does not depend on R. Thus, $\sqrt{n}\bar{x}/s_x \sim t_{n-1}$ *since this is the case when* $\mathbf{x} \sim N_n(\mathbf{0}, \mathbf{I})$*.*

5.6.2 Linear estimation

Consider now the problem of linear estimation in the so-called multiple regression model. Let $\mathcal{V} \subset \mathbb{R}^n$ be any k-dimensional vector subspace and

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}, \; E\,\mathbf{e} = \mathbf{0}, \; \text{var}\,\mathbf{e} = \sigma^2\mathbf{I}, \; \text{and} \; \boldsymbol{\mu} \in \mathcal{V}.$$

Let $\boldsymbol{\theta} = \mathbf{T}\boldsymbol{\mu}$, where $\mathbf{T} \in \mathbb{R}_n^m$, and consider the estimate $\hat{\boldsymbol{\theta}} = \mathbf{T}\hat{\boldsymbol{\mu}}$, where $\hat{\boldsymbol{\mu}} = \mathbf{P}\mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ on $\mathcal{V}$ (v. Section 1.6). We prove that among all possible unbiased linear estimates of $\boldsymbol{\theta}$, the regression estimate $\hat{\boldsymbol{\theta}}$ has the minimum variance. In this sense, $\hat{\boldsymbol{\theta}}$ is the "best" linear unbiased estimate (*blue*).

Proposition 5.7 (Gauss-Markov) $\hat{\boldsymbol{\theta}} = \text{blue}(\boldsymbol{\theta})$*.*

Proof. $\tilde{\boldsymbol{\theta}} = \mathbf{B}\mathbf{y}$ is unbiased for $\boldsymbol{\theta} \iff \mathbf{BP} = \mathbf{T}\mathbf{P}$. But then,

$$\text{var}\,\hat{\boldsymbol{\theta}} = \sigma^2\mathbf{T}\mathbf{P}\mathbf{T}' = \sigma^2\mathbf{B}\mathbf{P}\mathbf{B}' \leq \sigma^2\mathbf{B}\mathbf{B}' = \text{var}\,\tilde{\boldsymbol{\theta}}$$

with equality iff

$$\mathbf{B}\mathbf{Q}\mathbf{B}' = \mathbf{0} \iff \mathbf{B}\mathbf{Q} = \mathbf{0} \iff \mathbf{B} = \mathbf{B}\mathbf{P} \iff \tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}},$$

where $\mathbf{Q} = \mathbf{I} - \mathbf{P}$. □

For example, $\hat{\boldsymbol{\mu}} = \mathbf{P}\mathbf{y} = \text{blue}(\boldsymbol{\mu})$ with var $\hat{\boldsymbol{\mu}} = \sigma^2\mathbf{P}$.

Now, expressing $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ with respect to any basis $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ for $\mathcal{V}$ and recalling the representation (1.3) for $\mathbf{P}$, the coefficients are uniquely determined as $\boldsymbol{\beta} = \mathbf{B}_0\boldsymbol{\mu}$, where $\mathbf{B}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. But then, $\hat{\boldsymbol{\beta}} = \mathbf{B}_0\hat{\boldsymbol{\mu}} = \mathbf{B}_0\mathbf{y} = \text{blue}(\boldsymbol{\beta})$, where var $\hat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Obviously, $\hat{\boldsymbol{\beta}}_i = \text{blue}(\boldsymbol{\beta}_i)$, $i = 1, \ldots, k$.

Another optimality property of the "Gauss-Markov" estimate was recently discovered [Berk and Hwang (1989), Eaton (1988), Ali and Ponnapalli (1990)]: The probability of the Gauss-Markov estimate of $\boldsymbol{\theta}$ falling inside any fixed ellipsoid centered at $\boldsymbol{\theta}$ is greater than or equal to the probability that any linear unbiased estimate of $\boldsymbol{\theta}$ falls inside the same ellipsoid. It is interesting to remark that the Gauss-Markov estimate $\hat{\boldsymbol{\mu}} = \mathbf{P}\mathbf{y}$ is also the least-squares estimate. This follows from a general property of orthogonal projections:

Proposition 5.8

$$\min_{\boldsymbol{\mu}\in\mathcal{V}} |\mathbf{y} - \boldsymbol{\mu}|^2 = |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2,$$

where $\hat{\boldsymbol{\mu}} = \mathbf{P}\mathbf{y}$ *is the orthogonal projection of* $\mathbf{y}$ *on* $\mathcal{V}$.

Proof. For all $\boldsymbol{\mu} \in \mathcal{V}$,

$$\begin{aligned} |\mathbf{y} - \boldsymbol{\mu}|^2 &= |(\mathbf{y} - \hat{\boldsymbol{\mu}}) + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})|^2 \\ &= |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2 + |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|^2 + 2(\mathbf{y} - \hat{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\ &= |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2 + |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|^2 \end{aligned}$$

since $\mathbf{y} - \hat{\boldsymbol{\mu}} \in \mathcal{V}^\perp$ and $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \in \mathcal{V}$. Hence,

$$|\mathbf{y} - \boldsymbol{\mu}|^2 \geq |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2, \ \forall \boldsymbol{\mu} \in \mathcal{V},$$

with equality if $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$. □

Finally, since $\mathbf{Q} = \mathbf{I} - \mathbf{P}$ gives the orthogonal projection on $\mathcal{V}^\perp$, we find

$$\mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{Q}\mathbf{y} = \mathbf{Q}\mathbf{e} \Longrightarrow |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2 = \mathbf{e}'\mathbf{Q}\mathbf{e},$$

so that

$$E\ |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2 = E\ \mathbf{e}'\mathbf{Q}\mathbf{e} = E\ \text{tr}\ \mathbf{Q}\mathbf{e}\mathbf{e}' = \text{tr}\ \mathbf{Q}E\ \mathbf{e}\mathbf{e}' = (n-k)\sigma^2.$$

Thus, we determine the unbiased estimate $\hat{s}^2$ of σ^2 by

$$(n-k)\hat{s}^2 = |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2.$$

It is also clear that $\text{cov}(\hat{\boldsymbol{\mu}}, \mathbf{y} - \hat{\boldsymbol{\mu}}) = \text{cov}(\mathbf{P}\mathbf{y}, \mathbf{Q}\mathbf{y}) = \sigma^2\mathbf{P}\mathbf{Q} = \mathbf{0}$. Before stating the joint distribution under normality of our estimates $\hat{\boldsymbol{\mu}}$ and $\hat{s}^2$, we prove the following lemma on quadratic forms.

Lemma 5.1 *Let* $\mathbf{z} \sim N_n(\boldsymbol{\mu}, \mathbf{I})$ *and* $\mathbf{Q} \in \mathbb{R}^n_n$ *be an orthogonal projection of* rank $\mathbf{Q} = m$. *Then,* $\mathbf{z}'\mathbf{Q}\mathbf{z} \sim \chi^2_m(\delta)$, *where* $\delta = \boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu}/2$.

Proof. Let $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_m)$ be an orthonormal basis for Im $\mathbf{Q}$ and write $\mathbf{Q} = \mathbf{H}\mathbf{H}'$, where $\mathbf{H}'\mathbf{H} = \mathbf{I}_m$. Then, $\mathbf{z}'\mathbf{Q}\mathbf{z} = (\mathbf{H}'\mathbf{z})'(\mathbf{H}'\mathbf{z}) = |\mathbf{e}|^2$, where $\mathbf{e} = \mathbf{H}'\mathbf{z} \sim N_m(\mathbf{H}'\boldsymbol{\mu}, \mathbf{I})$. Hence, $|\mathbf{e}|^2 \sim \chi^2_m(\delta)$ with $\delta = |\mathbf{H}'\boldsymbol{\mu}|^2/2$. □

If, in addition, we assume normality,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},\ \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}),$$

then since

$$\hat{\boldsymbol{\beta}} = \mathbf{B}_0\mathbf{y} = \mathbf{B}_0\hat{\boldsymbol{\mu}},\ (n-k)\hat{s}^2 = |\mathbf{y} - \hat{\boldsymbol{\mu}}|^2 = \mathbf{e}'\mathbf{Q}\mathbf{e}, \text{ and } \hat{\boldsymbol{\mu}} \perp\!\!\!\perp \mathbf{y} - \hat{\boldsymbol{\mu}},$$

we have the general result

$$\hat{\boldsymbol{\beta}} \sim N_k\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right),\ (n-k)\hat{s}^2 \sim \sigma^2\chi^2_{n-k}, \text{ and } \hat{\boldsymbol{\beta}} \perp\!\!\!\perp s$$

with corollary

$$\frac{|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})|^2}{k\hat{s}^2} \sim F(k, n-k).$$

We close this section with a slight generalization of Lemma 5.1.

Corollary 5.1 *Assume* $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$, *and* $\mathbf{A}$ *is symmetric such that* $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}$ *and* rank $\boldsymbol{\Sigma}\mathbf{A} = m$. *Then,* $\mathbf{x}'\mathbf{A}\mathbf{x} \sim \chi^2_m(\delta)$, *where* $\delta = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}/2$.

Proof. Letting $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$ and $\mathbf{B} = \boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}$, then $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{z}'\mathbf{B}\mathbf{z}$, where $\mathbf{z} \sim N_n(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \mathbf{I})$, and the conclusion follows from Lemma 5.1 since $\mathbf{B}$ is an orthogonal projection of rank m. □

5.6.3 Simple correlation

Let (x_i, y_i) i.i.d. (x, y), $i = 1, \ldots, n$, be any "bivariate" sample. The correlation coefficient

$$\begin{aligned} \rho &= \operatorname{cor}(x, y) \\ &= \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)}\sqrt{\operatorname{var}(y)}} \end{aligned}$$

is usually estimated by the sample correlation coefficient

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^{1/2}\left[\sum_{i=1}^n (y_i - \bar{y})^2\right]^{1/2}} \\ &= \frac{(\mathbf{x} - \bar{x}\mathbf{1})'(\mathbf{y} - \bar{y}\mathbf{1})}{|\mathbf{x} - \bar{x}\mathbf{1}|\ |\mathbf{y} - \bar{y}\mathbf{1}|}. \end{aligned}$$

Note that r is just the cosine of the angle between the residual vectors $\mathbf{x} - \bar{x}\mathbf{1}$ and $\mathbf{y} - \bar{y}\mathbf{1}$. The main (nonparametric) reason for using r as an estimate of ρ is its (strong) consistency: $r \overset{\text{w.p.1}}{\to} \rho$ as $n \to \infty$.

If, in addition, we assume normality, a "pivotal statistic" may be derived. First, notice that since r is invariant with respect to relocation and rescaling in both x and y, we may suppose at the outset that

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Now, suppose that $\rho = 0$ so that $x \perp\!\!\!\perp y$. Then,

$$r = \frac{(\mathbf{x} - \bar{x}\mathbf{1})'(\mathbf{y} - \bar{y}\mathbf{1})}{|\mathbf{x} - \bar{x}\mathbf{1}|\ |\mathbf{y} - \bar{y}\mathbf{1}|} = \frac{(\mathbf{Q}\mathbf{x})'(\mathbf{Q}\mathbf{y})}{|\mathbf{Q}\mathbf{x}||\mathbf{Q}\mathbf{y}|},$$

where $\mathbf{Q} = \mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}'$ is an orthogonal projection of rank $\mathbf{Q} = n - 1$. We can write (v. Section 1.6) $\mathbf{Q} = \mathbf{H}\mathbf{H}'$ with $\mathbf{H}'\mathbf{H} = \mathbf{I}_{n-1}$. Then,

$$r = \frac{(\mathbf{H}'\mathbf{x})'(\mathbf{H}'\mathbf{y})}{|\mathbf{H}'\mathbf{x}||\mathbf{H}'\mathbf{y}|} = \frac{\mathbf{z}'\mathbf{w}}{|\mathbf{z}||\mathbf{w}|},$$

where $\mathbf{z} = \mathbf{H}'\mathbf{x}$ and $\mathbf{w} = \mathbf{H}'\mathbf{y}$ are independent $N_{n-1}(\mathbf{0}, \mathbf{I})$. Finally, letting $\mathbf{u} = \mathbf{z}/|\mathbf{z}|$ and $\mathbf{v} = \mathbf{w}/|\mathbf{w}|$, we have $\mathbf{u} \perp\!\!\!\perp \mathbf{v}$, and from Corollary 4.3, $\mathbf{u} \overset{d}{=} \mathbf{v} \sim \text{unif}(S^{n-2})$. Therefore, using Proposition 4.8,

$$r = \mathbf{u}'\mathbf{v} \overset{d}{=} u_1 = \frac{z_1}{|\mathbf{z}|}.$$

Thus,

$$\frac{r}{\sqrt{1 - r^2}} \overset{d}{=} \frac{z_1}{\left(z_2^2 + \cdots + z_{n-1}^2\right)^{1/2}},$$

where the z_i's are i.i.d. $N(0, 1)$ and

$$\sqrt{n-2}\frac{r}{\sqrt{1 - r^2}} \overset{d}{=} t_{n-2}.$$

We have proved:

Proposition 5.9 *If $(x_i, y_i)'$, $i = 1, \ldots, n$, are i.i.d. as a bivariate normal with $\rho = 0$, then*

$$\sqrt{n-2}\frac{r}{\sqrt{1 - r^2}} \overset{d}{=} t_{n-2}.$$

However, if $\rho \neq 0$ and

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}\right),$$

then we may apply this result to the linear transformation

$$\begin{pmatrix} x/\sigma - \rho y/\tau \\ y/\tau \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 - \rho^2 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

using

$$\tilde{r} = \frac{\sum_{i=1}^{n} \left((x_i - \bar{x})/\sigma - \rho(y_i - \bar{y})/\tau\right)(y_i - \bar{y})}{\left[\sum_{i=1}^{n} \left((x_i - \bar{x})/\sigma - \rho(y_i - \bar{y})/\tau\right)^2\right]^{1/2} \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}.$$

We find

$$\frac{\tilde{r}}{\sqrt{1-\tilde{r}^2}} \stackrel{d}{=} \frac{z_1}{(z_2^2 + \cdots + z_{n-1}^2)^{1/2}},$$

where the z_i's are i.i.d. $N(0,1)$ and, by direct computation,

$$\frac{\tilde{r}}{\sqrt{1-\tilde{r}^2}} = \frac{r - \rho c}{\sqrt{1-r^2}}, \text{ where } c = \frac{s_y/\tau}{s_x/\sigma}.$$

Thus, we obtain the result

$$\sqrt{n-2}\frac{(r-\rho c)}{\sqrt{1-r^2}} \stackrel{d}{=} t_{n-2}.$$

This is actually a pivotal for $\beta = \rho\sigma/\tau$. Later, the reader will be able to prove that

$$\sqrt{n}(r-\rho) \stackrel{d}{\to} (1-\rho^2)z, \ z \sim N(0,1)$$

(v. Problem 6.4.8), which can be used to obtain an approximate confidence interval for ρ. The exact distribution of r is treated in Section 8.4 in the more general context of multiple correlation coefficient.

5.7 Problems

1. Plot the contours of the $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution when

$$\begin{aligned} \boldsymbol{\mu} &= \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \\ \boldsymbol{\Sigma} &= \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}. \end{aligned}$$

2. Let $\mathbf{x} \sim N_n(\mu\mathbf{1}, \sigma^2\mathbf{I})$, $\mathbf{y} \sim N_n(\nu\mathbf{1}, \tau^2\mathbf{I})$, and $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ and consider

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}.$$

   (i) Determine the distribution of r.
   (ii) Determine $E\, r$ and var r.

3. Prove Proposition 5.4 with characteristic functions.

4. Obtain the integral moments of the t_p distribution.

5. Let $\mathbf{x}$ be such that $E\,\mathbf{x} = \boldsymbol{\mu}$ and $\operatorname{var}\mathbf{x} = \boldsymbol{\Sigma}$. Show that

$$\min_{\mathbf{c}} E\,|\mathbf{x}-\mathbf{c}|^2 = \operatorname{tr}\boldsymbol{\Sigma}$$

and that the minimum is attained at $\mathbf{c} = \boldsymbol{\mu}$.

6. Assume

$$\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix} \sim N_n\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right).$$

Demonstrate that

$$\min_{\mathbf{C},\mathbf{d}} E\,|\mathbf{x}_1 - (\mathbf{C}\mathbf{x}_2+\mathbf{d})|^2 = \operatorname{tr}\boldsymbol{\Sigma}_{11.2}$$

is attained at $\mathbf{C} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-$ and $\mathbf{d} = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-\boldsymbol{\mu}_2$.

7. Assume that $\mathbf{z} \sim N_n(\mathbf{0},\mathbf{I})$ and let $\bar{z} = \sum_{i=1}^n z_i/n$, $(n-1)s^2 = \sum_{i=1}^n (z_i-\bar{z})^2$.

(i) Prove $\bar{z}$, s, $(z_1-\bar{z})/s$ are mutually independent.
(ii) Determine the distribution of $(z_1-\bar{z})/s$.

Hint: Let $\mathbf{H} = (\mathbf{1}/\sqrt{n}, (\mathbf{e}_1 - n^{-1}\mathbf{1})/\sqrt{(n-1)/n}, \boldsymbol{\Gamma}) \in \mathbf{O}_n$, for some matrix $\boldsymbol{\Gamma}$, $\mathbf{w} = \mathbf{H}'\mathbf{z}$, and note that $(w_2,\dots,w_n)'$ is rotationally invariant.

8. Assume $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{e}$, $\mathbf{e} \sim N_n(\mathbf{0},\sigma^2\mathbf{I})$, where, as usual, the columns of $\mathbf{X} \in \mathbb{R}_k^n$ are linearly independent and let $\mathbf{C} \in \mathbb{R}_k^r$ be of rank r. Show that

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}}-\mathbf{d})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}-\mathbf{d})}{r\hat{s}^2} \sim F(r, n-k; \delta),$$

where

$$\delta = \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{d})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{d})}{2\sigma^2}.$$

9. Let $\mathbf{x} \sim N_n(\mu\mathbf{1},\sigma^2\mathbf{I})$.

(i) Assume $\mathbf{y}$ is fixed, $\mathbf{y} \notin \operatorname{span}\{\mathbf{1}\}$. Find the distribution of

$$r = \frac{(\mathbf{x}-\bar{x}\mathbf{1})'(\mathbf{y}-\bar{y}\mathbf{1})}{|\mathbf{x}-\bar{x}\mathbf{1}|\;|\mathbf{y}-\bar{y}\mathbf{1}|}.$$

(ii) This time assume $\mathbf{y}$ has any distribution satisfying

$$P(\mathbf{y} \notin \operatorname{span}\{\mathbf{1}\}) = 1$$

and $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$, and determine the distribution of r.

10. **Angular gaussian distribution.** The angular gaussian distribution is obtained by the projection of $\mathbf{x} \sim N_n(\mathbf{0},\boldsymbol{\Lambda})$ onto the unit sphere S^{n-1}; i.e., the angular gaussian density is that of $\mathbf{u} = \mathbf{x}/|\mathbf{x}|$.

(i) Prove that the angular gaussian density is

$$f(\mathbf{u}) = \frac{\Gamma(\frac{1}{2}n)}{2\pi^{n/2}}|\mathbf{\Lambda}|^{-1/2}(\mathbf{u}'\mathbf{\Lambda}^{-1}\mathbf{u})^{-n/2},\ \mathbf{u} \in S^{n-1}.$$

(ii) What is the special case $\mathbf{\Lambda} = \mathbf{I}$?

(iii) Prove that the angular gaussian distribution can also be obtained by projecting (onto S^{n-1}) $\mathbf{x}$ with density

$$f_{\mathbf{x}}(\mathbf{x}) = |\mathbf{\Lambda}|^{-1/2}g(\mathbf{x}'\mathbf{\Lambda}^{-1}\mathbf{x}).$$

The word gaussian is misleading here; symmetry is the key.

11. **Rotationally symmetric distributions on spheres** [Saw (1978)]. This class of distributions will be those for which the density is constant on those points $\mathbf{u} \in S^{n-1}$ satisfying $\mathbf{u}'\boldsymbol{\theta} = \delta$, $\forall \delta \in [-1, 1]$ and some fixed $\boldsymbol{\theta} \in S^{n-1}$.

(i) For some fixed $\lambda \geq 0$, consider the function $g(\lambda, \cdot) : [-1, 1] \to [0, \infty)$. Prove

$$\int_{S^{n-1}} \omega_n^{-1} g(\lambda, \mathbf{u}'\boldsymbol{\theta}) d\mathbf{u} = \int_{-1}^{1} g(\lambda, t) \frac{(1-t^2)^{(n-3)/2}}{B(\frac{1}{2}, \frac{1}{2}(n-1))} dt,$$

where $B(\cdot, \cdot)$ denotes the beta function and $\omega_n = 2\pi^{n/2}/\Gamma(\frac{1}{2}n)$ is the "area" of S^{n-1}.
Hint:

$$\int_{S^{n-1}} \omega_n^{-1} g(\lambda, \mathbf{u}'\boldsymbol{\theta}) d\mathbf{u} = E\ g(\lambda, \mathbf{u}'\boldsymbol{\theta}) = E\ g(\lambda, u_1),$$

where $\mathbf{u} = (u_1, \ldots, u_n)' \sim \text{unif}(S^{n-1})$ and use Problem 4.6.11.

(ii) Deduce that $f(\mathbf{u}) = \omega_n^{-1} g(\lambda, \mathbf{u}'\boldsymbol{\theta})$ is a density on S^{n-1} if

$$\int_{-1}^{1} g(\lambda, t) \frac{(1-t^2)^{(n-3)/2}}{B(\frac{1}{2}, \frac{1}{2}(n-1))} dt = 1.$$

Denote this distribution $\mathbf{u} \sim G_n(\lambda, \boldsymbol{\theta})$.

(iii) What are the "contours" of a $G_n(\lambda, \boldsymbol{\theta})$ distribution?

(iv) If $g(\lambda, t)$ is an increasing function of t, prove $G_n(\lambda, \boldsymbol{\theta})$ is unimodal. What is the mode?

(v) Prove: $\mathbf{u} \sim G_n(\lambda, \boldsymbol{\theta}) \Longrightarrow \mathbf{H}\mathbf{u} \sim G_n(\lambda, \mathbf{H}\boldsymbol{\theta})$, $\forall \mathbf{H} \in \mathbf{O}_n$.

(vi) Obtain the first two moments of $\mathbf{u} \sim G_n(\lambda, \boldsymbol{\theta})$,

$$\begin{aligned} E\ \mathbf{u} &= \rho_1 \boldsymbol{\theta}, \\ E\ \mathbf{u}\mathbf{u}' &= \{(1-\rho_2)\mathbf{I} + (n\rho_2 - 1)\boldsymbol{\theta}\boldsymbol{\theta}'\}/(n-1), \end{aligned}$$

where

$$\rho_i = \int_{-1}^{1} t^i g(\lambda, t) \frac{(1-t^2)^{(n-3)/2}}{B(\frac{1}{2}, \frac{1}{2}(n-1))} dt < \infty,\ i = 1, 2.$$

Hint: Use the representation $\mathbf{u} \stackrel{d}{=} t\boldsymbol{\theta}+(1-t^2)^{1/2}\boldsymbol{\zeta}$, where $t = \mathbf{u}'\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ is distributed uniformly on the sphere orthogonal to $\boldsymbol{\theta}$, $t \perp\!\!\!\perp \boldsymbol{\zeta}$ [Watson (1983), p. 44].

(vii) Prove

$$f_{\mathbf{x}}(\mathbf{x}) = g(\lambda, \boldsymbol{\theta}'\mathbf{x}/|\mathbf{x}|)(2\pi)^{-n/2} \exp\left(-\tfrac{1}{2}\mathbf{x}'\mathbf{x}\right)$$

is a density on $\mathbb{R}^n$ by transforming to polar coordinates $\mathbf{x} \mapsto (r, \mathbf{u})$, $r \geq 0$, $\mathbf{u} \in S^{n-1}$.

(viii) Demonstrate that the distribution $G_n(\lambda, \boldsymbol{\theta})$ can be obtained by projecting the distribution for $\mathbf{x} \sim f_{\mathbf{x}}$ onto S^{n-1}; i.e., if $\mathbf{x} \sim f_{\mathbf{x}}$, then $\mathbf{u} = \mathbf{x}/|\mathbf{x}| \sim G_n(\lambda, \boldsymbol{\theta})$.

Remark: The very special case $g(\lambda, t) = \exp(\lambda t)$ yields the Langevin distribution also known, for $n = 2$ and 3, as the Fisher-von Mises distribution on the circle and sphere [Fisher (1953), von Mises (1918)]. Tests for the mean direction, $\boldsymbol{\theta}$, of the Langevin distribution are discussed by Fujikoshi and Watamori (1992). Robust estimators of $(\lambda, \boldsymbol{\theta})$ for the Langevin distribution include the circular median [Mardia (1972)], the normalized spatial median [Ducharme and Milasevic (1987)], and the M-estimator on spheres [Ko and Chang (1993)]. Goodness-of-fit for directional data using smooth tests was considered by Boulerice and Ducharme (1997). Asymptotic behavior of sample mean direction on spheres, without symmetry condition on the p.d.f., was recently derived by Hendriks at al. (1996).

6
Multivariate sampling

6.1 Introduction

The basic tools for manipulating random samples from a multivariate distribution are developed in this chapter. We introduce random matrices in Section 6.2 and show the usefulness of the "vec operator" and Kronecker product in this regard. Also, the matrix variate normal distribution is defined and its basic properties are explained. Section 6.3 deals with theorems in the "asymptotic world" as the sample size goes to infinity. These are the central limit theorem, a general Slutsky theorem, and the so-called delta method.

6.2 Random matrices and multivariate sample

For $\mathbf{A} = (a_{ij}) = (\mathbf{a}_1, \ldots, \mathbf{a}_q) \in \mathbb{R}_q^p$, we may always regard $\mathbf{A}$ as a vector in $\mathbb{R}^{pq}$ where we define

$$\operatorname{vec}(\mathbf{A}) = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_q \end{pmatrix}.$$

This operation is obviously linear $\mathbb{R}_q^p \to \mathbb{R}^{pq}$ and we may regard $\mathbf{A}$ and $\operatorname{vec}(\mathbf{A})$ as synonymous.

For

$$\mathbf{X} = (x_{ij}) = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_p' \end{pmatrix}$$

random on $\mathbb{R}_q^p$, we may denote the mean of $\mathbf{X}$ by $E\ \mathbf{X} = \mathbf{M} = (\mu_{ij})$. However, the variance of $\mathbf{X}$ is a quadruply indexed array consisting of all covariances of the individual entries

$$\text{var } \mathbf{X} = (\sigma_{ijkl}) = (\text{cov}(x_{ij}, x_{kl}))\,.$$

Since there is no inherent order to this array, we find it convenient to impose one by equating

$$\text{var } \mathbf{X} = \text{var vec}(\mathbf{X}') = \mathbf{\Omega} = (\mathbf{\Omega}_{ij}) = (\text{cov}(\mathbf{x}_i, \mathbf{x}_j))\,.$$

The element in position (k, l) of the block $\mathbf{\Omega}_{ij}$ is $\text{cov}(x_{ik}, x_{jl})$. One must be very careful to remember that $\mathbf{\Omega}$ is $pq \times pq$. For instance, if we write $\mathbf{X} \sim N_q^p(\mathbf{M}, \mathbf{\Omega})$, we really mean that $\text{vec}(\mathbf{X}') \sim N_{pq}(\text{vec}(\mathbf{M}'), \mathbf{\Omega})$. In fact, this will be the definition. Moments of a multivariate normal matrix, $N_q^p(\mathbf{M}, \mathbf{\Omega})$, were given by Wong and Liu (1994). Characterization of a multivariate normal matrix distribution via conditioning is discussed by Gupta and Varga (1992) and Nguyen (1997).

The Kronecker product will be very handy for manipulating random matrices. The Kronecker product of $\mathbf{A} \in \mathbb{R}_q^p$ and $\mathbf{B} \in \mathbb{R}_s^r$ is a block-matrix with the block in position (i, j) being $a_{ij}\mathbf{B}$,

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B}) \in \mathbb{R}_{qs}^{pr}.$$

One can verify the basic properties.

Lemma 6.1 *The Kronecker product satisfies the following:*

(i) $(a\mathbf{A}) \otimes (b\mathbf{B}) = ab(\mathbf{A} \otimes \mathbf{B})$, $a, b \in \mathbb{R}$

(ii) $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = (\mathbf{A} \otimes \mathbf{C}) + (\mathbf{B} \otimes \mathbf{C})$

(iii) $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$

(iv) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$,

(v) $(\mathbf{AB}) \otimes (\mathbf{CD}) = (\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D})$

(vi) $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$, *whenever* $\mathbf{A}$ *and* $\mathbf{B}$ *are nonsingular.*

(vii) *If* $\mathbf{v} \neq \mathbf{0}$ *and* $\mathbf{u} \neq \mathbf{0}$ *are eigenvectors of* $\mathbf{A}$ *and* $\mathbf{B}$*, respectively,* $\mathbf{Av} = \lambda\mathbf{v}$*, and* $\mathbf{Bu} = \gamma\mathbf{u}$*, then* $\mathbf{v} \otimes \mathbf{u}$ *is an eigenvector of* $\mathbf{A} \otimes \mathbf{B}$ *corresponding to the eigenvalue* $\lambda\gamma$*.*

(viii) $\text{tr } (\mathbf{A} \otimes \mathbf{B}) = (\text{tr } \mathbf{A})(\text{tr } \mathbf{B})$

(ix) $|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^q\ |\mathbf{B}|^p$, $\mathbf{A} \in \mathbb{R}_p^p$, $\mathbf{B} \in \mathbb{R}_q^q$

(x) *If* $\mathbf{A} > \mathbf{0}$ *and* $\mathbf{B} > \mathbf{0}$, *then* $\mathbf{A} \otimes \mathbf{B} > \mathbf{0}$.

The following lemma will also be useful for handling random matrices. Its proof is left as an exercise.

Lemma 6.2 $\mathbf{A} \in \mathbb{R}^r_p$, $\mathbf{X} \in \mathbb{R}^p_q$, *and* $\mathbf{B} \in \mathbb{R}^q_s \implies \text{vec}(\mathbf{AXB}) = (\mathbf{B}' \otimes \mathbf{A})\text{vec}(\mathbf{X})$.

As a corollary useful for densities (v. Problem 6.4.4) we also have:

Corollary 6.1 *Let* $\mathbf{A} \in \mathbb{R}^p_p$, $\mathbf{X} \in \mathbb{R}^p_q$, *and* $\mathbf{B} \in \mathbb{R}^q_q$. *If* $\mathbf{Y} = \mathbf{AXB}$, *then* $J(\mathbf{Y} \to \mathbf{X}) = |\mathbf{A}|^q_+ \ |\mathbf{B}|^p_+$.

Proof. Since $\text{vec}(\mathbf{Y}) = \text{vec}(\mathbf{AXB}) = (\mathbf{B}' \otimes \mathbf{A})\text{vec}(\mathbf{X})$, then

$$J(\mathbf{Y} \to \mathbf{X}) = J(\text{vec}(\mathbf{Y}) \to \text{vec}(\mathbf{X})) = |\mathbf{B}' \otimes \mathbf{A}|_+ = |\mathbf{A}|^q_+ \ |\mathbf{B}|^p_+.$$

□

Example 6.1 *Consider a sample* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *i.i.d.* $\mathbf{x}$, *where* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and forms the "sample matrix"*

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix}.$$

Then, we see that

$$\mathbf{X} \sim N^n_p(\mathbf{1}\boldsymbol{\mu}', \mathbf{I}_n \otimes \boldsymbol{\Sigma}).$$

Example 6.2 *As another example, suppose that* $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$ *and form the "outer product" matrix*

$$\mathbf{W} = \mathbf{z}\mathbf{z}' = (z_1\mathbf{z}, \ldots, z_p\mathbf{z}).$$

Then, obviously,

$$E\ \mathbf{W} = \text{var}\ \mathbf{z} = \mathbf{I},$$

but the variance of $\mathbf{W}$ *depends on the fourth-order moments of* $\mathbf{z}$. *Since* $E\ z_i = E\ z_i^3 = 0$, $E\ z_i^2 = 1$, *and* $E\ z_i^4 = 3$, *it follows easily that*

$$\begin{aligned} E\ z_i\mathbf{z} &= \mathbf{e}_i, \\ E\ z_i z_j \mathbf{z}\mathbf{z}' &= \delta_{ij}\mathbf{I} + \mathbf{e}_i\mathbf{e}'_j + \mathbf{e}_j\mathbf{e}'_i, \end{aligned}$$

from which

$$\text{cov}(z_i\mathbf{z}, z_j\mathbf{z}) = \delta_{ij}\mathbf{I} + \mathbf{e}_j\mathbf{e}'_i.$$

At this point it becomes useful to define the "commutation matrix" $\mathbf{K}_p$, *a block-matrix whose block in position* (i, j) *is* $\mathbf{e}_j\mathbf{e}'_i \in \mathbb{R}^p_p$,

$$\mathbf{K}_p = (\mathbf{e}_j\mathbf{e}'_i) \in \mathbb{R}^{p^2}_{p^2}.$$

For example, for $p = 2$, we have

$$\mathbf{K}_2 = \begin{pmatrix} 1 & 0 & \vdots & 0 & 0 \\ 0 & 0 & \vdots & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & \vdots & 0 & 0 \\ 0 & 0 & \vdots & 0 & 1 \end{pmatrix}.$$

This enables one to write succinctly [Magnus and Neudecker (1979)*]*

$$\text{var } \mathbf{W} = (\mathbf{I} + \mathbf{K}_p).$$

To generalize slightly, suppose that $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ and let $\mathbf{W} = \mathbf{x}\mathbf{x}'$. Since $\mathbf{W} = \mathbf{x}\mathbf{x}' \stackrel{d}{=} \mathbf{A}\mathbf{z}\mathbf{z}'\mathbf{A}'$, where $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$ and $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}'$, the variance of $\mathbf{W}$ becomes

$$\begin{aligned} \text{var } \mathbf{W} &= \text{var } \mathbf{A}\mathbf{z}\mathbf{z}'\mathbf{A}' = \text{var } (\mathbf{A} \otimes \mathbf{A})\text{vec}(\mathbf{z}\mathbf{z}') \\ &= (\mathbf{A} \otimes \mathbf{A})(\mathbf{I} + \mathbf{K}_p)(\mathbf{A}' \otimes \mathbf{A}'). \end{aligned}$$

However, since $\mathbf{K}_p$ commutes with $\mathbf{A} \otimes \mathbf{A}$ (why?) (v. Problem 6.4.2*), then, finally,*

$$\text{var } \mathbf{W} = (\mathbf{I} + \mathbf{K}_p)(\mathbf{\Sigma} \otimes \mathbf{\Sigma}).$$

We can also write this expression componentwise as

$$\text{cov}(w_{ik}, w_{jl}) = \sigma_{ij}\sigma_{kl} + \sigma_{kj}\sigma_{il}, \tag{6.1}$$

where $\mathbf{\Sigma} = (\sigma_{ij})$. Suppose that

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

We may use the above result to determine the variance of (x^2, y^2, xy). This is needed later in obtaining the asymptotic distribution of the sample correlation coefficient.

For $\mathbf{x} \sim \text{unif}(B^n)$, $\mathbf{W} = \mathbf{x}\mathbf{x}'$, the above method may be adapted to help determine $E\ \mathbf{W}$ and var $\mathbf{W}$.

The distribution for linear transformations of multivariate normal matrices is straightforward with Lemma 6.2.

Proposition 6.1 *If $\mathbf{A} \in \mathbb{R}_p^r$, $\mathbf{X} \sim N_q^p(\mathbf{M}, \mathbf{\Omega})$, and $\mathbf{B} \in \mathbb{R}_s^q$, then*

$$\mathbf{A}\mathbf{X}\mathbf{B} \sim N_s^r\left(\mathbf{A}\mathbf{M}\mathbf{B}, (\mathbf{A} \otimes \mathbf{B}')\mathbf{\Omega}(\mathbf{A}' \otimes \mathbf{B})\right).$$

Proof. Since $\text{vec}(\mathbf{X}') \sim N_{pq}(\text{vec}(\mathbf{M}'), \mathbf{\Omega})$, then

$$\begin{aligned} \text{vec}\left((\mathbf{A}\mathbf{X}\mathbf{B})'\right) &= (\mathbf{A} \otimes \mathbf{B}')\text{vec}(\mathbf{X}') \\ &\sim N_{rs}\left((\mathbf{A} \otimes \mathbf{B}')\text{vec}(\mathbf{M}'), (\mathbf{A} \otimes \mathbf{B}')\mathbf{\Omega}(\mathbf{A}' \otimes \mathbf{B})\right). \end{aligned}$$

The proof is complete as $(\mathbf{A} \otimes \mathbf{B}')\text{vec}(\mathbf{M}') = \text{vec}\left((\mathbf{A}\mathbf{M}\mathbf{B})'\right)$. □

Example 6.3 *Assuming* $\mathbf{X} \sim N_q^p(\mathbf{M}, \mathbf{A} \otimes \mathbf{B})$, $\mathbf{A} \geq \mathbf{0}$ *is in* $\mathbb{R}_p^p$ *and* $\mathbf{B} \geq \mathbf{0}$ *is in* $\mathbb{R}_q^q$. *We evaluate* $E\ \mathbf{X}\mathbf{X}'$. *Let* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$ *and observe, with the choice* $\mathbf{A} = \mathbf{I}_p$ *and* $\mathbf{B} = \mathbf{e}_i$, *that* $\mathbf{x}_i \sim N_p(\mathbf{m}_i, b_{ii}\mathbf{A})$, *where* $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_q)$. *Then,*

$$E\ \mathbf{X}\mathbf{X}' = \sum_{i=1}^{q} E\ \mathbf{x}_i\mathbf{x}_i' = \sum_{i=1}^{q}(b_{ii}\mathbf{A} + \mathbf{m}_i\mathbf{m}_i')$$

leads to the expression

$$E\ \mathbf{X}\mathbf{X}' = (\text{tr } \mathbf{B})\mathbf{A} + \mathbf{M}\mathbf{M}'.$$

We now turn to considerations of convergence. For the general sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}$, where $E\ \mathbf{x} = \boldsymbol{\mu}$ and var $\mathbf{x} = \boldsymbol{\Sigma}$, the strong law of large numbers (S.L.L.N.) provides the *sample mean* as a natural estimate $\bar{\mathbf{x}} = \sum_{j=1}^{n} \mathbf{x}_j/n$ for $\boldsymbol{\mu}$:

$$\bar{\mathbf{x}} \overset{\text{w.p.1}}{\to} \boldsymbol{\mu}.$$

Of course, $\mathbf{W}_i = \mathbf{x}_i\mathbf{x}_i'$, $i = 1, \ldots, n$, are i.i.d. $\mathbf{x}\mathbf{x}'$, where $E\ \mathbf{x}\mathbf{x}' = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'$ and the S.L.L.N. applies to $\overline{\mathbf{W}} = \sum_{j=1}^{n} \mathbf{x}_j\mathbf{x}_j'/n$ so that

$$\overline{\mathbf{W}} \overset{\text{w.p.1}}{\to} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'.$$

Then, obviously, if we let $\hat{\boldsymbol{\Sigma}} = \overline{\mathbf{W}} - \bar{\mathbf{x}}\bar{\mathbf{x}}'$, we find

$$\hat{\boldsymbol{\Sigma}} \overset{\text{w.p.1}}{\to} \boldsymbol{\Sigma}.$$

However, $E\ \bar{\mathbf{x}}\bar{\mathbf{x}}' = \boldsymbol{\Sigma}/n + \boldsymbol{\mu}\boldsymbol{\mu}'$, so that

$$E\ \frac{n}{n-1}\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$$

and it has become customary to use this "unbiased" estimate.

The reader should have no particular difficulty in showing that as explicit functions of the sample matrix, these (unbiased and consistent) estimates may be expressed by

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1} \quad \text{and} \quad \mathbf{S} \equiv \frac{n}{n-1}\hat{\boldsymbol{\Sigma}} = \frac{1}{(n-1)}\mathbf{X}'\mathbf{Q}\mathbf{X},$$

where $\mathbf{Q} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$. The estimate $\mathbf{S}$ is the *sample variance*, which is often written as

$$\mathbf{S} = \frac{1}{(n-1)}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

As an expression of "pythagorus," we find

$$\mathbf{X} = \mathbf{Q}\mathbf{X} + \mathbf{P}\mathbf{X}, \text{ where } \mathbf{P} = \mathbf{I} - \mathbf{Q} = n^{-1}\mathbf{1}\mathbf{1}'$$

and, thus,

$$\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{Q}\mathbf{X} + \mathbf{X}'\mathbf{P}\mathbf{X} = (n-1)\mathbf{S} + n\bar{\mathbf{x}}\bar{\mathbf{x}}'.$$

6.3 Asymptotic distributions

The central limit theorem (C.L.T.) states that for any sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}$, where $E\,\mathbf{x} = \boldsymbol{\mu}$ and $\operatorname{var}\mathbf{x} = \boldsymbol{\Sigma}$,

$$\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{z}, \text{ where } \mathbf{z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Now, recall the very general fact that if $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ on $\mathbb{R}^p$ and $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^q$ is any continuous (with $P_{\mathbf{x}}$ probability 1)[1] function, then $\mathbf{g}(\mathbf{x}_n) \xrightarrow{d} \mathbf{g}(\mathbf{x})$ on $\mathbb{R}^q$. Note that since matrices in $\mathbb{R}^p_q$ are really only vectors in $\mathbb{R}^{pq}$, this result is considerably more general than it might appear at first.

Thus, if $\boldsymbol{\Sigma}$ is nonsingular (the singular case goes through as well; v. Problem 6.4.10),

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_p.$$

There is another very basic fact that derives from the Cramér-Wold theorem and the (univariate) Slutsky theorem.

Lemma 6.3 (Multivariate Slutsky) *If* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *on* $\mathbb{R}^p_q$ *and* $\mathbf{Y}_n \xrightarrow{d} \mathbf{C}$ *on* $\mathbb{R}^r_s$ *where* $\mathbf{C}$ *is any constant matrix, then*

$$(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{d} (\mathbf{X}, \mathbf{C}) \text{ on } \mathbb{R}^p_q \times \mathbb{R}^r_s.$$

Proof. From Cramér-Wold Proposition 2.10, for any linear combination

$$\begin{aligned} \sum_{i,j} t_{ij} x_{n,ij} &\xrightarrow{d} \sum_{i,j} t_{ij} x_{ij}, \\ \sum_{k,l} s_{kl} y_{n,kl} &\xrightarrow{d} \sum_{k,l} s_{kl} c_{kl}, \end{aligned}$$

and from the univariate Slutsky theorem,

$$\sum_{i,j} t_{ij} x_{n,ij} + \sum_{k,l} s_{kl} y_{n,kl} \xrightarrow{d} \sum_{i,j} t_{ij} x_{ij} + \sum_{k,l} s_{kl} c_{kl}.$$

Using Cramér-Wold again, the conclusion is reached. □

A more general statement on metric spaces can be found in Billingsley (1968, p. 27). It follows, of course, that for any continuous function,

$$\mathbf{g}(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{d} \mathbf{g}(\mathbf{X}, \mathbf{C}).$$

[1] Let $C_{\mathbf{g}} = \{\mathbf{t} \in \mathbb{R}^p : \mathbf{g} \text{ is continuous at } \mathbf{t}\}$. Then, $\mathbf{g}$ is continuous with $P_{\mathbf{x}}$ probability 1 means that $P_{\mathbf{x}}(C_{\mathbf{g}}) = P(\mathbf{x} \in C_{\mathbf{g}}) = 1$.

As a simple example

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_p.$$

One more general proposition:

Proposition 6.2 (Delta method) *If* $\sqrt{n}(\mathbf{x}_n - \mathbf{c}) \xrightarrow{d} \mathbf{z}$ *on* $\mathbb{R}^p$ *and* $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^q$ *is differentiable at* $\mathbf{c}$*, then*

$$\sqrt{n}\,(\mathbf{g}(\mathbf{x}_n) - \mathbf{g}(\mathbf{c})) \xrightarrow{d} \mathbf{Dg}(\mathbf{c})\ \mathbf{z}.$$

Proof. This is simply because by the very definition of the derivative at $\mathbf{c}$, the function

$$\mathbf{k}(\mathbf{t}) = \begin{cases} \mathbf{h}(\mathbf{t})/|\mathbf{t} - \mathbf{c}|, & \mathbf{t} \neq \mathbf{c} \\ \mathbf{0}, & \mathbf{t} = \mathbf{c}, \end{cases}$$

where

$$\mathbf{h}(\mathbf{t}) = (\mathbf{g}(\mathbf{t}) - \mathbf{g}(\mathbf{c})) - \mathbf{Dg}(\mathbf{c})\ (\mathbf{t} - \mathbf{c}),$$

is continuous at $\mathbf{c}$, and we may, therefore, write

$$\sqrt{n}\,(\mathbf{g}(\mathbf{x}_n) - \mathbf{g}(\mathbf{c})) = \mathbf{Dg}(\mathbf{c})\ \sqrt{n}(\mathbf{x}_n - \mathbf{c}) + \mathbf{k}(\mathbf{x}_n)|\sqrt{n}(\mathbf{x}_n - \mathbf{c})|.$$

Using Slutsky's theorem, we may conclude that since $\mathbf{k}(\mathbf{x}_n) \xrightarrow{d} \mathbf{0}$ and $|\sqrt{n}(\mathbf{x}_n - \mathbf{c})| \xrightarrow{d} |\mathbf{z}|$,

$$\sqrt{n}\,(\mathbf{g}(\mathbf{x}_n) - \mathbf{g}(\mathbf{c})) \xrightarrow{d} \mathbf{Dg}(\mathbf{c})\ \mathbf{z}.$$

□

This, of course, applies directly to the C.L.T. to give

$$\sqrt{n}\,(\mathbf{g}(\bar{\mathbf{x}}) - \mathbf{g}(\boldsymbol{\mu})) \xrightarrow{d} N_q\left(\mathbf{0}, \mathbf{Dg}(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{Dg}(\boldsymbol{\mu})'\right).$$

However, consider a more elaborate application: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. $\mathbf{x}$ as before with $E\,\mathbf{x} = \mathbf{0}$ and $\operatorname{var}\mathbf{x} = \boldsymbol{\Sigma}$. Then, let $\mathbf{W}_i = \mathbf{x}_i\mathbf{x}_i'$, $i = 1, \ldots, n$, and $\mathbf{W} = \mathbf{x}\mathbf{x}'$ so that

$$\begin{pmatrix} \mathbf{W}_1 \\ \mathbf{x}_1' \end{pmatrix}, \ldots, \begin{pmatrix} \mathbf{W}_n \\ \mathbf{x}_n' \end{pmatrix} \text{ are i.i.d. } \begin{pmatrix} \mathbf{W} \\ \mathbf{x}' \end{pmatrix}$$

with

$$E\begin{pmatrix} \mathbf{W} \\ \mathbf{x}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma} \\ \mathbf{0}' \end{pmatrix}$$

and

$$\operatorname{var}\begin{pmatrix} \mathbf{W} \\ \mathbf{x}' \end{pmatrix} = \begin{pmatrix} \operatorname{var}\mathbf{W} & \operatorname{cov}(\operatorname{vec}(\mathbf{W}), \mathbf{x}) \\ \operatorname{cov}(\mathbf{x}, \operatorname{vec}(\mathbf{W})) & \boldsymbol{\Sigma} \end{pmatrix} \equiv \boldsymbol{\Omega}.$$

By the C.L.T.,

$$\sqrt{n}\begin{pmatrix} \overline{\mathbf{W}} - \boldsymbol{\Sigma} \\ \bar{\mathbf{x}}' \end{pmatrix} \xrightarrow{d} N_p^{p+1}(\mathbf{0}, \boldsymbol{\Omega})$$

and the reader may then use Lemma 6.3 to find that

$$\sqrt{n}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) = \sqrt{n}(\overline{\mathbf{W}} - \mathbf{\Sigma}) - \frac{1}{\sqrt{n}}(\sqrt{n}\bar{\mathbf{x}})(\sqrt{n}\bar{\mathbf{x}})' \xrightarrow{d} N_p^p(\mathbf{0}, \text{var } \mathbf{W})$$

and, of course,

$$\sqrt{n}(\mathbf{S} - \mathbf{\Sigma}) \xrightarrow{d} N_p^p(\mathbf{0}, \text{var } \mathbf{W}).$$

Note that since the function $\mathbf{S}$ is unchanged if $\mathbf{x}$ is replaced by $\mathbf{x} - \boldsymbol{\mu}$, this result is automatically valid for the more general case where $E\,\mathbf{x} = \boldsymbol{\mu}$. The expression for var $\mathbf{W}$ was given in Example 6.2 for the normal case, and the elliptical case is treated in the sequel in Example 13.6.

Unfortunately, var $\mathbf{W}$ is seldom of a particular tractable form. It depends on the fourth-order multivariate cumulants of $\mathbf{x}$. The relation between product-moments and multivariate cumulants is rather technical and is relegated to Appendix B. There, it is proven generally for $\mathbf{W} = (w_{ij}) = \mathbf{x}\mathbf{x}'$ that

$$\text{cov}(w_{ik}, w_{jl}) = \mu_{1111}^{ijkl} - \mu_{11}^{ik}\mu_{11}^{jl} = k_{1111}^{ijkl} + k_{11}^{kl}k_{11}^{ij} + k_{11}^{il}k_{11}^{jk},$$

where the μ's are the product-moments and the k's are the cumulants of $\mathbf{x}$.

Example 6.4 *For a sample of size n from a bivariate distribution with finite fourth-order moments, we find the asymptotic distribution*

$$\sqrt{n}\left[\begin{pmatrix} s_1^2 \\ s_{12} \\ s_2^2 \end{pmatrix} - \begin{pmatrix} \sigma_1^2 \\ \sigma_{12} \\ \sigma_2^2 \end{pmatrix}\right] \xrightarrow{d} N_3(\mathbf{0}, \mathbf{\Omega}),$$

where

$$\mathbf{\Omega} = \begin{pmatrix} \mu_4^1 - (\mu_2^1)^2 & \mu_{31}^{12} - \mu_{11}^{12}\mu_2^1 & \mu_{22}^{12} - \mu_2^1\mu_2^2 \\ \cdot & \mu_{22}^{12} - (\mu_{11}^{12})^2 & \mu_{13}^{12} - \mu_{11}^{12}\mu_2^2 \\ \cdot & \cdot & \mu_4^2 - (\mu_2^2)^2 \end{pmatrix}.$$

The product-moments are

$$\begin{aligned}
\mu_4^1 &= E\,(x_1 - \mu_1)^4, \\
\mu_2^1 &= E\,(x_1 - \mu_1)^2 = \sigma_1^2, \\
\mu_{31}^{12} &= E\,(x_1 - \mu_1)^3(x_2 - \mu_2), \\
\mu_{11}^{12} &= E\,(x_1 - \mu_1)(x_2 - \mu_2) = \sigma_{12}, \\
\mu_2^2 &= E\,(x_2 - \mu_2)^2 = \sigma_2^2, \\
\mu_{22}^{12} &= E\,(x_1 - \mu_1)^2(x_2 - \mu_2)^2, \ \textit{etc.}
\end{aligned}$$

In general, $\bar{\mathbf{x}}$ and $\mathbf{S}$ will not be asymptotically independent unless all third-order product-moments of $\mathbf{x}$ in cov $(\text{vec}(\mathbf{W}), \mathbf{x})$ are null. But this is exactly the case when $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{x}$ has a spherical distribution since

$$\text{cov}\,(\text{vec}(\mathbf{W}), \mathbf{x}) = \text{cov}\,(\text{vec}(\mathbf{x}\mathbf{x}'), \mathbf{x})$$

$$= \quad (\mathbf{\Sigma}^{1/2} \otimes \mathbf{\Sigma}^{1/2})\ \mathrm{cov}\,(\mathrm{vec}(\mathbf{z}\mathbf{z}'), \mathbf{z})\ \mathbf{\Sigma}^{1/2}$$

and all third-order product-moments of $\mathbf{z}$ are null (v. Problem 4.6.5). However, if the underlying random vector $\mathbf{x}$ is already normal, then things reduce considerably.

For $p = 2$, the correlation coefficient, r, is a very simple function of $\mathbf{S}$ and, thus, it should be straightforward for the reader to obtain the asymptotic distribution of r (v. Problems 6.4.8-6.4.9). In fact, since this function is unchanged if the individual coordinates are normalized, we may assume at the outset that

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

6.4 Problems

1. Prove Lemma 6.2: If $\mathbf{A} \in \mathbb{R}^r_p$, $\mathbf{X} \in \mathbb{R}^p_q$, and $\mathbf{B} \in \mathbb{R}^q_s$, then
$$\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}' \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}).$$

2. Let $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^p_p$ and $\mathbf{K}_p$ be the "commutation matrix." Show the following:
(i) $\mathbf{K}_p = \sum_{i=1}^p \sum_{j=1}^p \mathbf{e}_i \mathbf{e}_j' \otimes \mathbf{e}_j \mathbf{e}_i'$,
(ii) $\mathbf{K}_p \mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}')$,
(iii) $\mathbf{K}_p(\mathbf{A} \otimes \mathbf{B}) = (\mathbf{B} \otimes \mathbf{A})\mathbf{K}_p$,
(iv) $\mathrm{tr}\ \mathbf{A}'\mathbf{B} = [\mathrm{vec}(\mathbf{A})]'\mathrm{vec}(\mathbf{B})$,
(v) If $\mathbf{A}$ is symmetric, $\mathrm{tr}\ \mathbf{A}^2 = [\mathrm{vec}(\mathbf{A})]'\frac{1}{2}(\mathbf{I} + \mathbf{K}_p)\mathrm{vec}(\mathbf{A})$.

3. Show that if $\mathbf{Z} \sim N^p_q(\mathbf{0}, \mathbf{I})$ and $\mathbf{P}$ and $\mathbf{Q}$ in $\mathbb{R}^p_p$ are orthogonal projections such that $\mathbf{PQ} = \mathbf{0}$, then $\mathbf{PZ} \perp\!\!\!\perp \mathbf{QZ}$.
Hint: Obtain
$$\mathrm{var} \begin{pmatrix} \mathbf{PZ} \\ \mathbf{QZ} \end{pmatrix}.$$

4. Obtain the p.d.f. of $\mathbf{X} \sim N^p_q(\mathbf{M}, \mathbf{A} \otimes \mathbf{B})$, where $\mathbf{A} > \mathbf{0}$ is in $\mathbb{R}^p_p$ and $\mathbf{B} > \mathbf{0}$ is in $\mathbb{R}^q_q$:
$$f(\mathbf{X}) = (2\pi)^{-\frac{pq}{2}} |\mathbf{A}|^{-\frac{q}{2}} |\mathbf{B}|^{-\frac{p}{2}} \mathrm{etr}\left[-\tfrac{1}{2}\mathbf{A}^{-1}(\mathbf{X} - \mathbf{M})\mathbf{B}^{-1}(\mathbf{X} - \mathbf{M})'\right],$$
where $\mathrm{etr}(\cdot) \equiv \exp[\mathrm{tr}(\cdot)]$.
Hint: Let $\mathbf{X} = \mathbf{A}^{1/2}\mathbf{Z}\mathbf{B}^{1/2} + \mathbf{M}$, where $\mathbf{Z} \sim N^p_q(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_q)$, and use Corollary 6.1.

5. Assume $E\ \mathbf{E} = \mathbf{0}$ and $\mathrm{var}\ \mathbf{E} = \mathbf{I}_n \otimes \mathbf{\Sigma}$, $\mathbf{\Sigma} \geq \mathbf{0}$ is in $\mathbb{R}^p_p$. Show that
(i) $\mathrm{var}\ \mathbf{E}' = \mathbf{\Sigma} \otimes \mathbf{I}_n$,
(ii) $E\ \mathbf{E}'\mathbf{A}\mathbf{E} = (\mathrm{tr}\ \mathbf{A})\mathbf{\Sigma}$.

6. Assume $\mathbf{\Sigma}_{22}$ is nonsingular and

$$(\mathbf{X}_1, \mathbf{X}_2) \sim N^n_{p_1+p_2}\left(\mathbf{1}(\boldsymbol{\mu}_1', \boldsymbol{\mu}_2'), \mathbf{I}_n \otimes \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}\right),$$

where $\mathbf{X}_i \in \mathbb{R}^n_{p_i}$, $\boldsymbol{\mu}_i \in \mathbb{R}^{p_i}$, and $\mathbf{\Sigma}_{ij} \in \mathbb{R}^{p_i}_{p_j}$, $i, j = 1, 2$. Prove

$$\mathbf{X}_1 \mid \mathbf{X}_2 \sim N^n_{p_1}\left(\mathbf{1}\boldsymbol{\mu}_1' + (\mathbf{X}_2 - \mathbf{1}\boldsymbol{\mu}_2')\mathbf{B}', \mathbf{I}_n \otimes \mathbf{\Sigma}_{11.2}\right),$$

with $\mathbf{B} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$.

7. For $\mathbf{W} = \mathbf{x}\mathbf{x}'$, in each case determine $E\ \mathbf{W}$ and var $\mathbf{W}$:

(i) $\mathbf{x} \sim N_2(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$,
(ii) $\mathbf{x} \sim \text{unif}(S^1)$,
(iii) $\mathbf{x} \sim \text{unif}(B^2)$.

8. Assume $(x_i, y_i)'$, $i = 1, \ldots, n$, are i.i.d.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

and let r be the sample correlation coefficient. Prove the asymptotic result $\sqrt{n}(r - \rho) \xrightarrow{d} N\left(0, (1 - \rho^2)^2\right)$.

9. Fisher's z-transform is

$$\begin{aligned} z &= \tanh^{-1}(r) = \tfrac{1}{2}\log\frac{1+r}{1-r}, \\ \zeta &= \tanh^{-1}(\rho) = \tfrac{1}{2}\log\frac{1+\rho}{1-\rho}. \end{aligned}$$

(i) Show that it is a "variance stabilizing transformation" for the correlation coefficient: $\sqrt{n-3}(z - \zeta) \xrightarrow{d} N(0, 1)$.
(ii) Use the fact that z is a monotone function of r to obtain an approximate $(1 - \alpha)100\%$ confidence interval for ρ,

$$\left[\tanh\left(z - \frac{z_{\alpha/2}}{(n-3)^{1/2}}\right), \tanh\left(z + \frac{z_{\alpha/2}}{(n-3)^{1/2}}\right)\right],$$

where $P(N(0, 1) > z_{\alpha/2}) = \alpha/2$.

10. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}$, where $E\ \mathbf{x} = \boldsymbol{\mu}$ and var $\mathbf{x} = \mathbf{\Sigma}$. Prove that

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{\Sigma}^-(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_r,$$

where $r = \text{rank } \mathbf{\Sigma}$.

11. Demonstrate the following representation of Mahalanobis distance:

$$\sup_{|\mathbf{h}|=1} \frac{\left|\mathbf{h}'\mathbf{x}_i - \frac{1}{n}\sum_{j=1}^n \mathbf{h}'\mathbf{x}_j\right|}{\left[\frac{1}{(n-1)}\sum_{k=1}^n \left(\mathbf{h}'\mathbf{x}_k - \frac{1}{n}\sum_{j=1}^n \mathbf{h}'\mathbf{x}_j\right)^2\right]^{1/2}} = d_i,$$

where $d_i = [(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})]^{1/2}$ is the Mahalanobis distance from $\mathbf{x}_i$ to $\bar{\mathbf{x}}$.

Remark: This was used by Stahel (1981) and Donoho (1982) to suggest the robust estimate of location as a weighted average

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^n w(u_i)\mathbf{x}_i}{\sum_{i=1}^n w(u_i)},$$

where $w(\cdot)$ is a positive and strictly decreasing function and

$$u_i = \sup_{|\mathbf{h}|=1} \frac{|\mathbf{h}'\mathbf{x}_i - \mathrm{med}_j(\mathbf{h}'\mathbf{x}_j)|}{\mathrm{med}_k\,|\mathbf{h}'\mathbf{x}_k - \mathrm{med}_j(\mathbf{h}'\mathbf{x}_j)|}.$$

The notation "med" refers to the ordinary median.

12. **Multivariate familial data** [Konishi and Khatri (1990)].
Suppose a random sample of n families on $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ with $E\,\mathbf{x} = \boldsymbol{\mu}$ and var $\mathbf{x} = \boldsymbol{\Sigma}$. Let

$$\mathbf{Z}_i = \begin{pmatrix} \mathbf{x}'_{1i} \\ \vdots \\ \mathbf{x}'_{k_i,i} \end{pmatrix},\ i = 1, \ldots, n,$$

denote the measurements on the ith family with $k_i \geq 1$ siblings, where $\mathbf{x}_{ji} = (x_{1ji}, \ldots, x_{pji})'$, $j = 1, \ldots, k_i$, is the score of the jth child on p characteristics. It is assumed that $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are mutually independent and

$$\begin{aligned} E\,\mathbf{Z}_i &= \mathbf{1}_{k_i}\boldsymbol{\mu}', \\ \mathrm{var}\,\mathbf{Z}_i &= (\mathbf{I}_{k_i} \otimes \boldsymbol{\Sigma}) + (\mathbf{1}_{k_i}\mathbf{1}'_{k_i} - \mathbf{I}_{k_i}) \otimes \boldsymbol{\Sigma}_s. \end{aligned}$$

The matrix $\boldsymbol{\Sigma}_s$ reflects the dependence among siblings. For the estimation of $\boldsymbol{\Sigma}$, let

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{\mathbf{x}}'_1 \\ \vdots \\ \bar{\mathbf{x}}'_n \end{pmatrix},\ \mathbf{V}_i = \sum_{j=1}^{k_i} (\mathbf{x}_{ji} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ji} - \bar{\mathbf{x}}_i)',$$

where $\bar{\mathbf{x}}_i = (\mathbf{x}_{1i} + \cdots + \mathbf{x}_{k_i,i})/k_i$. Further, let $\mathbf{B} \in \mathbb{R}^n_n$, $\mathbf{B} \geq \mathbf{0}$, such that $\mathbf{B}\mathbf{1}_n = \mathbf{0}$.

(i) Prove that

$$\hat{\boldsymbol{\Sigma}} = (\mathrm{tr}\ \mathbf{B})^{-1}\left(\bar{\mathbf{X}}'\mathbf{B}\bar{\mathbf{X}} + \sum_{i=1}^n \omega_i \mathbf{V}_i\right),$$

where the weights $\omega_1, \ldots, \omega_n$ are non-negative constants, satisfies

$$E\,\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + (\mathrm{tr}\,\mathbf{B})^{-1}\left\{\sum_{i=1}^n \omega_i(k_i - 1) - \mathrm{tr}[\mathbf{B}(\mathbf{I}_n - \mathbf{D}_n^{-1})]\right\}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_s),$$

where $\mathbf{D}_n = \text{diag}(k_1, \ldots, k_n)$.

(ii) Find a condition on the weights so that $\hat{\boldsymbol{\Sigma}}$ is unbiased for $\boldsymbol{\Sigma}$.

(iii) The corresponding estimate of $\boldsymbol{\Sigma}_s$ is given by

$$\hat{\boldsymbol{\Sigma}}_s = (\text{tr } \mathbf{B})^{-1} \left(\bar{\mathbf{X}}' \mathbf{B} \bar{\mathbf{X}} + \sum_{i=1}^{n} \nu_i \mathbf{V}_i \right),$$

where $\nu_1, \ldots, \nu_n$ are constants. Prove that for weights satisfying the condition

$$\sum_{i=1}^{n} \nu_i (k_i - 1) + \text{tr}(\mathbf{B}\mathbf{D}_n^{-1}) = 0,$$

$\hat{\boldsymbol{\Sigma}}_s$ is unbiased for $\boldsymbol{\Sigma}_s$.

A multivariate familial model for interclass correlation, with a "mother" for each family, was considered earlier by Srivastava et al. (1988). Principal component analysis for the model described here was developed by Konishi and Rao (1992). A general description of principal components is given in Chapter 10.

7
Wishart distributions

7.1 Introduction

As before,

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

denotes the sample matrix from which $\bar{\mathbf{x}}$ and $\mathbf{S}$,

$$\begin{aligned} n\bar{\mathbf{x}} &= \mathbf{X}'\mathbf{1}, \\ (n-1)\mathbf{S} &= \mathbf{X}'\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}', \end{aligned}$$

provide consistent unbiased estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. In Section 7.2, the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are derived assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}$ with $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$. The fundamental result about the joint distribution of $\bar{\mathbf{x}}$ and $\mathbf{S}$ is proved in Proposition 7.1. The basic properties of Wishart distributions are studied in Section 7.3. Section 7.4 presents the Box-Cox transformation to enhance the multivariate normality of the data.

7.2 Joint distribution of $\bar{\mathbf{x}}$ and $\mathbf{S}$

With underlying normality, $\bar{\mathbf{x}}$ and $\mathbf{S}$ are "optimal" in some respects. Denote $\mathbf{V} = (n-1)\mathbf{S}$. Using the notation $\exp[\mathrm{tr}(\cdot)] = \mathrm{etr}(\cdot)$, the p.d.f. for $\mathbf{X}$ can

be written in various ways:

$$\begin{aligned} f(\mathbf{X}) &= (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp\left[-\tfrac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right] \\ &= (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}e^{-\frac{n}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\exp\left[-\tfrac{1}{2}\text{ tr }\boldsymbol{\Sigma}^{-1}\mathbf{X}'\mathbf{X}+n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right] \\ &= (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\text{etr}\left\{-\tfrac{1}{2}\left[\mathbf{V}+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right]\boldsymbol{\Sigma}^{-1}\right\}. \qquad (7.1) \end{aligned}$$

By general properties of exponential families [Fraser (1976), pp. 339, 342, 406, or Casella and Berger (1990), pp. 254-255, 263], it is plain that $(\mathbf{X}'\mathbf{X},\bar{\mathbf{x}})$ (or any one-to-one function such as $(\mathbf{S},\bar{\mathbf{x}})$) is minimal sufficient and complete for $(\boldsymbol{\Sigma},\boldsymbol{\mu})$, so that by the Rao-Blackwell/Lehmann-Scheffé theorems, among all unbiased estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, $\bar{\mathbf{x}}$ and $\mathbf{S}$ have minimum variance. We say that $(\mathbf{S},\bar{\mathbf{x}})$ is the MVUE (Minimum Variance Unbiased Estimate) of $(\boldsymbol{\Sigma},\boldsymbol{\mu})$.

Furthermore, to obtain the maximum likelihood estimates (MLE) $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ when $n-1\geq p$, we minimize

$$\ln|\boldsymbol{\Sigma}|+\text{tr }\frac{1}{n}\mathbf{V}\boldsymbol{\Sigma}^{-1}+(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \qquad (7.2)$$

and (since the last term is ≥ 0) it is clear that $\hat{\boldsymbol{\mu}}=\bar{\mathbf{x}}$, so we need only minimize

$$\ln|n\mathbf{V}^{-1}\boldsymbol{\Sigma}|+\text{tr }\frac{1}{n}\mathbf{V}\boldsymbol{\Sigma}^{-1},$$

where the constant, $\ln|n\mathbf{V}^{-1}|$, was added. The condition $n-1\geq p$ ensures that $\mathbf{V}$ is nonsingular w.p.1. This is proved later in Corollary 7.2. But then, letting $\mathbf{T}=n\mathbf{V}^{-1}\boldsymbol{\Sigma}$, we need only determine the $\mathbf{T}$ that minimizes

$$\ln|\mathbf{T}|+\text{tr }\mathbf{T}^{-1}.$$

However, this is accomplished when all the eigenvalues of $\mathbf{T}$ are 1 so that $\mathbf{T}=\mathbf{I}$ and we conclude altogether

$$\hat{\boldsymbol{\mu}}=\bar{\mathbf{x}}\ \text{ and }\ \hat{\boldsymbol{\Sigma}}=\frac{1}{n}\mathbf{V}.$$

Remark: It is a well-known result, which can be traced back to Gauss, that the only location family, $f(x-\theta)$, of p.d.f. on $\mathbb{R}$ for which $\bar{x}$ is a MLE of θ originates from the normal density. This MLE characterization of normal density also holds on $\mathbb{R}^p$ [Stadje (1993)].

Let us consider the exact distribution of $\bar{\mathbf{x}}$ and $\mathbf{S}$. It is obvious that

$$\bar{\mathbf{x}}\sim N_p\left(\boldsymbol{\mu},\frac{1}{n}\boldsymbol{\Sigma}\right).$$

We begin by representing $\mathbf{x}\overset{d}{=}\mathbf{A}\mathbf{z}+\boldsymbol{\mu}$, for any $\mathbf{A}\mathbf{A}'=\boldsymbol{\Sigma}$, which, for the sample matrix, means that $\mathbf{X}\overset{d}{=}\mathbf{Z}\mathbf{A}'+\mathbf{1}\boldsymbol{\mu}'$. Thus,

$$(\bar{\mathbf{x}},\mathbf{S}_{\mathbf{x}})\overset{d}{=}(\mathbf{A}\bar{\mathbf{z}}+\boldsymbol{\mu},\mathbf{A}\mathbf{S}_{\mathbf{z}}\mathbf{A}').$$

However, in $\mathbf{Z} = (z_{ij})$, all the components are i.i.d. $N(0,1)$ so that even the columns are mutually independent. Thus, with orthogonal projections $\mathbf{P} = n^{-1}\mathbf{1}\mathbf{1}'$ and $\mathbf{Q} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$, it is clear that $\mathbf{PZ} \perp\!\!\!\perp \mathbf{QZ}$ (v. Problem 6.4.3), and since $n\bar{\mathbf{z}} = \mathbf{Z}'\mathbf{P}\mathbf{1}$ and $(n-1)\mathbf{S}_\mathbf{z} = \mathbf{Z}'\mathbf{Q}\mathbf{Q}\mathbf{Z}$, we see that $\bar{\mathbf{z}} \perp\!\!\!\perp \mathbf{S}_\mathbf{z}$, hence $\bar{\mathbf{x}} \perp\!\!\!\perp \mathbf{S}_\mathbf{x}$.

If next we express $\mathbf{Q} = \mathbf{H}\mathbf{H}'$, where $\mathbf{H}$ gives an orthonormal basis for $\mathbf{1}^\perp$ (of dimension $(n-1)$), it is made plain that

$$(n-1)\mathbf{S}_\mathbf{z} = \mathbf{Z}'\mathbf{H}\mathbf{H}'\mathbf{Z} = \mathbf{U}'\mathbf{U},$$

where $\mathbf{Z}'\mathbf{H} = \mathbf{U}' = (\mathbf{u}_1, \ldots, \mathbf{u}_{n-1})$, $\mathbf{u}_i$ i.i.d. $N_p(\mathbf{0}, \mathbf{I})$. Accordingly, we make the following definition.

Definition 7.1 Wishart distribution*:*

$$\mathbf{W} \sim W_p(m) \text{ iff } \mathbf{W} \stackrel{d}{=} \sum_{i=1}^{m} \mathbf{z}_i\mathbf{z}_i', \ \mathbf{z}_i \text{ i.i.d. } N_p(\mathbf{0}, \mathbf{I}).$$

$$\mathbf{V} \sim W_p(m, \mathbf{\Sigma}) \text{ iff } \mathbf{V} \stackrel{d}{=} \mathbf{A}\mathbf{W}\mathbf{A}', \ \mathbf{\Sigma} = \mathbf{A}\mathbf{A}', \ \mathbf{W} \sim W_p(m).$$

Thus, we have the fundamental statistical result:

Proposition 7.1 *For* $\mathbf{x}_i$ *i.i.d.* $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, $i = 1, \ldots, n$,

$$\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma}/n), \ (n-1)\mathbf{S} \sim W_p(n-1, \mathbf{\Sigma}), \text{ and } \bar{\mathbf{x}} \perp\!\!\!\perp \mathbf{S}.$$

One may, of course, go to some trouble to obtain an explicit density for the Wishart. However, one needs primarily to understand some of its basic properties and the density will not really reveal very much.

7.3 Properties of Wishart distributions

The distribution of the trace of $\mathbf{W} \sim W_p(m)$ follows almost immediately from the definition.

Proposition 7.2 $\mathbf{W} \sim W_p(m) \Longrightarrow \text{tr } \mathbf{W} \sim \chi^2_{mp}$.

Proof. By definition of $W_p(m)$,

$$\text{tr } \mathbf{W} \stackrel{d}{=} \text{tr } \sum_{i=1}^{m} \mathbf{z}_i\mathbf{z}_i' = \sum_{i=1}^{m} \mathbf{z}_i'\mathbf{z}_i,$$

where $\mathbf{z}_i$ are i.i.d. $N_p(\mathbf{0}, \mathbf{I})$. From Proposition 4.4, $\mathbf{z}_i'\mathbf{z}_i \sim \chi^2_p$. Corollary 3.1 then gives tr $\mathbf{W} \sim \chi^2_{mp}$. □

Now, a useful lemma to determine when $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$ is nonsingular w.p.1. is the following:

Lemma 7.1 $\mathbf{Z} = (z_{ij}) \in \mathbb{R}^n_n$ *with* z_{ij} *i.i.d.* $N(0,1) \Longrightarrow P(|\mathbf{Z}| = 0) = 0$.

Proof. The proof proceeds by induction. The result is true for $n = 1$, as z_{11} has an absolutely continuous distribution. Next, partition

$$\mathbf{Z} = \begin{pmatrix} z_{11} & \mathbf{z}'_{12} \\ \mathbf{z}_{21} & \mathbf{Z}_{22} \end{pmatrix}$$

and assume the result holds for $\mathbf{Z}_{22} \in \mathbb{R}^{n-1}_{n-1}$. Then,

$$\begin{aligned} P(|\mathbf{Z}| = 0) &= P(|\mathbf{Z}| = 0, |\mathbf{Z}_{22}| \neq 0) + P(|\mathbf{Z}| = 0, |\mathbf{Z}_{22}| = 0) \\ &= P(z_{11} = \mathbf{z}'_{12}\mathbf{Z}_{22}^{-1}\mathbf{z}_{21}, |\mathbf{Z}_{22}| \neq 0) \\ &= E\ \left[P(z_{11} = \mathbf{z}'_{12}\mathbf{Z}_{22}^{-1}\mathbf{z}_{21}, |\mathbf{Z}_{22}| \neq 0 \mid \mathbf{z}_{12}, \mathbf{z}_{21}, \mathbf{Z}_{22})\right] \\ &= 0. \end{aligned}$$

□

A slight generalization is contained in

Corollary 7.1 $\mathbf{Z} = (z_{ij}) \in \mathbb{R}^n_n$ *with* z_{ij} *i.i.d.* $N(0,1) \Longrightarrow P(|\mathbf{Z}| = t) = 0,\ \forall t.$

Proof.

$$\begin{aligned} &P(|\mathbf{Z}| = t) \\ &\quad = E\left[P\left(z_{11} = \mathbf{z}'_{12}\mathbf{Z}_{22}^{-1}\mathbf{z}_{21} + \frac{t}{|\mathbf{Z}_{22}|}, |\mathbf{Z}_{22}| \neq 0 \mid \mathbf{z}_{12}, \mathbf{z}_{21}, \mathbf{Z}_{22}\right)\right] \\ &\quad = 0. \end{aligned}$$

□

It should be observed that Lemma 7.1 and Corollary 7.1 remain valid if $\mathbf{Z}$ has any absolutely continuous distribution. We can now prove [Stein (1969), Dykstra (1970)]:

Proposition 7.3 $\mathbf{W} \sim W_p(m),\ m \geq p \Longrightarrow \mathbf{W}$ *is nonsingular w.p.*1.

Proof. The representation $\mathbf{W} \stackrel{d}{=} \mathbf{Z}'\mathbf{Z}$, where $\mathbf{Z}' = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$ and $\mathbf{z}_i$'s are i.i.d. $N_p(\mathbf{0}, \mathbf{I})$, gives

$$\text{rank } \mathbf{W} \stackrel{d}{=} \text{rank } \mathbf{Z}'\mathbf{Z} = \text{rank } \mathbf{Z} \geq \text{rank } (\mathbf{z}_1, \ldots, \mathbf{z}_p) \stackrel{\text{w.p.1}}{=} p$$

whence rank $\mathbf{W} \stackrel{\text{w.p.1}}{=} p$. □

Its corollary gives a condition on the sample size and the population variance for the sample variance matrix $\mathbf{S}$ to be nonsingular w.p.1.

Corollary 7.2 $\mathbf{V} \sim W_p(m, \mathbf{\Sigma}),\ m \geq p,\ |\mathbf{\Sigma}| \neq 0 \Longrightarrow |\mathbf{V}| \neq 0$ *w.p.*1.

Eaton and Perlman (1973) established that the sample variance matrix $\mathbf{S}$ is nonsingular w.p.1 for independent observations, which are not necessarily normal or identically distributed.

Concerning linear transformations of Wishart matrices, we have

Proposition 7.4 $\mathbf{V} \sim W_p(m, \mathbf{\Sigma}),\ \mathbf{B} \in \mathbb{R}_p^q \Longrightarrow \mathbf{BVB}' \sim W_q(m, \mathbf{B\Sigma B}')$.

Proof. Let $\mathbf{W} \sim W_p(m)$. Since $\mathbf{V} \stackrel{d}{=} \mathbf{AWA}'$, for any $\mathbf{AA}' = \mathbf{\Sigma}$, then

$$\mathbf{BVB}' \stackrel{d}{=} (\mathbf{BA})\mathbf{W}(\mathbf{BA})' \sim W_q(m, \mathbf{BAA}'\mathbf{B}').$$

□

Example 7.1 *Suppose* $\mathbf{W} \sim W_p(m)$. *What is* $E\ \mathbf{WAW}$ *for a fixed* $\mathbf{A} \geq \mathbf{0}$? *Since* $\mathbf{HWH}' \stackrel{d}{=} \mathbf{W}$, *for all* $\mathbf{H} \in \mathbf{O}_p$, *we see that*

$$\begin{aligned} \mathbf{WAW} &\stackrel{d}{=} \mathbf{HWH}'\mathbf{AHWH}',\ \forall \mathbf{H} \in \mathbf{O}_p \\ &\stackrel{d}{=} \mathbf{HWDWH}', \end{aligned}$$

where $\mathbf{H}$ *was chosen to diagonalize* $\mathbf{A}$, $\mathbf{H}'\mathbf{AH} = \mathbf{D} = \text{diag}(\lambda_i)$. *Thus,* $E\ \mathbf{WAW} = \mathbf{H}(E\ \mathbf{WDW})\mathbf{H}'$. *But using* $\mathbf{W} \stackrel{d}{=} \sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i$, *where* $\mathbf{z}_i \sim N_p(\mathbf{0}, \mathbf{I})$ *are independent, we find*

$$\begin{aligned} E\ \mathbf{WDW} &= \sum_{i,j} E\ \mathbf{z}_i\mathbf{z}_i'\mathbf{D}\mathbf{z}_j\mathbf{z}_j' \\ &= \sum_i E\ \mathbf{z}_i\mathbf{z}_i'\mathbf{D}\mathbf{z}_i\mathbf{z}_i' + \sum_{i \neq j} E\ \mathbf{z}_i\mathbf{z}_i'\mathbf{D}\mathbf{z}_j\mathbf{z}_j' \\ &= mE\ \mathbf{xx}'\mathbf{Dxx}' + m(m-1)E\ \mathbf{xx}'\mathbf{Dyy}', \end{aligned}$$

where $\mathbf{x}$ *and* $\mathbf{y}$ *are i.i.d.* $N_p(\mathbf{0}, \mathbf{I})$. *However,*

$$\begin{aligned} E\ \mathbf{xx}'\mathbf{Dxx}' &= \sum_i \lambda_i E\ x_i^2\mathbf{xx}' \\ &= \sum_i \lambda_i(\mathbf{I} + 2\mathbf{e}_i\mathbf{e}_i') \\ &= (\text{tr}\ \mathbf{A})\mathbf{I} + 2\mathbf{D} \end{aligned}$$

and

$$\begin{aligned} E\ \mathbf{xx}'\mathbf{Dyy}' &= \sum_i \lambda_i E\ x_iy_i\mathbf{xy}' \\ &= \sum_i \lambda_i\mathbf{e}_i\mathbf{e}_i' \\ &= \mathbf{D}. \end{aligned}$$

Hence,

$$E\ \mathbf{WDW} = m(\text{tr}\ \mathbf{A})\mathbf{I} + m(m+1)\mathbf{D}$$

and, finally, we obtain

$$E\ \mathbf{WAW} = m(\text{tr}\ \mathbf{A})\mathbf{I} + m(m+1)\mathbf{A}.$$

The characteristic function of Wishart distributions also follows from basic principles.

Example 7.2 *The characteristic function of* $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$*, evaluated at* $\mathbf{S}$ *symmetric, is defined by* $c_{\mathbf{V}}(\mathbf{S}) = E\ \exp(i\ \mathrm{tr}\ \mathbf{SV})$*. Write* $\mathbf{V} \stackrel{d}{=} \mathbf{A}\left(\sum_{j=1}^m \mathbf{z}_j\mathbf{z}_j'\right)\mathbf{A}'$*, for any* $\mathbf{AA}' = \mathbf{\Sigma}$*, and diagonalize* $\mathbf{A}'\mathbf{SA} = \mathbf{HDH}'$ *to obtain*

$$\begin{aligned}
c_{\mathbf{V}}(\mathbf{S}) &= E\ \exp\left[i\ \mathrm{tr}\ \mathbf{SA}\sum_{j=1}^m \mathbf{z}_j\mathbf{z}_j'\mathbf{A}'\right] \\
&= E\ \exp\left[i\ \mathrm{tr}\ \mathbf{HDH}'\sum_{j=1}^m \mathbf{z}_j\mathbf{z}_j'\right] \\
&= E\ \exp\left[i\ \mathrm{tr}\ \mathbf{D}\sum_{j=1}^m (\mathbf{H}'\mathbf{z}_j)(\mathbf{H}'\mathbf{z}_j)'\right] \\
&= E\ \exp\left[i\ \mathrm{tr}\ \mathbf{D}\sum_{j=1}^m \mathbf{z}_j\mathbf{z}_j'\right] \ \textit{since}\ \mathbf{H}'\mathbf{z}_j \stackrel{d}{=} \mathbf{z}_j \\
&= E\ \exp\left[i\ \sum_{j=1}^m\sum_{k=1}^p z_{jk}^2 d_k\right],\ \textit{where}\ \mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p) \\
&= \prod_{j=1}^m\prod_{k=1}^p c_{\chi_1^2}(d_k) \\
&= \prod_{j=1}^m\prod_{k=1}^p (1 - 2id_k)^{-1/2} \\
&= |\mathbf{I} - 2i\mathbf{D}|^{-m/2} \\
&= |\mathbf{I} - 2i\mathbf{S\Sigma}|^{-m/2}.
\end{aligned}$$

Hence, the characteristic function is given by

$$c_{\mathbf{V}}(\mathbf{S}) = |\mathbf{I} - 2i\mathbf{\Sigma S}|^{-m/2}.$$

Now, consider some results concerning the marginals of a Wishart. For this reason, partition $\mathbf{V} \in \mathbb{R}_p^p$ as

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

where $\mathbf{V}_{11} \in \mathbb{R}_r^r$ and $\mathbf{V}_{22} \in \mathbb{R}_s^s$, $r + s = p$. The matrix $\mathbf{\Sigma}$ is partitioned similarly.

Proposition 7.5 $\mathbf{V} \sim W_p(m, \mathbf{\Sigma}) \Longrightarrow \mathbf{V}_{11} \sim W_r(m, \mathbf{\Sigma}_{11})$.

Proof. Choose $\mathbf{B} = (\,\mathbf{I}_r \quad \mathbf{0}\,) \in \mathbb{R}_p^r$ in Proposition 7.4. □

Concerning independence, we have:

Proposition 7.6 $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$ *and* $\mathbf{\Sigma}_{12} = \mathbf{0} \Longrightarrow \mathbf{V}_{11} \perp\!\!\!\perp \mathbf{V}_{22}$.

Proof. By the very definition,

$$\mathbf{V} \stackrel{d}{=} \mathbf{U}'\mathbf{U} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{Y}' \end{pmatrix} (\, \mathbf{X} \quad \mathbf{Y} \,) = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix},$$

where $\mathbf{U}' = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and the $\mathbf{u}_i$'s are i.i.d $N_p(\mathbf{0}, \mathbf{\Sigma})$. Then, it suffices to recall (v. Problem 6.4.6) that for a multivariate normal, $\mathbf{\Sigma}_{12} = \mathbf{0}$ implies $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$. □

The previous two propositions are not surprising if we consider their statistical interpretation. First, the distribution of $\mathbf{V}$ is associated with the sample variance based on all p components, whereas that of $\mathbf{V}_{11}$ corresponds to a sample variance but considering only the first r components. Second, if $\mathbf{\Sigma}_{12} = \mathbf{0}$, then the distributions of $\mathbf{V}_{11}$ and $\mathbf{V}_{22}$ are associated with sample variances based on two independent subvectors of dimension r and s, $r + s = p$.

The next proposition, the proof of which is left as an exercise, relates to sums of independently distributed Wishart matrices. It has to do with the way one would pool information from independent samples to estimate the population variance (v. Problem 8.9.1).

Proposition 7.7 *If* $\mathbf{V}_i \stackrel{indep}{\sim} W_p(m_i, \mathbf{\Sigma})$, $i = 1, \ldots, k$, *then*

$$\sum_{i=1}^{k} \mathbf{V}_i \sim W_p\left(\sum_{i=1}^{k} m_i, \mathbf{\Sigma}\right).$$

Lemma 7.2 *Let* $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_r)$, *where the* $\mathbf{h}_i$*'s are orthonormal in* $\mathbb{R}^n$ *and* $\mathbf{Z} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{I}_p)$. *Then,*

1. $\mathbf{H}'\mathbf{Z} \sim N_p^r(\mathbf{0}, \mathbf{I}_r \otimes \mathbf{I}_p)$,

2. $\mathbf{Z}'\mathbf{H}\mathbf{H}'\mathbf{Z} \sim W_p(r)$.

Proof. Using Proposition 6.1, $\mathbf{H}'\mathbf{Z} \sim N_p^r(\mathbf{0}, (\mathbf{H}'\mathbf{H}) \otimes \mathbf{I}_p)$. This proves part 1 because $\mathbf{H}'\mathbf{H} = \mathbf{I}_r$. Since $\mathbf{Z}'\mathbf{h}_i$, $i = 1, \ldots, r$, are i.i.d. $N_p(\mathbf{0}, \mathbf{I})$, part 2 follows from the Wishart definition: $\mathbf{Z}'\mathbf{H}\mathbf{H}'\mathbf{Z} = \sum_{i=1}^{r} (\mathbf{Z}'\mathbf{h}_i)(\mathbf{Z}'\mathbf{h}_i)' \sim W_p(r)$. □

Proposition 7.8 *Let*

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma}).$$

If $\mathcal{V} \subset \mathbb{R}^n$ *is a linear subspace,* $\dim\ \mathcal{V} = r$, *and* $\mathbf{P}$ *is the orthogonal projection on* $\mathcal{V}$, *then* $\mathbf{X}'\mathbf{P}\mathbf{X} \sim W_p(r, \mathbf{\Sigma})$.

Proof. Choose an orthonormal basis $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_r)$ for $\mathcal{V}$ and observe that $\mathbf{P} = \mathbf{H}\mathbf{H}'$ and $r = \text{rank } \mathbf{P} = \dim\ \mathcal{V}$. Write $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}'$ and $\mathbf{X} \stackrel{d}{=} \mathbf{Z}\mathbf{A}'$, where $\mathbf{Z} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{I}_p)$. Therefore, using Lemma 7.2,

$$\begin{aligned} \mathbf{X}'\mathbf{P}\mathbf{X} &= \mathbf{X}'\mathbf{H}\mathbf{H}'\mathbf{X} \\ &\stackrel{d}{=} \mathbf{A}\mathbf{Z}'\mathbf{H}\mathbf{H}'\mathbf{Z}\mathbf{A}' \\ &\stackrel{d}{=} \mathbf{A}\mathbf{W}\mathbf{A}', \ \textit{where } \mathbf{W} \sim W_p(r). \end{aligned}$$

Hence, $\mathbf{X}'\mathbf{P}\mathbf{X} \sim W_p(r, \mathbf{\Sigma})$. □

General results on Wishart and chi-square distributions associated with matrix quadratic forms are available in Mathew and Nordström (1997). A fundamental result on marginals useful in the sequel is now stated, but first recall the notation $\mathbf{V}_{11.2} = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$, where $\mathbf{V}$ was partitioned as on page 90.

Proposition 7.9 *If* $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, $m \geq p$, $\mathbf{\Sigma} > \mathbf{0}$, *then*

$$\begin{aligned} \mathbf{V}_{11.2} &\sim W_r(m - s, \mathbf{\Sigma}_{11.2}), \\ \mathbf{V}_{21} \mid \mathbf{V}_{22} &\sim N_r^s(\mathbf{V}_{22}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}, \mathbf{V}_{22} \otimes \mathbf{\Sigma}_{11.2}), \\ \mathbf{V}_{22} &\sim W_s(m, \mathbf{\Sigma}_{22}), \end{aligned}$$

and $\mathbf{V}_{11.2} \perp\!\!\!\perp (\mathbf{V}_{21}, \mathbf{V}_{22})$.

Proof. As before, write

$$\mathbf{V} \stackrel{d}{=} \mathbf{U}'\mathbf{U} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{Y}' \end{pmatrix} (\, \mathbf{X} \quad \mathbf{Y}\,) = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix},$$

where $\mathbf{U} \sim N_p^m(\mathbf{0}, \mathbf{I}_m \otimes \mathbf{\Sigma})$. Thus, $\mathbf{X} \mid \mathbf{Y} \sim N_r^m(\mathbf{Y}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}, \mathbf{I}_m \otimes \mathbf{\Sigma}_{11.2})$ (v. Problem 6.4.6). Let $\mathbf{P} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'$ be the orthogonal projection on the column space of $\mathbf{Y}$ and $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, rank $\mathbf{Q} = m - s$. It is clear, since $\mathbf{Y} = \mathbf{P}\mathbf{Y}$, that

$$\begin{aligned} \mathbf{V}_{11.2} &= \mathbf{X}'[\mathbf{I} - \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}']\mathbf{X} = (\mathbf{Q}\mathbf{X})'(\mathbf{Q}\mathbf{X}), \\ \mathbf{V}_{21} &= \mathbf{Y}'\mathbf{X} = (\mathbf{P}\mathbf{Y})'(\mathbf{P}\mathbf{X}), \\ \mathbf{V}_{22} &= \mathbf{Y}'\mathbf{Y} = (\mathbf{P}\mathbf{Y})'(\mathbf{P}\mathbf{Y}). \end{aligned}$$

Since $\mathbf{Y}'\mathbf{X} \mid \mathbf{Y} \sim N_r^s((\mathbf{Y}'\mathbf{Y})\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}, (\mathbf{Y}'\mathbf{Y}) \otimes \mathbf{\Sigma}_{11.2})$ depends only on $\mathbf{Y}'\mathbf{Y}$, then $\mathbf{V}_{21} \mid \mathbf{V}_{22} \sim N_r^s(\mathbf{V}_{22}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}, \mathbf{V}_{22} \otimes \mathbf{\Sigma}_{11.2})$. From Proposition 7.8 and $\mathbf{Q}\mathbf{Y} = \mathbf{0}$, $\mathbf{V}_{11.2} \mid \mathbf{Y} \sim W_r(m - s, \mathbf{\Sigma}_{11.2})$, which does not depend on $\mathbf{Y}$; hence, $\mathbf{V}_{11.2} \perp\!\!\!\perp \mathbf{Y}$ and $\mathbf{V}_{11.2} \sim W_r(m - s, \mathbf{\Sigma}_{11.2})$, unconditionally. It is clear $\mathbf{V}_{22} \sim W_s(m, \mathbf{\Sigma}_{22})$. Only independence remains to be shown. However, conditionally on $\mathbf{Y}$, $\mathbf{P}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}\mathbf{X}$. To see this, note $\mathbf{P}\mathbf{Q} = \mathbf{0}$ and

$$\begin{aligned} \text{var}\left[\begin{pmatrix} \mathbf{P} \\ \mathbf{Q} \end{pmatrix} \mathbf{X} \mid \mathbf{Y}\right] &= \left[\begin{pmatrix} \mathbf{P} \\ \mathbf{Q} \end{pmatrix} (\mathbf{P}', \mathbf{Q}')\right] \otimes \mathbf{\Sigma}_{11.2} \\ &= \begin{pmatrix} \mathbf{P} \otimes \mathbf{\Sigma}_{11.2} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \otimes \mathbf{\Sigma}_{11.2} \end{pmatrix}. \end{aligned}$$

Hence, given $\mathbf{Y}$, $\mathbf{V}_{11.2} \perp\!\!\!\perp \mathbf{V}_{21}$. Finally, using Proposition 2.13,

$$\begin{aligned} E\left[f(\mathbf{V}_{11.2}) \cdot g(\mathbf{V}_{21}, \mathbf{V}_{22})\right] &= E\left[E\ f(\mathbf{V}_{11.2})\ g(\mathbf{V}_{21}, \mathbf{V}_{22}) \mid \mathbf{Y}\right] \\ &= E\left\{E\ \left[f(\mathbf{V}_{11.2}) \mid \mathbf{Y}\right]\ E\left[g(\mathbf{V}_{21}, \mathbf{V}_{22}) \mid \mathbf{Y}\right]\right\} \\ &= E\left\{Ef(\mathbf{V}_{11.2})\ E\left[g(\mathbf{V}_{21}, \mathbf{V}_{22}) \mid \mathbf{Y}\right]\right\} \\ &= Ef(\mathbf{V}_{11.2}) \cdot Eg(\mathbf{V}_{21}, \mathbf{V}_{22}), \end{aligned}$$

which proves independence. □

Proposition 7.9 with $r = 1$ and $s = p-1$ can be used to prove inductively several results concerning Wishart distributions. Here are two corollaries: the distribution of the *generalized variance*, $|\mathbf{V}|$, and the Wishart density.

Corollary 7.3 *If* $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, $m \geq p$, $\mathbf{\Sigma} > \mathbf{0}$, *then*

$$|\mathbf{V}| \sim |\mathbf{\Sigma}| \prod_{i=1}^{p} \chi^2_{m-p+i};$$

i.e., $|\mathbf{V}|/|\mathbf{\Sigma}|$ *is distributed as a product of* p *mutually independent chi-square variables.*

Proof. The result obviously holds for $p = 1$. Assume it holds for $p-1$. Let $r = 1$ and $s = p-1$ in Proposition 7.9. Then, $|\mathbf{V}| = v_{11.2}|\mathbf{V}_{22}|$, where $v_{11.2} \sim \sigma_{11.2}\chi^2_{m-p+1}$, $\mathbf{V}_{22} \sim W_{p-1}(m, \mathbf{\Sigma}_{22})$, and $v_{11.2} \perp\!\!\!\perp \mathbf{V}_{22}$. From the induction hypothesis,

$$|\mathbf{V}_{22}| \sim |\mathbf{\Sigma}_{22}| \prod_{i=1}^{p-1} \chi^2_{m-p+1+i} \stackrel{d}{=} |\mathbf{\Sigma}_{22}| \prod_{i=2}^{p} \chi^2_{m-p+i}$$

and the conclusion follows. □

Corollary 7.4 *If* $\mathbf{W} \sim W_p(m)$, $m \geq p$, *then the p.d.f. of* $\mathbf{W}$ *is*

$$f_{\mathbf{W}}(\mathbf{W}) = \frac{1}{2^{mp/2}\Gamma_p\left(\frac{1}{2}m\right)} |\mathbf{W}|^{(m-p-1)/2}\ \mathrm{etr}\left(-\tfrac{1}{2}\mathbf{W}\right),\ \mathbf{W} > \mathbf{0}, \qquad (7.3)$$

where $\Gamma_p(u) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left[u - \frac{1}{2}(i-1)\right]$, $u > \frac{1}{2}(p-1)$.

Proof. The result holds for $p = 1$, as the density reduces to a chi-square density. Let $r = 1$ and $s = p-1$ in Proposition 7.9, then

$$\begin{aligned} w_{11.2} &\sim \chi^2_{m-p+1}, \\ \mathbf{w}_{21} \mid \mathbf{W}_{22} &\sim N_{p-1}(\mathbf{0}, \mathbf{W}_{22}), \\ \mathbf{W}_{22} &\sim W_{p-1}(m). \end{aligned}$$

Thus, the joint p.d.f. of $(w_{11.2}, \mathbf{w}_{21}, \mathbf{W}_{22})$ is

$$\begin{aligned} &\frac{1}{2^{(m-p+1)/2}\Gamma[\frac{1}{2}(m-p+1)]} w_{11.2}^{(m-p+1)/2-1} \exp(-\tfrac{1}{2}w_{11.2}) \\ &\cdot\ (2\pi)^{-(p-1)/2} |\mathbf{W}_{22}|^{-1/2} \exp(-\tfrac{1}{2}\mathbf{w}_{21}'\mathbf{W}_{22}^{-1}\mathbf{w}_{21}) \end{aligned}$$

$$\cdot \quad \frac{1}{2^{m(p-1)/2}\Gamma_{p-1}(\frac{1}{2}m)}|\mathbf{W}_{22}|^{(m-p)/2}\text{etr}(-\tfrac{1}{2}\mathbf{W}_{22}).$$

Make the change of variables

$$(w_{11.2}, \mathbf{w}_{21}, \mathbf{W}_{22}) \mapsto (w_{11}, \mathbf{w}_{21}, \mathbf{W}_{22})$$

with jacobian $J(w_{11.2}, \mathbf{w}_{21}, \mathbf{W}_{22} \to w_{11}, \mathbf{w}_{21}, \mathbf{W}_{22}) = J(w_{11.2} \to w_{11}) = 1$ while using the relations $|\mathbf{W}| = w_{11.2}|\mathbf{W}_{22}|$ and $w_{11.2} = w_{11} - \mathbf{w}_{21}'\mathbf{W}_{22}^{-1}\mathbf{w}_{21}$ to get the result. □

The reader should check that $\Gamma_p(u)$ is a *generalized gamma function* in the sense that $\Gamma_1(u) = \Gamma(u)$, $u > 0$. The density of $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, $\mathbf{\Sigma} > \mathbf{0}$, $m \geq p$, follows directly from the transformation $\mathbf{V} = \mathbf{AWA}'$, for any $\mathbf{AA}' = \mathbf{\Sigma}$, and the jacobian in Proposition 2.19 (v. Problem 7.5.7). James (1954) and Olkin and Roy (1954) proposed a constructive proof by jacobians of transformations on k-surfaces (manifolds). It requires a knowledge of differential forms and integration on k-surfaces which goes beyond the scope of this book. The theory of singular Wishart distributions ($m < p$) is available in Uhlig (1994).

The function (7.3) is a density function even when the number of degrees of freedom $m \in \mathbb{R}$, possibly noninteger, satisfies $m > p - 1$ [Muirhead (1982), p. 62].

7.4 Box-Cox transformations

A method [Andrews et al. (1971)] that is an extension of the technique of Box and Cox (1964) is described for obtaining data-based transformations of multivariate observations to enhance the normality of their distribution. Specifically, power transformations of the original variables are estimated to effect both marginal and joint normality. The likelihood method, used by Box and Cox (1964) for the univariate problem, is the one adopted here for the multivariate case. The simple family of power transformations defined by

$$x_j^{(\lambda_j)} = \begin{cases} (x_j^{\lambda_j} - 1)/\lambda_j, & \lambda_j \neq 0, \\ \ln x_j, & \lambda_j = 0, \end{cases}$$

$j = 1, \ldots, p$, will be considered. Each variable x_j must be non-negative, otherwise, with a known lower bound, we may add a constant sufficiently large, a_j, and consider $x_j + a_j$ as the original variable. Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = (x_{ij}) \in \mathbb{R}_p^n,$$

$$\mathbf{X}^{(\boldsymbol{\lambda})} = \begin{pmatrix} \mathbf{x}_1^{(\boldsymbol{\lambda})'} \\ \vdots \\ \mathbf{x}_n^{(\boldsymbol{\lambda})'} \end{pmatrix} = (x_{ij}^{(\lambda_j)}) \in \mathbb{R}_p^n,$$

be the sample matrices of the original and transformed data, respectively, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ is the unknown vector of power transformation parameters. If $\boldsymbol{\lambda}$ is the vector of parameters yielding joint normality, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the density of $\mathbf{X}^{(\boldsymbol{\lambda})}$ is from (7.1):

$$f(\mathbf{X}^{(\boldsymbol{\lambda})}) = (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \cdot \operatorname{etr}\left\{-\tfrac{1}{2}\left[\mathbf{V}^{(\boldsymbol{\lambda})} + n(\bar{\mathbf{x}}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})(\bar{\mathbf{x}}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\right]\boldsymbol{\Sigma}^{-1}\right\},$$

where

$$\bar{\mathbf{x}}^{(\boldsymbol{\lambda})} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^{(\boldsymbol{\lambda})},$$

$$\mathbf{V}^{(\boldsymbol{\lambda})} = \sum_{i=1}^{n} (\mathbf{x}_i^{(\boldsymbol{\lambda})} - \bar{\mathbf{x}}^{(\boldsymbol{\lambda})})(\mathbf{x}_i^{(\boldsymbol{\lambda})} - \bar{\mathbf{x}}^{(\boldsymbol{\lambda})})'.$$

The jacobian of the transformation, $J(\mathbf{X}^{(\boldsymbol{\lambda})} \to \mathbf{X})$, is

$$J = \prod_{j=1}^{p}\prod_{i=1}^{n} x_{ij}^{\lambda_j - 1}.$$

Hence, the density of the genuine data $\mathbf{X}$ is

$$\begin{aligned} f(\mathbf{X}) &= f(\mathbf{X}^{(\boldsymbol{\lambda})}) \cdot J \\ &= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \operatorname{etr}\left\{-\tfrac{1}{2}\left[\mathbf{V}^{(\boldsymbol{\lambda})} + n(\bar{\mathbf{x}}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})(\bar{\mathbf{x}}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\right]\boldsymbol{\Sigma}^{-1}\right\} \cdot J. \end{aligned}$$

The log-likelihood of $(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ is, up to an additive constant,

$$\begin{aligned} l(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = &-\frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\operatorname{tr}(\mathbf{V}^{(\boldsymbol{\lambda})}\boldsymbol{\Sigma}^{-1}) \\ &-\frac{n}{2}(\bar{\mathbf{x}}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(\boldsymbol{\lambda})} - \boldsymbol{\mu}) + \sum_{j=1}^{p}(\lambda_j - 1)\sum_{i=1}^{n}\ln x_{ij}. \end{aligned} \tag{7.4}$$

For a specified $\boldsymbol{\lambda}$, the maximum likelihood estimate of $(\boldsymbol{\Sigma}, \boldsymbol{\mu})$, exactly as for (7.2), is given by

$$\left(\mathbf{V}^{(\boldsymbol{\lambda})}/n, \bar{\mathbf{x}}^{(\boldsymbol{\lambda})}\right).$$

If these estimates are substituted in (7.4), the maximized log-likelihood function is, up to an additive constant,

$$l_{\max}(\boldsymbol{\lambda}) = -\frac{n}{2}\ln|\mathbf{V}^{(\boldsymbol{\lambda})}| + \sum_{j=1}^{p}(\lambda_j - 1)\sum_{i=1}^{n}\ln x_{ij}, \tag{7.5}$$

a function of p parameters which can be computed and studied. The maximum likelihood estimate $\hat{\boldsymbol{\lambda}}$ may be obtained by numerically maximizing (7.5). Also, confidence regions for $\boldsymbol{\lambda}$ may be obtained. One such $(1-\alpha)100\%$ confidence region for $\boldsymbol{\lambda}$ based on asymptotic considerations [Fraser (1976), p. 357] is

$$\{\boldsymbol{\lambda}:\ l_{\max}(\hat{\boldsymbol{\lambda}}) - l_{\max}(\boldsymbol{\lambda}) \leq \tfrac{1}{2}\chi^2_{1-\alpha,p}\},$$

where $\chi^2_{1-\alpha,p}$ is the $(1-\alpha)$-quantile of a χ^2_p distribution. The likelihood criterion used here specifies joint normality rather than marginal normality as the goal of the transformation.

7.5 Problems

1. The maximum likelihood estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ were derived by minimizing

$$\ln|\boldsymbol{\Sigma}| + \operatorname{tr}\frac{1}{n}\mathbf{V}\boldsymbol{\Sigma}^{-1} + (\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}).$$

   Derive the MLE, this time by calculus, using the vector and matrix differentiation rules of Problems 1.8.9-1.8.10.

2. Prove Proposition 7.7.

3. Show that if $\mathbf{V} \sim W_p(m, \boldsymbol{\Sigma})$, then:
   (i) $\operatorname{var}\mathbf{V} = m[\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma} + (\boldsymbol{\sigma}_j\boldsymbol{\sigma}_i')] = m(\mathbf{I}+\mathbf{K}_p)(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})$ where $\mathbf{K}_p$ is the "commutation matrix."
   (ii) Prove Proposition 7.7 again, but using characteristic functions this time.

4. Let $\mathbf{W} \sim W_p(m)$, $m \geq p$. Prove:
   (i) $1/(\mathbf{t}'\mathbf{W}^{-1}\mathbf{t}) \sim \chi^2_{m-p+1}$, for any $\mathbf{t}$, $|\mathbf{t}| = 1$.
   (ii) If $\mathbf{x} \perp\!\!\!\perp \mathbf{W}$ and $p_{\mathbf{x}}(\mathbf{0}) = 0$, then $\mathbf{x}$ is independent of

$$\frac{(\mathbf{x}'\mathbf{x})}{(\mathbf{x}'\mathbf{W}^{-1}\mathbf{x})} \sim \chi^2_{m-p+1}.$$

   Hint: $\mathbf{HWH}' \stackrel{d}{=} \mathbf{W}$, $\forall \mathbf{H} \in \mathbf{O}_p$.

5. Assume $\mathbf{W} \sim W_p(m)$, $m \geq p$, and $\mathbf{A} \geq \mathbf{0}$. Prove:
   (i) $E\ \mathbf{W} = m\mathbf{I}$,
   (ii) $E\ \mathbf{W}^{-1} = \mathbf{I}/(m-p-1)$.

6. **Moments of generalized variance.**
   (i) Let $\mathbf{W} \sim W_p(m)$, $m \geq p$. Prove

$$E\ |\mathbf{W}|^h = 2^{ph}\frac{\Gamma_p(\frac{1}{2}m+h)}{\Gamma_p(\frac{1}{2}m)},\ h > \tfrac{1}{2}(p-m-1).$$

Hint: $E\ |\mathbf{W}|^h$ has an integrand of the form of a $W_p(m+2h)$ density. Use the normalizing constant $c_{p,m} = [2^{mp/2}\Gamma_p(\frac{1}{2}m)]^{-1}$.

(ii) If $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, $m \geq p$, $\mathbf{\Sigma} > \mathbf{0}$, then

$$E\ |\mathbf{V}|^h = |\mathbf{\Sigma}|^h 2^{ph} \frac{\Gamma_p(\frac{1}{2}m + h)}{\Gamma_p(\frac{1}{2}m)},\ h > \tfrac{1}{2}(p - m - 1).$$

7. **Wishart density.**
Obtain the p.d.f. of $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, $m \geq p$, $\mathbf{\Sigma} > \mathbf{0}$:

$$f_{\mathbf{V}}(\mathbf{V}) = \frac{1}{2^{mp/2}\Gamma_p\left(\frac{1}{2}m\right)|\mathbf{\Sigma}|^{m/2}} |\mathbf{V}|^{(m-p-1)/2}\ \text{etr}\left(-\tfrac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{V}\right),$$

$\mathbf{V} > \mathbf{0}$.

8. Let $\mathbf{W} \sim W_p(m)$ and consider the *correlation matrix* $\mathbf{R} = (r_{ij})$, where

$$r_{ij} = \frac{w_{ij}}{w_{ii}^{1/2} w_{jj}^{1/2}}.$$

Demonstrate that the density of $\mathbf{R}$ is

$$f(\mathbf{R}) = \frac{[\Gamma(\frac{1}{2}m)]^p}{\Gamma_p(\frac{1}{2}m)} |\mathbf{R}|^{(m-p-1)/2}.$$

Hint: Use the transformation $\mathbf{W} \mapsto w_{11}, \ldots, w_{pp}, \mathbf{R}$.

9. **Inverted Wishart distribution.**
Derive the p.d.f. of $\mathbf{U} = \mathbf{V}^{-1}$, where $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, $m \geq p$, $\mathbf{\Sigma} > \mathbf{0}$:

$$f_{\mathbf{U}}(\mathbf{U}) = \frac{1}{2^{mp/2}\Gamma_p\left(\frac{1}{2}m\right)|\mathbf{\Sigma}|^{m/2}} |\mathbf{U}|^{-(m+p+1)/2} \text{etr}\left(-\tfrac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{U}^{-1}\right),$$

$\mathbf{U} > \mathbf{0}$.

10. Assume $\mathbf{W} \sim W_p(m)$ and define $\mathbf{W} = \mathbf{T}\mathbf{T}'$ for a unique $\mathbf{T} \in \mathbf{L}_p^+$.
(i) Prove $t_{ij} \sim N(0,1)$, $1 \leq i < j \leq p$, and $t_{ii}^2 \sim \chi_{m-i+1}^2$, $1 \leq i \leq p$, are all mutually independent.
(ii) Using (i) prove tr $\mathbf{W} \sim \chi_{pm}^2$.
(iii) Using (i) again, prove that if $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, $m \geq p$, $\mathbf{\Sigma} > \mathbf{0}$, then $|\mathbf{V}| \sim |\mathbf{\Sigma}| \prod_{i=1}^p \chi_{m-p+i}^2$.

8
Tests on mean and variance

8.1 Introduction

Having laid the distribution of $(\mathbf{S}, \bar{\mathbf{x}})$ on a good footing in Chapter 7, we now present inference problems such as the Hotelling-T^2 test on the mean vector, the simultaneous confidence intervals on means, the inference about multiple and partial correlation coefficients, the test of sphericity, and the test of equality of variances. In some cases, the tests are optimal in some sense. This is the case of the Hotelling-T^2 test and the test of multiple correlation, which are shown to be uniformly most powerful invariant (UMPI). The asymptotic distribution of eigenvalues, both in the one-sample and two-sample cases, is treated in Section 8.8. Tables of critical points with references to applications for most multivariate tests are available in Kres (1983). The approach adopted here rests mainly on likelihood ratio tests, although other general and valid testing procedures based on minimization of divergence measures exist in the literature [Wakaki et al. (1990)].

8.2 Hotelling-T^2

Now, assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. $\mathbf{x}$ with $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} > \mathbf{0}$. The properties of Wishart distributions in Chapter 7 provide an easy way to obtain the distribution of the Hotelling-T^2 statistic

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \tag{8.1}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the unbiased estimate for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This is needed to test the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against all alternatives or to build a confidence ellipsoid for $\boldsymbol{\mu}$. In fact, the following proposition shows that the Hotelling-T^2 statistic is a monotone function of the likelihood ratio test (LRT) statistic. As usual, let

$$\mathbf{V} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

be the matrix of sums of squares and cross-products.

Proposition 8.1 *The likelihood ratio statistic for* $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ *against* $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ *is*

$$\Lambda = \left(1 + \frac{1}{(n-1)} T^2\right)^{-n/2}.$$

Proof. The unrestricted MLE of $(\boldsymbol{\Sigma}, \boldsymbol{\mu})$ is $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{V}$. However, the MLE of $\boldsymbol{\Sigma}$ under H_1 is obtained from (7.2) by minimizing

$$\begin{aligned} &\ln|\boldsymbol{\Sigma}| + \operatorname{tr} \frac{1}{n}\mathbf{V}\boldsymbol{\Sigma}^{-1} + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \\ &\quad = \ln|\boldsymbol{\Sigma}| + \operatorname{tr} \frac{1}{n}[\mathbf{V} + n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)']\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

Using the same technique as on page 86 we find

$$\begin{aligned} \hat{\hat{\boldsymbol{\Sigma}}} &= \frac{1}{n}[\mathbf{V} + n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'] \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_0)(\mathbf{x}_i - \boldsymbol{\mu}_0)'. \end{aligned}$$

Thus, with (7.1), the LRT becomes

$$\begin{aligned} \Lambda &= \frac{L(\hat{\hat{\boldsymbol{\Sigma}}}, \boldsymbol{\mu}_0)}{L(\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}})} \\ &= \frac{|\hat{\hat{\boldsymbol{\Sigma}}}|^{-n/2} \exp(-\frac{1}{2}np)}{|\hat{\boldsymbol{\Sigma}}|^{-n/2} \exp(-\frac{1}{2}np)} \\ &= \frac{|\frac{1}{n}[\mathbf{V} + n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)']|^{-n/2}}{|\frac{1}{n}\mathbf{V}|^{-n/2}} \\ &= |\mathbf{I} + n\mathbf{V}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'|^{-n/2} \\ &= \left(1 + \frac{1}{(n-1)} T^2\right)^{-n/2}, \end{aligned}$$

where the last equality made use of Problem 1.8.3. □

The distribution of T^2 is a direct consequence of the following proposition.

Proposition 8.2 *If* $\mathbf{z} \sim N_p(\boldsymbol{\delta}, \mathbf{I})$, $\mathbf{W} \sim W_p(m)$, $m \geq p$, *and* $\mathbf{z} \perp\!\!\!\perp \mathbf{W}$, *then*

$$\mathbf{z}'\mathbf{W}^{-1}\mathbf{z} \sim F_c(p, m-p+1; \boldsymbol{\delta}'\boldsymbol{\delta}/2).$$

Proof. Using an orthogonal transformation $\mathbf{H} = (\mathbf{z}/|\mathbf{z}|, \boldsymbol{\Gamma})' \in \mathbf{O}_p$, we get immediately

$$\begin{aligned} \mathbf{z}'\mathbf{W}^{-1}\mathbf{z} &= (\mathbf{H}\mathbf{z})'(\mathbf{H}\mathbf{W}\mathbf{H}')^{-1}(\mathbf{H}\mathbf{z}) \\ &= |\mathbf{z}|^2 \mathbf{e}_1' \mathbf{V}^{-1} \mathbf{e}_1, \end{aligned}$$

where $\mathbf{V} = \mathbf{H}\mathbf{W}\mathbf{H}'$. Since the conditional distribution $\mathbf{V} \mid \mathbf{z} \sim W_p(m)$ does not depend on $\mathbf{z}$, then $\mathbf{V} \sim W_p(m)$, unconditionally, and $\mathbf{V} \perp\!\!\!\perp \mathbf{z}$. Letting

$$\mathbf{V}^{-1} = \begin{pmatrix} v^{11} & \mathbf{v}^{21\prime} \\ \mathbf{v}^{21} & \mathbf{V}^{22} \end{pmatrix},$$

$$\mathbf{z}'\mathbf{W}^{-1}\mathbf{z} = |\mathbf{z}|^2 v^{11} = \mathbf{z}'\mathbf{z}/v_{11.2},$$

where the last equality made use of Problem 1.8.1. The conclusion follows since $\mathbf{z}'\mathbf{z} \sim \chi^2_p(\boldsymbol{\delta}'\boldsymbol{\delta}/2)$ and, by Proposition 7.9, $v_{11.2} \sim \chi^2_{m-p+1}$. □

As a corollary, we obtain the distribution of Hotelling-T^2.

Corollary 8.1 *The non-null distribution of* T^2 *for* $n \geq p+1$ *is*

$$T^2/(n-1) \sim F_c(p, n-p; \delta), \text{ with } \delta = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)/2.$$

Proof. In terms of the sample matrix, as on page 86, $\mathbf{X} \stackrel{d}{=} \mathbf{Z}\mathbf{A}' + \mathbf{1}\boldsymbol{\mu}'$, where $\mathbf{Z} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{I}_p)$ and $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$, and, thus, $(\bar{\mathbf{x}}, \mathbf{S}_\mathbf{x}) \stackrel{d}{=} (\mathbf{A}\bar{\mathbf{z}} + \boldsymbol{\mu}, \mathbf{A}\mathbf{S}_\mathbf{z}\mathbf{A}')$. Therefore,

$$\begin{aligned} \frac{T^2}{(n-1)} &= n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'[(n-1)\mathbf{S}]^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \\ &\stackrel{d}{=} n\left(\mathbf{A}\bar{\mathbf{z}} + \boldsymbol{\mu} - \boldsymbol{\mu}_0\right)' \left[(n-1)\mathbf{A}\mathbf{S}_\mathbf{z}\mathbf{A}'\right]^{-1} \left(\mathbf{A}\bar{\mathbf{z}} + \boldsymbol{\mu} - \boldsymbol{\mu}_0\right) \\ &= n\left[\bar{\mathbf{z}} + \mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right]' \left[(n-1)\mathbf{S}_\mathbf{z}\right]^{-1} \left[\bar{\mathbf{z}} + \mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right]. \end{aligned}$$

The proof follows from Proposition 8.2, as

$$\begin{aligned} n^{1/2}\left[\bar{\mathbf{z}} + \mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right] &\sim N_p\left(n^{1/2}\mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0), \mathbf{I}\right) \\ (n-1)\mathbf{S}_\mathbf{z} &\sim W_p(n-1) \end{aligned}$$

are independent. □

Example 8.1 *The power function of an* α *significance level Hotelling-*T^2 *test may now be evaluated as a function of*

$$\delta = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)/2$$

in the following manner:

$$\beta = P\left(F_c(p, n-p; \delta) \geq t_\alpha\right),$$

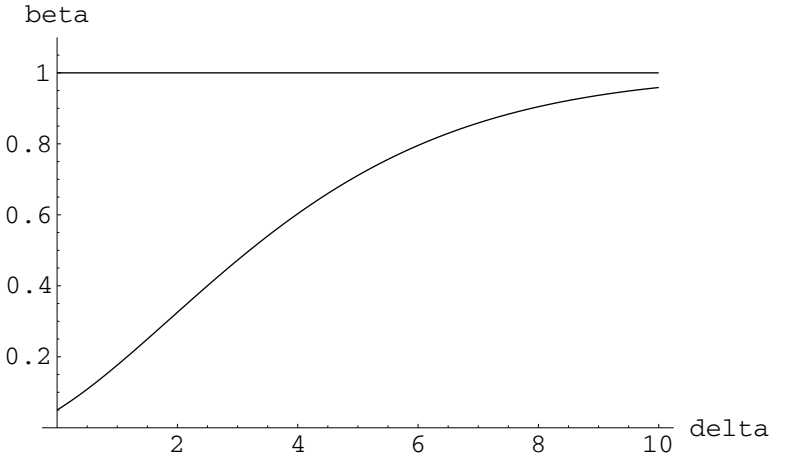

Figure 8.1. Power function of Hotelling-T^2 when $p = 3$ and $n = 40$ at a level of significance $\alpha = 0.05$.

where $t_\alpha = [p/(n-p)]F_\alpha(p, n-p)$ *is the critical point. Proposition* 4.7 *and Problem* 3.5.6 *yields*

$$\begin{aligned}\beta &= \sum_{k=0}^{\infty} e^{-\delta}\frac{\delta^k}{k!}P\left(F_c(p+2k, n-p) \geq t_\alpha\right) \\ &= \sum_{k=0}^{\infty} e^{-\delta}\frac{\delta^k}{k!}\int_{t_\alpha}^{\infty} B\left(\tfrac{1}{2}(p+2k), \tfrac{1}{2}(n-p)\right)^{-1}\frac{F^{(p+2k)/2-1}}{(1+F)^{(n+2k)/2}}dF.\end{aligned}$$

Numerical evaluation in Mathematica for $p = 3$, $n = 40$ *and* $\alpha = 0.05$ *produced the plot in Figure* 8.1.

The robustness of Hotelling-T^2 is easily established. Without normality, assuming $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. $\mathbf{x}$, $E\ \mathbf{x} = \boldsymbol{\mu}_0$ and var $\mathbf{x} = \boldsymbol{\Sigma}$, the asymptotic distributions in Section 6.3 gave $T^2 \xrightarrow{d} \chi^2_p$. This asymptotic distribution is the same regardless of the underlying distribution of $\mathbf{x}$.

A different situation arises in presence of "contamination." Assume the simple situation $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ are i.i.d. $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, but there is one (or more) contaminated observation $\mathbf{x}_n \sim N_p(\boldsymbol{\mu}_0 + \boldsymbol{\gamma}, \boldsymbol{\Sigma})$. We assume $\boldsymbol{\Sigma}$ known for the sake of simplicity. It is easily checked that $n^{1/2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim N_p(\boldsymbol{\gamma}/n^{1/2}, \boldsymbol{\Sigma})$ and, thus, from Corollary 5.1,

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim \chi^2_p(\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}/2n).$$

Since $P(\chi^2_p(\delta) \geq c)$ is monotone increasing in δ [Ghosh (1970), p. 302], all other parameters being fixed, it follows that T^2 will reject $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ with probability converging to 1 as $|\boldsymbol{\gamma}| \to \infty$ (for fixed n) even though all observations, but one, have mean $\boldsymbol{\mu}_0$. A procedure which is insensitive to

contamination of the data consists of building an Hotelling-T^2 test

$$T^2 = n(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)$$

from a robust estimate $(\boldsymbol{\Sigma}_n, \boldsymbol{\mu}_n)$ such as an M-estimate or S-estimate. These truly robust tests are studied in Chapter 13.

We end this section with a discussion of invariant tests on the mean vector. Consider the canonical problem of testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$. The group of transformations $\mathbf{G}_p$ acts on the observations as $\mathbf{x}_i \mapsto \mathbf{A}\mathbf{x}_i$, where $\mathbf{A} \in \mathbf{G}_p$. This transformation induces the following transformations on the minimal sufficient statistic $(\bar{\mathbf{x}}, \mathbf{S})$ and parameters, $\bar{\mathbf{x}} \mapsto \mathbf{A}\bar{\mathbf{x}}$, $\mathbf{S} \mapsto \mathbf{A}\mathbf{S}\mathbf{A}'$, and $\boldsymbol{\mu} \mapsto \mathbf{A}\boldsymbol{\mu}$, $\boldsymbol{\Sigma} \mapsto \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$. Note that the hypotheses are preserved because $\boldsymbol{\mu} = \mathbf{0}$ iff $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$, for any $\mathbf{A} \in \mathbf{G}_p$. We define a test function $f(\bar{\mathbf{x}}, \mathbf{S})$ to be invariant iff it yields the same value on the original as on the transformed data, i.e.,

$$f(\mathbf{y}, \mathbf{W}) = f(\mathbf{A}\mathbf{y}, \mathbf{A}\mathbf{W}\mathbf{A}'),\ \forall \mathbf{A} \in \mathbf{G}_p, \forall (\mathbf{y}, \mathbf{W}) \in \mathbb{R}^p \times \mathcal{P}_p.$$

This has important implications. First, the choice $\mathbf{A} = \mathbf{S}^{-1/2}$ yields

$$f(\bar{\mathbf{x}}, \mathbf{S}) = f(\mathbf{S}^{-1/2}\bar{\mathbf{x}}, \mathbf{I}).$$

Now, there exists an orthogonal transformation $\mathbf{H} \in \mathbf{O}_p$ (v. Problem 1.8.14) such that $\mathbf{H}\mathbf{S}^{-1/2}\bar{\mathbf{x}} = (\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}})^{1/2}\mathbf{e}_1$. Choosing now $\mathbf{A} = \mathbf{H}$, we find

$$\begin{aligned} f(\bar{\mathbf{x}}, \mathbf{S}) &= f(\mathbf{H}\mathbf{S}^{-1/2}\bar{\mathbf{x}}, \mathbf{H}\mathbf{H}') \\ &= f((\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}})^{1/2}\mathbf{e}_1, \mathbf{I}), \end{aligned}$$

which shows that any invariant test function depends on the data only through $T^2 = n\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}$.

Second, selecting $\mathbf{A} = \boldsymbol{\Sigma}^{-1/2}$ gives

$$f(\bar{\mathbf{x}}, \mathbf{S}) = f(\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}}, \boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2}),$$

where

$$\begin{aligned} \boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}} &\sim N_p(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, n^{-1}\mathbf{I}) \\ (n-1)\boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2} &\sim W_p(n-1). \end{aligned}$$

Using the same argument, there exists an orthogonal transformation $\mathbf{H} \in \mathbf{O}_p$ such that $\mathbf{H}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu} = (\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{1/2}\mathbf{e}_1$. Choosing this time $\mathbf{A} = \mathbf{H}$, we find

$$f(\bar{\mathbf{x}}, \mathbf{S}) = f(\mathbf{H}\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}}, \mathbf{H}\boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2}\mathbf{H}'),$$

where

$$\begin{aligned} \mathbf{H}\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}} &\sim N_p((\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{1/2}\mathbf{e}_1, n^{-1}\mathbf{I}) \\ (n-1)\mathbf{H}\boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2}\mathbf{H}' &\sim W_p(n-1), \end{aligned}$$

and, thus, the power function of any invariant test depends on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ only through the parameter function $\delta = n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2$. These results are summarized.

Proposition 8.3 *For testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$, any invariant test with respect to the group $\mathbf{G}_p$ depends on the minimal sufficient statistic $(\bar{\mathbf{x}}, \mathbf{S})$ only through $T^2 = n\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}$. Moreover, the power function of any invariant test depends on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ only through the parameter function $\delta = n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2$.*

The LRT is obviously invariant. In the class of invariant tests, it is possible to show that the Hotelling-T^2 is uniformly most powerful. We say that T^2 is the UMPI test.

Proposition 8.4 *For testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$, the Hotelling-$T^2$ test, $T^2 = n\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}$, is UMPI.*

Proof. It has already been established that T^2 is invariant and that all invariant tests, depending on $(\bar{\mathbf{x}}, \mathbf{S})$, are a function of T^2. The problem thus reduces to finding the UMP test for $H_0 : \delta = 0$ based on one observation from $x \equiv T^2/(n-1) \sim F_c(p, n-p; \delta)$, where $\delta = n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2$. The density of x was given in Problem 4.6.3:

$$f(x;\delta) = \sum_{k=0}^{\infty} e^{-\delta}\frac{\delta^k}{k!}\frac{\Gamma\left(\frac{1}{2}(n+2k)\right)}{\Gamma\left(\frac{1}{2}(p+2k)\right)\ \Gamma\left(\frac{1}{2}(n-p)\right)}\frac{x^{(p+2k)/2-1}}{(1+x)^{(n+2k)/2}},\ x>0.$$

From the Neyman-Pearson lemma, the most powerful test for $H_0 : \delta = 0$ rejects H_0 for large values of the ratio

$$\frac{f(x;\delta)}{f(x;0)} = c_1\sum_{k=0}^{\infty} e^{-\delta}\frac{\delta^k}{k!}\frac{\Gamma\left(\frac{1}{2}(n+2k)\right)}{\Gamma\left(\frac{1}{2}(p+2k)\right)}\left[\frac{x}{(1+x)}\right]^k \geq c_2.$$

Since this ratio is monotone increasing in x, this is equivalent to rejecting H_0 for large values of x, $x \geq c_3$. This rejection region does not depend on δ and, thus, the test is uniformly most powerful. □

An asymptotic expansion of the distribution function of T^2 was obtained whose first term is χ^2_p under the elliptical distribution [Iwashita (1997)] and for general non-normality [Fujikoshi (1997), Kano (1995)]. Improvement to the chi-square approximation by monotone transformation of T^2 is also possible [Fujisawa (1997)]. For a modification to T^2 with the same chi-square asymptotic distribution but in the case of infinite second moment, refer to Sepanski (1994).

Kudô (1963) was the first to propose a multivariate analogue, when $\boldsymbol{\Sigma}$ is known, to the one-sided t-test. The multivariate problem is to test the null hypothesis, $H_0 : \boldsymbol{\mu} = \mathbf{0}$, against the one-sided alternative hypothesis, $H_1 : \boldsymbol{\mu} \geq \mathbf{0}$, where $\boldsymbol{\mu} \geq \mathbf{0}$ is interpreted componentwise. It can be stated even more generally in terms of cone. The LRT for the one-sided problem, with unknown $\boldsymbol{\Sigma}$, was obtained by Perlman (1969). Tang (1994, 1996) discussed unbiasedness and invariance of tests in the one-sided multivariate problem. Silvapulle (1995) derived the null distribution of a Hotelling-T^2

type statistic. There is a conditional test by Fraser, Guttman and Srivastava (1991); v. also Wang and McDermott (1998a, 1998b).

8.3 Simultaneous confidence intervals on means

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For any preassigned level $\beta = 1-\alpha$, define the quantile $F_\alpha(p_1, p_2)$ by the equation $P\left(F(p_1, p_2) \geq F_\alpha(p_1, p_2)\right) = \alpha$. By applying Hotelling's result, we have exactly

$$P\left(n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{(n-p)} F_\alpha(p, n-p)\right) = \beta,$$

whereby we see that in $\beta \times 100\%$ of such experiments, the "true" $\boldsymbol{\mu}$ lies in the random ellipsoid

$$\{\boldsymbol{\mu} \in \mathbb{R}^p : \ n(\boldsymbol{\mu} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}}) \leq c_\alpha\},$$

where we have simply let

$$c_\alpha = \frac{(n-1)p}{(n-p)} F_\alpha(p, n-p).$$

We are $\beta \times 100\%$ "confident" that our particular observed ellipsoid,

$$CR(\boldsymbol{\mu}; \beta) = \{\boldsymbol{\mu} \in \mathbb{R}^p : \ n(\boldsymbol{\mu} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}}) \leq c_\alpha\},$$

contains $\boldsymbol{\mu}$ since $P\left(\boldsymbol{\mu} \in CR(\boldsymbol{\mu}; \beta)\right) = \beta$.

Sequential fixed-size confidence regions for the mean vector were investigated by Srivastava (1967) and Datta and Mukhopadhyay (1997).

8.3.1 Linear hypotheses

In many experiments, one simply wishes to compare the various components of $\boldsymbol{\mu}$ to each other. For instance, one may ask: "Is μ_1 equal to μ_3?" "Is the difference between μ_2 and the average of μ_1 and μ_3 equal to 3.1?" Answering the first question amounts to testing the hypothesis

$$H_0 : \mu_1 - \mu_3 = 0,$$

while the second question is equivalent to testing the hypothesis

$$H_0 : 2\mu_2 - (\mu_1 + \mu_3) = 6.2.$$

Questions like these are said to be linear in $\boldsymbol{\mu}$, and in the general case, there would be a certain specified vector $\mathbf{a} \in \mathbb{R}^p$ and constant $c \in \mathbb{R}$ for which we would wish to test

$$H_0 : \mathbf{a}'\boldsymbol{\mu} = c.$$

Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if we let $y_i = \mathbf{a}'\mathbf{x}_i$, $i = 1, \ldots, n$, then clearly $y_1, \ldots, y_n$ are i.i.d. $N_1(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. Obviously, one may apply the univariate results directly to the y data so that

$$H_0 \text{ is correct iff } \sqrt{n}(\bar{y} - c)/s_y \stackrel{d}{=} t_{n-1}$$

and, of course,

$$CI(\mathbf{a}'\boldsymbol{\mu}; \beta) = \left[\bar{y} - \frac{s_y}{\sqrt{n}} t_{\alpha/2, n-1},\ \bar{y} + \frac{s_y}{\sqrt{n}} t_{\alpha/2, n-1}\right].$$

One should notice that, conveniently, $\bar{y} = \mathbf{a}'\bar{\mathbf{x}}$ and $s_y^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$.

Realistically, one would have more than one such question to consider, and so that there would be r specified vectors $\mathbf{a}_i \in \mathbb{R}^p$, $i = 1, \ldots, r$, with corresponding constants $c_i \in \mathbb{R}$, $i = 1, \ldots, r$, for which we would wish simultaneously to test

$$H_0 : \mathbf{a}_1'\boldsymbol{\mu} = c_1, \ldots, \mathbf{a}_r'\boldsymbol{\mu} = c_r.$$

This is clearly equivalent to testing

$$H_0 : \mathbf{A}'\boldsymbol{\mu} = \mathbf{c},$$

where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$ and $\mathbf{c} = (c_1, \ldots, c_r)'$. Letting $\mathbf{y}_i = \mathbf{A}'\mathbf{x}_i$, $i = 1, \ldots, n$, then, clearly, $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are i.i.d. $N_r(\mathbf{A}'\boldsymbol{\mu}, \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})$. Under the assumption that the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_r$ are linearly independent, one may apply the multivariate results above directly to the $\mathbf{y}$ data so that

$$H_0 \text{ is correct iff } n(\bar{\mathbf{y}} - \mathbf{c})'\mathbf{S}_{\mathbf{y}}^{-1}(\bar{\mathbf{y}} - \mathbf{c}) \stackrel{d}{=} \frac{(n-1)r}{n-r} F(r, n-r)$$

and a confidence ellipsoid for $\boldsymbol{\nu} = \mathbf{A}'\boldsymbol{\mu}$ is

$$CR(\boldsymbol{\nu}; \beta) = \{\boldsymbol{\nu} \in \mathbb{R}^r : n(\boldsymbol{\nu} - \bar{\mathbf{y}})'\mathbf{S}_{\mathbf{y}}^{-1}(\boldsymbol{\nu} - \bar{\mathbf{y}}) \leq k_\alpha\},$$

where

$$k_\alpha = \frac{(n-1)r}{(n-r)} F_\alpha(r, n-r).$$

Notice that $\bar{\mathbf{y}} = \mathbf{A}'\bar{\mathbf{x}}$ and $\mathbf{S}_{\mathbf{y}} = \mathbf{A}'\mathbf{S}\mathbf{A}$.

For obvious pragmatic reasons, one might in practice wish to have individual confidence intervals for each component ν_i, $i = 1, \ldots, r$. Thus, we would like to specify r intervals for these quantities in which we are *simultaneously* confident. The following lemma is needed.

Lemma 8.1 *Assume* $\mathbf{S} \in \mathcal{P}_p$ *and* $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r) \in \mathbb{R}_r^p$ *is of* rank r. *Then*

$$\frac{(\mathbf{a}_i'\mathbf{x})^2}{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i} \leq \mathbf{x}'\mathbf{A}(\mathbf{A}'\mathbf{S}\mathbf{A})^{-1}\mathbf{A}'\mathbf{x} \leq \mathbf{x}'\mathbf{S}^{-1}\mathbf{x},\ \forall \mathbf{x} \in \mathbb{R}^p.$$

Proof. Using Rayleigh's quotient (v. Problem 1.8.12),

$$\sup_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}(\mathbf{A}'\mathbf{S}\mathbf{A})^{-1}\mathbf{A}'\mathbf{x}}{\mathbf{x}'\mathbf{S}^{-1}\mathbf{x}} = \lambda_1\left(\mathbf{A}(\mathbf{A}'\mathbf{S}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}\right) = 1$$

and the right inequality follows. Let $\mathbf{y} = \mathbf{A}'\mathbf{x} = (y_1, \mathbf{y}_2')'$ and $\mathbf{B} = \mathbf{A}'\mathbf{S}\mathbf{A} \in \mathcal{P}_r$ partitioned as

$$\mathbf{B} = \begin{pmatrix} b_{11} & \mathbf{b}_{21}' \\ \mathbf{b}_{21} & \mathbf{B}_{22} \end{pmatrix}$$

with inverse (v. Problem 1.8.1)

$$\mathbf{B}^{-1} = \begin{pmatrix} b_{11}^{-1} + b_{11}^{-2}\mathbf{b}_{21}'\mathbf{B}_{22.1}^{-1}\mathbf{b}_{21} & -b_{11}^{-1}\mathbf{b}_{21}'\mathbf{B}_{22.1}^{-1} \\ -b_{11}^{-1}\mathbf{B}_{22.1}^{-1}\mathbf{b}_{21} & \mathbf{B}_{22.1}^{-1} \end{pmatrix}.$$

Then,

$$\mathbf{y}'\mathbf{B}^{-1}\mathbf{y} = \frac{y_1^2}{b_{11}} + ||y_1 b_{11}\mathbf{B}_{22.1}^{-1/2}\mathbf{b}_{21} - \mathbf{B}_{22.1}^{-1/2}\mathbf{y}_2||^2 \geq \frac{y_1^2}{b_{11}} = \frac{(\mathbf{a}_1'\mathbf{x})^2}{\mathbf{a}_1'\mathbf{S}\mathbf{a}_1},$$

which is the left inequality. □

Since by simple algebra in Lemma 8.1, we have the inequalities

$$\frac{(\bar{y}_i - \nu_i)^2}{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i} \leq (\bar{\mathbf{y}} - \boldsymbol{\nu})'\mathbf{S}_{\mathbf{y}}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\nu}) \leq (\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}),$$

we find that

$$P\left(n^{1/2}\frac{\mid \bar{y}_i - \nu_i \mid}{(\mathbf{a}_i'\mathbf{S}\mathbf{a}_i)^{1/2}} \leq k_\alpha^{1/2}, i = 1, \ldots, r\right)$$
$$\geq P\left(n(\bar{\mathbf{y}} - \boldsymbol{\nu})'\mathbf{S}_{\mathbf{y}}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\nu}) \leq k_\alpha\right) = \beta.$$

Therefore, we are at least $\beta \times 100\%$ confident in simultaneously presenting the r observed "Roy-Bose" intervals

$$\bar{y}_i - \frac{k_\alpha^{1/2}}{n^{1/2}}(\mathbf{a}_i'\mathbf{S}\mathbf{a}_i)^{1/2} \leq \nu_i \leq \bar{y}_i + \frac{k_\alpha^{1/2}}{n^{1/2}}(\mathbf{a}_i'\mathbf{S}\mathbf{a}_i)^{1/2},\ i = 1, \ldots, r. \tag{8.2}$$

One should note that $k_\alpha \leq c_\alpha$ (why?), so we do somewhat better using k_α. The constant c_α, however, allows all possible linear combinations since

$$\sup_{\mathbf{a}\neq\mathbf{0}} \frac{(\mathbf{a}'\bar{\mathbf{x}} - \mathbf{a}'\boldsymbol{\mu})^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} = (\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}).$$

Therefore, we are at least $\beta \times 100\%$ confident in simultaneously presenting all of the observed "Scheffé" intervals

$$\mathbf{a}'\bar{\mathbf{x}} - \frac{c_\alpha^{1/2}}{n^{1/2}}(\mathbf{a}'\mathbf{S}\mathbf{a})^{1/2} \leq \mathbf{a}'\boldsymbol{\mu} \leq \mathbf{a}'\bar{\mathbf{x}} + \frac{c_\alpha^{1/2}}{n^{1/2}}(\mathbf{a}'\mathbf{S}\mathbf{a})^{1/2},\ \forall \mathbf{a} \in \mathbb{R}^p. \tag{8.3}$$

Although the "Scheffé" intervals are wider, they can be useful in making a great number of unplanned comparisons between means.

Actually, if the number, r, of questions that one asks is very small, we can sometimes even improve on k_α. Let

$$T_i = n^{1/2} \frac{(\bar{y}_i - \nu_i)}{(\mathbf{a}_i' \mathbf{S} \mathbf{a}_i)^{1/2}}, \ i = 1, \ldots, r.$$

Note $T_i \stackrel{d}{=} t_{n-1}$, $i = 1, \ldots, r$, but they are not independent. Then, if we define the event $A_i = \{|T_i| \leq t_0\}$,

$$\begin{aligned} P\left(|T_i| \leq t_0, \ i = 1, \ldots, r\right) &= P\left(\cap_{i=1}^r A_i\right) \\ &= 1 - P\left(\cup_{i=1}^r A_i^c\right) \\ &\geq 1 - \sum_{i=1}^r P\left(A_i^c\right) \qquad (8.4) \\ &= 1 - r \ P\left(|t_{n-1}| > t_0\right). \end{aligned}$$

The inequality (8.4) is the Bonferroni inequality. If we deliberately equate the final term to $\beta = 1 - \alpha$ and then solve for t_0, we find that

$$P(t_{n-1} > t_0) = \frac{\alpha}{2r},$$

or, equivalently,

$$t_0 = t_{\alpha/2r, n-1}.$$

Therefore, one can see that if we let

$$b_\alpha = t^2_{\alpha/2r, n-1},$$

we will still be *at least* $\beta \times 100\%$ confident if, instead of the "Roy-Bose" intervals (8.2), we present the r "Bonferroni" intervals

$$\bar{y}_i - \frac{b_\alpha^{1/2}}{n^{1/2}} (\mathbf{a}_i' \mathbf{S} \mathbf{a}_i)^{1/2} \leq \nu_i \leq \bar{y}_i + \frac{b_\alpha^{1/2}}{n^{1/2}} (\mathbf{a}_i' \mathbf{S} \mathbf{a}_i)^{1/2}, \ i = 1, \ldots, r.$$

Note that the relative length of "Roy-Bose" to "Bonferroni" is obviously $\sqrt{b_\alpha / k_\alpha}$, and in a particular application, one would use the method with the shorter intervals.

For non-normal data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}$ with $E\ \mathbf{x} = \boldsymbol{\mu}$ and var $\mathbf{x} = \boldsymbol{\Sigma}$, large sample "Roy-Bose" and "Scheffé" simultaneous confidence intervals can be constructed similarly (v. Problems 8.9.5 and 8.9.6) by appealing first to the central limit theorem.

8.3.2 Nonlinear hypotheses

In certain experiments, one might wish to compare the various components of $\boldsymbol{\mu}$ to each other in ways that are plainly nonlinear. For instance, one may ask: "Is μ_1 equal to μ_3^2?" "Is the difference between μ_2 and the product of μ_1 and μ_3 equal to 5.7?" The first question corresponds to the hypothesis

$$H_0 : \mu_1 - \mu_3^2 = 0$$

and the second to

$$H_0 : \mu_2 - \mu_1\mu_3 = 5.7.$$

In each case, there is a certain function $g : \mathbb{R}^p \to \mathbb{R}$ and constant $c \in \mathbb{R}$ for which we are entertaining the hypothesis

$$H_0 : g(\boldsymbol{\mu}) = c.$$

If we actually had r such hypotheses to consider in the same experiment, then there would, of course, be r specified real-valued functions g_i, $i = 1, \ldots, r$, with constants $c_i \in \mathbb{R}$, $i = 1, \ldots, r$, for which we would wish simultaneously to test

$$H_0 : g_1(\boldsymbol{\mu}) = c_1, \ldots, g_r(\boldsymbol{\mu}) = c_r.$$

By letting $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^r$ defined by $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_r(\mathbf{x}))$ be continuously differentiable at $\boldsymbol{\mu}$, and $\mathbf{c} \in \mathbb{R}^r$, this is simply equivalent to

$$H_0 : \mathbf{g}(\boldsymbol{\mu}) = \mathbf{c}.$$

The results in this section are asymptotic, so we assume possibly non-normal data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x}$ with $E\ \mathbf{x} = \boldsymbol{\mu}$ and var $\mathbf{x} = \boldsymbol{\Sigma}$. One may not apply the specific results in Section 8.3.1 directly to this situation, but one may yet apply the same essential logic as formulated in that section by appealing to the central limit theorem, $\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and the delta method in Proposition 6.2, $\sqrt{n}\,(\mathbf{g}(\bar{\mathbf{x}}) - \mathbf{g}(\boldsymbol{\mu})) \xrightarrow{d} N_r\,(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{g})$, where $\boldsymbol{\Sigma}_\mathbf{g} = [\mathbf{D}\mathbf{g}(\boldsymbol{\mu})]\boldsymbol{\Sigma}[\mathbf{D}\mathbf{g}(\boldsymbol{\mu})]'$. This time, if we let $\mathbf{y} = \mathbf{g}(\bar{\mathbf{x}})$, $\boldsymbol{\nu} = \mathbf{g}(\boldsymbol{\mu})$, and $\mathbf{S}_\mathbf{g} = [\mathbf{D}\mathbf{g}(\bar{\mathbf{x}})]\mathbf{S}[\mathbf{D}\mathbf{g}(\bar{\mathbf{x}})]'$ then, as on page 79,

$$n(\mathbf{y} - \boldsymbol{\nu})'\mathbf{S}_\mathbf{g}^{-1}(\mathbf{y} - \boldsymbol{\nu}) \xrightarrow{d} \chi^2_r.$$

Thus, with

$$d_\alpha = \chi^2_{\alpha,r},$$

the confidence ellipsoid for $\boldsymbol{\nu}$,

$$CR(\boldsymbol{\nu}; \beta) = \{\boldsymbol{\nu} \in \mathbb{R}^r :\ n(\boldsymbol{\nu} - \mathbf{y})'\mathbf{S}_\mathbf{g}^{-1}(\boldsymbol{\nu} - \mathbf{y}) \le d_\alpha\},$$

has an asymptotic coverage probability of β, i.e., $P\,(\boldsymbol{\nu} \in CR(\boldsymbol{\nu}; \beta)) \to \beta$ as $n \to \infty$.

To have individual confidence intervals on each component ν_i, $i = 1, \ldots, r$, we have the inequality in Lemma 8.1:

$$\frac{(y_i - \nu_i)^2}{\mathbf{S}_{\mathbf{g},ii}} \le (\mathbf{y} - \boldsymbol{\nu})'\mathbf{S}_\mathbf{g}^{-1}(\mathbf{y} - \boldsymbol{\nu}).$$

From this purely algebraic fact, it follows that

$$\lim_{n\to\infty} P\left(n^{1/2}\frac{|\ y_i - \nu_i\ |}{\mathbf{S}_{\mathbf{g},ii}^{1/2}} \le d_\alpha^{1/2},\ i = 1, \ldots, r\right) \ge \beta.$$

Thus, asymptotically, we are *at least* $\beta \times 100\%$ confident in simultaneously presenting the r observed intervals

$$y_i - \frac{d_\alpha^{1/2}}{n^{1/2}} \mathbf{S}_{\mathbf{g},ii}^{1/2} \leq \nu_i \leq y_i + \frac{d_\alpha^{1/2}}{n^{1/2}} \mathbf{S}_{\mathbf{g},ii}^{1/2}, \ i = 1, \ldots, r.$$

If r is quite small, one might try to improve on d_α using a Bonferroni approach.

The construction of simultaneous confidence intervals on functions $\phi(\mathbf{\Sigma})$ is treated quite generally in Dümbgen (1998). Asymptotic considerations for the Wishart model show that the resulting confidence bounds are substantially smaller than those obtained by inverting likelihood ratio tests.

8.4 Multiple correlation

The multiple correlation coefficient R is the maximum correlation possible between a variable x_1 and a linear combination, $\mathbf{t}'\mathbf{x}_2$, of a vector $\mathbf{x}_2$. Not surprisingly, with underlying normality, the likelihood ratio test of $H_0 : x_1 \perp\!\!\!\perp \mathbf{x}_2$ will be a function of the sample multiple correlation coefficient $\hat{R}$. Assume $x_1 \in \mathbb{R}$ and $\mathbf{x}_2 \in \mathbb{R}^{p-1}$ have a joint normal distribution,

$$\begin{pmatrix} x_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_p \left(\mathbf{0}, \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{21}' \\ \boldsymbol{\sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \right),$$

where $\mathbf{\Sigma}_{22} = \mathbf{A}^2 > \mathbf{0}$. We have set the mean to $\mathbf{0}$ without any loss of generality. Since the simple correlation coefficient is invariant to rescaling of each variable, we can assume at the outset that var $\mathbf{t}'\mathbf{x}_2 = \mathbf{t}'\mathbf{\Sigma}_{22}\mathbf{t} = 1$ and solve

$$\max_{\mathbf{t}'\mathbf{\Sigma}_{22}\mathbf{t}=1} \operatorname{cor}(x_1, \mathbf{t}'\mathbf{x}_2).$$

For any $\mathbf{t}$ such that $\mathbf{t}'\mathbf{\Sigma}_{22}\mathbf{t} = 1$,

$$\operatorname{cor}^2(x_1, \mathbf{t}'\mathbf{x}_2) = (\boldsymbol{\sigma}_{21}'\mathbf{t})^2/\sigma_{11} = \langle \mathbf{A}^{-1}\boldsymbol{\sigma}_{21}, \mathbf{A}\mathbf{t} \rangle^2/\sigma_{11} \leq \boldsymbol{\sigma}_{21}'\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}/\sigma_{11}.$$

The last inequality follows from the Cauchy-Schwarz inequality given in Proposition 1.1. It is an equality iff $\mathbf{A}\mathbf{t} \propto \mathbf{A}^{-1}\boldsymbol{\sigma}_{21}$, or, equivalently, $\mathbf{t} \propto \mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$. The maximum correlation possible is $R \geq 0$, where $R^2 = \boldsymbol{\sigma}_{21}'\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}/\sigma_{11}$, and is called the *multiple correlation coefficient* between x_1 and $\mathbf{x}_2$. It should be noted immediately that the maximum correlation is achieved by $\mathbf{t}'\mathbf{x}_2 = E\,(x_1 \mid \mathbf{x}_2) = \boldsymbol{\sigma}_{21}'\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_2$, i.e., by the conditional mean of x_1 given $\mathbf{x}_2$. In order to test $H_0 : R = 0$ (equivalently, $H_0 : \boldsymbol{\sigma}_{21} = \mathbf{0}$ or $H_0 : x_1 \perp\!\!\!\perp \mathbf{x}_2$), the sample variance, based on a random sample of size n,

$$(n-1)\mathbf{S} \equiv \mathbf{V} = \begin{pmatrix} v_{11} & \mathbf{v}_{21}' \\ \mathbf{v}_{21} & \mathbf{V}_{22} \end{pmatrix},$$

is partitioned and is distributed as $\mathbf{V} \sim W_p(n-1, \mathbf{\Sigma})$. In the obvious manner, the sample version is $\hat{R}^2 = \mathbf{v}_{21}' \mathbf{V}_{22}^{-1} \mathbf{v}_{21} / v_{11}$ and $\hat{R} \geq 0$ is called the *sample multiple correlation coefficient.*

Proposition 8.5 *The likelihood ratio test Λ rejects H_0 for small values of $\Lambda = (1 - \hat{R}^2)^{n/2}$.*

Proof. Based on the likelihood (7.1) from $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, $\mathbf{\Sigma} > \mathbf{0}$, the MLE of $\boldsymbol{\mu}$ is always $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. Without constraints, the MLE of $\mathbf{\Sigma}$ is $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{V}$, but when $\boldsymbol{\sigma}_{21} = \mathbf{0}$, the constrained MLE becomes

$$\hat{\hat{\mathbf{\Sigma}}} = \frac{1}{n} \begin{pmatrix} v_{11} & \mathbf{0}' \\ \mathbf{0} & \mathbf{V}_{22} \end{pmatrix}.$$

Thus,

$$\begin{aligned} \Lambda = \frac{L(\hat{\hat{\mathbf{\Sigma}}}, \hat{\boldsymbol{\mu}})}{L(\hat{\mathbf{\Sigma}}, \hat{\boldsymbol{\mu}})} &= \frac{|\hat{\hat{\mathbf{\Sigma}}}|^{-n/2}}{|\hat{\mathbf{\Sigma}}|^{-n/2}} \frac{\operatorname{etr}(-\frac{1}{2}\mathbf{V}\hat{\hat{\mathbf{\Sigma}}}^{-1})}{\operatorname{etr}(-\frac{1}{2}\mathbf{V}\hat{\mathbf{\Sigma}}^{-1})} \\ &= \left[\frac{v_{11}|\mathbf{V}_{22}|}{|\mathbf{V}|}\right]^{-n/2} \frac{\exp(-\frac{1}{2}np)}{\exp(-\frac{1}{2}np)} \\ &= \left(\frac{v_{11.2}}{v_{11}}\right)^{n/2} = (1 - \hat{R}^2)^{n/2}, \end{aligned}$$

where the last equality made use of $|\mathbf{V}| = v_{11.2}|\mathbf{V}_{22}|$. □

Of greater interest is the distribution of $\hat{R}^2$ in which negative binomial probabilities intervene. The reader should recall at this point that a negative binomial variable represents the number of failures, k, before the rth success in a sequence of independent bernoulli trials.

Definition 8.1 Negative binomial*: $x \sim nb(r, p)$, $r > 0$ and $0 \leq p \leq 1$, iff the probability function of x is given by*

$$p_k = P(x = k) = \binom{r+k-1}{k} p^r (1-p)^k, \ k = 0, 1, \ldots.$$

In Definition 8.1, r need not be an integer. In that case, the combination factor is calculated via the gamma function:

$$\binom{r+k-1}{k} = \frac{\Gamma(r+k)}{k!\Gamma(r)} = \frac{(r)_k}{k!},$$

where $(r)_0 = 1$ and $(r)_k = r(r+1)\cdots(r+k-1)$ for $k = 1, 2, \ldots$. Recall that $F_c(s_1, s_2)$ denotes the canonical F_c distribution (v. Definition 3.7).

Proposition 8.6

$$P\left(\frac{\hat{R}^2}{1-\hat{R}^2} \leq t\right) = \sum_{k=0}^{\infty} p_k \cdot P\left(F_c(p-1+2k, n-p) \leq t\right),$$

$$P\left(\hat{R}^2 \le t\right) = \sum_{k=0}^{\infty} p_k \cdot P\left(\text{beta}\left(\tfrac{1}{2}(p-1+2k); \tfrac{1}{2}(n-p)\right) \le t\right),$$

where p_k are the negative binomial probabilities

$$p_k = \frac{\left(\frac{1}{2}(n-1)\right)_k}{k!}(1-R^2)^{(n-1)/2}R^{2k},\ k = 0, 1, \ldots.$$

Proof. Since $v_{11.2} = v_{11}(1-\hat{R}^2)$, then $\hat{R}^2/(1-\hat{R}^2) = \mathbf{v}_{21}'\mathbf{V}_{22}^{-1}\mathbf{v}_{21}/v_{11.2}$. With the help of Proposition 7.9, $v_{11.2} \perp\!\!\!\perp (\mathbf{v}_{21}, \mathbf{V}_{22})$ and $v_{11.2} \sim \sigma_{11.2}\chi^2_{n-p}$. We also have

$$\mathbf{v}_{21} \mid \mathbf{V}_{22} \sim N_{p-1}(\mathbf{V}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \sigma_{11.2}\mathbf{V}_{22})$$

from which

$$\sigma_{11.2}^{-1/2}\mathbf{V}_{22}^{-1/2}\mathbf{v}_{21} \mid \mathbf{V}_{22} \sim N_{p-1}(\sigma_{11.2}^{-1/2}\mathbf{V}_{22}^{1/2}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \mathbf{I})$$

and, therefore, $\mathbf{v}_{21}'\mathbf{V}_{22}^{-1}\mathbf{v}_{21} \mid \mathbf{V}_{22} \sim \sigma_{11.2}\chi^2_{p-1}(\delta)$, where

$$\delta = \boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1}\mathbf{V}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}/(2\sigma_{11.2}).$$

Hence, conditional on $\mathbf{V}_{22}$,

$$\hat{R}^2/(1-\hat{R}^2) \sim F_c(p-1, n-p; \delta).$$

Using Proposition 4.7,

$$P\left(\frac{\hat{R}^2}{1-\hat{R}^2} \le t \mid \mathbf{V}_{22}\right) = \sum_{k=0}^{\infty} e^{-\delta}\frac{\delta^k}{k!}P\left(F_c(p-1+2k, n-p) \le t\right).$$

To obtain the unconditional distribution, take expectations on both sides with respect to the distribution of $\mathbf{V}_{22}$. First, we need the distribution of δ. Since $\mathbf{V}_{22} \sim W_{p-1}(n-1, \boldsymbol{\Sigma}_{22})$, then

$$\delta \sim \frac{R^2}{(1-R^2)}\tfrac{1}{2}\chi^2_{n-1} \stackrel{d}{=} G\left(\tfrac{1}{2}(n-1), \frac{R^2}{(1-R^2)}\right).$$

The expectation computation is immediate (v. Problem 8.9.10) if we use a result well known in bayesian inference [Johnson et al. (1992), p. 204] that if K given δ is Poisson(δ) and $\delta \sim G(p, \theta)$, then the marginal of K is negative binomial, $K \sim \text{nb}(p, (1+\theta)^{-1})$. Hence,

$$p_k = P(K=k) = E\ P(K=k \mid \delta) = E\ e^{-\delta}\frac{\delta^k}{k!},$$

completing the proof of the first result. The second result follows with the obvious monotone transformation. □

Thus, $\hat{R}^2/(1-\hat{R}^2)$ is distributed as a negative binomial mixture of canonical F_c distributions, whereas that of $\hat{R}^2$ is a negative binomial mixture of beta distributions. The moments of $\hat{R}$ (v. Problem 8.9.9) follow directly from the later characterization. The null distribution is just a special case.

Proposition 8.7 *Assuming* $R = 0$, $\hat{R}^2/(1-\hat{R}^2) \sim F_c(p-1, n-p)$.

The exact distribution of the simple correlation coefficient, introduced earlier in Section 5.6.3, when $\rho \neq 0$ is just another special case when $p = 2$. The invariance of the multiple correlation coefficient is discussed in Problem 8.9.13.

Proposition 8.8 *For testing* $H_0 : R = 0$ *against* $H_1 : R > 0$, *the test which rejects for large values of* $\hat{R}$ *is UMPI.*

Proof. The statistic $\hat{R}$ is clearly invariant and it was established in Problem 8.9.13 that all invariant tests, depending on $(\bar{\mathbf{x}}, \mathbf{V})$, are a function of $\hat{R}$. The problem thus reduces to finding the UMP test based on one observation from $\hat{R}$. The density of $x \equiv \hat{R}^2$ follows from Proposition 8.6,

$$f(x; R^2) = \sum_{k=0}^{\infty} p_k \frac{1}{B(\frac{1}{2}(p-1+2k), \frac{1}{2}(n-p))} x^{\frac{1}{2}(p-1+2k)-1}(1-x)^{\frac{1}{2}(n-p)-1},$$

$0 < x < 1$. From the Neyman-Pearson lemma, the most powerful test rejects H_0 for large values of the ratio

$$\frac{f(x; R^2)}{f(x; 0)} = c_1 \sum_{k=0}^{\infty} p_k \frac{\Gamma\left(\frac{1}{2}(n-1+2k)\right)}{\Gamma\left(\frac{1}{2}(p-1+2k)\right)} x^k \geq c_2.$$

Since this ratio is monotone increasing in x this is equivalent to rejecting H_0 for large values of x, $x \geq c_3$. This rejection region does not depend on R and, thus, the test is uniformly most powerful. □

Example 8.2 *The power function of the likelihood ratio test for* $H_0 : R = 0$ *may be evaluated with Proposition* 8.6*:*

$$\begin{aligned} \beta &= \sum_{k=0}^{\infty} p_k \int_{t_\alpha}^{1} B\left(\tfrac{1}{2}(p-1+2k), \tfrac{1}{2}(n-p)\right)^{-1} \\ &\quad \cdot x^{(p-1+2k)/2-1}(1-x)^{(n-p)/2-1} dx, \end{aligned}$$

where $t_\alpha = \text{beta}_\alpha\left(\frac{1}{2}(p-1); \frac{1}{2}(n-p)\right)$ *is the critical point. A numerical evaluation in Mathematica of* β *for* $p = 3$ *and* $n = 20$ *at the significance level* $\alpha = 0.05$ *gave the plot in Figure* 8.2.

For large samples, the asymptotic distribution provides a simpler distribution. By the delta method in Proposition 6.2, $\hat{R}^2$ is asymptotically normal since it is a function of the sample variance $\mathbf{S}$, which is itself asymptotically normal. However, rather than calculating the derivatives, it is somewhat easier to use $o_p(n^{-1/2})$ asymptotic expansions. This technique is illustrated in the following proof.

Proposition 8.9 *The null and alternative asymptotic distributions of* $\hat{R}^2$, *when sampling from a multivariate normal distribution, are given by*

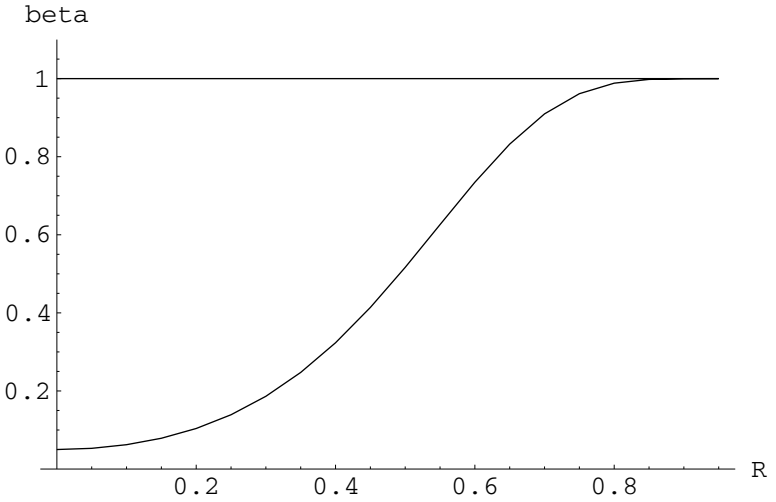

Figure 8.2. Power function of the likelihood ratio test for $H_0 : R = 0$ when $p = 3$, and $n = 20$ at a level of significance $\alpha = 0.05$.

(i) $n^{1/2}(\hat{R}^2 - R^2) \xrightarrow{d} N\left(0, 4R^2(1-R^2)^2\right)$,

(ii) *If* $R = 0$, *then* $n\hat{R}^2 \xrightarrow{d} \chi^2_{p-1}$.

Proof. By invariance arguments (v. Problem 8.9.13), assume without loss of generality,

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & R\mathbf{e}_1' \\ R\mathbf{e}_1 & \mathbf{I}_{p-1} \end{pmatrix},$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^{p-1}$. Since

$$n^{1/2}(\mathbf{S} - \mathbf{\Sigma}) \xrightarrow{d} \mathbf{Z} = \begin{pmatrix} z_{11} & \mathbf{z}_{21}' \\ \mathbf{z}_{21} & \mathbf{Z}_{22} \end{pmatrix},$$

where $\mathbf{Z} \sim N_p^p\left(\mathbf{0}, (\mathbf{I} + \mathbf{K}_p)(\mathbf{\Sigma} \otimes \mathbf{\Sigma})\right)$, then we can write the $o_p(n^{-1/2})$ expansions

$$\begin{aligned} s_{11} &= 1 + n^{-1/2} z_{11} + o_p(n^{-1/2}), \\ \mathbf{s}_{21} &= R\mathbf{e}_1 + n^{-1/2}\mathbf{z}_{21} + o_p(n^{-1/2}), \\ \mathbf{S}_{22} &= \mathbf{I} + n^{-1/2}\mathbf{Z}_{22} + o_p(n^{-1/2}), \end{aligned}$$

where $o_p(n^{-1/2})$ is such that $n^{1/2} \cdot o_p(n^{-1/2}) \xrightarrow{p} 0$ [Serfling (1980), p. 9]. Straightforward algebra, with the aid of Problem 1.8.15, then gives

$$\begin{aligned} &\frac{\mathbf{s}_{21}'\mathbf{S}_{22}^{-1}\mathbf{s}_{21}}{s_{11}} \\ &= [1 + n^{-1/2} z_{11} + o_p(n^{-1/2})]^{-1} \cdot [R\mathbf{e}_1 + n^{-1/2}\mathbf{z}_{21} + o_p(n^{-1/2})]' \end{aligned}$$

$$\begin{aligned} & \cdot[\mathbf{I} + n^{-1/2}\mathbf{Z}_{22} + o_p(n^{-1/2})]^{-1} \cdot [R\mathbf{e}_1 + n^{-1/2}\mathbf{z}_{21} + o_p(n^{-1/2})] \\ = \; & [1 - n^{-1/2}z_{11} + o_p(n^{-1/2})] \cdot [R\mathbf{e}_1 + n^{-1/2}\mathbf{z}_{21} + o_p(n^{-1/2})]' \\ & \cdot[\mathbf{I} - n^{-1/2}\mathbf{Z}_{22} + o_p(n^{-1/2})] \cdot [R\mathbf{e}_1 + n^{-1/2}\mathbf{z}_{21} + o_p(n^{-1/2})] \\ = \; & R^2 + 2n^{-1/2}Rz_{21} - n^{-1/2}R^2z_{22} - n^{-1/2}R^2z_{11} + o_p(n^{-1/2}). \end{aligned}$$

Thus, $n^{1/2}(\hat{R}^2 - R^2) \xrightarrow{d} 2Rz_{21} - R^2z_{22} - R^2z_{11}$, but since (v. equation (6.1)) $(z_{21}, z_{11}, z_{22})' \sim N_3(\mathbf{0}, \mathbf{\Omega})$, where

$$\mathbf{\Omega} = \begin{pmatrix} 1 + R^2 & 2R & 2R \\ 2R & 2 & 2R^2 \\ 2R & 2R^2 & 2 \end{pmatrix},$$

the linear combination with $\mathbf{a} = (2R, -R^2, -R^2)'$ yields

$$2Rz_{21} - R^2z_{22} - R^2z_{11} \sim N(0, \mathbf{a}'\mathbf{\Omega}\mathbf{a}),$$

whereby a direct evaluation provides $\mathbf{a}'\mathbf{\Omega}\mathbf{a} = 4R^2(1-R^2)^2$. This proves (i). To prove (ii), note that when $R = 0$, $n^{1/2}\mathbf{s}_{21} \xrightarrow{d} \mathbf{z}_{21}$, where $\mathbf{z}_{21} \sim N_{p-1}(\mathbf{0}, \mathbf{I})$. However, since $\mathbf{S}_{22} \xrightarrow{p} \mathbf{I}$ and $s_{11} \xrightarrow{p} 1$, then $n^{1/2}s_{11}^{-1/2}\mathbf{S}_{22}^{-1/2}\mathbf{s}_{21} \xrightarrow{d} \mathbf{z}_{21}$ and $n\hat{R}^2 \xrightarrow{d} |\mathbf{z}_{21}|^2 \stackrel{d}{=} \chi^2_{p-1}$. □

As a corollary, we get the asymptotic distribution of $\hat{R}$ and of Fisher's z-transform.

Corollary 8.2 *The asymptotic distributions of $\hat{R}$ and of its Fisher's z transform, when sampling from a multivariate normal distribution, are given by*

(i) $n^{1/2}(\hat{R} - R) \xrightarrow{d} N\left(0, (1 - R^2)^2\right)$,

(ii) $n^{1/2}\left(\tanh^{-1}(\hat{R}) - \tanh^{-1}(R)\right) \xrightarrow{d} N(0, 1)$.

Proof. It follows directly from the delta method applied to the square root transformation and to the $\tanh^{-1}$ transformation. □

More general results on the asymptotic distributions of correlation coefficients obtained from any asymptotically normal equivariant estimate of variance, not necessarily $\mathbf{S}$, will be given in Chapter 13 for a sample from an elliptical distribution.

8.4.1 Asymptotic moments

The mixture beta characterization of $\hat{R}^2$ in Proposition 8.6 can be used to obtain immediately the exact moments of $\hat{R}^2$ in terms of those of beta distributions (v. Problem 8.9.9). Simple approximations for large n are, however, possible as is now shown.

Using Proposition 8.6, we have

$$P\left(1-\hat{R}^2 \le t\right) = \sum_{k=0}^{\infty} p_k \cdot P\left(\text{beta}\left(\tfrac{1}{2}(n-p); \tfrac{1}{2}(p-1+2k)\right) \le t\right).$$

From the moments of $x \sim \text{beta}(a,b)$ given by

$$E\ x^h = \frac{\Gamma(a+b)\Gamma(a+h)}{\Gamma(a)\Gamma(a+b+h)} = \frac{(a)_h}{(a+b)_h},\ h=1,2,\ldots,$$

we can write

$$E\ (1-\hat{R}^2)^h = \sum_{k=0}^{\infty} \frac{\left(\frac{1}{2}(n-1)\right)_k}{k!}(1-R^2)^{(n-1)/2}R^{2k}\frac{\left(\frac{1}{2}(n-p)\right)_h}{\left(\frac{1}{2}(n-1+2k)\right)_h}.$$

The hypergeometric function

$${}_2F_1(a,b;c;z) \equiv \sum_{k=0}^{\infty} \frac{(a)_k(b)_k}{(c)_k}\frac{z^k}{k!}$$

after some simple algebra allows to write

$$E\ (1-\hat{R}^2)^h = \frac{\left(\frac{1}{2}(n-p)\right)_h}{\left(\frac{1}{2}(n-1)\right)_h}$$
$$\cdot(1-R^2)^{(n-1)/2}\ {}_2F_1(\tfrac{1}{2}(n-1), \tfrac{1}{2}(n-1); \tfrac{1}{2}(n-1)+h; R^2).$$

Then, upon using Kummer's formula [Erdélyi et al. (1953), p. 105]

$${}_2F_1(a,b;c;z) = (1-z)^{(c-a-b)}\ {}_2F_1(c-a, c-b; c; z),$$

we finally find

$$E\ (1-\hat{R}^2)^h = \frac{\left(\frac{1}{2}(n-p)\right)_h}{\left(\frac{1}{2}(n-1)\right)_h}(1-R^2)^h\ {}_2F_1(h,h;\tfrac{1}{2}(n-1)+h;R^2).$$

For $h=1$, we then obtain

$$\begin{aligned} E\ (1-\hat{R}^2) &= \frac{(n-p)}{(n-1)}(1-R^2)\ {}_2F_1(1,1;\tfrac{1}{2}(n+1);R^2) \\ &= \frac{(n-p)}{(n-1)}(1-R^2)\left[1+\frac{2R^2}{(n+1)}+O(n^{-2})\right] \end{aligned}$$

and

$$E\ \hat{R}^2 = R^2 + \frac{(p-1)}{(n-1)}(1-R^2) - 2\frac{(n-p)}{(n^2-1)}R^2(1-R^2) + O(n^{-2}).$$

This expression shows that $\hat{R}^2$ is biased, as it overestimates R^2. The MVUE of R^2 [Olkin and Pratt (1958)] is

$$U(\hat{R}^2) = 1 - \frac{(n-3)}{(n-p)}(1-\hat{R}^2)\ {}_2F_1(1,1;\tfrac{1}{2}(n-p+2);1-\hat{R}^2).$$

The MVUE has the drawback of taking negative values when $\hat{R}$ is close to 0. In fact, using the relation [Erdélyi et al. (1953), p. 61]

$$ {}_2F_1(a,b;c;1) = \frac{\Gamma(c)\,\Gamma(c-a-b)}{\Gamma(c-a)\,\Gamma(c-b)}, $$

it is easily established that $U(0) = -(p-1)/(n-p-2)$ and, of course, $U(1) = 1$. A similar expansion for $h = 2$ can be done to obtain an asymptotic expansion for var $\hat{R}^2$.

8.5 Partial correlation

Assume two subsets of variables $\mathbf{x}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{p_2}$ have a joint normal distribution,

$$ \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right). $$

The *partial correlation coefficient* between variables x_i and x_j, in the subset $\mathbf{x}_1$, is just the ordinary simple correlation ρ between x_i and x_j but with the variables in the subset $\mathbf{x}_2$ held fixed. This will be denoted by $\rho_{ij|\mathbf{x}_2}$. It can be expressed in terms of $\boldsymbol{\Sigma}$ if one recalls the conditional normal of Section 5.5:

$$ \mathbf{x}_1 \mid \mathbf{x}_2 \sim N_{p_1}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11.2}). $$

Writing $\boldsymbol{\Sigma}_{11.2} = (\sigma_{ij|\mathbf{x}_2})$, where $\sigma_{ij|\mathbf{x}_2}$ denotes the (i,j) element of $\boldsymbol{\Sigma}_{11.2}$, then

$$ \rho_{ij|\mathbf{x}_2} = \frac{\sigma_{ij|\mathbf{x}_2}}{\sigma_{ii|\mathbf{x}_2}^{1/2}\sigma_{jj|\mathbf{x}_2}^{1/2}}. $$

Using Proposition 7.9, we already know that since $(n-1)\mathbf{S} = \mathbf{V} \sim W_p(n-1, \boldsymbol{\Sigma})$, then

$$ \mathbf{V}_{11.2} \sim W_{p_1}(n-1-p_2, \boldsymbol{\Sigma}_{11.2}), $$

where $\mathbf{V}$ was partitioned in conformity as

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

Since $\mathbf{V}$ is proportional to the MLE $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$, it is clear that the MLE of $\rho_{ij|\mathbf{x}_2}$ is just

$$ r_{ij|\mathbf{x}_2} = \frac{v_{ij|\mathbf{x}_2}}{v_{ii|\mathbf{x}_2}^{1/2}v_{jj|\mathbf{x}_2}^{1/2}}, $$

where $\mathbf{V}_{11.2} = (v_{ij|\mathbf{x}_2})$ and $v_{ij|\mathbf{x}_2}$ denotes the (i,j) element of $\mathbf{V}_{11.2}$. Considering the distribution of $\mathbf{V}_{11.2}$, the distribution of $r_{ij|\mathbf{x}_2}$ is the same as

for a simple correlation coefficient but with $n - p_2$ in place of n. We have proved:

Proposition 8.10

$$P\left(\frac{r^2_{ij|\mathbf{x}_2}}{1-r^2_{ij|\mathbf{x}_2}} \le t\right) = \sum_{k=0}^{\infty} p_k \cdot P\left(F_c(1+2k, n-p_2-2) \le t\right),$$

$$P\left(r^2_{ij|\mathbf{x}_2} \le t\right) = \sum_{k=0}^{\infty} p_k \cdot P\left(\text{beta}\left(\tfrac{1}{2}(1+2k); \tfrac{1}{2}(n-p_2-2)\right) \le t\right),$$

where p_k are the negative binomial probabilities

$$p_k = \frac{\left(\frac{1}{2}(n-p_2-1)\right)_k}{k!}(1-\rho^2_{ij|\mathbf{x}_2})^{(n-p_2-1)/2}\rho^{2k}_{ij|\mathbf{x}_2},\ k = 0, 1, \ldots.$$

For large samples, as for the simple correlation coefficient, it follows from Problem 6.4.8 that

$$n^{1/2}\left(r_{ij|\mathbf{x}_2} - \rho_{ij|\mathbf{x}_2}\right) \xrightarrow{d} N\left(0, (1-\rho^2_{ij|\mathbf{x}_2})^2\right).$$

A Fisher's z-transform as for the simple correlation coefficient in Problem 6.4.9 is definitely possible for a partial correlation coefficient.

8.6 Test of sphericity

Assume $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$, and consider testing the hypothesis that the p variables in $\mathbf{x} = (x_1, \ldots, x_p)'$ are independent and have the same variance:

$$H_0 : \boldsymbol{\Sigma} = \gamma\mathbf{I},\ \gamma > 0.$$

Based on a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, regardless of the hypothesis H_0, as long as $\boldsymbol{\Sigma} > \mathbf{0}$, the MLE of $\boldsymbol{\mu}$ is always $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. Now, without constraint, the MLE of $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{V}$, where, as usual,

$$\mathbf{V} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

However, under H_0, the MLE is obtained by solving

$$\max_{\gamma>0} |\gamma\mathbf{I}|^{-n/2}\ \text{etr}\left(-\tfrac{1}{2}\gamma^{-1}\mathbf{V}\right).$$

Taking logarithms, the function to maximize is

$$-\tfrac{1}{2}np\ln\gamma - \tfrac{1}{2}\gamma^{-1}\ \text{tr}\ \mathbf{V},$$

and the solution is easily calculated, $\hat{\gamma} = \text{tr } \mathbf{V}/np$. Therefore, the likelihood ratio, first derived by Mauchly (1940), becomes

$$\begin{aligned}\Lambda &= \frac{L(\hat{\gamma}\mathbf{I}, \hat{\boldsymbol{\mu}})}{L(\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}})} = \frac{|\hat{\gamma}\mathbf{I}|^{-n/2}}{|\hat{\boldsymbol{\Sigma}}|^{-n/2}} \frac{\text{etr}\left(-\frac{1}{2}\hat{\gamma}^{-1}\mathbf{V}\right)}{\text{etr}\left(-\frac{1}{2}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{V}\right)} \\ &= \left[\frac{|\frac{1}{n}\mathbf{V}|}{|(\frac{1}{np}\text{tr } \mathbf{V})\mathbf{I}|}\right]^{n/2} \frac{\exp(-\frac{1}{2}np)}{\exp(-\frac{1}{2}np)}.\end{aligned}$$

Thus,

$$\tilde{\Lambda} \equiv \Lambda^{2/n} = \frac{|\mathbf{V}|}{(\frac{1}{p}\text{tr } \mathbf{V})^p} = \left(\frac{\prod_{i=1}^p l_i^{1/p}}{\frac{1}{p}\sum_{i=1}^p l_i}\right)^p,$$

where $l_1 \geq \cdots \geq l_p$ are the ordered eigenvalues of $\mathbf{V}$. The LRT compares the geometric and arithmetic means of those eigenvalues; they coincide when $\mathbf{V}$ has the structure as in H_0.

Proposition 8.11 *The LRT for testing* $H_0 : \boldsymbol{\Sigma} = \gamma\mathbf{I}$, $\gamma > 0$ *against* $H_1 : \boldsymbol{\Sigma} > \mathbf{0}$ *rejects* H_0 *for small values of* $\tilde{\Lambda} = |\mathbf{V}|/(\frac{1}{p}\text{ tr } \mathbf{V})^p$.

At this point, we remind the reader about the general expression for the asymptotic degrees of freedom for likelihood ratio tests. In general, for testing $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ against $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$ under regularity conditions, then, under H_0, $-2\ln\Lambda \xrightarrow{d} \chi_f^2$ as the sample size $n \to \infty$. The degrees of freedom f is the difference between the number of free parameters in $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_0^c$ and the number of free parameters in $\boldsymbol{\Theta}_0$.

From the general theory of LRT, it is clear that the asymptotic null distribution is

$$-2\ln\Lambda \xrightarrow{d} \chi_f^2, \;\; f = \tfrac{1}{2}p(p+1) - 1.$$

Better approximations can be obtained by calculating the moments of Λ (or $\tilde{\Lambda}$) as in Section 12.3. The moments are easily calculated with the following lemma.

Lemma 8.2 *When* $\boldsymbol{\Sigma} = \gamma\mathbf{I}$, $\gamma > 0$, tr $\mathbf{V} \perp\!\!\!\perp |\mathbf{V}|/(\text{tr } \mathbf{V})^p$.

Proof. When $\boldsymbol{\Sigma} = \gamma\mathbf{I}$, clearly the distribution of $|\mathbf{V}|/(\text{tr } \mathbf{V})^p$ does not depend on γ. From the likelihood for $(\boldsymbol{\mu}, \gamma)$, which forms an exponential family, the minimal sufficient and complete statistic is $(\bar{\mathbf{x}}, \text{tr } \mathbf{V})$. The conclusion follows using Basu's[1] theorem. □

[1]Basu's theorem: If T is complete and sufficient for the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, then $T \perp\!\!\!\perp A$, for any ancillary statistic A. By definition, a statistic A is ancillary iff its distribution does not depend on θ.

But then, since

$$\tilde{\Lambda}\left(\frac{1}{p}\ \mathrm{tr}\ \mathbf{V}\right)^p = |\mathbf{V}|,$$

then

$$E\ \tilde{\Lambda}^h \cdot E\ \left(\frac{1}{p}\ \mathrm{tr}\ \mathbf{V}\right)^{ph} = E\ |\mathbf{V}|^h,$$

from which

$$E\ \tilde{\Lambda}^h = \frac{E\ |\mathbf{V}|^h}{E\ (\frac{1}{p}\ \mathrm{tr}\ \mathbf{V})^{ph}}.$$

Proposition 8.12 *When* $\mathbf{\Sigma} = \gamma\mathbf{I}$, $\gamma > 0$, *then*

$$E\ \tilde{\Lambda}^h = p^{ph}\frac{\Gamma\left(\frac{1}{2}(n-1)p\right)}{\Gamma\left(\frac{1}{2}(n-1)p+ph\right)}\frac{\Gamma_p\left(\frac{1}{2}(n-1)+h\right)}{\Gamma_p\left(\frac{1}{2}(n-1)\right)}.$$

Proof. We have $\mathbf{V} \sim \gamma W_p(n-1)$. The proof follows directly from the above remark in conjunction with Corollary 7.3, Proposition 7.2, and the definition of $\Gamma_p(\cdot)$ on page 93. Moments of chi-square distributions can be obtained from Section 3.2 (v. Problem 3.5.1). □

Finally, the exact distribution of $\tilde{\Lambda}$ can be characterized as a product of independent beta variables [Srivastava and Khatri (1979), p. 209].

Proposition 8.13 *The exact null distribution of* $\tilde{\Lambda}$ *is*

$$\tilde{\Lambda} \stackrel{d}{=} \prod_{i=1}^{p-1} \mathrm{beta}[\tfrac{1}{2}(n-1-i), i(\tfrac{1}{2}+\tfrac{1}{p})]; \tag{8.5}$$

i.e., $\tilde{\Lambda}$ *is distributed as the product of* $p-1$ *mutually independent beta variables.*

Proof. We make use of the multiplicative formula of Gauss [Erdélyi et al. (1953), p. 4]

$$\Gamma(mz) = (2\pi)^{-(m-1)/2} m^{mz-\frac{1}{2}} \prod_{r=0}^{m-1} \Gamma(z+\tfrac{r}{m}),\ m = 2, 3, \ldots,$$

with $m = p$ and $z = \frac{1}{2}(n-1), \frac{1}{2}(n-1)+h$. We can then rewrite the moments as

$$\begin{aligned} E\ \tilde{\Lambda}^h &= \frac{\prod_{i=1}^{p}\Gamma[\frac{1}{2}(n-1)+h-\frac{1}{2}(i-1)]}{\prod_{i=1}^{p}\Gamma[\frac{1}{2}(n-1)-\frac{1}{2}(i-1)]}\frac{\prod_{r=0}^{p-1}\Gamma[\frac{1}{2}(n-1)+\frac{r}{p}]}{\prod_{r=0}^{p-1}\Gamma[\frac{1}{2}(n-1)+h+\frac{r}{p}]} \\ &= \frac{\prod_{i=1}^{p-1}\Gamma[\frac{1}{2}(n-1)+h-\frac{1}{2}i]\ \Gamma[\frac{1}{2}(n-1)+\frac{i}{p}]}{\prod_{i=1}^{p-1}\Gamma[\frac{1}{2}(n-1)-\frac{1}{2}i]\ \Gamma[\frac{1}{2}(n-1)+h+\frac{i}{p}]}. \end{aligned}$$

It is then straightforward to check that all moments of order $h > 0$ on the left and right sides of $\stackrel{d}{=}$ in (8.5) are the same. Since the domain is the bounded interval $[0, 1]$, there is a unique distribution with these moments [Serfling (1980), p. 46]. □

The group $\mathbf{O}_p \times \mathbb{R}^p \times (\mathbb{R}\backslash\{0\})$ transforms the data as $\mathbf{x}_i \mapsto a\mathbf{H}\mathbf{x}_i + \mathbf{b}$, for any $\mathbf{H} \in \mathbf{O}_p$, $\mathbf{b} \in \mathbb{R}^p$, and $a \neq 0$. It preserves normality and induces transformations on the minimal sufficient statistic $(\bar{\mathbf{x}}, \mathbf{V})$ and parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\bar{\mathbf{x}} \mapsto a\mathbf{H}\bar{\mathbf{x}} + \mathbf{b}$, $\mathbf{V} \mapsto a^2\mathbf{H}\mathbf{V}\mathbf{H}'$, $\boldsymbol{\mu} \mapsto a\mathbf{H}\boldsymbol{\mu} + \mathbf{b}$, and $\boldsymbol{\Sigma} \mapsto a^2\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'$. Thus, the transformation also preserves the sphericity. A test function $f(\bar{\mathbf{x}}, \mathbf{V})$ is said to be invariant with respect to this group of transformations when it takes the same value on the original data as on the transformed data, i.e.,

$$f(\mathbf{y}, \mathbf{W}) = f(a\mathbf{H}\mathbf{y} + \mathbf{b}, a^2\mathbf{H}\mathbf{W}\mathbf{H}'),$$

$\forall(\mathbf{H}, \mathbf{b}, a) \in \mathbf{O}_p \times \mathbb{R}^p \times (\mathbb{R}\backslash\{0\})$, $\forall(\mathbf{y}, \mathbf{W}) \in \mathbb{R}^p \times \mathcal{P}_p$. This invariance property yields formidable simplifications.

First, if we diagonalize $\mathbf{V} = \mathbf{H}\mathbf{D}\mathbf{H}'$, where $\mathbf{D} = \text{diag}(l_1, \ldots, l_p)$, then choosing $a = l_p^{-1/2}$ and $\mathbf{b} = -a\mathbf{H}'\bar{\mathbf{x}}$, we find

$$\begin{aligned} f(\bar{\mathbf{x}}, \mathbf{V}) &= f(a\mathbf{H}'\bar{\mathbf{x}} + \mathbf{b}, a^2\mathbf{H}'\mathbf{V}\mathbf{H}) \\ &= f(\mathbf{0}, \tilde{\mathbf{D}}), \end{aligned}$$

where $\tilde{\mathbf{D}} = \text{diag}(l_1/l_p, \ldots, l_{p-1}/l_p, 1)$ depends on the sample only through the ratios $l_1/l_p, \ldots, l_{p-1}/l_p$. So, any invariant test can be written as a function of l_i/l_p, $i = 1, \ldots, p-1$.

Second, if we diagonalize $\boldsymbol{\Sigma} = \mathbf{G}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{G}'$, where $\mathbf{D}_{\boldsymbol{\lambda}}$ lists the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ on its diagonal, then choosing $a = \lambda_p^{-1/2}$ and $\mathbf{b} = -a\mathbf{G}'\boldsymbol{\mu}$, we find

$$\begin{aligned} a\mathbf{G}'\bar{\mathbf{x}} + \mathbf{b} &\sim N_p(\mathbf{0}, n^{-1}\tilde{\mathbf{D}}_{\boldsymbol{\lambda}}), \\ a^2\mathbf{G}'\mathbf{V}\mathbf{G} &\sim W_p(n-1, \tilde{\mathbf{D}}_{\boldsymbol{\lambda}}), \end{aligned}$$

where $\tilde{\mathbf{D}}_{\boldsymbol{\lambda}} = \text{diag}(\lambda_1/\lambda_p, \ldots, \lambda_{p-1}/\lambda_p, 1)$. Thus, the non-null distribution of any invariant test depends on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ only through the ratios $\lambda_1/\lambda_p, \ldots, \lambda_{p-1}/\lambda_p$. These invariance results are summarized in a proposition.

Proposition 8.14 *With respect to the above group of transformations, any invariant test depends on the minimal sufficient statistic* $(\bar{\mathbf{x}}, \mathbf{V})$ *only through the ratios* $l_1/l_p, \ldots, l_{p-1}/l_p$ *of eigenvalues of* $\mathbf{V}$. *The power function of any invariant test depends on* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *only through the ratios* $\lambda_1/\lambda_p, \ldots, \lambda_{p-1}/\lambda_p$ *of eigenvalues of* $\boldsymbol{\Sigma}$.

The LRT is obviously invariant. There is no uniformly most powerful invariant (UMPI) test for the sphericity hypothesis, but John (1971) showed the test based on $J = \text{tr } \mathbf{V}^2/(\text{tr } \mathbf{V})^2$ is locally most powerful (best)

invariant (LBI). The null distribution of the LBI test is given in John (1972).

8.7 Test of equality of variances

The data consist of a independent samples $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ i.i.d. $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $\boldsymbol{\Sigma}_i > \mathbf{0}$, $i = 1, \ldots, a$. Let $n = \sum_{i=1}^a n_i$ be the total number of observations. The hypothesis in question here is the equality of variances

$$H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a$$

which is being tested against all alternatives. Since the samples are independent, the likelihood function can be built immediately from (7.1),

$$\begin{aligned} &L(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_a, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_a) \\ &\quad \propto \prod_{i=1}^a |\boldsymbol{\Sigma}_i|^{-\frac{n_i}{2}} \text{ etr}\left\{-\tfrac{1}{2}\left[\mathbf{V}_i + n_i(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i)(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i)'\right]\boldsymbol{\Sigma}_i^{-1}\right\}, \end{aligned}$$

where, as usual,

$$\begin{aligned} \bar{\mathbf{x}}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}, \\ \mathbf{V}_i &= \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)', \ i = 1, \ldots, a. \end{aligned}$$

Without the restriction specified in H_0, the parameters are unrelated and, thus, the unrestricted MLE is just the usual $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i$ and $\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}\mathbf{V}_i$. Under H_0, however, we have $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}$, for some unknown $\boldsymbol{\Sigma}$, and, thus, $\hat{\hat{\boldsymbol{\mu}}}_i = \bar{\mathbf{x}}_i$ and $\hat{\hat{\boldsymbol{\Sigma}}} = \frac{1}{n}\mathbf{V}$, where $\mathbf{V} = \sum_{i=1}^a \mathbf{V}_i$ pools all the variances together. Thus, the LRT becomes

$$\begin{aligned} \Lambda &= \frac{L(\frac{1}{n}\mathbf{V}, \ldots, \frac{1}{n}\mathbf{V}, \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_a)}{L(\frac{1}{n_1}\mathbf{V}_1, \ldots, \frac{1}{n_a}\mathbf{V}_a, \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_a)} \\ &= \frac{\prod_{i=1}^a |\frac{1}{n}\mathbf{V}|^{-n_i/2}}{\prod_{i=1}^a |\frac{1}{n_i}\mathbf{V}_i|^{-n_i/2}} \frac{\exp(-\frac{1}{2}np)}{\exp(-\frac{1}{2}np)} \\ &= \frac{\prod_{i=1}^a |\mathbf{V}_i|^{n_i/2}}{|\mathbf{V}|^{n/2}} \frac{n^{pn/2}}{\prod_{i=1}^a n_i^{pn_i/2}}. \end{aligned}$$

Proposition 8.15 *The LRT for testing $H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a$ rejects the hypothesis for small values of*

$$\Lambda = \frac{\prod_{i=1}^a |\mathbf{V}_i|^{n_i/2}}{|\mathbf{V}|^{n/2}} \frac{n^{pn/2}}{\prod_{i=1}^a n_i^{pn_i/2}}.$$

The group $\mathbf{G}_p \times (\mathbb{R}^p)^a$ transforms the observations as $\mathbf{x}_{ij} \mapsto \mathbf{A}\mathbf{x}_{ij} + \mathbf{b}_i$, for any $\mathbf{A} \in \mathbf{G}_p$ and $\mathbf{b}_i \in \mathbb{R}^p$, $i = 1, \ldots, a$. This obviously preserves the normality and transforms the statistics as $\bar{\mathbf{x}}_i \mapsto \mathbf{A}\bar{\mathbf{x}}_i + \mathbf{b}_i$, $\mathbf{V}_i \mapsto \mathbf{A}\mathbf{V}_i\mathbf{A}'$, and $\mathbf{V} \mapsto \mathbf{A}\mathbf{V}\mathbf{A}'$, and induces the parameter transformation $\boldsymbol{\Sigma}_i \mapsto \mathbf{A}\boldsymbol{\Sigma}_i\mathbf{A}'$. The hypothesis H_0 is thus also preserved by this group of transformations. Therefore, the LRT statistic evaluated at the transformed data $\mathbf{A}\mathbf{x}_{ij} + \mathbf{b}_i$ is

$$\begin{aligned}\Lambda &= \frac{\prod_{i=1}^a |\mathbf{A}\mathbf{V}_i\mathbf{A}'|^{n_i/2}}{|\mathbf{A}\mathbf{V}\mathbf{A}'|^{n/2}} \frac{n^{pn/2}}{\prod_{i=1}^a n_i^{pn_i/2}} \\ &= \frac{\prod_{i=1}^a |\mathbf{V}_i|^{n_i/2}}{|\mathbf{V}|^{n/2}} \frac{n^{pn/2}}{\prod_{i=1}^a n_i^{pn_i/2}},\end{aligned}$$

which is identical to the LRT statistic evaluated at the original data $\mathbf{x}_{ij}$. We say that the LRT is invariant with respect to this group of transformations. In general, a test function $f(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_a, \mathbf{V}_1, \ldots, \mathbf{V}_a)$ is termed invariant iff

$$\begin{aligned}&f(\mathbf{y}_1, \ldots, \mathbf{y}_a, \mathbf{W}_1, \ldots, \mathbf{W}_a) \\ &\quad = f(\mathbf{A}\mathbf{y}_1 + \mathbf{b}_1, \ldots, \mathbf{A}\mathbf{y}_a + \mathbf{b}_a, \mathbf{A}\mathbf{W}_1\mathbf{A}', \ldots, \mathbf{A}\mathbf{W}_a\mathbf{A}'),\end{aligned}$$

$\forall(\mathbf{A}, \mathbf{b}_1, \ldots, \mathbf{b}_a) \in \mathbf{G}_p \times (\mathbb{R}^p)^a, \forall(\mathbf{y}_1, \ldots, \mathbf{y}_a, \mathbf{W}_1, \ldots, \mathbf{W}_a) \in (\mathbb{R}^p)^a \times (\mathcal{P}_p)^a$. This has important consequences.

First, by deliberately choosing $\mathbf{A} = \boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}$ under H_0, and $\mathbf{b}_i = -\mathbf{A}\bar{\mathbf{x}}_i$, it is clear that $\mathbf{A}\mathbf{V}_i\mathbf{A}' \sim W_p(n_i - 1)$ do not involve any unknown parameters. Thus, the null distribution of any invariant test function

$$\begin{aligned}&f(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_a, \mathbf{V}_1, \ldots, \mathbf{V}_a) \\ &\quad = f(\mathbf{A}\bar{\mathbf{x}}_1 + \mathbf{b}_1, \ldots, \mathbf{A}\bar{\mathbf{x}}_a + \mathbf{b}_a, \mathbf{A}\mathbf{V}_1\mathbf{A}', \ldots, \mathbf{A}\mathbf{V}_a\mathbf{A}') \\ &\quad = f(\mathbf{0}, \ldots, \mathbf{0}, \mathbf{A}\mathbf{V}_1\mathbf{A}', \ldots, \mathbf{A}\mathbf{V}_a\mathbf{A}')\end{aligned}$$

such as Λ is parameter free. Note that we need only consider test functions of this form since $(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_a, \mathbf{V}_1, \ldots, \mathbf{V}_a)$ is sufficient for $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_a)$.

Second, in the special case $a = 2$, diagonalize $\mathbf{V}_1^{-1/2}\mathbf{V}_2\mathbf{V}_1^{-1/2} = \mathbf{H}\mathbf{D}\mathbf{H}'$, where $\mathbf{H} \in \mathbf{O}_p$ and $\mathbf{D} = \operatorname{diag}(l_1, \ldots, l_p)$ contains the eigenvalues of $\mathbf{V}_1^{-1}\mathbf{V}_2$. This time by deliberately choosing $\mathbf{A} = \mathbf{H}'\mathbf{V}_1^{-1/2}$, $\mathbf{b}_i = -\mathbf{A}\bar{\mathbf{x}}_i$, we find that for any invariant test

$$\begin{aligned}f(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{V}_1, \mathbf{V}_2) &= f(\mathbf{0}, \mathbf{0}, \mathbf{A}\mathbf{V}_1\mathbf{A}', \mathbf{A}\mathbf{V}_2\mathbf{A}') \\ &= f(\mathbf{0}, \mathbf{0}, \mathbf{I}, \mathbf{D})\end{aligned}$$

is a function of $l_1, \ldots, l_p$ only. Thus, any invariant test function depends on the data only through $l_1, \ldots, l_p$. Similarly, diagonalizing $\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2} = \mathbf{G}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{G}'$, where $\mathbf{G} \in \mathbf{G}_p$ and $\mathbf{D}_{\boldsymbol{\lambda}}$ contains the eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$, we have after choosing $\mathbf{A} = \mathbf{G}'\boldsymbol{\Sigma}_1^{-1/2}$, $\mathbf{A}\mathbf{V}_1\mathbf{A}' \sim W_p(n_1 - 1)$ and

$\mathbf{A}\mathbf{V}_2\mathbf{A}' \sim W_p(n_2-1, \mathbf{D}_{\boldsymbol{\lambda}})$. Thus, the non-null distribution of any invariant test function depends on $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ only through the eigenvalues of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$.

Proposition 8.16 *With respect to the group of transformations* $\mathbf{G}_p \times (\mathbb{R}^p)^2$*, any invariant test for testing* $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ *depends on* $(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{V}_1, \mathbf{V}_2)$ *only through the eigenvalues* $l_1, \ldots, l_p$ *of* $\mathbf{V}_1^{-1}\mathbf{V}_2$*. The power function of any invariant test depends on* $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ *only through the eigenvalues* $\lambda_1, \ldots, \lambda_p$ *of* $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$*.*

For example, the LRT when $a = 2$ can be written

$$\Lambda = \frac{n^{pn/2}}{n_1^{pn_1/2} n_2^{pn_2/2}} \prod_{i=1}^{p} \frac{l_i^{n_2/2}}{(1+l_i)^{n/2}}.$$

An alternative invariant test function [Nagao (1973)] is

$$N = \tfrac{1}{2} \sum_{i=1}^{a} n_i \ \mathrm{tr} \left(\frac{n}{n_i} \mathbf{V}_i \mathbf{V}^{-1} - \mathbf{I} \right)^2.$$

Continuing now with the moments of the null distribution of the LRT, we comment first on a result of unbiasedness. Although the LRT is a biased test, Perlman (1980) proved that the slight modification

$$\Lambda^* = \frac{\prod_{i=1}^{a} |\mathbf{V}_i|^{m_i/2}}{|\mathbf{V}|^{m/2}} \frac{m^{pm/2}}{\prod_{i=1}^{a} m_i^{pm_i/2}},$$

where the sample sizes n_i are replaced by the corresponding degrees of freedom $m_i = n_i - 1$ and $m = \sum_{i=1}^{a} m_i = n - a$, yields an unbiased test. We will, thus, concentrate on the latter. It was Bartlett (1937) who first proposed the use of the modified LRT, Λ^*. For $a = 2$, unbiasedness of Λ^* was established earlier by Sugiura and Nagao (1968), whereas Srivastava, Khatri, and Carter (1978) proved a monotonicity property stronger than unbiasedness.

The null moments of Λ^* is a simple consequence of invariance coupled with the normalizing constant $c_{p,m} = [2^{mp/2}\Gamma_p(\frac{1}{2}m)]^{-1}$ of a $W_p(m)$ p.d.f.

Proposition 8.17 *Under* H_0*, the moments of the modified LRT* Λ^* *are given by*

$$E\ \Lambda^{*h} = \frac{m^{pmh/2}}{\prod_{i=1}^{a} m_i^{pm_i h/2}} \frac{\Gamma_p(\frac{1}{2}m)}{\Gamma_p[\frac{1}{2}m(1+h)]} \prod_{i=1}^{a} \frac{\Gamma_p[\frac{1}{2}m_i(1+h)]}{\Gamma_p(\frac{1}{2}m_i)}.$$

Proof. Under H_0, by invariance, we can assume $\boldsymbol{\Sigma}_i = \mathbf{I}$ and $\mathbf{V}_i \sim W_p(m_i)$ are independently distributed. Thus, from the $W_p(m_i)$ densities, we have

$$E\ \Lambda^{*h} = \prod_{i=1}^{a} c_{p,m_i} \int_{\mathbf{V}_1 > \mathbf{0}} \cdots \int_{\mathbf{V}_a > \mathbf{0}} |\mathbf{V}|^{-mh/2}$$

$$\cdot \prod_{i=1}^{a} |\mathbf{V}_i|^{[m_i(1+h)-p-1]/2} \text{etr}(-\tfrac{1}{2}\mathbf{V}_i) d\mathbf{V}_1 \cdots d\mathbf{V}_a.$$

The integrand is seen to contain the p.d.f. of $\mathbf{V}_i \sim W_p(m_i(1+h))$ independently distributed. However, when this is the case $\mathbf{V} \sim W_p(m(1+h))$. Thus, we find

$$E\ \Lambda^{*h} = \frac{\prod_{i=1}^{a} c_{p,m_i}}{\prod_{i=1}^{a} c_{p,m_i(1+h)}}\ E\ |\mathbf{V}|^{-mh/2},$$

where $\mathbf{V} \sim W_p(m(1+h))$. Using the moments of the generalized variance in Problem 7.5.6 and simplifying, the conclusion is reached. □

An accurate approximation to the null distribution of the modified LRT Λ^* by asymptotic expansion of high order is discussed in Example 12.4.

8.8 Asymptotic distributions of eigenvalues

Based on a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, several tests on the variance $\boldsymbol{\Sigma}$ are a function of the eigenvalues of $\mathbf{V} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \sim W_p(n-1, \boldsymbol{\Sigma})$. It was seen that an invariant test for sphericity, $H_0 : \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, depends only on $(l_1/l_p, \ldots, l_{p-1}/l_p)'$ where $l_1 \geq \cdots \geq l_p$ are the eigenvalues of $\mathbf{V}$. Also, in the two independent samples problem,

$$\begin{aligned} &\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1} \quad \text{i.i.d.} \quad N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \\ &\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2} \quad \text{i.i.d.} \quad N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \end{aligned}$$

an invariant test for the equality of variances, $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, depends only on the eigenvalues of $\mathbf{V}_1^{-1}\mathbf{V}_2$, where

$$\mathbf{V}_i = \sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \sim W_p(n_i - 1, \boldsymbol{\Sigma}_i),\ i = 1, 2.$$

The distribution of eigenvalues of various random matrices thus plays an important role in testing hypotheses.

8.8.1 The one-sample problem

We investigate the asymptotic distribution of the eigenvalues $l_1, \ldots, l_p$ of $\mathbf{V} \sim W_p(m, \boldsymbol{\Sigma})$. We already know there exists $\mathbf{H} \in \mathbf{O}_p$ such that $\mathbf{H}'\boldsymbol{\Sigma}\mathbf{H} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$, and since $\mathbf{V}$ and $\mathbf{H}'\mathbf{V}\mathbf{H}$ have the same eigenvalues, we can assume at the outset that $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$ is diagonal. An effective method for such problems is to write

$$\mathbf{S} = \frac{\mathbf{V}}{m} = \boldsymbol{\Lambda} + m^{-1/2}\mathbf{V}^{(1)},$$

where $\mathbf{V}^{(1)} = m^{1/2}(\mathbf{S} - \mathbf{\Lambda})$ is $O_p(1)$, and expand the eigenvalues of $\mathbf{S}$, l_i/m, around λ_i in powers of $m^{-1/2}$. This is called the *perturbation method* [Bellman (1960), Kato (1982)]. We now clearly outline the steps to obtain an approximation, with remainder of the order $O(m^{-1})$, to the distribution function of a nearly arbitrary function $f(\mathbf{l}/m)$ of $\mathbf{l} = (l_1, \ldots, l_p)'$.

Step 1: Perturbation method

More generally, consider a diagonal matrix $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and assume that the perturbation of $\mathbf{\Lambda}$ can be expressed as a power series in ϵ as follows:

$$\mathbf{R} = \mathbf{\Lambda} + \epsilon \mathbf{V}^{(1)} + \epsilon^2 \mathbf{V}^{(2)} + O(\epsilon^3),$$

where $\mathbf{V}^{(j)}$, $j = 1, 2$, are symmetric and ϵ is a small real number. We shall discuss the case when λ_α is *distinct* from the other $p-1$ eigenvalues. Let l_α be the αth eigenvalue of $\mathbf{R}$ and $\mathbf{c}_\alpha = (c_{1\alpha}, \ldots, c_{p\alpha})'$ the corresponding normalized eigenvector with $c_{\alpha\alpha} > 0$. The quantities l_α and $\mathbf{c}_\alpha$ can be assumed of the form [Bellman (1960), p. 61]

$$l_\alpha = \lambda_\alpha + \epsilon \lambda_\alpha^{(1)} + \epsilon^2 \lambda_\alpha^{(2)} + O(\epsilon^3), \tag{8.6}$$

$$\mathbf{c}_\alpha = \mathbf{e}_\alpha + \epsilon \sum_{i=1}^{p} a_{i\alpha}^{(1)} \mathbf{e}_i + \epsilon^2 \sum_{i=1}^{p} a_{i\alpha}^{(2)} \mathbf{e}_i + O(\epsilon^3), \tag{8.7}$$

where $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)'$ is the ith canonical basis vector. We determine the unknown coefficients $\lambda_\alpha^{(1)}$, $\lambda_\alpha^{(2)}$, $a_{i\alpha}^{(1)}$, and $a_{i\alpha}^{(2)}$ by substituting (8.6) and (8.7) into the equation

$$\mathbf{R}\mathbf{c}_\alpha = l_\alpha \mathbf{c}_\alpha$$

and equating the coefficients of the powers of ϵ. This gives

$$[\mathbf{\Lambda} + \epsilon \mathbf{V}^{(1)} + \epsilon^2 \mathbf{V}^{(2)} + O(\epsilon^3)][\mathbf{e}_\alpha + \epsilon \sum_{i=1}^{p} a_{i\alpha}^{(1)} \mathbf{e}_i + \epsilon^2 \sum_{i=1}^{p} a_{i\alpha}^{(2)} \mathbf{e}_i + O(\epsilon^3)]$$

$$= [\lambda_\alpha + \epsilon \lambda_\alpha^{(1)} + \epsilon^2 \lambda_\alpha^{(2)} + O(\epsilon^3)][\mathbf{e}_\alpha + \epsilon \sum_{i=1}^{p} a_{i\alpha}^{(1)} \mathbf{e}_i + \epsilon^2 \sum_{i=1}^{p} a_{i\alpha}^{(2)} \mathbf{e}_i + O(\epsilon^3)],$$

and equating the coefficients, we obtain the equations

$$\lambda_\alpha \mathbf{e}_\alpha = \lambda_\alpha \mathbf{e}_\alpha, \tag{8.8}$$

$$\sum_{i=1}^{p} a_{i\alpha}^{(1)} \lambda_i \mathbf{e}_i + \mathbf{v}_\alpha^{(1)} = \sum_{i=1}^{p} a_{i\alpha}^{(1)} \lambda_\alpha \mathbf{e}_i + \lambda_\alpha^{(1)} \mathbf{e}_\alpha, \tag{8.9}$$

and

$$\sum_{i=1}^{p} a_{i\alpha}^{(2)} \lambda_i \mathbf{e}_i + \sum_{i=1}^{p} a_{i\alpha}^{(1)} \mathbf{v}_i^{(1)} + \mathbf{v}_\alpha^{(2)}$$

$$= \sum_{i=1}^{p} a_{i\alpha}^{(2)} \lambda_\alpha \mathbf{e}_i + \sum_{i=1}^{p} a_{i\alpha}^{(1)} \lambda_\alpha^{(1)} \mathbf{e}_i + \lambda_\alpha^{(2)} \mathbf{e}_\alpha, \tag{8.10}$$

where $\mathbf{V}^{(j)} = (\mathbf{v}_1^{(j)}, \ldots, \mathbf{v}_p^{(j)})$, $j = 1, 2$. The αth component of (8.9) yields

$$a_{\alpha\alpha}^{(1)}\lambda_\alpha + v_{\alpha\alpha}^{(1)} = a_{\alpha\alpha}^{(1)}\lambda_\alpha + \lambda_\alpha^{(1)}$$

from which $\lambda_\alpha^{(1)} = v_{\alpha\alpha}^{(1)}$. The component $i \neq \alpha$ of the same equation yields

$$a_{i\alpha}^{(1)}\lambda_i + v_{i\alpha}^{(1)} = a_{i\alpha}^{(1)}\lambda_\alpha,$$

from which $a_{i\alpha}^{(1)} = -v_{i\alpha}^{(1)}\lambda_{i\alpha}$, $i \neq \alpha$, where

$$\lambda_{i\alpha} = 1/(\lambda_i - \lambda_\alpha).$$

Note that $a_{\alpha\alpha}^{(1)}$ can be chosen arbitrarily and we set $a_{\alpha\alpha}^{(1)} = 0$ here. The unknown quantities $\lambda_\alpha^{(2)}$ and $a_{i\alpha}^{(2)}$ can be determined similarly using (8.10). The expansions (8.6)-(8.7) from the perturbation analysis thus take the final form

$$\begin{aligned} l_\alpha &= \lambda_\alpha + \epsilon v_{\alpha\alpha}^{(1)} + \epsilon^2 \left[v_{\alpha\alpha}^{(2)} + \sum_{\beta\neq\alpha} \lambda_{\alpha\beta} {v_{\alpha\beta}^{(1)}}^2 \right] + O(\epsilon^3), \qquad (8.11) \\ c_{i\alpha} &= -\lambda_{i\alpha} \left\{ \epsilon v_{i\alpha}^{(1)} + \epsilon^2 \left[\lambda_{i\alpha} v_{i\alpha}^{(1)} v_{\alpha\alpha}^{(1)} + \sum_{\beta\neq\alpha} \lambda_{\alpha\beta} v_{i\beta}^{(1)} v_{\beta\alpha}^{(1)} \right] \right\} \\ &\quad + O(\epsilon^3),\ i \neq \alpha, \\ c_{\alpha\alpha} &= 1 + \epsilon^2 \left[-\tfrac{1}{2} \sum_{\beta\neq\alpha} \lambda_{\alpha\beta}^2 {v_{\alpha\beta}^{(1)}}^2 \right] + O(\epsilon^3). \end{aligned}$$

Returning to our one-sample problem, assuming λ_α is distinct, the eigenvalue l_α/m of $\mathbf{S}$ can be expanded by setting $\mathbf{V}^{(2)} = \mathbf{0}$ as

$$l_\alpha/m = \lambda_\alpha + m^{-1/2} v_{\alpha\alpha}^{(1)} + m^{-1} \sum_{\beta\neq\alpha} \lambda_{\alpha\beta} {v_{\alpha\beta}^{(1)}}^2 + O_p(m^{-3/2}). \qquad (8.12)$$

Step 2: Taylor series of $f(\mathbf{l}/m)$
Assuming $f(\cdot)$ is continuously differentiable in a neighborhood of

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)',$$

we can write the Taylor series around $\boldsymbol{\lambda}$,

$$\begin{aligned} f(\mathbf{l}/m) &= f(\boldsymbol{\lambda}) + Df(\boldsymbol{\lambda})(\mathbf{l}/m - \boldsymbol{\lambda}) \\ &\quad + \tfrac{1}{2}(\mathbf{l}/m - \boldsymbol{\lambda})' D^2 f(\boldsymbol{\lambda})(\mathbf{l}/m - \boldsymbol{\lambda}) + O_p(m^{-3/2}). \end{aligned}$$

Upon using (8.12), this becomes

$$\begin{aligned} f(\mathbf{l}/m) &= f(\boldsymbol{\lambda}) + m^{-1/2} \sum_{i=1}^p f_i v_{ii}^{(1)} + m^{-1} \sum_{i=1}^p f_i \sum_{\beta\neq i} \lambda_{i\beta} {v_{i\beta}^{(1)}}^2 \\ &\quad + \tfrac{1}{2} m^{-1} \sum_{i=1}^p \sum_{j=1}^p f_{ij} v_{ii}^{(1)} v_{jj}^{(1)} + O_p(m^{-3/2}), \qquad (8.13) \end{aligned}$$

where

$$\begin{aligned} Df(\boldsymbol{\lambda}) &= (f_1, \ldots, f_p)' = (\partial f(\boldsymbol{\lambda})/\partial\lambda_i)\,, \\ D^2 f(\boldsymbol{\lambda}) &= (f_{ij}) = \left(\partial^2 f(\boldsymbol{\lambda})/\partial\lambda_i \partial\lambda_j\right). \end{aligned}$$

Step 3: Expansion of the characteristic function
The characteristic function of $m^{1/2}\,[f(\mathbf{l}/m) - f(\boldsymbol{\lambda})]$ thus becomes

$$\begin{aligned} & E\ \exp\{itm^{1/2}[f(\mathbf{l}/m) - f(\boldsymbol{\lambda})]\} \\ &= E\ \exp\left(it\sum_{i=1}^{p} f_i v_{ii}^{(1)}\right) \exp\left[\frac{it}{\sqrt{m}}\left(\sum_{i=1}^{p} f_i \sum_{\beta\neq i} \lambda_{i\beta} {v_{i\beta}^{(1)}}^2\right.\right. \\ &\quad \left.\left. +\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} f_{ij} v_{ii}^{(1)} v_{jj}^{(1)}\right) + O_p(m^{-1})\right] \\ &= E\ \exp\left(it\sum_{i=1}^{p} f_i v_{ii}^{(1)}\right)\left[1 + \frac{it}{\sqrt{m}}\left(\sum_{i=1}^{p} f_i \sum_{\beta\neq i} \lambda_{i\beta} {v_{i\beta}^{(1)}}^2\right.\right. \\ &\quad \left.\left. +\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} f_{ij} v_{ii}^{(1)} v_{jj}^{(1)}\right) + O_p(m^{-1})\right]. \end{aligned} \tag{8.14}$$

We need then to evaluate the following expectations in (8.14):

$$E\ \exp\left(it\sum_{i=1}^{p} f_i v_{ii}^{(1)}\right), \tag{8.15}$$

$$E\ \exp\left(it\sum_{i=1}^{p} f_i v_{ii}^{(1)}\right) \cdot {v_{i\beta}^{(1)}}^2,\ \beta \neq i, \tag{8.16}$$

$$E\ \exp\left(it\sum_{i=1}^{p} f_i v_{ii}^{(1)}\right) \cdot v_{ii}^{(1)} v_{jj}^{(1)}. \tag{8.17}$$

Step 4: Sugiura's lemma [Sugiura (1973)]
Let $\mathbf{V} \sim W_p(m, \boldsymbol{\Sigma})$ and $\mathbf{S} = \mathbf{V}/m$.

Lemma 8.3 *Let $g(\mathbf{S})$ be an analytic function at $\mathbf{S} = \boldsymbol{\Sigma}$ and put $\mathbf{T} = m^{1/2}(\mathbf{S} - \boldsymbol{\Sigma})$. Define a matrix of differential operators by*

$$\boldsymbol{\partial} = \left(\tfrac{1}{2}(1 + \delta_{ij})\partial/\partial_{ij}\right)$$

applied to the function $g(\boldsymbol{\Gamma})$ of a symmetric matrix $\boldsymbol{\Gamma} = (\gamma_{ij})$. Then, for any symmetric matrix $\mathbf{A}$ and sufficiently large m,

$$E\ g(\mathbf{S})\mathrm{etr}(it\mathbf{AT}) = \mathrm{etr}[-t^2(\mathbf{A}\boldsymbol{\Sigma})^2] \cdot \left[1 + m^{-1/2}\sum_{j=1}^{2} d_{2j-1}(it)^{2j-1}\right.$$

$$+m^{-1}\sum_{j=1}^{3} g_{2j}(it)^{2j} + O(m^{-3/2})\Bigg] g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Sigma}},$$

where each coefficient is given by

$$\begin{aligned}
d_1 &= 2 \text{ tr } (\Sigma\mathbf{A}\mathbf{\Sigma}\boldsymbol{\partial}),\\
d_3 &= \frac{4}{3} \text{ tr } (\mathbf{\Sigma}\mathbf{A})^3,\\
g_0 &= \text{tr } (\mathbf{\Sigma}\boldsymbol{\partial})^2,\\
g_2 &= 4 \text{ tr } (\mathbf{\Sigma}\mathbf{A})^2\mathbf{\Sigma}\boldsymbol{\partial} + \frac{1}{2}d_1^2,\\
g_4 &= 2 \text{ tr } (\mathbf{\Sigma}\mathbf{A})^4 + d_1 d_3,\\
g_6 &= \frac{1}{2}d_3^2.
\end{aligned}$$

Before presenting the proof, we comment on Taylor series and differential operators. An analytic function $g(x)$, at x_0, of a real variable x can be written as a Taylor series

$$\begin{aligned}
g(x) &= g(x_0) + \sum_{j=1}^{\infty} \frac{g^{(j)}(x_0)}{j!}(x-x_0)^j\\
&= e^{(x-x_0)\partial}\cdot g(x)|_{x=x_0}\\
&= [1 + (x-x_0)\partial + \tfrac{1}{2}(x-x_0)^2\partial^2 + \cdots]g(x)|_{x=x_0},
\end{aligned}$$

where $\partial^j g(x)|_{x=x_0} = \partial^j g(x_0)/\partial x^j$ is the jth derivative of g evaluated at x_0. In the same way, for a function $g(\mathbf{S})$ analytic at $\mathbf{\Sigma}$, of a symmetric matrix $\mathbf{S}$, we have

$$g(\mathbf{S}) = \{\text{etr}(\mathbf{S}-\mathbf{\Sigma})\boldsymbol{\partial}\}\, g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Sigma}}.$$

Proof. Note that $\mathbf{S} \xrightarrow{p} \mathbf{\Sigma}$. Taylor series expansion of $g(\mathbf{S})$ at $\mathbf{\Sigma}$ gives

$$g(\mathbf{S}) = \{\text{etr}(\mathbf{S}-\mathbf{\Sigma})\boldsymbol{\partial}\}\, g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Sigma}}.$$

After multiplying by $\text{etr}(it\mathbf{A}\mathbf{T})$ and taking expectations with respect to $\mathbf{V} \sim W_p(m, \mathbf{\Sigma})$, we get

$$\begin{aligned}
E\ \text{etr}(it\mathbf{A}\mathbf{T})g(\mathbf{S}) &= |\mathbf{I} - 2m^{-1/2}it\mathbf{A}\mathbf{\Sigma} - 2m^{-1}\mathbf{\Sigma}\boldsymbol{\partial}|^{-m/2}\\
&\quad \cdot \Big\{\text{etr}(-m^{1/2}it\mathbf{A}\mathbf{\Sigma} - \mathbf{\Sigma}\boldsymbol{\partial})\Big\}\, g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Sigma}}.
\end{aligned}$$

The above determinant can be arranged according to powers of m as in Sugiura and Nagao (1971). □

Evaluation of (8.15)

Let $\mathbf{A} = \text{diag}(f_1, \ldots, f_p)$, $g(\mathbf{\Gamma}) \equiv 1$, and $\mathbf{\Sigma} = \mathbf{\Lambda}$, and note that

$$\begin{aligned} \text{tr } \mathbf{A}\mathbf{V}^{(1)} &= \sum_{i=1}^{p} f_i v_{ii}^{(1)}, \\ \text{tr}(\mathbf{A}\mathbf{\Lambda})^2 &= \sum_{i=1}^{p} f_i^2 \lambda_i^2 \equiv \tau^2/2 \text{ (say)}. \end{aligned}$$

Since $d_1 g(\mathbf{\Gamma}) = 0$, the lemma yields

$$E \exp\left(it \sum_{i=1}^{p} f_i v_{ii}^{(1)}\right) = \exp(-\tfrac{1}{2}t^2\tau^2)[1 + m^{-1/2} d_3 (it)^3 + O(m^{-1})],$$

where $d_3 = (4/3)\sum_{i=1}^{p} f_i^3 \lambda_i^3$.

Evaluation of (8.16)

Let $\mathbf{A} = \text{diag}(f_1, \ldots, f_p)$ and $g(\mathbf{\Gamma}) = m\gamma_{i\beta}^2$, $i \neq \beta$. Note that $g(\mathbf{\Lambda}) = 0$ and the differential operator d_1,

$$d_1 = 2 \sum_{k=1}^{p} f_k \lambda_k^2 \partial_{kk},$$

is a linear combination of $\partial_{kk} \equiv \partial/\partial_{kk}$ and, thus, $d_1 g(\mathbf{\Gamma}) = 0$. For similar reasons, we also have $g_2 g(\mathbf{\Gamma}) = 0$ and $d_3 g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Lambda}} = g_4 g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Lambda}} = g_6 g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Lambda}} = 0$, which implies

$$E \exp\left(it \sum_{i=1}^{p} f_i v_{ii}^{(1)}\right) \cdot {v_{i\beta}^{(1)}}^2 = \exp(-\tfrac{1}{2}t^2\tau^2)[m^{-1} g_0 + O(m^{-3/2})] g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Lambda}}.$$

However, g_0 is the differential operator

$$g_0 = \text{tr}(\mathbf{\Lambda}\boldsymbol{\partial})^2 = \sum_{k=1}^{p}\sum_{l=1}^{p} \lambda_k \lambda_l \partial_{kl}^2$$

and, thus,

$$g_0 g(\mathbf{\Gamma})|_{\mathbf{\Gamma}=\mathbf{\Lambda}} = \sum_{k=1}^{p}\sum_{l=1}^{p} \lambda_k \lambda_l \partial_{kl}^2 (m\gamma_{i\beta}^2)|_{\mathbf{\Gamma}=\mathbf{\Lambda}} = m\lambda_i \lambda_\beta.$$

Hence, we get

$$E \exp\left(it \sum_{i=1}^{p} f_i v_{ii}^{(1)}\right) \cdot {v_{i\beta}^{(1)}}^2 = \exp(-\tfrac{1}{2}t^2\tau^2)[\lambda_i \lambda_\beta + O(m^{-1/2})].$$

Evaluation of (8.17)

Similarly, letting $\mathbf{A} = \text{diag}(f_1, \ldots, f_p)$ and $g(\mathbf{\Gamma}) = m(\gamma_{ii} - \lambda_i)(\gamma_{jj} - \lambda_j)$, we find

$$E \exp\left(it \sum_{i=1}^{p} f_i v_{ii}^{(1)}\right) \cdot v_{ii}^{(1)} v_{jj}^{(1)} = \exp(-\tfrac{1}{2}t^2\tau^2)[2\lambda_i^2 \delta_{ij} + O(m^{-1/2})].$$

We now return to the expansion of the characteristic function in (8.14). Hence, altogether, the expansion of the characteristic function becomes
$E\ \exp\{itm^{1/2}[f(\mathbf{l}/m) - f(\boldsymbol{\lambda})]/\tau\}$

$$= \exp(-\tfrac{1}{2}t^2\tau^2)[1 + m^{-1/2}\sum_{j=1}^{2}\frac{a_{2j-1}}{\tau^{2j-1}}(it)^{2j-1} + O(m^{-1})],$$

where

$$a_1 = \sum_{i=1}^{p}\sum_{\beta\neq i} f_i\lambda_{i\beta}\lambda_i\lambda_\beta + \sum_{i=1}^{p} f_{ii}\lambda_i^2, \tag{8.18}$$

$$a_3 = \frac{4}{3}\sum_{i=1}^{p} f_i^3\lambda_i^3. \tag{8.19}$$

Step 5: Inversion of the characteristic function
Using the inversion formula (2.2), an expansion for the density function of

$$s = m^{1/2}[f(\mathbf{l}/m) - f(\boldsymbol{\lambda})]/\tau$$

is

$$\begin{aligned} f(s) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-its}e^{-\frac{1}{2}t^2}dt \\ &\quad + m^{-1/2}\sum_{j=1}^{2}\frac{a_{2j-1}}{\tau^{2j-1}}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-its}e^{-\frac{1}{2}t^2}(it)^{2j-1}dt + O(m^{-1}) \\ &= \phi(s) - m^{-1/2}\sum_{j=1}^{2}\frac{a_{2j-1}}{\tau^{2j-1}}\phi^{(2j-1)}(s) + O(m^{-1}), \end{aligned}$$

and similarly for the distribution function of s,

$$F(s) = \Phi(s) - m^{-1/2}\sum_{j=1}^{2}\frac{a_{2j-1}}{\tau^{2j-1}}\Phi^{(2j-1)}(s) + O(m^{-1}),$$

where ϕ and Φ are respectively the density function and distribution function of the standard normal distribution. We have proved:

Proposition 8.18 *Let $f(\cdot)$ be continuously differentiable in a neighborhood of $\boldsymbol{\lambda}$. If the population eigenvalues λ_α are all distinct and $\tau^2 = 2\sum_{i=1}^{p} f_i^2\lambda_i^2 \neq 0$, then the distribution function of*

$$s = m^{1/2}[f(\mathbf{l}/m) - f(\boldsymbol{\lambda})]/\tau$$

can be expanded for large m as

$$\Phi(s) - m^{-1/2}\sum_{j=1}^{2}\frac{a_{2j-1}}{\tau^{2j-1}}\Phi^{(2j-1)}(s) + O(m^{-1}).$$

Corollary 8.3 *Let $f(\cdot)$ be continuously differentiable in a neighborhood of $\boldsymbol{\lambda}$. If the population eigenvalues λ_α are all distinct and $\tau^2 = 2\sum_{i=1}^p f_i^2\lambda_i^2 \neq 0$, then the limiting distribution is given by*

$$s = m^{1/2}[f(\mathbf{l}/m) - f(\boldsymbol{\lambda})]/\tau \xrightarrow{d} N(0,1).$$

For an individual eigenvalue the expansion follows immediately.

Corollary 8.4 *Let l_α be the αth largest eigenvalue of $\mathbf{V} \sim W_p(m, \boldsymbol{\Lambda})$. If λ_α is distinct from all other $p-1$ eigenvalues, the distribution function of*

$$s = m^{1/2}(l_\alpha/m - \lambda_\alpha)/(\sqrt{2}\lambda_\alpha)$$

can be expanded for large m as

$$\Phi(s) - m^{-1/2}\sum_{j=1}^{2}\frac{a_{2j-1}}{\tau^{2j-1}}\Phi^{(2j-1)}(s) + O(m^{-1}),$$

where

$$\begin{aligned} a_1 &= \sum_{\beta\neq\alpha}\lambda_\alpha\lambda_\beta/(\lambda_\alpha - \lambda_\beta), \\ a_3 &= \frac{4}{3}\lambda_\alpha^3. \end{aligned}$$

The sample eigenvalues are asymptotically independent, as the following corollary shows.

Corollary 8.5 *Let $\mathbf{l} = (l_1, \ldots, l_p)'$ be the eigenvalues of $\mathbf{V} \sim W_p(m, \boldsymbol{\Lambda})$. If the population eigenvalues λ_α are all distinct, then the joint limiting distribution is given by*

$$(m/2)^{1/2}\boldsymbol{\Lambda}^{-1}(\mathbf{l}/m - \boldsymbol{\lambda}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}).$$

Proof. From (8.12), we can write

$$m^{1/2}(\mathbf{l}/m - \boldsymbol{\lambda}) = (v_{11}^{(1)}, \ldots, v_{pp}^{(1)})' + O_p(m^{-1/2}) \equiv \mathbf{v}^{(1)} + O_p(m^{-1/2}).$$

The asymptotic distribution of $\mathbf{V}^{(1)}$ was derived in Section 6.3, and for the marginal $\mathbf{v}^{(1)}$, we find, using (6.1), $\mathbf{v}^{(1)} \xrightarrow{d} N_p(\mathbf{0}, 2\boldsymbol{\Lambda}^2)$. □

The asymptotic expansion in Corollary 8.4 gives the first two terms of a more accurate approximation, with remainder $O(m^{-3/2})$,

$$\Phi(s) - m^{-1/2}\sum_{j=1}^{2}\frac{a_{2j-1}}{\tau^{2j-1}}\Phi^{(2j-1)}(s) + m^{-1}\sum_{j=1}^{3}\frac{b_{2j}}{\tau^{2j}}\Phi^{(2j)}(s) + O(m^{-3/2}),$$

where a_1 and a_3 are given in Corollary 8.4 and

$$b_2 = 2\lambda_\alpha^2\sum_{\beta\neq\alpha}\lambda_\beta/(\lambda_\alpha - \lambda_\beta) - 2\lambda_\alpha^3\sum_{\beta\neq\alpha}\lambda_\beta/(\lambda_\alpha - \lambda_\beta)^2$$

$$+\frac{3}{2}\lambda_\alpha^2 \sum_{\beta\neq\alpha} \lambda_\beta^2/(\lambda_\alpha-\lambda_\beta)^2 + \lambda_\alpha^2 \sum_{\substack{\gamma<\beta \\ \gamma,\beta\neq\alpha}} \lambda_\gamma\lambda_\beta/(\lambda_\alpha-\lambda_\gamma)(\lambda_\alpha-\lambda_\beta),$$

$$b_4 = 2\lambda_\alpha^4 + \frac{4}{3}\lambda_\alpha^4 \sum_{\beta\neq\alpha} \lambda_\beta/(\lambda_\alpha-\lambda_\beta),$$

$$b_6 = \frac{8}{9}\lambda_\alpha^6,$$

provided by Sugiura (1973). The $O(m^{-3/2})$ asymptotic expansion of the joint distribution function of $(m/2)^{1/2}\mathbf{\Lambda}^{-1}(\mathbf{l}/m-\boldsymbol{\lambda})$ can be found in Sugiura (1976).

This expansion was independently obtained by Muirhead and Chikuse (1975) by a different technique based on G.A. Anderson (1965). Waternaux (1976) gave the asymptotic distributions of the sample eigenvalues for a non-normal population. Kollo and Neudecker (1993) presented the result of Waternaux (1976) in matrix form with other results on the the eigenstructure of the sample correlation matrix. The case of multiple population eigenvalues is considerably more complicated; the interested reader is referred to Fujikoshi (1977, 1978). However, in the case of multiple eigenvalues, the leading term of the asymptotic distribution is easier to obtain using a method of Eaton and Tyler (1991). This is done in Section 8.8.3. The asymptotic distribution in Corollary 8.5 for sampling from an elliptical distribution is given in Problem 13.6.18. Asymptotic distributions of eigenvalues of the sample correlation matrix are treated in Section 10.5 on principal components. Seminal papers on asymptotic expansions for the distribution of eigenvalues are those of T.W. Anderson (1963, 1965).

8.8.2 The two-sample problem

The asymptotic distribution of the eigenvalues $l_1,\ldots,l_p$ of $\mathbf{S}_1^{-1}\mathbf{S}_2$, where

$$m_1\mathbf{S}_1 \sim W_p(m_1,\mathbf{\Sigma}_1),\ m_2\mathbf{S}_2 \sim W_p(m_2,\mathbf{\Sigma}_2),\ \mathbf{S}_1 \perp\!\!\!\perp \mathbf{S}_2,$$

is now derived. Let $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_p)'$ be the eigenvalues of $\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2$. There exists $\mathbf{A}\in\mathbf{G}_p$ such that

$$\begin{aligned}\mathbf{A}\mathbf{\Sigma}_1\mathbf{A}' &= \mathbf{I},\\ \mathbf{A}\mathbf{\Sigma}_2\mathbf{A}' &= \operatorname{diag}(\lambda_1,\ldots,\lambda_p),\end{aligned}$$

and since the eigenvalues of $\mathbf{S}_1^{-1}\mathbf{S}_2$ and $(\mathbf{A}\mathbf{S}_1\mathbf{A}')^{-1}\mathbf{A}\mathbf{S}_2\mathbf{A}'$ are identical, we can assume at the outset without any loss of generality that $\mathbf{\Sigma}_1=\mathbf{I}$ and $\mathbf{\Sigma}_2\equiv\mathbf{\Lambda}=\operatorname{diag}(\lambda_1,\ldots,\lambda_p)$. Let $m=m_1+m_2$, $m_i/m\to\rho_i$ $(i=1,2)$ and note that

$$\begin{aligned}\mathbf{S}_1 &= \mathbf{I}+m^{-1/2}\mathbf{W}_1,\\ \mathbf{S}_2 &= \mathbf{\Lambda}+m^{-1/2}\mathbf{W}_2,\end{aligned}$$

where $\mathbf{W}_1 = m^{1/2}(\mathbf{S}_1 - \mathbf{I})$ and $\mathbf{W}_2 = m^{1/2}(\mathbf{S}_2 - \mathbf{\Lambda})$ are $O_p(1)$. The perturbation method is now adapted to this situation. Using Problem 1.8.15, the inverse of $\mathbf{S}_1$ is expanded as

$$\mathbf{S}_1^{-1} = \mathbf{I} - m^{-1/2}\mathbf{W}_1 + O_p(m^{-1}).$$

Hence, the expansion

$$\mathbf{S}_1^{-1}\mathbf{S}_2 = \mathbf{\Lambda} + m^{-1/2}(\mathbf{W}_2 - \mathbf{\Lambda}\mathbf{W}_1) + O_p(m^{-1})$$

is obtained. From (8.11), the eigenvalue of $\mathbf{S}_1^{-1}\mathbf{S}_2$, l_α, is expanded as

$$l_\alpha = \lambda_\alpha + m^{-1/2}[w_{2,\alpha\alpha} - \lambda_\alpha w_{1,\alpha\alpha}], \tag{8.20}$$

where $\mathbf{W}_k = (w_{k,ij})$, $k = 1, 2$. From (8.20), $\mathbf{l} = (l_1, \ldots, l_p)'$ can be expanded as

$$m^{1/2}(\mathbf{l} - \boldsymbol{\lambda}) = (\mathbf{w}_2 - \mathbf{\Lambda}\mathbf{w}_1) + O_p(m^{-1/2}),$$

where $\mathbf{w}_k = (w_{k,11}, \ldots, w_{k,pp})'$, $k = 1, 2$. From the asymptotic results of Chapter 6,

$$\begin{aligned} \mathbf{w}_1 &\xrightarrow{d} N_p(\mathbf{0}, 2\mathbf{I}/\rho_1), \\ \mathbf{w}_2 &\xrightarrow{d} N_p(\mathbf{0}, 2\mathbf{\Lambda}^2/\rho_2), \end{aligned}$$

and, hence,

$$\mathbf{w}_2 - \mathbf{\Lambda}\mathbf{w}_1 \xrightarrow{d} N_p\left(\mathbf{0}, 2\left(\frac{1}{\rho_1} + \frac{1}{\rho_2}\right)\mathbf{\Lambda}^2\right).$$

Finally, with the suitable norming constant, we find

$$(m\rho_1\rho_2/2)^{1/2}\mathbf{\Lambda}^{-1}(\mathbf{l} - \boldsymbol{\lambda}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}).$$

We have proved:

Proposition 8.19 *Let $\mathbf{l} = (l_1, \ldots, l_p)'$ be the eigenvalues of $\mathbf{S}_1^{-1}\mathbf{S}_2$, where $\mathbf{S}_1 = \mathbf{V}_1/m_1$, $\mathbf{S}_2 = \mathbf{V}_2/m_2$, $m = m_1 + m_2$, $m_i/m \to \rho_i$ $(i = 1, 2)$, $\mathbf{V}_1 \sim W_p(m_1)$, $\mathbf{V}_2 \sim W_p(m_2, \mathbf{\Lambda})$, and $\mathbf{V}_1 \perp\!\!\!\perp \mathbf{V}_2$. If the population eigenvalues λ_α are all distinct, then the joint limiting distribution can be expressed as*

$$(m\rho_1\rho_2/2)^{1/2}\mathbf{\Lambda}^{-1}(\mathbf{l} - \boldsymbol{\lambda}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}).$$

An expansion of the joint distribution function of $(m\rho_1\rho_2/2)^{1/2}\mathbf{\Lambda}^{-1}(\mathbf{l} - \boldsymbol{\lambda})$, with remainder of the order $O(m^{-3/2})$, can be found in Sugiura (1976).

8.8.3 The case of multiple eigenvalues

The asymptotic distribution, with remainder of the order $O_p(n^{-1/2})$, of the eigenvalues of random symmetric matrices is derived using the Wielandt's

inequality method introduced in Eaton and Tyler (1991). Consider a symmetric matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D} \end{pmatrix},$$

where $\mathbf{A} \in \mathbb{R}_p^p$, $\mathbf{B} \in \mathbb{R}_q^q$, $\mathbf{D} \in \mathbb{R}_r^r$, and $p = q + r$. Let $\rho^2(\mathbf{C})$ denote the largest eigenvalue of $\mathbf{C}\mathbf{C}'$ and let

$$\begin{aligned} &\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p, \\ &\beta_1 \geq \beta_2 \geq \cdots \geq \beta_q, \\ &\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r, \end{aligned}$$

be the ordered eigenvalues of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$, respectively.

Proposition 8.20 (Wielandt (1967)) *If $\beta_q > \delta_1$, then*

$$\begin{aligned} &0 \leq \alpha_j - \beta_j \leq \rho^2(\mathbf{C})/(\beta_j - \delta_1), \ j = 1, \ldots, q, \\ &0 \leq \delta_{r-i} - \alpha_{p-i} \leq \rho^2(\mathbf{C})/(\beta_q - \delta_{r-i}), \ i = 0, \ldots, r-1. \end{aligned}$$

A proof of Wielandt's inequality can be found in Eaton and Tyler (1991), who used it to find the asymptotic distribution of the eigenvalues of symmetric random matrices in the case of multiple eigenvalues. The matrix $\mathbf{A}$ can be viewed as a perturbation of a block-diagonal matrix, namely $\mathbf{A} = \mathbf{A}_0 + \mathbf{E}$, where

$$\mathbf{A}_0 = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \text{ and } \mathbf{E} = \begin{pmatrix} \mathbf{0} & \mathbf{C} \\ \mathbf{C}' & \mathbf{0} \end{pmatrix}.$$

By Wielandt's inequality, the eigenvalues of $\mathbf{A}_0$ are perturbed quadratically in $\mathbf{E}$ when $\mathbf{A}_0$ is perturbed linearly in $\mathbf{E}$. Generally, eigenvalues are only perturbed linearly when the matrix is perturbed linearly. The quadratic perturbation of eigenvalues in Wielandt's inequality is due to the special structure of $\mathbf{E}$ relative to $\mathbf{A}_0$.

Let $\mathcal{S}_p$ be the set of $p \times p$ real symmetric matrices. Consider a sequence of random matrices $\mathbf{S}_n \in \mathcal{S}_p$, and assume that

$$\mathbf{W}_n = n^{1/2}(\mathbf{S}_n - \boldsymbol{\Sigma}) \xrightarrow{d} \mathbf{W}, \tag{8.21}$$

for some $\boldsymbol{\Sigma} \in \mathcal{S}_p$, and hence $\mathbf{W} \in \mathcal{S}_p$. Given $\boldsymbol{\Sigma} \in \mathcal{S}_p$, let

$$\boldsymbol{\phi}(\boldsymbol{\Sigma}) = (\phi_1(\boldsymbol{\Sigma}), \ldots, \phi_p(\boldsymbol{\Sigma}))'$$

be the vector of ordered eigenvalues

$$\phi_1(\boldsymbol{\Sigma}) \geq \phi_2(\boldsymbol{\Sigma}) \geq \cdots \geq \phi_p(\boldsymbol{\Sigma}).$$

The asymptotic distribution of

$$n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}_n) - \boldsymbol{\phi}(\boldsymbol{\Sigma})\right)$$

is studied.

We consider in the first place the case $\mathbf{\Sigma} = \text{diag}(d_1\mathbf{I}_{p_1}, \ldots, d_k\mathbf{I}_{p_k})$, $p = p_1 + \cdots + p_k$, where $d_1 > d_2 > \cdots > d_k$ represent the distinct eigenvalues of $\mathbf{\Sigma}$ with the multiplicity of d_i being p_i, $i = 1, \ldots, k$. To reflect the block structure of $\mathbf{\Sigma}$, consider the partitioned matrix

$$\mathbf{S}_n = \begin{pmatrix} \mathbf{S}_{n,11} & \mathbf{S}_{n,12} & \cdots & \mathbf{S}_{n,1k} \\ \mathbf{S}_{n,21} & \mathbf{S}_{n,22} & \cdots & \mathbf{S}_{n,2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{n,k1} & \mathbf{S}_{n,k2} & \cdots & \mathbf{S}_{n,kk} \end{pmatrix},$$

where $\mathbf{S}_{n,ij} \in \mathbb{R}^{p_i}_{p_j}$.

Lemma 8.4 *For $k = 2$,*

$$\boldsymbol{\phi}(\mathbf{S}_n) - \begin{pmatrix} \boldsymbol{\phi}(\mathbf{S}_{n,11}) \\ \boldsymbol{\phi}(\mathbf{S}_{n,22}) \end{pmatrix} \text{ is } O_p(n^{-1}).$$

Proof. Let $\mathcal{A}_n = \{\mathbf{S}_n \mid \phi_{p_1}(\mathbf{S}_{n,11}) > \phi_1(\mathbf{S}_{n,22})\}$. Since $\boldsymbol{\phi}$ is continuous and, from (8.21), $\mathbf{S}_{n,11} \xrightarrow{p} d_1\mathbf{I}_{p_1}$ and $\mathbf{S}_{n,22} \xrightarrow{p} d_2\mathbf{I}_{p_2}$, it follows that $\phi_{p_1}(\mathbf{S}_{n,11}) \xrightarrow{p} d_1$ and $\phi_1(\mathbf{S}_{n,22}) \xrightarrow{p} d_2$. Thus, $P(\mathcal{A}_n) \to 1$, so attention can be restricted to $\mathcal{A}_n$, $n = 1, 2, \ldots$. For $\mathbf{S}_n \in \mathcal{A}_n$, Wielandt's inequality implies for $1 \leq i \leq p_1$,

$$0 \leq \phi_i(\mathbf{S}_n) - \phi_i(\mathbf{S}_{n,11}) \leq \rho^2(\mathbf{S}_{n,12}) / \left(\phi_i(\mathbf{S}_{n,11}) - \phi_1(\mathbf{S}_{n,22})\right).$$

By (8.21), $\mathbf{S}_{n,12}$ is $O_p(n^{-1/2})$, and since ρ is continuous, it follows that $\rho^2(\mathbf{S}_{n,12})$ is $O_p(n^{-1})$. Since

$$\phi_i(\mathbf{S}_{n,11}) - \phi_1(\mathbf{S}_{n,11}) \xrightarrow{p} d_1 - d_2 > 0,$$

then $\phi_i(\mathbf{S}_n) - \phi_i(\mathbf{S}_{n,11})$ is $O_p(n^{-1})$, $i = 1, \ldots, p_1$. The proof of $\phi_{p-j}(\mathbf{S}_n) - \phi_{p_2-j}(\mathbf{S}_{n,22})$ is $O_p(n^{-1})$, $j = 0, \ldots, p_2 - 1$, is analogous. □

By applying Lemma 8.4, $k - 1$ times, the following asymptotic equivalence result is obtained. The vector $\mathbf{1}_{p_i} \in \mathbb{R}^{p_i}$ is the vector of ones.

Proposition 8.21

$$n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}_n) - \boldsymbol{\phi}(\mathbf{\Sigma})\right) = \mathbf{Z}_n + \mathbf{R}_n,$$

where

$$\mathbf{Z}_n = n^{1/2} \begin{pmatrix} \boldsymbol{\phi}(\mathbf{S}_{n,11}) - d_1\mathbf{1}_{p_1} \\ \vdots \\ \boldsymbol{\phi}(\mathbf{S}_{n,kk}) - d_k\mathbf{1}_{p_k} \end{pmatrix}$$

and the remainder term $\mathbf{R}_n$ is $O_p(n^{-1/2})$.

Since $\mathbf{R}_n \xrightarrow{p} \mathbf{0}$, using Slutsky's theorem the asymptotic distribution of

$$n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}_n) - \boldsymbol{\phi}(\mathbf{\Sigma})\right)$$

is that of the leading term $\mathbf{Z}_n$. Considering the partitioned $\mathbf{W} = (\mathbf{W}_{ij})$, $\mathbf{W}_{ij} \in \mathbb{R}^{p_i}_{p_j}$, $i, j = 1, \ldots, k$, we have immediately from (8.21) that

$$n^{1/2} \begin{pmatrix} \mathbf{S}_{n,11} - d_1 \mathbf{I}_{p_1} \\ \vdots \\ \mathbf{S}_{n,kk} - d_k \mathbf{I}_{p_k} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathbf{W}_{11} \\ \vdots \\ \mathbf{W}_{kk} \end{pmatrix}.$$

Now, because the function

$$\mathbf{G}(\mathbf{W}) = \begin{pmatrix} \boldsymbol{\phi}(\mathbf{W}_{11}) \\ \vdots \\ \boldsymbol{\phi}(\mathbf{W}_{kk}) \end{pmatrix}$$

is continuous and since

$$\boldsymbol{\phi}\left(n^{1/2}(\mathbf{S}_{n,11} - d_1 \mathbf{I}_{p_1})\right) = n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}_{n,11}) - d_1 \mathbf{1}_{p_1}\right),$$

it follows that

$$\mathbf{Z}_n = \mathbf{G}(\mathbf{W}_n) \xrightarrow{d} \mathbf{G}(\mathbf{W}).$$

We have proved:

Proposition 8.22 *If* $n^{1/2}(\mathbf{S}_n - \boldsymbol{\Sigma}) \xrightarrow{d} \mathbf{W}$ *and* $\boldsymbol{\Sigma}$ *is diagonal, then*

$$n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}_n) - \boldsymbol{\phi}(\boldsymbol{\Sigma})\right) \xrightarrow{d} \mathbf{G}(\mathbf{W}).$$

In the general case where $\boldsymbol{\Sigma}$ is not diagonal, there exists $\mathbf{H} \in \mathbf{O}_p$ such that

$$\boldsymbol{\Sigma} = \mathbf{H}\ \mathrm{diag}\left(d_1 \mathbf{I}_{p_1}, \ldots, d_k \mathbf{I}_{p_k}\right) \mathbf{H}' \equiv \mathbf{H}\mathbf{D}\mathbf{H}'.$$

From (8.21), $n^{1/2}(\mathbf{H}\mathbf{S}_n\mathbf{H}' - \mathbf{D}) \xrightarrow{d} \mathbf{H}\mathbf{W}\mathbf{H}'$. Since $\boldsymbol{\phi}(\mathbf{H}\mathbf{S}_n\mathbf{H}') = \boldsymbol{\phi}(\mathbf{S}_n)$ and $\boldsymbol{\phi}(\boldsymbol{\Sigma}) = \boldsymbol{\phi}(\mathbf{D})$, we obtain the general result

$$n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}_n) - \boldsymbol{\phi}(\boldsymbol{\Sigma})\right) \xrightarrow{d} \mathbf{G}(\mathbf{H}\mathbf{W}\mathbf{H}').$$

This general result is summarized.

Proposition 8.23 *If* $n^{1/2}(\mathbf{S}_n - \boldsymbol{\Sigma}) \xrightarrow{d} \mathbf{W}$, *then*

$$n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}_n) - \boldsymbol{\phi}(\boldsymbol{\Sigma})\right) \xrightarrow{d} \mathbf{G}(\mathbf{H}\mathbf{W}\mathbf{H}'),$$

where $\mathbf{H}$ *diagonalizes* $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = \mathbf{H}\ \mathrm{diag}\left(d_1 \mathbf{I}_{p_1}, \ldots, d_k \mathbf{I}_{p_k}\right) \mathbf{H}'$.

An important special case is when $\mathbf{W}$ in (8.21) is a multivariate normal matrix and all eigenvalues of $\boldsymbol{\Sigma}$ are distinct. In that case, $\mathbf{H}\mathbf{W}\mathbf{H}'$ is also a multivariate normal matrix. Also, since all eigenvalues of $\boldsymbol{\Sigma}$ have multiplicity 1, then $\mathbf{G}(\mathbf{H}\mathbf{W}\mathbf{H}')$ is just a p-dimensional marginal of $\mathbf{H}\mathbf{W}\mathbf{H}'$ and, hence, has a p-dimensional normal distribution.

Example 8.3 *It was seen in Chapter* 6 *that when sampling from a* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *distribution, the asymptotic distribution of the sample variance* $\mathbf{S}$ *is* $n^{1/2}(\mathbf{S} - \boldsymbol{\Sigma}) \xrightarrow{d} \mathbf{W}$, *where*

$$\mathbf{W} \sim N_p^p(\mathbf{0}, (\mathbf{I} + \mathbf{K})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})).$$

We derive the asymptotic distribution of the eigenvalues of $\mathbf{S}$ *when all eigenvalues of* $\boldsymbol{\Sigma}$ *are distinct. Using Proposition* 8.23, *we have* $n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}) - \boldsymbol{\phi}(\boldsymbol{\Sigma})\right) \xrightarrow{d} \mathbf{G}(\mathbf{HWH}')$, *where* $\mathbf{H}$ *diagonalizes* $\boldsymbol{\Sigma}$, $\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' \equiv \mathbf{D}$ *(say). But from Proposition* 6.1, $\mathbf{HWH}' \sim N_p^p(\mathbf{0}, (\mathbf{I} + \mathbf{K})(\mathbf{D} \otimes \mathbf{D}))$. *From* (6.1), $(\mathbf{HWH}')_{ii} \sim N(0, 2d_i^2)$, $i = 1, \ldots, p$, *and are independently distributed. Since all eigenvalues of* $\boldsymbol{\Sigma}$ *are distinct, then*

$$n^{1/2}\left(\boldsymbol{\phi}(\mathbf{S}) - \boldsymbol{\phi}(\boldsymbol{\Sigma})\right) \xrightarrow{d} N_p(\mathbf{0}, 2\mathbf{D}^2),$$

which is the result proven previously in Corollary 8.5.

Example 8.4 *We now derive the asymptotic distribution of the* r *smallest eigenvalues of* $\mathbf{S}$ *when the smallest eigenvalue of* $\boldsymbol{\Sigma}$ *has multiplicity* r,

$$\boldsymbol{\phi}(\boldsymbol{\Sigma}) = (\phi_1(\boldsymbol{\Sigma}), \ldots, \phi_{p-r}(\boldsymbol{\Sigma}), \lambda, \ldots, \lambda)'.$$

As in Example 8.3, $\mathbf{HWH}' \sim N_p^p(\mathbf{0}, (\mathbf{I} + \mathbf{K}_p)(\mathbf{D} \otimes \mathbf{D}))$. *Hence, the lower right* $r \times r$ *block of* $\mathbf{HWH}'$ *is distributed as* $N_r^r(\mathbf{0}, \lambda^2(\mathbf{I} + \mathbf{K}_r))$. *Using Proposition* 8.23, *we have finally*

$$n^{1/2}\lambda^{-1}\left(\phi_{p-r+1}(\mathbf{S}) - \lambda, \ldots, \phi_p(\mathbf{S}) - \lambda\right) \xrightarrow{d} \mathbf{w},$$

where $\mathbf{w}$ *is distributed as the eigenvalues of a* $N_r^r(\mathbf{0}, (\mathbf{I} + \mathbf{K}_r))$ *distribution.*

An application of Wielandt's inequality to bootstrapping eigenvalues can be found in Eaton and Tyler (1991), who extend the work of Beran and Srivastava (1985, 1987). Earlier papers on the case of multiple eigenvalues include those of James (1969), Chattopadhyay and Pillai (1973), Chikuse (1976), Khatri and Srivastava (1978), and Srivastava and Carter (1980).

8.9 Problems

1. **Two-sample** T^2**.**
   Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y}_1, \ldots, \mathbf{y}_m$ i.i.d. $N_p(\boldsymbol{\tau}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$, be two independent samples. Define the sample variances

$$\begin{aligned} \mathbf{S}_{\mathbf{x}} &= \frac{1}{(n-1)} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \\ \mathbf{S}_{\mathbf{y}} &= \frac{1}{(m-1)} \sum_{i=1}^{m} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})', \end{aligned}$$

and

$$\mathbf{S}_{pool} = \frac{1}{(n+m-2)}[(n-1)\mathbf{S}_{\mathbf{x}} + (m-1)\mathbf{S}_{\mathbf{y}}].$$

Determine

(i) the distribution of $\mathbf{S}_{pool}$,
(ii) the distribution of

$$T^2 = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})' \mathbf{S}_{pool}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$$

used for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\tau}$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\tau}$.

2. **Invariance of two-sample T^2.**
This is a continuation of Problem 8.9.1.

(i) Prove $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \mathbf{S}_{pool})$ is minimal sufficient for $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\Sigma})$.
(ii) Consider $(\mathbf{A}, \mathbf{b})$ in the group of transformations $\mathbf{G}_p \times \mathbb{R}^p$ acting as $\bar{\mathbf{x}} \mapsto \mathbf{A}\bar{\mathbf{x}} + \mathbf{b}$, $\bar{\mathbf{y}} \mapsto \mathbf{A}\bar{\mathbf{y}} + \mathbf{b}$, and $\mathbf{S}_{pool} \mapsto \mathbf{A}\mathbf{S}_{pool}\mathbf{A}'$. Prove that this group of transformations leaves the testing problem invariant and that any invariant test depends on $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \mathbf{S}_{pool})$ only through T^2.
(iii) Prove that any invariant test has a power function depending on $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\Sigma})$ only through $(\boldsymbol{\mu} - \boldsymbol{\tau})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\tau})$.

3. **Common mean vector.**
For independent samples

$$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}, \text{ i.i.d. } N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i), \ i = 1, \ldots, a,$$

from distributions with a common mean vector $\boldsymbol{\mu}$, let $(\mathbf{S}_i, \bar{\mathbf{x}}_i)$ be the MVUE from each sample. Consider estimating the common mean $\boldsymbol{\mu}$ with respect to the weighted least-squares criterion

$$\min_{\boldsymbol{\mu}} \sum_{i=1}^{a} c_i n_i (\bar{\mathbf{x}}_i - \boldsymbol{\mu})' \mathbf{S}_i^{-1} (\bar{\mathbf{x}}_i - \boldsymbol{\mu})$$

for some constants $c_i > 0$, $\sum_{i=1}^{a} c_i = 1$. Establish the estimate of $\boldsymbol{\mu}$ is given by

$$\tilde{\boldsymbol{\mu}} = \left(\sum_{i=1}^{a} c_i n_i \mathbf{S}_i^{-1}\right)^{-1} \left(\sum_{i=1}^{a} c_i n_i \mathbf{S}_i^{-1} \bar{\mathbf{x}}_i\right).$$

Remark: Jordan and Krishnamoorthy (1995) built an exact $\beta \times 100\%$ confidence region centered at $\tilde{\boldsymbol{\mu}}$.

4. **Test of symmetry.**
Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$. Choose $\mathbf{C} \in \mathbb{R}_p^{p-1}$ of rank $\mathbf{C} = p - 1$ such that $\mathbf{C1} = \mathbf{0}$. Prove the following:

(i) ker $\mathbf{C} = \text{span}\{\mathbf{1}\}$ and, therefore,

$$H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0} \Longleftrightarrow H_0 : \mu_1 = \cdots = \mu_p.$$

(ii) Any $p-1$ columns of $\mathbf{C}$ are linearly independent, which implies that $\mathbf{C} = \mathbf{A}(\mathbf{I}_{p-1}, -\mathbf{1})$, for some nonsingular $\mathbf{A} \in \mathbb{R}^{p-1}_{p-1}$. Conclude that, thereafter, the value of

$$T^2 = n(\mathbf{C}\bar{\mathbf{x}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{x}}),$$

where $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$ and $(n-1)\mathbf{S} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, does not depend on the choice of $\mathbf{C}$.

(iii) The null (under H_0) distribution of T^2 is

$$T^2/(n-1) \sim F_c(p-1, n-p+1).$$

5. Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x} \in \mathbb{R}^p$ (possibly non-normal) with $E\,\mathbf{x} = \boldsymbol{\mu}$ and var $\mathbf{x} = \boldsymbol{\Sigma}$. Establish that, asymptotically, we are at least $(1-\alpha)\times 100\%$ confident in simultaneously presenting all of the observed "Scheffé" intervals:

$$\mathbf{a}'\bar{\mathbf{x}} - \left(\frac{\chi^2_{\alpha,p}}{n}\right)^{1/2} (\mathbf{a}'\mathbf{Sa})^{1/2} \le \mathbf{a}'\boldsymbol{\mu} \le \mathbf{a}'\bar{\mathbf{x}} + \left(\frac{\chi^2_{\alpha,p}}{n}\right)^{1/2} (\mathbf{a}'\mathbf{Sa})^{1/2},$$

$\forall \mathbf{a} \in \mathbb{R}^p$.

6. Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x} \in \mathbb{R}^p$ (possibly non-normal) with $E\,\mathbf{x} = \boldsymbol{\mu}$ and var $\mathbf{x} = \boldsymbol{\Sigma}$. Let $\mathbf{a}_1, \ldots, \mathbf{a}_r$ be linearly independent in $\mathbb{R}^p$. Establish that, asymptotically, we are at least $(1-\alpha)\times 100\%$ confident in simultaneously presenting the observed "Roy-Bose" intervals:

$$\mathbf{a}_i'\bar{\mathbf{x}} - \left(\frac{\chi^2_{\alpha,r}}{n}\right)^{1/2} (\mathbf{a}_i'\mathbf{Sa}_i)^{1/2} \le \mathbf{a}_i'\boldsymbol{\mu} \le \mathbf{a}_i'\bar{\mathbf{x}} + \left(\frac{\chi^2_{\alpha,r}}{n}\right)^{1/2} (\mathbf{a}_i'\mathbf{Sa}_i)^{1/2},$$

$i = 1, \ldots, r$.

7. **Test of proportionality.**
Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$ a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$, obtain the likelihood ratio test Λ for

$$H_0 : \boldsymbol{\Sigma} = \gamma\boldsymbol{\Sigma}_0,\ \gamma > 0 \quad \text{versus} \quad H_1 : \boldsymbol{\Sigma} > \mathbf{0},$$

where $\boldsymbol{\Sigma}_0 > \mathbf{0}$ is a known matrix:

$$\Lambda^{2/n} = |\boldsymbol{\Sigma}_0^{-1}\mathbf{V}| / \left(\frac{1}{p}\ \text{tr}\ \boldsymbol{\Sigma}_0^{-1}\mathbf{V}\right)^p,$$

with $\mathbf{V} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ as usual.

8. **Test for a given variance.**
Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$.

(i) Prove that the LRT for $H_0 : \boldsymbol{\Sigma} = \mathbf{I}$ versus $H_1 : \boldsymbol{\Sigma} \neq \mathbf{I}$ is given by

$$\Lambda = (e/n)^{pn/2} |\mathbf{V}|^{n/2} \operatorname{etr}(-\tfrac{1}{2}\mathbf{V}),$$

where $\mathbf{V} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$.
Remark: This test is biased, but Sugiura and Nagao (1968) have shown the slight modification

$$\Lambda^* = (e/m)^{pm/2} |\mathbf{V}|^{m/2} \operatorname{etr}(-\tfrac{1}{2}\mathbf{V}),$$

where $m = n - 1$, gives an unbiased test.

(ii) Using the Wishart density in Problem 7.5.7, prove that

$$E\ \Lambda^{*h} = \left(\frac{2e}{m}\right)^{mph/2} \frac{|\boldsymbol{\Sigma}|^{mh/2}}{|\mathbf{I} + h\boldsymbol{\Sigma}|^{m(1+h)/2}} \frac{\Gamma_p[\frac{1}{2}m(1+h)]}{\Gamma_p(\frac{1}{2}m)}.$$

Hint: The integrand has the form of a Wishart density. Simply find the normalizing constant.

(iii) Under H_0,

$$E\ \Lambda^{*h} = \left(\frac{2e}{m}\right)^{mph/2} (1+h)^{-mp(1+h)/2} \frac{\Gamma_p[\frac{1}{2}m(1+h)]}{\Gamma_p(\frac{1}{2}m)}.$$

Remark: An accurate approximation to the null distribution of Λ^* using those moments is given in Example 12.5.

9. Use Proposition 8.6 to obtain the moments of $\hat{R}^2$:

$$E\ \hat{R}^{2h} = \sum_{k=0}^{\infty} p_k \frac{\Gamma\left(\frac{1}{2}(p-1+2k)+h\right)}{\Gamma\left(\frac{1}{2}(p-1+2k)\right)} \frac{\Gamma\left(\frac{1}{2}(n-1+2k)\right)}{\Gamma\left(\frac{1}{2}(n-1+2k)+h\right)},$$

where p_k are the negative binomial probabilities given in Proposition 8.6.

10. Demonstrate that if K given δ is Poisson(δ) and $\delta \sim G(p, \theta)$, then the marginal of K is the negative binomial $K \sim \text{nb}(p, (1+\theta)^{-1})$.

11. Write Nagao's test for the equality of two variances, $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, as a function of the eigenvalues $l_1, \ldots, l_p$ of $\mathbf{V}_1^{-1}\mathbf{V}_2$, where as usual $\mathbf{V}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$, $i = 1, 2$.

12. Write the LBI test for sphericity as a function of $l_1/l_p, \ldots, l_{p-1}/l_p$, where $l_1 \geq \cdots \geq l_p$ are the ordered eigenvalues of $\mathbf{V} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$.

13. **Invariance of multiple correlation.**
Let $\mathbf{x}_i = (x_{i1}, \mathbf{x}_{i2}')'$, $i = 1, \ldots, n$, be i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$. Consider the group of transformations

$$\mathbf{x}_i \mapsto \begin{pmatrix} ax_{i1} + b \\ \mathbf{A}\mathbf{x}_{i2} + \mathbf{b} \end{pmatrix},$$

for any $a \neq 0$, $b \in \mathbb{R}$, $\mathbf{A} \in \mathbf{G}_{p-1}$, and $\mathbf{b} \in \mathbb{R}^{p-1}$.

(i) Show that this transformation induces the following transformations on the sufficient statistics and parameters:

$$\begin{aligned}
\bar{\mathbf{x}} &\mapsto \begin{pmatrix} a\bar{x}_1 + b \\ \mathbf{A}\bar{\mathbf{x}}_2 + \mathbf{b} \end{pmatrix}, \\
\mathbf{V} &\mapsto \begin{pmatrix} a^2 v_{11} & a\mathbf{v}'_{21}\mathbf{A}' \\ a\mathbf{A}\mathbf{v}_{21} & \mathbf{A}\mathbf{V}_{22}\mathbf{A}' \end{pmatrix}, \\
\boldsymbol{\mu} &\mapsto \begin{pmatrix} a\mu_1 + b \\ \mathbf{A}\boldsymbol{\mu}_2 + \mathbf{b} \end{pmatrix}, \\
\boldsymbol{\Sigma} &\mapsto \begin{pmatrix} a^2\sigma_{11} & a\boldsymbol{\sigma}'_{21}\mathbf{A}' \\ a\mathbf{A}\boldsymbol{\sigma}_{21} & \mathbf{A}\boldsymbol{\Sigma}_{22}\mathbf{A}' \end{pmatrix}.
\end{aligned}$$

(ii) Choose $a = v_{11}^{-1/2}$, $b = -a\bar{x}_1$, $\mathbf{A} = \mathbf{V}_{22}^{-1/2}$, and $\mathbf{b} = -\mathbf{A}\bar{\mathbf{x}}_2$ to prove that any invariant test $f(\bar{\mathbf{x}}, \mathbf{V})$ depends on the data only through $\mathbf{u} = \mathbf{V}_{22}^{-1/2}\mathbf{v}_{21}/v_{11}^{1/2}$.

(iii) Prove that there exists an orthogonal transformation $\mathbf{H} \in \mathbf{O}_{p-1}$ such that $\mathbf{H}\mathbf{u} = \hat{R}\mathbf{e}_1$ (v. Problem 1.8.14).

(iv) Choosing further $a = 1$, $b = 0$, $\mathbf{A} = \mathbf{H}$, and $\mathbf{b} = \mathbf{0}$, prove that any invariant test is necessarily a function of $\hat{R}$.

(v) Prove that the non-null distribution of any invariant test depends on the parameters only through R.

14. **Test of equality of means and variances.**
The data consists of a independent samples $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$ i.i.d. $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $\boldsymbol{\Sigma}_i > \mathbf{0}$, $i = 1, \ldots, a$. Let $n = \sum_{i=1}^a n_i$ be the total number of observations. The hypothesis is the equality of the distributions

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_a;\ \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a$$

which is being tested against all alternatives. Let

$$\begin{aligned}
\bar{\mathbf{x}}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}, \\
\bar{\mathbf{x}} &= \frac{1}{n}\sum_{i=1}^{a} n_i \bar{\mathbf{x}}_i,
\end{aligned}$$

be the ith sample mean and overall mean, respectively, and

$$\begin{aligned}
\mathbf{V}_i &= \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)',\ i = 1, \ldots, a, \\
\mathbf{B} &= \sum_{i=1}^{a} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'
\end{aligned}$$

be the usual "within" and "between" sums of squares, respectively.

(i) Prove that the LRT is

$$\Lambda = \frac{\prod_{i=1}^{a} |\mathbf{V}_i|^{n_i/2}}{|\sum_{i=1}^{a} \mathbf{V}_i + \mathbf{B}|^{n/2}} \frac{n^{np/2}}{\prod_{i=1}^{a} n_i^{n_i p/2}}.$$

Remark: Perlman (1980) proved this LRT yields an unbiased test.

(ii) Use the group of transformations $\mathbf{x}_{ij} \mapsto \mathbf{A}\mathbf{x}_{ij} + \mathbf{a}$, for any $\mathbf{A} \in \mathbf{G}_p$, $\mathbf{a} \in \mathbb{R}^p$, to argue that the null distribution of Λ (or any other invariant test) can be obtained by setting $\boldsymbol{\mu}_i = \mathbf{0}$ and $\boldsymbol{\Sigma}_i = \mathbf{I}$ without loss of generality.

(iii) Establish that $\mathbf{V}_1, \ldots, \mathbf{V}_a, \mathbf{B}$ are mutually independent whenever $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a$ holds. Moreover, verify that, under H_0, $\mathbf{V}_i \sim W_p(n_i - 1)$ and $\mathbf{B} \sim W_p(a - 1)$.
Hint: Let

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{in_i} \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_a \end{pmatrix}$$

be the sample matrices. For appropriately chosen orthogonal projections $\mathbf{Q}_i$, $i = 1, \ldots, a$, and $\mathbf{Q}$, so that $\mathbf{V}_i = \mathbf{X}'_i \mathbf{Q}_i \mathbf{X}_i$ and $\mathbf{B} = \mathbf{X}'\mathbf{Q}\mathbf{X}$, write

$$\begin{pmatrix} \mathbf{Q}_1\mathbf{X}_1 \\ \vdots \\ \mathbf{Q}_a\mathbf{X}_a \\ \mathbf{Q}\mathbf{X} \end{pmatrix} = \left(\begin{array}{cccc} \mathbf{Q}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_a \\ \hline \multicolumn{4}{c}{\mathbf{Q}} \end{array}\right) \mathbf{X}$$
$$\equiv \mathbf{C}\mathbf{X}.$$

Then use Proposition 6.1 and verify that $\mathbf{C}\mathbf{C}'$ is block-diagonal.

(iv) Obtain the null moments of Λ:

$$E\,\Lambda^h = \frac{n^{nph/2}}{\prod_{i=1}^{a} n_i^{n_i ph/2}} \frac{\Gamma_p[\frac{1}{2}(n-1)]}{\Gamma_p[\frac{1}{2}n(1+h) - \frac{1}{2}]} \prod_{i=1}^{a} \frac{\Gamma_p[\frac{1}{2}n_i(1+h) - \frac{1}{2}]}{\Gamma_p[\frac{1}{2}(n_i - 1)]}.$$

Hint: Recall the normalizing constant $c_{p,m}$ of a $W_p(m)$ p.d.f. and use a similar argument as in the proof of Proposition 8.17 to establish

$$E\,\Lambda^h = \frac{n^{nph/2}}{\prod_{i=1}^{a} n_i^{n_i ph/2}} \prod_{i=1}^{a} \frac{c_{p,n_i-1}}{c_{p,n_i(1+h)-1}} E\,|\mathbf{W}|^{-nh/2},$$

where $\mathbf{W} \sim W_p(n(1+h) - 1)$.
Remark: The null moments are invoked in Problem 12.4.1 to develop an accurate approximation to the null distribution of Λ.

15. Assume the population eigenvalue λ_α is distinct from all other $p-1$ eigenvalues. Establish that the logarithmic transformation is a variance stabilizing transformation for the sample eigenvalue l_α of $\mathbf{V} \sim W_p(m, \mathbf{\Lambda})$.

9
Multivariate regression

9.1 Introduction

Multivariate regression with p responses as opposed to p multiple regressions is getting increasingly more attention, especially in the context of prediction. In this chapter, we generalize the multiple regression model of Section 5.6.2 to the multivariate case. The estimation method of Section 9.2 relies also on orthogonal projections. The model considered is

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

where $\mathbf{Y} \in \mathbb{R}^n_p$, $\mathbf{B} \in \mathbb{R}^k_p$, and $\mathbf{X} \in \mathbb{R}^n_k$ of rank $\mathbf{X} = k$ is fixed. The error term $\mathbf{E}$ is such that $E\ \mathbf{E} = \mathbf{0}$ and var $\mathbf{E} = \mathbf{I}_n \otimes \mathbf{\Sigma}$ with $\mathbf{\Sigma} > \mathbf{0}$ in $\mathbb{R}^p_p$. The observation vectors consisting of the rows of $\mathbf{Y}$ are thus uncorrelated. The Gauss-Markov estimate is derived first. Then, assuming normality, the maximum likelihood estimates of $\mathbf{B}$ and $\mathbf{\Sigma}$ are obtained together with the fundamental result about their joint distribution. Section 9.3 derives the likelihood ratio test for the general linear hypothesis

$$H_0 : \mathbf{CB} = \mathbf{0}$$

against all alternatives where $\mathbf{C} \in \mathbb{R}^r_k$ of rank $\mathbf{C} = r$ in the above model. In the last sections, we discuss the practical and more commonly encountered situation of k random (observed) predictors and the problem of prediction of p responses from the same set of k predictors. Finally, an application to the MANOVA one-way classification model is treated as a special case.

The multivariate regression model can be seen as a set of p correlated multiple regression models of Section 5.6.2. With the partition

$$\begin{aligned} \mathbf{Y} &= (\mathbf{y}_1, \ldots, \mathbf{y}_p), \\ \mathbf{B} &= (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p), \\ \mathbf{E} &= (\mathbf{e}_1, \ldots, \mathbf{e}_p), \end{aligned} \tag{9.1}$$

we can rewrite $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ as

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{e}_i, \ i = 1, \ldots, p,$$

where $\mathbf{e}_i \sim N_n(\mathbf{0}, \sigma_{ii}\mathbf{I})$. However, the p multiple regression models are correlated since $\text{cov}(\mathbf{e}_i, \mathbf{e}_j) = \sigma_{ij}\mathbf{I}$. Testing a relationship between the various $\boldsymbol{\beta}_i$'s will require one to treat the p models as one multivariate regression model.

9.2 Estimation

First, observe that $\mathbb{R}_p^n$ is a linear space on which we define the usual inner product

$$\langle \mathbf{Y}, \mathbf{Z} \rangle = \text{tr}(\mathbf{Y}'\mathbf{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}z_{ij}, \text{ for any } \mathbf{Y}, \mathbf{Z} \in \mathbb{R}_p^n.$$

The mean of $\mathbf{Y}$ is in a subspace $\mathcal{V} = \{\mathbf{XA} : \ \mathbf{A} \in \mathbb{R}_p^k\}$. To define the orthogonal projection of $\mathbf{Y}$ on $\mathcal{V}$, a basis for $\mathcal{V}$ is needed. Partition $\mathbf{X}$ into columns

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k).$$

An element $\mathbf{XA} \in \mathcal{V}$, where

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

is partitioned into rows, is of the form

$$\begin{aligned} \mathbf{XA} &= \sum_{i=1}^{k} \mathbf{x}_i\mathbf{a}_i' = \sum_{i=1}^{k} \mathbf{x}_i \left(\sum_{j=1}^{p} a_{ij}\mathbf{e}_j' \right) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{p} a_{ij}(\mathbf{x}_i\mathbf{e}_j'); \end{aligned}$$

hence, $\{\mathbf{x}_i\mathbf{e}_j' \in \mathbb{R}_p^n : \ i = 1, \ldots, k; \ j = 1, \ldots, p\}$ spans $\mathcal{V}$. Moreover, linear independence holds since

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{p} a_{ij}(\mathbf{x}_i\mathbf{e}_j') = \mathbf{0} \quad &\Longrightarrow \quad \sum_{i=1}^{k} \mathbf{x}_i \left(\sum_{j=1}^{p} a_{ij}\mathbf{e}_j' \right) = \mathbf{0} \\
&\Longrightarrow \quad \sum_{j=1}^{p} a_{ij}\mathbf{e}_j' = \mathbf{0}, \ \forall i \\
&\Longrightarrow \quad a_{ij} = 0, \forall i, j.
\end{aligned}$$

Next, to calculate the orthogonal projection, say $\mathbf{X}\hat{\mathbf{B}}$, of $\mathbf{Y}$ on $\mathcal{V}$, note that since $\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} \in \mathcal{V}^\perp$, then $\langle \mathbf{x}_i\mathbf{e}_j', \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} \rangle = 0$, $\forall i, j$. But, this means that

$$\text{tr}\left[\mathbf{e}_j\mathbf{x}_i'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})\right] = \mathbf{x}_i'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})\mathbf{e}_j = 0, \ \forall i, j.$$

Therefore, $\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = \mathbf{0}$, which gives

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

With the partition in (9.1) we find $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p)$, where

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_i.$$

For the purpose of estimation of the regression coefficients, the p multiple regression models can be treated separately.

The orthogonal projection of $\mathbf{Y}$ on $\mathcal{V}$ becomes $\mathbf{X}\hat{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \equiv \mathbf{P}\mathbf{Y}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This provides the "orthogonal direct sum"

$$\mathbf{Y} = \mathbf{P}\mathbf{Y} + \mathbf{Q}\mathbf{Y}$$

with $\mathbf{Q} = \mathbf{I} - \mathbf{P}$ as usual.

The Gauss-Markov property is the subject of Proposition 9.1. Consider the parameter $\mathbf{C} = \mathbf{D}(\mathbf{X}\mathbf{B})\mathbf{F}$ and its natural estimate $\hat{\mathbf{C}} = \mathbf{D}(\mathbf{P}\mathbf{Y})\mathbf{F}$. Among all linear unbiased estimates, $\hat{\mathbf{C}}$ is the "best" (blue) in the sense that it has the minimum variance:

Proposition 9.1 (Gauss-Markov) $\hat{\mathbf{C}} = \text{blue}(\mathbf{C})$.

Proof. Let $\tilde{\mathbf{C}} = \mathbf{G}\mathbf{Y}\mathbf{H}$ be any linear unbiased estimate. Then,

$$\begin{aligned}
E\ \tilde{\mathbf{C}} = \mathbf{C}, \ \forall \mathbf{B} \in \mathbb{R}_p^k \quad &\Longleftrightarrow \quad \mathbf{G}\mathbf{X}\mathbf{B}\mathbf{H} = \mathbf{D}\mathbf{X}\mathbf{B}\mathbf{F}, \ \forall \mathbf{B} \in \mathbb{R}_p^k \\
&\Longleftrightarrow \quad \mathbf{G}\mathbf{P}\mathbf{Y}\mathbf{H} = \mathbf{D}\mathbf{P}\mathbf{Y}\mathbf{F}, \ \forall \mathbf{Y} \in \mathbb{R}_p^n \\
&\Longleftrightarrow \quad \mathbf{H}'\mathbf{Y}'\mathbf{P}\mathbf{G}' = \mathbf{F}'\mathbf{Y}'\mathbf{P}\mathbf{D}', \ \forall \mathbf{Y} \in \mathbb{R}_p^n \\
&\Longleftrightarrow \quad [(\mathbf{G}\mathbf{P}) \otimes \mathbf{H}']\text{vec}(\mathbf{Y}') = [(\mathbf{D}\mathbf{P}) \otimes \mathbf{F}']\text{vec}(\mathbf{Y}') \\
&\Longleftrightarrow \quad (\mathbf{G}\mathbf{P}) \otimes \mathbf{H}' = (\mathbf{D}\mathbf{P}) \otimes \mathbf{F}'.
\end{aligned}$$

Now, we have

$$\text{var}\ \hat{\mathbf{C}} \quad = \quad [(\mathbf{D}\mathbf{P}) \otimes \mathbf{F}'](\mathbf{I}_n \otimes \boldsymbol{\Sigma})[(\mathbf{D}\mathbf{P})' \otimes \mathbf{F}]$$

$$\begin{aligned} &= [(\mathbf{GP}) \otimes \mathbf{H}'](\mathbf{I}_n \otimes \mathbf{\Sigma})[(\mathbf{GP})' \otimes \mathbf{H}] \\ &= (\mathbf{GPG}') \otimes (\mathbf{H}'\mathbf{\Sigma}\mathbf{H}) \leq (\mathbf{GG}') \otimes (\mathbf{H}'\mathbf{\Sigma}\mathbf{H}) = \text{var } \tilde{\mathbf{C}}, \end{aligned}$$

with equality iff

$$\begin{aligned} \mathbf{GP} = \mathbf{G} &\iff [(\mathbf{GP}) \otimes \mathbf{H}']\text{vec}(\mathbf{Y}') = (\mathbf{G} \otimes \mathbf{H}')\text{vec}(\mathbf{Y}'),\ \forall \mathbf{Y} \in \mathbb{R}_p^n \\ &\iff [(\mathbf{DP}) \otimes \mathbf{F}']\text{vec}(\mathbf{Y}') = (\mathbf{G} \otimes \mathbf{H}')\text{vec}(\mathbf{Y}'),\ \forall \mathbf{Y} \in \mathbb{R}_p^n \\ &\iff \hat{\mathbf{C}} = \tilde{\mathbf{C}}. \end{aligned}$$

□

In Proposition 9.1, if $\mathbf{F} = \mathbf{I}$ and $\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\hat{\mathbf{B}} = \text{blue}(\mathbf{B})$; also, if $\mathbf{F} = \mathbf{e}_j$ and $\mathbf{D} = \mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\hat{b}_{ij} = \text{blue}(b_{ij})$, and so on.

The estimate $\hat{\mathbf{B}}$ is obviously unbiased for $\mathbf{B}$; furthermore, noting that $\mathbf{QY} = \mathbf{QE}$ and using Problem 6.4.5,

$$E\ \mathbf{Y}'\mathbf{QY} = E\ \mathbf{E}'\mathbf{QE} = (\text{tr } \mathbf{Q})\mathbf{\Sigma} = (n-k)\mathbf{\Sigma}.$$

Thus, $\mathbf{S} \equiv \mathbf{Y}'\mathbf{QY}/(n-k)$ is an unbiased estimate of $\mathbf{\Sigma}$.

Under normality, these estimates are optimal in the sense that they have minimum variance among all unbiased estimates. The likelihood for $(\mathbf{B}, \mathbf{\Sigma})$ from $\mathbf{Y}$ is

$$\begin{aligned} L(\mathbf{B}, \mathbf{\Sigma}) &\propto |\mathbf{\Sigma}|^{-n/2}\ \text{etr}\left[-\tfrac{1}{2}\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})\right] \\ &\propto |\mathbf{\Sigma}|^{-n/2}\ \text{etr}\left[-\tfrac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{B}'\mathbf{X}'\mathbf{XB}\right] \\ &\quad \cdot \exp\left\{-\tfrac{1}{2}\ \text{tr}\left[\mathbf{\Sigma}^{-1} \cdot (\mathbf{Y}'\mathbf{Y}) - 2\mathbf{\Sigma}^{-1}\mathbf{B}' \cdot (\mathbf{X}'\mathbf{Y})\right]\right\}. \end{aligned}$$

From general properties of exponential families [Fraser (1976), pp. 339, 342, 406 or Casella and Berger (1990), pp. 254-255, 263], the statistic $(\mathbf{Y}'\mathbf{Y}, \mathbf{X}'\mathbf{Y})$ is minimal sufficient and complete for $(\mathbf{B}, \mathbf{\Sigma})$. Of course, any one-to-one function such as $(\hat{\mathbf{B}}, \mathbf{S})$ is also minimal sufficient and complete. Thus, from Rao-Blackwell/Lehmann-Scheffé theorems, among all unbiased estimates, $\hat{\mathbf{B}}$ and $\mathbf{S}$ have minimum variance.

Using the decomposition $\mathbf{Y} = \mathbf{PY} + \mathbf{QY}$, where $\mathbf{PQ} = \mathbf{0}$, the log-likelihood can be written as

$$l(\mathbf{B}, \mathbf{\Sigma}) = cte - \frac{n}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\ \text{tr}\left\{\mathbf{\Sigma}^{-1}\left[\mathbf{Y}'\mathbf{QY} + (\mathbf{PY} - \mathbf{XB})'(\mathbf{PY} - \mathbf{XB})\right]\right\}.$$

Thus, to obtain the maximum likelihood estimates (MLE) $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Sigma}}$ when $n-k \geq p$ (for $\mathbf{V} \equiv (n-k)\mathbf{S} = \mathbf{Y}'\mathbf{QY}$ to be nonsingular w.p.1), we minimize

$$\ln|\mathbf{\Sigma}| + \text{tr}\ \frac{1}{n}\mathbf{V}\mathbf{\Sigma}^{-1} + \frac{1}{n}\ \text{tr}\ (\mathbf{PY} - \mathbf{XB})\mathbf{\Sigma}^{-1}(\mathbf{PY} - \mathbf{XB})',$$

and since the last term is ≥ 0, it is clear that $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, so we need only minimize

$$\ln|\mathbf{\Sigma}| + \text{tr}\frac{1}{n}\mathbf{V}\mathbf{\Sigma}^{-1}.$$

However, we already solved a similar problem in Chapter 7 when we derived the maximum likelihood estimates of the mean and variance of a multivariate normal distribution. Using the same result, we find that

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{V}$$

is the maximum likelihood estimate of $\mathbf{\Sigma}$. Proposition 9.2 gives the joint distribution of $\hat{\mathbf{B}}$ and $\mathbf{S}$.

Proposition 9.2 *With underlying normality, the joint distribution of* $\hat{\mathbf{B}}$ *and* $\mathbf{S}$ *is*

$$\begin{aligned} \hat{\mathbf{B}} &\sim N_p^k\left(\mathbf{B}, (\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{\Sigma}\right), \\ (n-k)\mathbf{S} &\sim W_p(n-k, \mathbf{\Sigma}). \end{aligned}$$

Moreover, $\hat{\mathbf{B}} \perp\!\!\!\perp \mathbf{S}$.

Proof. Since $\mathbf{Y} \sim N_p^n(\mathbf{XB}, \mathbf{I}_n \otimes \mathbf{\Sigma})$, the distribution of $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ follows from Proposition 6.1. Next, since $(n-k)\mathbf{S} = \mathbf{Y}'\mathbf{Q}\mathbf{Y} = \mathbf{E}'\mathbf{Q}\mathbf{E}$, the distribution of $(n-k)\mathbf{S}$ is a direct consequence of Proposition 7.8. Since $\mathbf{PQ} = \mathbf{0}$, we obtain immediately that

$$\begin{aligned} \text{var}\left[\begin{pmatrix}\mathbf{P}\\ \mathbf{Q}\end{pmatrix}\mathbf{Y}\right] &= \left(\begin{pmatrix}\mathbf{P}\\ \mathbf{Q}\end{pmatrix} \otimes \mathbf{I}_p\right)(\mathbf{I}_n \otimes \mathbf{\Sigma})\left((\mathbf{P}, \mathbf{Q}) \otimes \mathbf{I}_p\right) \\ &= \begin{pmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}\end{pmatrix} \otimes \mathbf{\Sigma} = \begin{pmatrix}\mathbf{P} \otimes \mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q} \otimes \mathbf{\Sigma}\end{pmatrix} \end{aligned}$$

and, thus, $\mathbf{PY} \perp\!\!\!\perp \mathbf{QY}$, which implies $\hat{\mathbf{B}} \perp\!\!\!\perp \mathbf{S}$. □

9.3 The general linear hypothesis

Consider now the problem of testing the general linear hypothesis

$$H_0 : \mathbf{CB} = \mathbf{0}$$

against all alternatives where $\mathbf{C} \in \mathbb{R}_k^r$ of rank $\mathbf{C} = r$ in the multivariate regression model

$$\mathbf{Y} \sim N_p^n(\mathbf{XB}, \mathbf{I}_n \otimes \mathbf{\Sigma})$$

with $\mathbf{X} \in \mathbb{R}_k^n$ of rank $\mathbf{X} = k$. The likelihood ratio test will be more easily expressed using a "canonical" form for this problem.

9.3.1 Canonical form

The canonical form is obtained by transforming the original response $\mathbf{Y}$, in two steps, so that in the new model $\mathbf{X}$ becomes $\mathbf{X}_0$ and $\mathbf{C}$ reduces to

$\mathbf{C}_0$, where

$$\mathbf{X}_0 = \begin{pmatrix} \mathbf{I}_k \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{C}_0 = (\mathbf{I}_r, \mathbf{0}).$$

Step 1: This step is to reduce $\mathbf{X}$ to $\mathbf{X}_0$. The Gram-Schmidt method applied to the columns of $\mathbf{X}$ (v. Problem 1.8.7) gives $\mathbf{X} = \mathbf{H}_1\mathbf{U}$, where $\mathbf{U} \in \mathbf{U}_k^+$ and $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{I}_k$. There exists $\mathbf{\Gamma}_1 \in \mathbb{R}_{n-k}^n$ such that $(\mathbf{H}_1, \mathbf{\Gamma}_1) \in \mathbf{O}_n$. Let

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{H}_1' \\ \mathbf{\Gamma}_1' \end{pmatrix} \mathbf{Y},$$

then $\tilde{\mathbf{Y}} \sim N_p^n(\mathbf{X}_0\tilde{\mathbf{B}}, \mathbf{I}_n \otimes \mathbf{\Sigma})$ with $\tilde{\mathbf{B}} = \mathbf{U}\mathbf{B}$. But since $\mathbf{C}\mathbf{B} = \mathbf{C}\mathbf{U}^{-1}\tilde{\mathbf{B}}$, the hypothesis $H_0 : \mathbf{C}\mathbf{B} = \mathbf{0}$ becomes $H_0 : \tilde{\mathbf{C}}\tilde{\mathbf{B}} = \mathbf{0}$, where $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{U}^{-1}$.

Step 2: The second step is to reduce $\tilde{\mathbf{C}}$ to $\mathbf{C}_0$. Once again, the Gram-Schmidt method applied to the rows of $\tilde{\mathbf{C}}$ yields $\tilde{\mathbf{C}} = \mathbf{L}\mathbf{H}_2$, where $\mathbf{H}_2\mathbf{H}_2' = \mathbf{I}_r$, $\mathbf{L} \in \mathbf{L}_r^+$. There exists $\mathbf{\Gamma}_2 \in \mathbb{R}_k^{k-r}$ such that

$$\begin{pmatrix} \mathbf{H}_2 \\ \mathbf{\Gamma}_2 \end{pmatrix} \in \mathbf{O}_k.$$

Let

$$\tilde{\tilde{\mathbf{Y}}} = \begin{pmatrix} \begin{pmatrix} \mathbf{H}_2 \\ \mathbf{\Gamma}_2 \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-k} \end{pmatrix} \tilde{\mathbf{Y}},$$

then $\tilde{\tilde{\mathbf{Y}}} \sim N_p^n(E\,\tilde{\tilde{\mathbf{Y}}}, \mathbf{I}_n \otimes \mathbf{\Sigma})$, where

$$E\,\tilde{\tilde{\mathbf{Y}}} = \begin{pmatrix} \mathbf{H}_2 \\ \mathbf{\Gamma}_2 \\ \mathbf{0} \end{pmatrix} \tilde{\mathbf{B}}$$

or

$$E \begin{pmatrix} \tilde{\tilde{\mathbf{Y}}}_1 \\ \tilde{\tilde{\mathbf{Y}}}_2 \\ \tilde{\tilde{\mathbf{Y}}}_3 \end{pmatrix} \equiv \begin{pmatrix} \tilde{\tilde{\mathbf{B}}}_1 \\ \tilde{\tilde{\mathbf{B}}}_2 \\ \mathbf{0} \end{pmatrix},$$

where $\tilde{\tilde{\mathbf{Y}}}$ was partitioned in conformity with $\tilde{\tilde{\mathbf{B}}}_1 = \mathbf{H}_2\tilde{\mathbf{B}} \in \mathbb{R}_p^r$ and $\tilde{\tilde{\mathbf{B}}}_2 = \mathbf{\Gamma}_2\tilde{\mathbf{B}} \in \mathbb{R}_p^{k-r}$. Now, to transform the hypothesis, note that

$$\tilde{\mathbf{C}}\tilde{\mathbf{B}} = \mathbf{L}\mathbf{H}_2\tilde{\mathbf{B}} = \mathbf{L}\tilde{\tilde{\mathbf{B}}}_1 = \mathbf{0} \Longleftrightarrow \tilde{\tilde{\mathbf{B}}}_1 = \mathbf{0}.$$

Hence, the hypothesis becomes $H_0 : \tilde{\tilde{\mathbf{B}}}_1 = \mathbf{0}$. Because the rows of $\tilde{\tilde{\mathbf{Y}}}$ are all independent, an equivalent problem in its canonical form, with the obvious change of notation, is to test

$$H_0 : \mathbf{M}_1 = \mathbf{0} \text{ against } H_1 : \mathbf{M}_1 \neq \mathbf{0}$$

based on

$$\begin{aligned}\mathbf{Z}_1 &\sim N_s^t(\mathbf{M}_1, \mathbf{I}_t \otimes \mathbf{\Sigma}),\\ \mathbf{Z}_2 &\sim N_s^u(\mathbf{M}_2, \mathbf{I}_u \otimes \mathbf{\Sigma}),\\ \mathbf{Z}_3 &\sim N_s^v(\mathbf{0}, \mathbf{I}_v \otimes \mathbf{\Sigma}),\ v \geq s,\end{aligned}$$

where $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are independent.

9.3.2 LRT for the canonical problem

We can now obtain a simple expression for the likelihood ratio test. In the canonical model, the likelihood function for $(\mathbf{M}_1, \mathbf{M}_2, \mathbf{\Sigma})$ is

$$\begin{aligned}L(\mathbf{M}_1, \mathbf{M}_2, \mathbf{\Sigma}) \quad \propto \quad & |\mathbf{\Sigma}|^{-n/2} \text{ etr}\left[-\tfrac{1}{2}\mathbf{\Sigma}^{-1}(\mathbf{Z}_1 - \mathbf{M}_1)'(\mathbf{Z}_1 - \mathbf{M}_1)\right]\\ & \cdot\text{etr}\left[-\tfrac{1}{2}\mathbf{\Sigma}^{-1}(\mathbf{Z}_2 - \mathbf{M}_2)'(\mathbf{Z}_2 - \mathbf{M}_2)\right]\\ & \cdot\text{etr}\left[-\tfrac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{Z}_3'\mathbf{Z}_3\right],\end{aligned}$$

where $n = t + u + v$. Note that in this form, the minimal and sufficient statistic is $(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3'\mathbf{Z}_3)$. For maximum likelihood estimates when $v \geq s$ (for $\mathbf{Z}_3'\mathbf{Z}_3$ to be nonsingular w.p.1), we minimize

$$\begin{aligned}-\frac{2}{n}l(\mathbf{M}_1, \mathbf{M}_2, \mathbf{\Sigma}) \quad = \quad & cte + \ln|\mathbf{\Sigma}| + \frac{1}{n}\ \text{tr}\ \mathbf{\Sigma}^{-1}(\mathbf{Z}_1 - \mathbf{M}_1)'(\mathbf{Z}_1 - \mathbf{M}_1)\\ & + \frac{1}{n}\ \text{tr}\ \mathbf{\Sigma}^{-1}(\mathbf{Z}_2 - \mathbf{M}_2)'(\mathbf{Z}_2 - \mathbf{M}_2) + \frac{1}{n}\ \text{tr}\ \mathbf{\Sigma}^{-1}\mathbf{Z}_3'\mathbf{Z}_3.\end{aligned}$$

Since each term is ≥ 0, it follows that the maximum likelihood estimates are

$$\hat{\mathbf{M}}_1 = \mathbf{Z}_1,\ \hat{\mathbf{M}}_2 = \mathbf{Z}_2,\ \text{and}\ \hat{\mathbf{\Sigma}} = \mathbf{Z}_3'\mathbf{Z}_3/n.$$

Also, when $\mathbf{M}_1 = \mathbf{0}$, the maximum likelihood estimates become

$$\hat{\hat{\mathbf{M}}}_1 = \mathbf{0},\ \hat{\hat{\mathbf{M}}}_2 = \mathbf{Z}_2,\ \text{and}\ \hat{\hat{\mathbf{\Sigma}}} = (\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{Z}_3)/n.$$

Therefore, the LRT is the test which rejects H_0 for small values of

$$\Lambda = \frac{L(\hat{\hat{\mathbf{M}}}_1, \hat{\hat{\mathbf{M}}}_2, \hat{\hat{\mathbf{\Sigma}}})}{L(\hat{\mathbf{M}}_1, \hat{\mathbf{M}}_2, \hat{\mathbf{\Sigma}})} = \frac{|\hat{\hat{\mathbf{\Sigma}}}|^{-n/2}}{|\hat{\mathbf{\Sigma}}|^{-n/2}} = \frac{|\mathbf{Z}_3'\mathbf{Z}_3|^{n/2}}{|\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{Z}_3|^{n/2}}.$$

Definition 9.1 *U-distribution: $U \sim U(p; m, n)$ iff $U \stackrel{d}{=} |\mathbf{W}_1|/|\mathbf{W}_1 + \mathbf{W}_2|$, where $\mathbf{W}_1 \sim W_p(n)$, $\mathbf{W}_2 \sim W_p(m)$, and $\mathbf{W}_1 \perp\!\!\!\perp \mathbf{W}_2$, $m + n > p$.*

Properties of U-distributions are deferred to Section 11.4.

Going back to the original model, the likelihood ratio test can be expressed in terms of $\hat{\mathbf{B}}$, $\hat{\mathbf{\Sigma}}$, and $\mathbf{X}$. Composing the two transformations

$\mathbf{Y} \mapsto \tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Y}} \mapsto \tilde{\tilde{\mathbf{Y}}}$, we obtain

$$\begin{pmatrix} \tilde{\tilde{\mathbf{Y}}}_1 \\ \tilde{\tilde{\mathbf{Y}}}_2 \\ \tilde{\tilde{\mathbf{Y}}}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{H}_2\mathbf{H}_1'\mathbf{Y} \\ \mathbf{\Gamma}_2\mathbf{H}_1'\mathbf{Y} \\ \mathbf{\Gamma}_1'\mathbf{Y} \end{pmatrix},$$

and after long but straightforward calculations, the LRT is expressed as

$$\begin{aligned} \Lambda^{2/n} &= \frac{|\tilde{\tilde{\mathbf{Y}}}_3'\tilde{\tilde{\mathbf{Y}}}_3|}{|\tilde{\tilde{\mathbf{Y}}}_1'\tilde{\tilde{\mathbf{Y}}}_1 + \tilde{\tilde{\mathbf{Y}}}_3'\tilde{\tilde{\mathbf{Y}}}_3|} \\ &= \frac{|n\hat{\mathbf{\Sigma}}|}{|n\hat{\mathbf{\Sigma}} + \hat{\mathbf{B}}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\mathbf{B}}|}. \end{aligned}$$

The null distribution of the LRT statistic follows directly from the canonical form of the model and the definition of U-distributions.

Proposition 9.3 *The null distribution of the LRT statistic* Λ *for testing* $H_0 : \mathbf{CB} = \mathbf{0}$ *against* $H_1 : \mathbf{CB} \neq \mathbf{0}$*, where* $\mathbf{C} \in \mathbb{R}_k^r$ *of* rank $\mathbf{C} = r$*, in the model* $\mathbf{Y} \sim N_p^n(\mathbf{XB}, \mathbf{I}_n \otimes \mathbf{\Sigma})$*, with* $\mathbf{X} \in \mathbb{R}_k^n$ *of* rank $\mathbf{X} = k$*, is* $\Lambda^{2/n} \sim U(p; r, n-k)$*.*

When n is large, a simple approximation can be used for the null distribution of Λ. From the LRT general theory, we can immediately write

$$-2\ln\Lambda \xrightarrow{d} \chi^2_{pr},\ n \to \infty.$$

9.3.3 Invariant tests

The problem in its canonical form is to test

$$H_0 : \mathbf{M}_1 = \mathbf{0} \text{ against } H_1 : \mathbf{M}_1 \neq \mathbf{0} \tag{9.2}$$

based on

$$\begin{aligned} \mathbf{Z}_1 &\sim N_s^t(\mathbf{M}_1, \mathbf{I}_t \otimes \mathbf{\Sigma}), \\ \mathbf{Z}_2 &\sim N_s^u(\mathbf{M}_2, \mathbf{I}_u \otimes \mathbf{\Sigma}), \\ \mathbf{Z}_3 &\sim N_s^v(\mathbf{0}, \mathbf{I}_v \otimes \mathbf{\Sigma}),\ v \geq s, \end{aligned}$$

where $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are independent. Since $v \geq s$, $\mathbf{Z}_3'\mathbf{Z}_3$ is nonsingular w.p.1 and let $m \equiv \min(s,t) = \text{rank } \mathbf{Z}_1$ w.p.1.

The group $\mathbf{G}_s \times \mathbb{R}_s^u \times \mathbf{O}_t \times \mathbf{O}_u \times \mathbf{O}_v$ transforms the variables as $\mathbf{Z}_1 \mapsto \mathbf{H}_1\mathbf{Z}_1\mathbf{A}$, $\mathbf{Z}_2 \mapsto \mathbf{H}_2\mathbf{Z}_2\mathbf{A} + \mathbf{B}$, and $\mathbf{Z}_3 \mapsto \mathbf{H}_3\mathbf{Z}_3\mathbf{A}$ for any $(\mathbf{A}, \mathbf{B}, \mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3) \in \mathbf{G}_s \times \mathbb{R}_s^u \times \mathbf{O}_t \times \mathbf{O}_u \times \mathbf{O}_v$. This induces the parameter transformations $\mathbf{M}_1 \mapsto \mathbf{H}_1\mathbf{M}_1\mathbf{A}$, $\mathbf{M}_2 \mapsto \mathbf{H}_2\mathbf{M}_2\mathbf{A} + \mathbf{B}$, and $\mathbf{\Sigma} \mapsto \mathbf{A}'\mathbf{\Sigma}\mathbf{A}$. Thus, we will say that a test function is invariant iff

$$f(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3) = f(\mathbf{H}_1\mathbf{Z}_1\mathbf{A}, \mathbf{H}_2\mathbf{Z}_2\mathbf{A} + \mathbf{B}, \mathbf{H}_3\mathbf{Z}_3\mathbf{A}),$$

$\forall(\mathbf{A},\mathbf{B},\mathbf{H}_1,\mathbf{H}_2,\mathbf{H}_3) \in \mathbf{G}_s \times \mathbb{R}_s^u \times \mathbf{O}_t \times \mathbf{O}_u \times \mathbf{O}_v$, $\forall(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3) \in \mathbb{R}_s^t \times \mathbb{R}_s^u \times \mathbb{R}_s^v$. The choice $\mathbf{B} = -\mathbf{H}_2\mathbf{Z}_2\mathbf{A}$ shows that any invariant test does not depend on $\mathbf{Z}_2$,

$$f(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3) = f(\mathbf{H}_1\mathbf{Z}_1\mathbf{A},\mathbf{0},\mathbf{H}_3\mathbf{Z}_3\mathbf{A}).$$

Since rank $\mathbf{Z}_3 = s$ w.p.1., then using Problem 1.8.7, there exists $\mathbf{U} \in \mathbf{U}_s^+$ and $\mathbf{H} \in \mathbb{R}_s^v$ satisfying $\mathbf{H}'\mathbf{H} = \mathbf{I}_s$ such that $\mathbf{Z}_3 = \mathbf{H}\mathbf{U}$. The choice $\mathbf{A} = \mathbf{U}^{-1}\mathbf{G}$, where $\mathbf{G} \in \mathbf{G}_s$ is arbitrary for now, yields

$$f(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3) = f(\mathbf{H}_1\mathbf{Z}_1\mathbf{U}^{-1}\mathbf{G},\mathbf{0},\mathbf{H}_3\mathbf{H}\mathbf{G}).$$

From the singular value decomposition (Proposition 1.11) there exists $\mathbf{G} \in \mathbf{O}_s$ and $\mathbf{H}_1 \in \mathbf{O}_t$ such that

$$\mathbf{H}_1(\mathbf{Z}_1\mathbf{U}^{-1})\mathbf{G} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $\mathbf{D}$ is diagonal and contains the square root of the nonzero eigenvalues of $(\mathbf{Z}_1\mathbf{U}^{-1})(\mathbf{Z}_1\mathbf{U}^{-1})' = \mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1'$. We thus have

$$f(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3) = f\left(\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},\mathbf{0},\mathbf{H}_3\mathbf{H}\mathbf{G}\right).$$

Finally, since $(\mathbf{H}\mathbf{G})'(\mathbf{H}\mathbf{G}) = \mathbf{I}_s$, the s columns of $\mathbf{H}\mathbf{G}$ are orthonormal in $\mathbb{R}^v$, and by completing to an orthonormal basis of $\mathbb{R}^v$, there exists $\mathbf{\Gamma}$ such that

$$\mathbf{H}_3 \equiv \begin{pmatrix} (\mathbf{H}\mathbf{G})' \\ \mathbf{\Gamma}' \end{pmatrix} \in \mathbf{O}_v$$

and

$$\mathbf{H}_3\mathbf{H}\mathbf{G} = \begin{pmatrix} \mathbf{I}_s \\ \mathbf{0} \end{pmatrix}.$$

Thus, altogether, we find

$$f(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3) = f\left(\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},\mathbf{0},\begin{pmatrix} \mathbf{I}_s \\ \mathbf{0} \end{pmatrix}\right),$$

which shows that any invariant test depends on $(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3)$ only through the nonzero eigenvalues of $\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1'$.

Invariance permits a reduction of the parameter space also. Since the transformed parameters are $\mathbf{M}_1 \mapsto \mathbf{H}_1\mathbf{M}_1\mathbf{A}$, $\mathbf{M}_2 \mapsto \mathbf{H}_2\mathbf{M}_2\mathbf{A}+\mathbf{B}$, and $\mathbf{\Sigma} \mapsto \mathbf{A}'\mathbf{\Sigma}\mathbf{A}$, the choice $\mathbf{B} = -\mathbf{H}_2\mathbf{M}_2\mathbf{A}$ shows that the non-null distribution of any invariant test is independent of $\mathbf{M}_2$. Similarly, using the singular value decomposition, there exists $\mathbf{G} \in \mathbf{O}_s$ and $\mathbf{H}_1 \in \mathbf{O}_t$ such that

$$\mathbf{H}_1(\mathbf{M}_1\mathbf{\Sigma}^{-1/2})\mathbf{G} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $\mathbf{D}$ is diagonal and contains the square root of the nonzero eigenvalues of

$$(\mathbf{M}_1\boldsymbol{\Sigma}^{-1/2})(\mathbf{M}_1\boldsymbol{\Sigma}^{-1/2})' = \mathbf{M}_1\boldsymbol{\Sigma}^{-1}\mathbf{M}_1'.$$

Thus, the choice $\mathbf{A} = \boldsymbol{\Sigma}^{-1/2}\mathbf{G}$ shows that the non-null distribution of any invariant test depends on $(\mathbf{M}_1, \mathbf{M}_2, \boldsymbol{\Sigma})$ only through the nonzero eigenvalues of

$$\mathbf{M}_1\boldsymbol{\Sigma}^{-1}\mathbf{M}_1'.$$

We have proved:

Proposition 9.4 *For the hypothesis testing situation* (9.2) *and group of transformations described above, any invariant test depends on* $(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3)$ *only through the nonzero eigenvalues of* $\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1'$. *Moreover, the non-null distribution of any invariant test depends on* $(\mathbf{M}_1, \mathbf{M}_2, \boldsymbol{\Sigma})$ *only through the nonzero eigenvalues of* $\mathbf{M}_1\boldsymbol{\Sigma}^{-1}\mathbf{M}_1'$.

The non-null distribution of those eigenvalues is complicated except in the case $m \equiv \min(s,t) = 1$, where there is only one such eigenvalue. From Proposition 9.4, assume without loss of generality that

$$\begin{aligned} \mathbf{Z}_1 &\sim N_s^t\left(\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \mathbf{I}_t \otimes \mathbf{I}_s\right), \\ \mathbf{Z}_3 &\sim N_s^v(\mathbf{0}, \mathbf{I}_v \otimes \mathbf{I}_s), \end{aligned}$$

where $\mathbf{D}$ contains the square root of the nonzero eigenvalues of $\mathbf{M}_1\boldsymbol{\Sigma}^{-1}\mathbf{M}_1'$.

(i) The case $t = 1$. Here, we have

$$\mathbf{Z}_1' \sim N_s((\mathbf{M}_1\boldsymbol{\Sigma}^{-1}\mathbf{M}_1')^{1/2}\mathbf{e}_1, \mathbf{I}_s)$$

and $\mathbf{Z}_3'\mathbf{Z}_3 \sim W_s(v)$. Hence, from Proposition 8.2 on Hotelling's test, the conclusion is

$$\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1' \sim F_c(s, v-s+1; \mathbf{M}_1\boldsymbol{\Sigma}^{-1}\mathbf{M}_1'/2).$$

(ii) The case $s = 1$. Here, the distributions are

$$\mathbf{Z}_1 \sim N_t((\mathbf{M}_1'\mathbf{M}_1/\sigma^2)^{1/2}\mathbf{e}_1, \mathbf{I}_t),$$

where $\boldsymbol{\Sigma} = \sigma^2$ was set, and $\mathbf{Z}_3'\mathbf{Z}_3 \sim \chi_v^2$. Thus, by definition,

$$\frac{\mathbf{Z}_1'\mathbf{Z}_1}{\mathbf{Z}_3'\mathbf{Z}_3} \sim F_c(t, v; \mathbf{M}_1'\mathbf{M}_1/2\sigma^2).$$

Another equivalent expression for the LRT is

$$\begin{aligned} \Lambda^{2/n} &= \frac{|\mathbf{Z}_3'\mathbf{Z}_3|}{|\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{Z}_3|} \\ &= |\mathbf{I}_s + \mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}|^{-1} \\ &= \prod_{i=1}^{m}(1+l_i)^{-1}, \end{aligned}$$

where $l_1 \geq \cdots \geq l_m$ are the ordered nonzero eigenvalues of

$$\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1'.$$

Thus, Λ takes on small values when those eigenvalues are large. Other possible tests could be used such as the following:

$$\text{Lawley-Hotelling:} \quad T_0^2 = \sum_{i=1}^m l_i = \text{tr } \mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}$$

$$\text{Pillai:} \quad V = \sum_{i=1}^m \frac{l_i}{1+l_i} = \text{tr } \mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{Z}_3)^{-1}$$

$$\text{Roy:} \quad l_1 = \text{largest eigenvalue of } \mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1'.$$

None of these tests has a power function which dominates the others over the whole parameter space or even locally [Fujikoshi (1988)]. However, it is easy to see that the asymptotic (as $v \to \infty$, s, t, and u are fixed) null distribution of the three tests $-2\ln\Lambda$, vT_0^2, and vV is χ_{st}^2. From the LRT general theory, we already know $-2\ln\Lambda \xrightarrow{d} \chi_{st}^2$. Under H_0, we can assume $\mathbf{\Sigma} = \mathbf{I}_s$ without loss of generality. From the law of large numbers, $\mathbf{Z}_3'\mathbf{Z}_3/v \xrightarrow{p} \mathbf{I}_s$; hence, we have from Lemma 6.3

$$v \text{ tr } \mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1} \xrightarrow{d} \text{tr } \mathbf{Z}_1'\mathbf{Z}_1 \overset{d}{=} \chi_{st}^2.$$

The same argument applies to vV. The asymptotic null distribution of Roy's test is quite different and is given as Problem 9.7.5 together with an interpretation as a union-intersection test.

In the very special case $m \equiv \min(s,t) = 1$, these three tests are equivalent to the LRT, which is uniformly most powerful invariant (UMPI). The proof of this UMPI property is the same as for the Hotelling-T^2 test (v. Proposition 8.4) since the non-null distribution in both cases (i) and (ii) above is a noncentral canonical F_c distribution.

Kariya et al. (1987) considered hypotheses related to selection and independence under multivariate regression models. Breiman and Friedman (1997) presented several methods of predicting responses in a multivariate regression model. The likelihood ratio test for detecting a single outlier (a shift in the mean) in a multivariate regression model was obtained by Srivastava and von Rosen (1998).

9.4 Random design matrix $\mathbf{X}$

When the prediction variables $\mathbf{X}$, just as the dependent variables $\mathbf{Y}$, are observed, then it is appropriate to consider $\mathbf{X}$ as a random matrix. The model most commonly encountered assumes

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \ \mathbf{E} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma}),$$

where the errors are independently distributed of the prediction variables, i.e., $\mathbf{E} \perp\!\!\!\perp \mathbf{X}$. When $\mathbf{X}$ has an absolutely continuous distribution, the argument in the proof of Proposition 7.5 shows that $\mathbf{X}'\mathbf{X}$ is nonsingular w.p.1. The conditional model

$$\mathbf{Y}|\mathbf{X} \sim N_p^n(\mathbf{XB}, \mathbf{I}_n \otimes \mathbf{\Sigma})$$

is thus identical to the case of a fixed $\mathbf{X}$. Using Proposition 9.2, we find the following properties of the same estimates:

(i) $\hat{\mathbf{B}}$ is unbiased,

$$E\ \hat{\mathbf{B}} = E\ E(\hat{\mathbf{B}}|\mathbf{X}) = E\ \mathbf{B} = \mathbf{B}.$$

(ii) $(n-k)\mathbf{S} \sim W_p(n-k, \mathbf{\Sigma})$. Indeed,

$$(n-k)\mathbf{S}|\mathbf{X} \sim W_p(n-k, \mathbf{\Sigma})$$

and this conditional distribution does not depend on $\mathbf{X}$.

(iii) $\hat{\mathbf{B}} \perp\!\!\!\perp \mathbf{S}$. With Proposition 2.13,

$$\begin{aligned} E\ g(\hat{\mathbf{B}}) \cdot h(\mathbf{S}) &= E\ E[g(\hat{\mathbf{B}})h(\mathbf{S})|\mathbf{X}] \\ &= E\{E[g(\hat{\mathbf{B}})|\mathbf{X}]E[h(\mathbf{S})|\mathbf{X}]\} \\ &= E\{E[g(\hat{\mathbf{B}})|\mathbf{X}]E[h(\mathbf{S})]\} \\ &= E\ g(\hat{\mathbf{B}}) \cdot E\ h(\mathbf{S}). \end{aligned}$$

Moreover, for testing the general linear hypothesis $H_0 : \mathbf{CB} = \mathbf{0}$, the conditional null distribution of $\Lambda^{2/n}$ does not depend on $\mathbf{X}$ and, thus,

$$\Lambda^{2/n} \sim U(p; r, n-k), \text{ unconditionally.}$$

The non-null distribution of $\Lambda^{2/n}$, however, will depend on the distribution of $\mathbf{X}$, as is the case for $p = 1$ as exemplified by Problem 5.7.8, where the noncentrality parameter of the distribution of the F-test depends on $\mathbf{X}$.

Example 9.1 *The variance of* $\hat{\mathbf{B}}$ *may be evaluated as*

$$\begin{aligned} \text{var } \hat{\mathbf{B}} &= \text{var vec}(\hat{\mathbf{B}}') \\ &= E\ \text{var}[\text{vec}(\hat{\mathbf{B}}')|\mathbf{X}] + \text{var } E[\text{vec}(\hat{\mathbf{B}}')|\mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{\Sigma}] + \text{var vec}(\mathbf{B}') \\ &= \left(E\ (\mathbf{X}'\mathbf{X})^{-1}\right) \otimes \mathbf{\Sigma}. \end{aligned}$$

The last expectation may be evaluated directly in some cases. For example, if $\mathbf{X} \sim N_k^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Omega})$, $\mathbf{\Omega} > \mathbf{0}$, *then with Problem* 7.5.5 *and since* $\mathbf{X}'\mathbf{X} \sim W_k(n, \mathbf{\Omega})$,

$$E\ (\mathbf{X}'\mathbf{X})^{-1} = (n-k-1)^{-1}\mathbf{\Omega}^{-1}.$$

9.5 Predictions

The problem of predicting several, possibly correlated, responses from the same set of predictors is becoming increasingly important. Applications by Breiman and Freidman (1997) include prediction of changes in the valuations of stocks in 60 industry groups by using over 100 econometric variables as predictors. Or, in chemometrics, the prediction of 6 output characteristics of the polymers produced as predicted by 22 predictor variables. Another example by Brown (1980, pp. 247-292) lists electoral results for all 71 Scottish constituencies in the Bristish general elections of February and October 1974. Data consist of total votes for each of the four parties (Conservative, Labour, Liberal, and Nationalist) in each election, together with a categorical variable listing the location of the constituency by six regions, and the size of the electorate in each constituency. The objective is to use the February and October results from part of the constituencies to predict the remaining October results from the corresponding February data. Research papers related to predictions include Stone (1974), van der Merwe and Zidek (1980), Bilodeau and Kariya (1989), and Breiman and Friedman (1997).

Assume the "centered" model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},\ \mathbf{E} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma}),\ \mathbf{X} \perp\!\!\!\perp \mathbf{E}.$$

For the sake of simplicity, we take $\mathbf{X}$ centered, $\mathbf{X} \sim N_k^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Omega})$. For given values of the prediction variables, $\mathbf{x}' = (x_1, \ldots, x_k)$, it is desired to obtain a prediction of the dependent variables, $\mathbf{y}' = (y_1, \ldots, y_p)$. Using the Gauss-Markov (GM) estimate $\hat{\mathbf{B}}$, an obvious prediction method is

$$\hat{\mathbf{y}}' = \mathbf{x}'\hat{\mathbf{B}} = (\mathbf{x}'\hat{\boldsymbol{\beta}}_1, \ldots, \mathbf{x}'\hat{\boldsymbol{\beta}}_p)$$

so that prediction of the ith variable is done considering only the ith multiple regression model. Assuming the "future" observation follows the same model, i.e., $\mathbf{y}' = \mathbf{x}'\mathbf{B} + \mathbf{e}'$, where $\mathbf{x} \sim N_k(\mathbf{0}, \mathbf{\Omega})$, $\mathbf{e} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, and $\mathbf{x} \perp\!\!\!\perp \mathbf{e}$, and is independent of the past, $(\mathbf{x}, \mathbf{e}) \perp\!\!\!\perp (\mathbf{X}, \mathbf{E})$, one can evaluate the *risk* of the GM prediction as

$$\begin{aligned} E\ (\hat{\mathbf{y}} - \mathbf{y})'\mathbf{\Sigma}^{-1}(\hat{\mathbf{y}} - \mathbf{y}) &= E\ [\mathbf{x}'(\hat{\mathbf{B}} - \mathbf{B}) - \mathbf{e}']\mathbf{\Sigma}^{-1}[(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{x} - \mathbf{e}] \\ &= E\ \mathrm{tr}\left[(\hat{\mathbf{B}} - \mathbf{B})\mathbf{\Sigma}^{-1}(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{\Omega}\right] + p \\ &= \mathrm{tr}\left[E\ (\hat{\mathbf{B}} - \mathbf{B})\mathbf{\Sigma}^{-1}(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{\Omega}\right] + p. \end{aligned}$$

The SPER (sum of Squares of Prediction Error when the independent variable is Random) risk is obtained on subtracting p from the above:

$$R_{\mathrm{SPER}}(\hat{\mathbf{B}}) = E\ \mathrm{tr}\ (\hat{\mathbf{B}} - \mathbf{B})\mathbf{\Sigma}^{-1}(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{\Omega}.$$

Letting $\mathbf{U} = (\hat{\mathbf{B}} - \mathbf{B})\mathbf{\Sigma}^{-1/2} \sim N_p^k(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1} \otimes \mathbf{I})$, Example 6.3 gives $E(\mathbf{U}\mathbf{U}'|\mathbf{X}) = p(\mathbf{X}'\mathbf{X})^{-1}$. Finally, with Example 9.1, we get

$$R_{\text{SPER}}(\hat{\mathbf{B}}) = pk/(n-k-1).$$

A closely related risk function, more tractable mathematically, is SPE defined by

$$R_{\text{SPE}}(\hat{\mathbf{B}}) = E \text{ tr } (\hat{\mathbf{B}} - \mathbf{B})\mathbf{\Sigma}^{-1}(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{X}'\mathbf{X}.$$

Smaller risk may be achieved with an estimate $\tilde{\mathbf{B}} = \hat{\mathbf{B}}\mathbf{A}$ for a certain $\mathbf{A} \in \mathbb{R}_p^p$. The corresponding prediction for each variable in $\tilde{\mathbf{y}} = \mathbf{A}'\hat{\mathbf{y}}$ is seen to be a linear combination (multivariate flattening) of the p prediction equations.

Example 9.2 Multivariate flattening. *Assuming* $(\mathbf{B}, \mathbf{\Sigma}, \mathbf{\Omega})$ *is known, then an optimal multivariate flattening [Breiman and Friedman* (1997)*] would be solution of*

$$\min_{\mathbf{A}} E\ (\mathbf{A}'\hat{\mathbf{y}} - \mathbf{y})'\mathbf{\Sigma}^{-1}(\mathbf{A}'\hat{\mathbf{y}} - \mathbf{y}).$$

Now, since

$$(\mathbf{A}'\hat{\mathbf{y}} - \mathbf{y})'\mathbf{\Sigma}^{-1}(\mathbf{A}'\hat{\mathbf{y}} - \mathbf{y}) = (\mathbf{\Sigma}^{-1/2}\mathbf{A}'\hat{\mathbf{y}} - \mathbf{\Sigma}^{-1/2}\mathbf{y})'(\mathbf{\Sigma}^{-1/2}\mathbf{A}'\hat{\mathbf{y}} - \mathbf{\Sigma}^{-1/2}\mathbf{y}),$$

letting $\mathbf{C} = \mathbf{\Sigma}^{-1/2}\mathbf{A}'$ *and* $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{y}$*, the optimization problem becomes equivalent to*

$$\min_{\mathbf{C}} E\ |\mathbf{z} - \mathbf{C}\hat{\mathbf{y}}|^2.$$

With Problem 5.7.6*, the solution is readily obtained:*

$$\mathbf{C} = \text{cov}(\mathbf{z}, \hat{\mathbf{y}}) \cdot [\text{var } \hat{\mathbf{y}}]^{-1}.$$

Each factor is evaluated as

$$\begin{aligned}
\text{cov}(\mathbf{z}, \hat{\mathbf{y}}) &= \mathbf{\Sigma}^{-1/2}\text{cov}(\mathbf{y}, \hat{\mathbf{y}}) \\
&= \mathbf{\Sigma}^{-1/2} E\ \mathbf{y}\hat{\mathbf{y}}' \\
&= \mathbf{\Sigma}^{-1/2} E\ (\mathbf{B}'\mathbf{x} + \mathbf{e})\mathbf{x}'\hat{\mathbf{B}} \\
&= \mathbf{\Sigma}^{-1/2} E\ \mathbf{B}'\mathbf{x}\mathbf{x}'\hat{\mathbf{B}} \\
&= \mathbf{\Sigma}^{-1/2}\mathbf{B}'\mathbf{\Omega}\mathbf{B}
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\text{var } \hat{\mathbf{y}} &= E\ \hat{\mathbf{y}}\hat{\mathbf{y}}' \\
&= E\ [\mathbf{B} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}]'\mathbf{x}\mathbf{x}'[\mathbf{B} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}] \\
&= \mathbf{B}'\mathbf{\Omega}\mathbf{B} + E\ \mathbf{E}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E} \\
&= \mathbf{B}'\mathbf{\Omega}\mathbf{B} + E\left(\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right)\mathbf{\Sigma} \quad (v.\ \textit{Problem}\ 6.4.5) \\
&= \mathbf{B}'\mathbf{\Omega}\mathbf{B} + [k/(n-k-1)]\mathbf{\Sigma} \quad (v.\ \textit{Problem}\ 7.5.4).
\end{aligned}$$

Hence, altogether, the optimal $\mathbf{A}$ *is given by*

$$\mathbf{A} = [(\mathbf{B}'\mathbf{\Omega}\mathbf{B}) + r\mathbf{\Sigma}]^{-1}(\mathbf{B}'\mathbf{\Omega}\mathbf{B}),\ r = k/(n-k-1).$$

Sample-based $\hat{\mathbf{A}}$ and modifications thereof are given in the above papers. In particular, for small r, $\mathbf{A} \approx \mathbf{I} - r(\mathbf{B}'\mathbf{\Omega}\mathbf{B})^{-1}\mathbf{\Sigma}$ (v. Problem 1.8.15) and van der Merwe and Zidek (1980) established that the sample-based

$$\hat{\mathbf{A}} = \mathbf{I} - r(n-k)(\hat{\mathbf{B}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}})^{-1}\mathbf{S}$$

which they called FICYREG (FIltered Canonical Y REGression) leads to smaller SPE risk than GM for

$$r = (k-p-1)/(n-k+p+1)$$

provided $n > k > p+1$. Bilodeau and Kariya (1989) proposed the modified Efron-Morris (1976)

$$\hat{\mathbf{A}} = \mathbf{I} - r(n-k)(\hat{\mathbf{B}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}})^{-1}\mathbf{S} - b(n-k)\mathbf{S}/\mathrm{tr}(\hat{\mathbf{B}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}})$$

and showed that it leads still to a smaller SPE risk than FICYREG for

$$r = (k-p-1)/(n-k+p+1) \quad \text{and} \quad b = (p-1)/(n-k+p+1).$$

Note that the choice $b = 0$ reduces to FICYREG. Breiman and Friedman (1997) considered $\hat{\mathbf{A}}$ built from cross-validation (CV) and generalized cross-validation (GCV). Their large-scale simulations point strongly toward the superiority of CV and GCV over other commonly used prediction techniques. The GCV in particular seems very promising since its evaluation is nearly as simple as GM. The CV, in contrast, is computationally intensive. The unbiased estimate of the SPE risk for the GCV predictions was recently obtained by Bilodeau (1998).

9.6 One-way classification

In this section, the one-factor univariate analysis of variance is generalized to test the equality of several means of multivariate normal populations. Let $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i}$ i.i.d. $N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, $i = 1, \ldots, a$, be a independent samples from multivariate normal distributions with common variance $\mathbf{\Sigma} > \mathbf{0}$. In matrix notation, let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_{11}' \\ \vdots \\ \mathbf{y}_{1n_1}' \\ \vdots \\ \mathbf{y}_{a1}' \\ \vdots \\ \mathbf{y}_{an_a}' \end{pmatrix},\ \mathbf{X} = \mathrm{diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_a}),\ \mathbf{B} = \begin{pmatrix} \boldsymbol{\mu}_1' \\ \vdots \\ \boldsymbol{\mu}_a' \end{pmatrix}.$$

Then, the a samples can be written as the multivariate regression model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},\ \mathbf{E} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma}),$$

where $n = \sum_{i=1}^a n_i$. The hypothesis of equality of means

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_a$$

can be translated into a general linear hypothesis. Define

$$\mathbf{C} = (\mathbf{I}_{a-1}, -\mathbf{1}_{a-1});$$

then the hypothesis becomes $H_0 : \mathbf{CB} = \mathbf{0}$, where $\mathbf{C} \in \mathbb{R}_a^{a-1}$ of rank $\mathbf{C} = a-1$. Using the canonical formulation of this problem, the reader can verify that the LRT is

$$\Lambda^{2/n} = \frac{|\mathbf{SS}_w|}{|\mathbf{SS}_w + \mathbf{SS}_b|},$$

where

$$\begin{aligned} \mathbf{SS}_w &= \sum_{i=1}^a \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)', \\ \mathbf{SS}_b &= \sum_{i=1}^a n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})' \end{aligned}$$

are the usual "within" and "between" sums of squares with

$$\bar{\mathbf{y}}_i = \sum_{j=1}^{n_i} \mathbf{y}_{ij}/n_i \quad \text{and} \quad \bar{\mathbf{y}} = \sum_{i=1}^a n_i \bar{\mathbf{y}}_i / n.$$

The other analysis-of-variance models such as the two-way classification model can be generalized similarly to test the effect of each factor or the presence of interactions between factors. We will not pursue this any further here.

9.7 Problems

1. Show that the likelihood ratio test statistic Λ for testing the general linear hypothesis can be written

$$\Lambda^{2/n} = \frac{|n\hat{\mathbf{\Sigma}}|}{|n\hat{\mathbf{\Sigma}} + \hat{\mathbf{B}}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\mathbf{B}}|}.$$

2. The estimate $\mathbf{X}\hat{\mathbf{B}} = \mathbf{PY}$ is the orthogonal projection of $\mathbf{Y}$ on $\mathcal{V} = \{\mathbf{XA} :\ \mathbf{A} \in \mathbb{R}_p^k\}$. Use this fact to prove that $\mathbf{X}\hat{\mathbf{B}}$ is also the solution of the least-squares problem

$$\min_{\mathbf{V} \in \mathcal{V}} \operatorname{tr} \mathbf{\Omega}(\mathbf{Y} - \mathbf{V})'(\mathbf{Y} - \mathbf{V}),$$

for any fixed $\mathbf{\Omega} \in \mathbb{R}_p^p$, $\mathbf{\Omega} > \mathbf{0}$.

3. Prove that if $\mathbf{Z}_1 \sim N_s^t(\mathbf{M}_1, \mathbf{I}_t \otimes \mathbf{\Sigma})$, $\mathbf{Z}_3 \sim N_s^v(\mathbf{0}, \mathbf{I}_v \otimes \mathbf{\Sigma})$, and $\mathbf{Z}_1 \perp\!\!\!\perp \mathbf{Z}_3$, where $\mathbf{\Sigma} > \mathbf{0}$ and $v \geq s+2$, then

$$E\ \mathbf{Z}_1'\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1} = \frac{t}{v-s-1}\mathbf{I}_s + \frac{1}{v-s-1}\mathbf{M}_1'\mathbf{M}_1\mathbf{\Sigma}^{-1}.$$

4. Using the canonical model, prove that the LRT Λ for the hypothesis of the equality of several multivariate means is

$$\Lambda^{2/n} = |\mathbf{SS}_w|/|\mathbf{SS}_w + \mathbf{SS}_b|,$$

as described in Section 9.6. What is the null distribution of this test?

5. The general linear hypothesis in its canonical form is to test

$$H_0 : \mathbf{M}_1 = \mathbf{0} \text{ against } H_1 : \mathbf{M}_1 \neq \mathbf{0}$$

based on

$$\begin{aligned} \mathbf{Z}_1 &\sim N_s^t(\mathbf{M}_1, \mathbf{I}_t \otimes \mathbf{\Sigma}), \\ \mathbf{Z}_2 &\sim N_s^u(\mathbf{M}_2, \mathbf{I}_u \otimes \mathbf{\Sigma}), \\ \mathbf{Z}_3 &\sim N_s^v(\mathbf{0}, \mathbf{I}_v \otimes \mathbf{\Sigma}),\ v \geq s, \end{aligned}$$

where $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are independent.

(i) Prove the asymptotic result as $v \to \infty$ concerning Roy's test: If $\mathbf{M}_1 = \mathbf{0}$ then $vl_1 \xrightarrow{d} \alpha_1$, where l_1 is the largest eigenvalue of $\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1'$ and α_1 is the largest eigenvalue of a random matrix $\mathbf{W} \sim W_s(t)$.

(ii) Union-intersection test.

(a) For a given $\mathbf{h} \in \mathbb{R}^t$, $|\mathbf{h}| = 1$, define $H_{\mathbf{h},0} : \mathbf{M}_1'\mathbf{h} = \mathbf{0}$ and $H_{\mathbf{h},1} : \mathbf{M}_1'\mathbf{h} \neq \mathbf{0}$. Prove

$$\begin{aligned} H_0 &= \cap_{\mathbf{h}}\{H_{\mathbf{h},0} : |\mathbf{h}| = 1\}, \\ H_1 &= \cup_{\mathbf{h}}\{H_{\mathbf{h},1} : |\mathbf{h}| = 1\}. \end{aligned}$$

(b) For a given $\mathbf{h}$, $|\mathbf{h}| = 1$, prove that the LRT for $H_{\mathbf{h},0}$ against $H_{\mathbf{h},1}$ accepts $H_{\mathbf{h},0}$ for small values of

$$R_{\mathbf{h}} = \mathbf{h}'\mathbf{Z}_1(\mathbf{Z}_3'\mathbf{Z}_3)^{-1}\mathbf{Z}_1'\mathbf{h}.$$

Demonstrate the null distribution $R_{\mathbf{h}} \sim F_c(s, v-s+1)$ does not depend on $\mathbf{h}$.

(c) The union-intersection test accepts H_0 iff $\sup_{|\mathbf{h}|=1} R_{\mathbf{h}} \leq c$ for some constant c. Demonstrate the union-intersection test statistic $\sup_{|\mathbf{h}|=1} R_{\mathbf{h}} = l_1$ is, in fact, Roy's test.

Remark: For a given $\mathbf{h}$, the test based on $R_{\mathbf{h}}$ is UMPI for testing $H_{\mathbf{h},0}$ against $H_{\mathbf{h},1}$ (the non-null distribution of $R_{\mathbf{h}}$ is a noncentral canonical F_c distribution just like Hotelling's-T^2; v. Proposition 8.4), but Roy's test is not generally UMPI for testing H_0 against H_1.

10
Principal components

10.1 Introduction

In this chapter we assume that $\mathbf{x} \in \mathbb{R}^p$ with $E\,\mathbf{x} = \boldsymbol{\mu}$ and $\operatorname{var}\mathbf{x} = \boldsymbol{\Sigma} = (\sigma_{ij})$. When the dimension p is large, the principal components method seeks to replace $\mathbf{x}$ by $\mathbf{y} \in \mathbb{R}^k$, where $k < p$ (and hopefully much smaller), without losing too much "information." This is sometimes particularly useful for a graphical description of the data since it is much easier to view vectors of low dimension. Section 10.2 defines principal components and gives their interpretation as normalized linear combinations with maximum variance. In Section 10.3, we explain an optimal property of principal components as best approximating subspace of dimension k in terms of squared prediction error. Section 10.4 introduces the sample principal components; they give the coordinates of the projected data which is closest, in terms of euclidian distance, to the original data. Section 10.5 treats the sample principal components calculated from the correlation matrix. Finally, Section 10.6 presents a simple test for multivariate normality which generalizes the univariate Shapiro and Wilk's statistic. A book entirely devoted to principal component analysis is that of Jolliffe (1986).

10.2 Definition and basic properties

The *total variance* of $\mathbf{x}$ is defined as

$$E\ |\mathbf{x}-\boldsymbol{\mu}|^2 = \sum_{i=1}^{p} \text{var}\ x_i = \sum_{i=1}^{p} \sigma_{ii} = \text{tr}\ \boldsymbol{\Sigma}.$$

Recall that $\boldsymbol{\Sigma} \geq \mathbf{0}$ can be written as $\boldsymbol{\Sigma} = \mathbf{HDH}'$, where

$$\begin{aligned} \mathbf{H} &= (\mathbf{h}_1, \ldots, \mathbf{h}_p) \in \mathbf{O}_p, \\ \mathbf{D} &= \text{diag}(\lambda_1, \ldots, \lambda_p), \end{aligned}$$

and $\lambda_1 \geq \cdots \geq \lambda_p$ are the ordered eigenvalues of $\boldsymbol{\Sigma}$. Since we are only interested in var $\mathbf{x}$, we will assume throughout this chapter that $\boldsymbol{\mu} = \mathbf{0}$. If we let

$$\mathbf{y} = \mathbf{H}'\mathbf{x} = \begin{pmatrix} \mathbf{h}_1'\mathbf{x} \\ \vdots \\ \mathbf{h}_p'\mathbf{x} \end{pmatrix},$$

var $\mathbf{y} = \mathbf{D}$. Then $\sum_{i=1}^{p} \text{var}\ y_i = \sum_{i=1}^{p} \lambda_i = \text{tr}\ \boldsymbol{\Sigma}$, so $\mathbf{x}$ and $\mathbf{y}$ have the same "total variance." Moreover, the variables y_i's are uncorrelated,

$$\text{cov}(\mathbf{h}_i'\mathbf{x}, \mathbf{h}_j'\mathbf{x}) = \mathbf{h}_i'\boldsymbol{\Sigma}\mathbf{h}_j = \lambda_j \mathbf{h}_i'\mathbf{h}_j = \lambda_j \delta_{ij}.$$

Definition 10.1 *The variables $y_i = \mathbf{h}_i'\mathbf{x}$, $i = 1, \ldots, p$, are, by definition, the* principal components *of* $\mathbf{x}$.

Since $\mathbf{HH}' = \mathbf{I}$, then $\mathbf{x} = (\sum_{i=1}^{p} \mathbf{h}_i\mathbf{h}_i')\mathbf{x} = \sum_{i=1}^{p} y_i\mathbf{h}_i$ and the principal components can be viewed as the coordinates of $\mathbf{x}$ with respect to the orthonormal basis $\{\mathbf{h}_1, \ldots, \mathbf{h}_p\}$ of $\mathbb{R}^p$. When the ratio $\sum_{i=1}^{k} \lambda_i / \text{tr}\ \boldsymbol{\Sigma}$ is close to 1, then $(y_1, \ldots, y_k)'$ can effectively replace $\mathbf{x}$ without losing much in terms of "total variance."

The principal components can also be got sequentially as follows. First a normalized linear combination $\mathbf{t}'\mathbf{x}$, $|\mathbf{t}| = 1$, is sought such that var $\mathbf{t}'\mathbf{x} = \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}$ is maximum. Since for all $\mathbf{t}$, $|\mathbf{t}| = 1$,

$$\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} = \sum_{i=1}^{p} \lambda_i (\mathbf{t}'\mathbf{h}_i)^2 \leq \lambda_1 \sum_{i=1}^{p} (\mathbf{t}'\mathbf{h}_i)^2 = \lambda_1 \mathbf{t}' \left(\sum_{i=1}^{p} \mathbf{h}_i\mathbf{h}_i' \right) \mathbf{t} = \lambda_1 |\mathbf{t}|^2 = \lambda_1;$$

hence, $\max_{\mathbf{t}'\mathbf{t}=1} \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} = \lambda_1$, which is attained for $\mathbf{t} = \mathbf{h}_1$. So, the first principal component $y_1 = \mathbf{h}_1'\mathbf{x}$ is the normalized linear combination with maximum variance. Now, given $y_i = \mathbf{h}_i'\mathbf{x}$, $i = 1, \ldots, k$, another linear combination $\mathbf{s}'\mathbf{x}$, $|\mathbf{s}| = 1$, is sought which maximizes the variance $\mathbf{s}'\boldsymbol{\Sigma}\mathbf{s}$ and is uncorrelated with $y_1, \ldots, y_k$. Note that $\text{cov}(\mathbf{s}'\mathbf{x}, y_i) = \lambda_i \mathbf{s}'\mathbf{h}_i$, $i = 1, \ldots, k$. As above, for all $\mathbf{s} \perp \mathbf{h}_1, \ldots, \mathbf{h}_k$, $|\mathbf{s}| = 1$, we have

$$\mathbf{s}'\boldsymbol{\Sigma}\mathbf{s} = \sum_{i=k+1}^{p} \lambda_i (\mathbf{s}'\mathbf{h}_i)^2 \leq \lambda_{k+1} \sum_{i=k+1}^{p} (\mathbf{s}'\mathbf{h}_i)^2 = \lambda_{k+1} \sum_{i=1}^{p} (\mathbf{s}'\mathbf{h}_i)^2 = \lambda_{k+1}.$$

Hence,

$$\max_{\substack{\mathbf{s}'\mathbf{s}=1 \\ \mathbf{s}\perp\mathbf{h}_1,\ldots,\mathbf{h}_k}} \mathbf{s}'\boldsymbol{\Sigma}\mathbf{s} = \lambda_{k+1}$$

is attained for $\mathbf{s} = \mathbf{h}_{k+1}$, which means that $y_{k+1} = \mathbf{h}'_{k+1}\mathbf{x}$ is the normalized linear combination with maximum variance among all those uncorrelated with $y_1, \ldots, y_k$.

10.3 Best approximating subspace

The orthogonal projection of $\mathbf{x}$ on the subspace spanned by the first k eigenvectors, $\mathbf{P}_k\mathbf{x}$, is

$$\mathbf{P}_k\mathbf{x} = \left(\sum_{i=1}^{k} \mathbf{h}_i\mathbf{h}'_i\right)\mathbf{x} = \sum_{i=1}^{k} y_i\mathbf{h}_i.$$

Proposition 10.1 shows that $\mathbf{P}_k\mathbf{x}$ gives the best approximation to $\mathbf{x}$ by a subspace of dimension at most k in terms of squared prediction error. Before stating the result, we present a lemma. Denote by $\mathcal{P}_k^{\perp}$ the set of all orthogonal projections $\mathbf{P} \in \mathbb{R}_p^p$ of rank $\mathbf{P} = k$.

Lemma 10.1 *Let $\boldsymbol{\Sigma} \geq \mathbf{0}$ in $\mathbb{R}_p^p$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$. Then,*

$$\begin{aligned} \max_{\mathbf{P}\in\mathcal{P}_k^{\perp}} \text{tr } \boldsymbol{\Sigma}\mathbf{P} &= \sum_{i=1}^{k} \lambda_i, \\ \min_{\mathbf{P}\in\mathcal{P}_k^{\perp}} \text{tr } \boldsymbol{\Sigma}(\mathbf{I}-\mathbf{P}) &= \sum_{i=k+1}^{p} \lambda_i \end{aligned}$$

are attained at $\mathbf{P} = \sum_{i=1}^{k} \mathbf{h}_i\mathbf{h}'_i$, where

$$\boldsymbol{\Sigma} = \mathbf{H}\mathbf{D}\mathbf{H}', \ \mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_p) \in \mathbf{O}_p, \ \mathbf{D} = \text{diag}(\lambda_1, \ldots, \lambda_p).$$

Proof. Take any $\mathbf{P} \in \mathcal{P}_k^{\perp}$. Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$ whose columns form an orthonormal basis for Im $\mathbf{P}$, then $\mathbf{P} = \mathbf{A}\mathbf{A}'$. Now,

$$\text{tr } \boldsymbol{\Sigma}\mathbf{P} = \text{tr } \mathbf{H}\mathbf{D}\mathbf{H}'\mathbf{A}\mathbf{A}' = \text{tr } \mathbf{D}(\mathbf{H}'\mathbf{A})(\mathbf{H}'\mathbf{A})',$$

and note that $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_k) \equiv \mathbf{H}'\mathbf{A}$ has orthonormal columns too, i.e., $\mathbf{G}'\mathbf{G} = \mathbf{I}_k$. Therefore,

$$\begin{aligned} \text{tr } \boldsymbol{\Sigma}\mathbf{P} &= \text{tr } \mathbf{D}\sum_{i=1}^{k} \mathbf{g}_i\mathbf{g}'_i = \sum_{i=1}^{k} \mathbf{g}'_i\mathbf{D}\mathbf{g}_i \\ &\leq \sum_{i=1}^{k} \max_{\substack{\mathbf{g}'\mathbf{g}=1 \\ \mathbf{g}\perp\mathbf{g}_1,\ldots,\mathbf{g}_{i-1}}} \mathbf{g}'\mathbf{D}\mathbf{g} = \sum_{i=1}^{k} \lambda_i \end{aligned}$$

(when $i = 1$ the orthogonality condition is void) with equality if $\mathbf{g}_i = \mathbf{e}_i$, which means $\mathbf{A} = \mathbf{HG} = (\mathbf{h}_1, \ldots, \mathbf{h}_k)$. This shows the first part related to the maximum. The second part is immediate. □

Proposition 10.1 *Assume* $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma} > \mathbf{0}$, *and let* $\mathbf{B} \in \mathbb{R}_p^k$ *of* rank $\mathbf{B} = k$, $\mathbf{C} \in \mathbb{R}_k^p$. *Then,*

$$\min_{\mathbf{B},\mathbf{C}} E\ |\mathbf{x} - \mathbf{CBx}|^2 = \sum_{i=k+1}^{p} \lambda_i$$

is attained when $\mathbf{CB} = \mathbf{P}_k$.

Proof. Fix $\mathbf{B}$. We have

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{Bx} \end{pmatrix} \sim N_{p+k}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma B}' \\ \mathbf{B\Sigma} & \mathbf{B\Sigma B}' \end{pmatrix}\right)$$

and $\mathbf{x} \mid \mathbf{Bx} \sim N_p(\mathbf{0}, \mathbf{\Sigma} - \mathbf{\Sigma B}'(\mathbf{B\Sigma B}')^{-1}\mathbf{B\Sigma})$. Using Problem 5.7.6,

$$\begin{aligned} \min_{\mathbf{C}} E\ |\mathbf{x} - \mathbf{CBx}|^2 &= \text{tr}\ \left[\mathbf{\Sigma} - \mathbf{\Sigma B}'(\mathbf{B\Sigma B}')^{-1}\mathbf{B\Sigma}\right] \\ &= \text{tr}\ \mathbf{\Sigma}\left[\mathbf{I} - \mathbf{AB}'(\mathbf{B\Sigma B}')^{-1}\mathbf{BA}\right] \\ &= \text{tr}\ \mathbf{\Sigma}(\mathbf{I} - \mathbf{P}), \end{aligned}$$

where $\mathbf{A} = \mathbf{HD}^{1/2}\mathbf{H}'$ and $\mathbf{P} = \mathbf{AB}'(\mathbf{B\Sigma B}')^{-1}\mathbf{BA}$, and the extremum is reached at $\mathbf{C} = (\mathbf{\Sigma B}')(\mathbf{B\Sigma B}')^{-1}$. Now, $\mathbf{P}$ is an orthogonal projection of rank k. From Lemma 10.1, tr $\mathbf{\Sigma}(\mathbf{I} - \mathbf{P})$ is minimized when

$$\mathbf{AB}'(\mathbf{B\Sigma B}')^{-1}\mathbf{BA} = \sum_{i=1}^{k} \mathbf{h}_i\mathbf{h}_i'$$

or

$$\mathbf{B}'(\mathbf{B\Sigma B}')^{-1}\mathbf{B} = \sum_{i=1}^{k} \lambda_i^{-1}\mathbf{h}_i\mathbf{h}_i'.$$

Finally, $\mathbf{CB} = \mathbf{\Sigma}[\mathbf{B}'(\mathbf{B\Sigma B}')^{-1}\mathbf{B}] = \sum_{i=1}^{k} \mathbf{h}_i\mathbf{h}_i' = \mathbf{P}_k$. □

Obviously, if $\boldsymbol{\mu} \neq \mathbf{0}$ in Proposition 10.1, the best approximation of rank k is $\mathbf{P}_k(\mathbf{x} - \boldsymbol{\mu}) + \boldsymbol{\mu}$, which represents the orthogonal projection of $\mathbf{x}$ on the affine subspace span$\{\mathbf{h}_1, \ldots, \mathbf{h}_k\} + \boldsymbol{\mu}$.

10.4 Sample principal components from S

The variance $\mathbf{\Sigma}$ is usually unknown. Sample principal components can be obtained from the estimate $\mathbf{S} = \mathbf{V}/m$, $m = n - 1$, where, as usual,

$$\mathbf{V} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

Since $\mathbf{S} \geq \mathbf{0}$, write

$$\mathbf{S} = \hat{\mathbf{H}} \operatorname{diag}(l_1/m, \ldots, l_p/m)\, \hat{\mathbf{H}}',$$

where

$$\hat{\mathbf{H}} = (\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_p) \in \mathbf{O}_p$$

and $l_1 \geq \cdots \geq l_p$ are the ordered eigenvalues of $\mathbf{V}$. The sample principal components of $\mathbf{x}$ are defined as $\hat{\mathbf{h}}_i'\mathbf{x}$, $i = 1, \ldots, p$.

Let $\mathcal{V} \subset \mathbb{R}^p$ be a k-dimensional subspace and denote by $\mathcal{V} + \mathbf{a} = \{\mathbf{x} + \mathbf{a} : \mathbf{x} \in \mathcal{V}\}$ the corresponding affine subspace. What is the affine subspace $\mathcal{V}+\mathbf{a}$ of dimension k such that the orthogonal projection of the data on $\mathcal{V} + \mathbf{a}$ is "closest" to the original data? First, we must specify what is meant by "closest." As a measure of distance, take the usual euclidian distance

$$d(\mathcal{V}, \mathbf{a}) = \sum_{i=1}^{n} |\mathbf{x}_i - \hat{\mathbf{x}}_i|^2,$$

where $\hat{\mathbf{x}}_i = \hat{\mathbf{P}}(\mathbf{x}_i - \mathbf{a}) + \mathbf{a}$ is the orthogonal projection of $\mathbf{x}_i$ on $\mathcal{V} + \mathbf{a}$.

Proposition 10.2 *Among all k-dimensional subspaces $\mathcal{V}$ and vectors $\mathbf{a} \in \mathbb{R}^p$, the distance $d(\mathcal{V}, \mathbf{a})$ is minimized for $\mathbf{a} = \bar{\mathbf{x}}$ and $\mathcal{V} = \operatorname{span}\{\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_k\}$.*

Proof. Define $\mathbf{V} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, $\hat{\mathbf{P}}$ the orthogonal projection on $\mathcal{V}$, and $\hat{\mathbf{Q}} = \mathbf{I} - \hat{\mathbf{P}}$. Then,

$$\begin{aligned}
d(\mathcal{V}, \mathbf{a}) &= \sum_{i=1}^{n} |\mathbf{x}_i - \hat{\mathbf{P}}(\mathbf{x}_i - \mathbf{a}) - \mathbf{a}|^2 \\
&= \sum_{i=1}^{n} |\hat{\mathbf{Q}}\mathbf{x}_i - \hat{\mathbf{Q}}\mathbf{a}|^2 \\
&= \sum_{i=1}^{n} |(\hat{\mathbf{Q}}\mathbf{x}_i - \hat{\mathbf{Q}}\bar{\mathbf{x}}) + (\hat{\mathbf{Q}}\bar{\mathbf{x}} - \hat{\mathbf{Q}}\mathbf{a})|^2 \\
&= \sum_{i=1}^{n} |\hat{\mathbf{Q}}\mathbf{x}_i - \hat{\mathbf{Q}}\bar{\mathbf{x}}|^2 + \sum_{i=1}^{n} |\hat{\mathbf{Q}}\bar{\mathbf{x}} - \hat{\mathbf{Q}}\mathbf{a}|^2 \\
&= \operatorname{tr} \hat{\mathbf{Q}}\mathbf{V} + n(\bar{\mathbf{x}} - \mathbf{a})'\hat{\mathbf{Q}}(\bar{\mathbf{x}} - \mathbf{a}).
\end{aligned}$$

The two terms in the last expression are non-negative; hence, $\mathbf{a} = \bar{\mathbf{x}}$. Also, from Lemma 10.1 and since $\mathbf{V} \propto \mathbf{S}$, $\hat{\mathbf{P}} = \sum_{i=1}^{k} \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i'$. □

The ratio $f(\boldsymbol{\lambda}) = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$ of total variance explained by the first k principal components is estimated by $f(\mathbf{l}/m) = \sum_{i=1}^{k} l_i / \sum_{i=1}^{p} l_i$. A large sample $(1 - \alpha) \times 100\%$ confidence interval on this ratio $f(\boldsymbol{\lambda})$, when all population eigenvalues λ_α are distinct, can be constructed with Proposition 8.18.

We end this section with a word of caution: Principal components are not invariant with respect to individual rescaling of the p variables in $\mathbf{x}$; that is, if $\mathbf{w} = \mathbf{\Phi x}$, where $\mathbf{\Phi} = \text{diag}(\phi_1, \ldots, \phi_p)$, then $\mathbf{\Phi \Sigma \Phi}$ does not have the same eigen-structure as $\mathbf{\Sigma}$. This means, for example, that the interesting projections of the data found with Proposition 10.2 may look entirely different after rescaling. Also, if the first variable x_1 has a variance much larger than the variances of all other variables, $x_2, \ldots, x_p$, then the first principal component y_1 will be approximately equivalent to x_1. Principal components are thus most meaningful when all variables are measured in the same units and have variances of the same magnitude. For this reason, principal components are often calculated from the sample correlation matrix $\mathbf{R}$ rather than the sample variance $\mathbf{S}$.

10.5 Sample principal components from $\mathbf{R}$

If we let

$$\begin{aligned} \mathbf{S}_0 &= \text{diag}(s_{11}, \ldots, s_{pp}), \\ \mathbf{\Sigma}_0 &= \text{diag}(\sigma_{11}, \ldots, \sigma_{pp}), \end{aligned}$$

then the population and sample correlation matrices are given by

$$\begin{aligned} \mathbf{R} &= \mathbf{S}_0^{-1/2} \mathbf{S} \mathbf{S}_0^{-1/2}, \\ \boldsymbol{\rho} &= \mathbf{\Sigma}_0^{-1/2} \mathbf{\Sigma} \mathbf{\Sigma}_0^{-1/2}. \end{aligned}$$

Then, as in the previous section, we can decompose

$$\begin{aligned} \boldsymbol{\rho} &= \mathbf{G}\ \text{diag}(\gamma_1, \ldots, \gamma_p)\ \mathbf{G}', \\ \mathbf{R} &= \hat{\mathbf{G}}\ \text{diag}(f_1, \ldots, f_p)\ \hat{\mathbf{G}}', \end{aligned}$$

and define the sample principal components from the standardized variables, $\mathbf{z} = \mathbf{S}_0^{-1/2}(\mathbf{x} - \bar{\mathbf{x}})$, and $\mathbf{R}$ as $\hat{\mathbf{g}}_i' \mathbf{z}$, $i = 1, \ldots, p$, where $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$ and similarly for $\hat{\mathbf{G}}$. The ratio of total variance (of the standardized variables $z_i = (x_i - \bar{x}_i)/\sqrt{s_{ii}}$) explained by the first k principal components becomes

$$f(\boldsymbol{\gamma}) = \sum_{i=1}^{k} \gamma_i / p.$$

The construction of a confidence interval on this ratio $f(\boldsymbol{\gamma})$ thus necessitates the asymptotic distribution of the eigenvalues f_i of the sample correlation matrix $\mathbf{R}$. This is now derived using the perturbation method of Section 8.8.

Using the Taylor series

$$x^{-1/2} = a^{-1/2} - \tfrac{1}{2} a^{-3/2}(x - a) + \cdots,$$

we have directly

$$\mathbf{S}_0^{-1/2} = [\mathbf{I} - \tfrac{1}{2}\boldsymbol{\Sigma}_0^{-1/2}(\mathbf{S}_0 - \boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}_0^{-1/2} + \cdots]\boldsymbol{\Sigma}_0^{-1/2}.$$

Define

$$\begin{aligned} \mathbf{V} &= (v_{ij}) = n^{1/2}(\boldsymbol{\Sigma}_0^{-1/2}\mathbf{S}\boldsymbol{\Sigma}_0^{-1/2} - \boldsymbol{\rho}), \\ \mathbf{V}_0 &= \operatorname{diag}(v_{11}, \ldots, v_{pp}), \end{aligned}$$

and note that $\mathbf{V}$ is $O_p(1)$. Then, we can write

$$\begin{aligned} \mathbf{R} &= [\mathbf{I} - \tfrac{1}{2}n^{-1/2}\mathbf{V}_0 + \cdots]\boldsymbol{\Sigma}_0^{-1/2}[\boldsymbol{\Sigma} + (\mathbf{S} - \boldsymbol{\Sigma})]\boldsymbol{\Sigma}_0^{-1/2}[\mathbf{I} - \tfrac{1}{2}n^{-1/2}\mathbf{V}_0 + \cdots] \\ &= [\mathbf{I} - \tfrac{1}{2}n^{-1/2}\mathbf{V}_0 + \cdots](\boldsymbol{\rho} + n^{-1/2}\mathbf{V})[\mathbf{I} - \tfrac{1}{2}n^{-1/2}\mathbf{V}_0 + \cdots] \\ &= \boldsymbol{\rho} + n^{-1/2}(\mathbf{V} - \tfrac{1}{2}\boldsymbol{\rho}\mathbf{V}_0 - \tfrac{1}{2}\mathbf{V}_0\boldsymbol{\rho}) + O_p(n^{-1}), \end{aligned}$$

from which

$$\mathbf{G}'\mathbf{R}\mathbf{G} = \boldsymbol{\Gamma} + n^{-1/2}\mathbf{V}^{(1)} + O_p(n^{-1}),$$

where

$$\begin{aligned} \boldsymbol{\Gamma} &= \operatorname{diag}(\gamma_1, \ldots, \gamma_p), \\ \mathbf{V}^{(1)} &= (v_{ij}^{(1)}) = \mathbf{G}'(\mathbf{V} - \tfrac{1}{2}\boldsymbol{\rho}\mathbf{V}_0 - \tfrac{1}{2}\mathbf{V}_0\boldsymbol{\rho})\mathbf{G}. \end{aligned} \tag{10.1}$$

Equation (8.11) in the perturbation method then leads, assuming γ_α to be a distinct eigenvalue, to the expansion

$$f_\alpha = \gamma_\alpha + n^{-1/2}v_{\alpha\alpha}^{(1)} + O_p(n^{-1}),$$

or, in vector form, assuming all eigenvalues γ_α to be distinct, to the expansion

$$\begin{aligned} n^{1/2}(\mathbf{f} - \boldsymbol{\gamma}) &= (v_{11}^{(1)}, \ldots, v_{pp}^{(1)})' + O_p(n^{-1/2}) \\ &\equiv \mathbf{v}^{(1)} + O_p(n^{-1/2}). \end{aligned}$$

Now, since $\mathbf{V}$ is asymptotically normal with mean $\mathbf{0}$, so is $\mathbf{V}^{(1)}$ and its marginal $\mathbf{v}^{(1)}$. We need only calculate the asymptotic variance of $\mathbf{v}^{(1)}$.

From (10.1) and the relation $\boldsymbol{\rho}\mathbf{G} = \mathbf{G}\boldsymbol{\Gamma}$, we have

$$\begin{aligned} v_{\alpha\alpha}^{(1)} &= \mathbf{g}_\alpha'\mathbf{V}\mathbf{g}_\alpha - \gamma_\alpha\mathbf{g}_\alpha'\mathbf{V}_0\mathbf{g}_\alpha \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p} g_{j\alpha}g_{k\alpha}v_{jk} - \gamma_\alpha\sum_{j=1}^{p} g_{j\alpha}^2 v_{jj}; \end{aligned}$$

hence,

$$\begin{aligned} \operatorname{cov}(v_{\alpha\alpha}^{(1)}, v_{\beta\beta}^{(1)}) &= \sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{p}\sum_{l=1}^{p} g_{j\alpha}g_{k\alpha}g_{i\beta}g_{l\beta}\operatorname{cov}(v_{jk}, v_{il}) \\ &\quad + \gamma_\alpha\gamma_\beta\sum_{j=1}^{p}\sum_{i=1}^{p} g_{j\alpha}^2 g_{i\beta}^2 \operatorname{cov}(v_{jj}, v_{ii}) \end{aligned}$$

$$-\gamma_\alpha \sum_{j=1}^{p}\sum_{i=1}^{p}\sum_{l=1}^{p} g_{j\alpha}^2 g_{i\beta} g_{l\beta}\ \mathrm{cov}(v_{jj}, v_{il})$$
$$-\gamma_\beta \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} g_{i\beta}^2 g_{j\alpha} g_{k\alpha}\ \mathrm{cov}(v_{ii}, v_{jk}).$$

Since $\mathbf{V} \xrightarrow{d} N_p^p(\mathbf{0}, (\mathbf{I}+\mathbf{K}_p)(\boldsymbol{\rho}\otimes\boldsymbol{\rho}))$, we find upon using (6.1) that

$$\begin{aligned}\lim_{n\to\infty} \mathrm{cov}(v_{\alpha\alpha}^{(1)}, v_{\beta\beta}^{(1)}) &= \sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{p}\sum_{l=1}^{p} g_{j\alpha} g_{k\alpha} g_{i\beta} g_{l\beta}(\rho_{kl}\rho_{ji} + \rho_{jl}\rho_{ki}) \\ &\quad +\gamma_\alpha\gamma_\beta \sum_{j=1}^{p}\sum_{i=1}^{p} g_{j\alpha}^2 g_{i\beta}^2 2\rho_{ji}^2 \\ &\quad -\gamma_\alpha \sum_{j=1}^{p}\sum_{i=1}^{p}\sum_{l=1}^{p} g_{j\alpha}^2 g_{i\beta} g_{l\beta} 2\rho_{jl}\rho_{ji} \\ &\quad -\gamma_\beta \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} g_{i\beta}^2 g_{j\alpha} g_{k\alpha} 2\rho_{ik}\rho_{ij}.\end{aligned}$$

Finally, with the simple relations

$$\begin{aligned}\sum_{k=1}^{p}\sum_{l=1}^{p} g_{k\alpha} g_{l\beta}\rho_{kl} &= \gamma_\alpha \delta_{\alpha\beta}, \\ \sum_{l=1}^{p} g_{l\beta}\rho_{jl} &= \gamma_\beta g_{j\beta},\end{aligned}$$

we obtain the simplification

$$\lim_{n\to\infty} \mathrm{cov}(v_{\alpha\alpha}^{(1)}, v_{\beta\beta}^{(1)}) = 2\gamma_\alpha\gamma_\beta \left[\delta_{\alpha\beta} - (\gamma_\alpha+\gamma_\beta)\sum_{j=1}^{p} g_{j\alpha}^2 g_{j\beta}^2 \right.$$
$$\left. + \sum_{j=1}^{p}\sum_{i=1}^{p} g_{j\alpha}^2 g_{i\beta}^2 \right.$$

We summarize the result.

Proposition 10.3 *Let* $\mathbf{f} = (f_1, \ldots, f_p)'$ *be the eigenvalues of the sample correlation matrix* $\mathbf{R}$. *If the eigenvalues* γ_α *of the population correlation matrix* $\boldsymbol{\rho}$ *are all distinct, then the joint limiting distribution is*

$$n^{1/2}(\mathbf{f}-\boldsymbol{\gamma}) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega} = (\omega_{\alpha\beta})$ *is given by*

$$\omega_{\alpha\beta} = 2\gamma_\alpha\gamma_\beta \left[\delta_{\alpha\beta} - (\gamma_\alpha + \gamma_\beta) \sum_{j=1}^{p} g_{j\alpha}^2 g_{j\beta}^2 + \sum_{j=1}^{p}\sum_{i=1}^{p} g_{j\alpha}^2 g_{i\beta}^2 \rho_{ji}^2 \right].$$

The limiting distribution of a function such as $f(\mathbf{f}) = \sum_{i=1}^{k} f_i/p$ for the ratio of total variance explained by the first k principal components is easily derived by the delta method [v. Problem 10.7.5]. Problem 13.6.19 provides the asymptotic distribution of $n^{1/2}(\mathbf{f} - \boldsymbol{\gamma})$ when sampling from an elliptical distribution. Konishi (1979) obtained, with Sugiura's lemma, a more accurate approximation with remainder $O(n^{-1})$, similar to that of Proposition 8.18, for the distribution function of

$$s = (n-1)^{1/2} \left(f(\mathbf{f}) - f(\boldsymbol{\gamma}) \right),$$

where $f(\cdot)$ is a continuously differentiable function in a neighborhood of $\boldsymbol{\gamma}$.

10.6 A test for multivariate normality

Shapiro and Wilk's (1965) W statistic has been found to be the best omnibus test for detecting departures from univariate normality. Royston (1983) extends the application of W to testing multivariate normality, but the procedure involves a certain approximation which needs to be justified. The procedure of Srivastava and Hui (1987) does not require such an approximation and has a simple asymptotic null distribution and the calculations are straightforward.

Srivastava and Hui (1987) proposed two test statistics for testing multivariate normality. These are based on principal components and may be considered as a generalization of the Shapiro-Wilk statistic. As in Section 10.4, write

$$\mathbf{S} = \hat{\mathbf{H}} \operatorname{diag}(l_1/m, \ldots, l_p/m)\, \hat{\mathbf{H}}', \; m = n-1,$$

where

$$\hat{\mathbf{H}} = (\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_p) \in \mathbf{O}_p.$$

The sample principal components of $\mathbf{x}_j$, $j = 1, \ldots, n$, are defined as $y_{ij} = \hat{\mathbf{h}}_i' \mathbf{x}_j$, $i = 1, \ldots, p$, $j = 1, \ldots, n$. Thus, under the null hypothesis of multivariate normality, we can treat $y_{i1}, \ldots, y_{in}$, $i = 1, \ldots, p$, as p approximately independent samples. For sample i, the univariate Shapiro-Wilk statistic is defined as

$$W(i) = \frac{m}{n l_i} \left[\sum_{j=1}^{n} a_j y_{i(j)} \right]^2, \; i = 1, \ldots, p,$$

where a_j's are the constants tabulated in Shapiro and Wilk (1965) and

$$y_{i(1)} \leq y_{i(2)} \leq \cdots \leq y_{i(n)}$$

are the ordered values of $y_{i1}, \ldots, y_{in}$. For $n > 50$, the values of a_j are given by Shapiro and Francia (1972) and up to 2000 by Royston (1982).

From Shapiro and Wilk (1968), we note that for each i, $W(i)$ can be transformed to an approximate standard normal variable $G(W(i))$ by using Johnson's (1949) S_B system,

$$G(W(i)) = \gamma + \delta \ln \left[\frac{W(i) - \epsilon}{1 - W(i)} \right],$$

where γ, δ, and ϵ can be found in Table 1 of Shapiro and Wilk (1968) up to $n = 50$. For $n > 50$, values of γ, δ, and ϵ can be obtained with the help of the results in Shapiro and Francia (1972) and Royston (1982). Let

$$M_1 = -2 \sum_{i=1}^{p} \ln \left[\Phi \left(G(W(i)) \right) \right],$$

where $\Phi(\cdot)$ is the distribution of a standard normal variable. Note that if $U \sim \text{unif}(0,1)$, then $-2 \ln U \sim \chi^2_2$. Srivastava and Hui (1987) proposed M_1 as their first test statistic for testing multivariate normality, where M_1 is approximately distributed as χ^2_{2p} under the hypothesis of multivariate normality. Large values of M_1 will indicate non-normality.

Next, they observed that small values of $W(i)$ indicate a departure from normality for variate i. Thus, they considered the minimum of all components and proposed

$$M_2 = \min_{1 \leq i \leq p} W(i)$$

as the second test statistic. The null distribution of M_2 is approximately given by

$$P(M_2 \leq t) = 1 - \left[1 - \Phi \left(G(t) \right) \right]^p. \tag{10.2}$$

For p =2, 4, and 6 and n =10, 25, and 50, a simulation study [Srivastava and Hui (1987)] found that the null distribution of both M_1 and M_2 are well approximated by χ^2_{2p} and (10.2), respectively. Examples of the use of M_1 and M_2 on data sets are provided by Looney (1995) with the necessary SAS procedures or FORTRAN subroutines.

Most tests for multivariate normality are functions of the squared radii (or squared Mahalanobis distances of $\mathbf{x}_i$ to $\bar{\mathbf{x}}$),

$$d_i^2 = |\mathbf{z}_i|^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \ i = 1, \ldots, n.$$

Some graphical procedures [Andrews et al. (1973), Cox and Small (1978), Gnanadesikan and Kettenring (1972)] are based on d_i^2. One such Q-Q plot is described in Section 11.4.1. Malkovich and Afifi (1973) considered

the supremum of the standardized skewness and kurtosis over all linear combinations $\mathbf{t}'\mathbf{x}$,

$$\begin{aligned}\beta_1^M &= \max_{\mathbf{t}\in S^{p-1}} \frac{\{E\ (\mathbf{t}'\mathbf{x}-\mathbf{t}'\boldsymbol{\mu})^3\}^2}{(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})^3},\\ \beta_2^M &= \max_{\mathbf{t}\in S^{p-1}} \left|\frac{E\ (\mathbf{t}'\mathbf{x}-\mathbf{t}'\boldsymbol{\mu})^4}{(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})^2}-3\right|.\end{aligned}$$

The tests are based on the sample versions

$$\begin{aligned}\beta_{1,n}^M &= \max_{\mathbf{t}\in S^{p-1}} b_{1,n}(\mathbf{t}),\\ \beta_{2,n}^M &= \max_{\mathbf{t}\in S^{p-1}} |b_{2,n}(\mathbf{t})-3|,\end{aligned}$$

respectively, where

$$\begin{aligned}b_{1,n}(\mathbf{t}) &= \frac{\{\frac{1}{n}\sum_{i=1}^n(\mathbf{t}'\mathbf{x}_i-\mathbf{t}'\bar{\mathbf{x}})^3\}^2}{(\mathbf{t}'\mathbf{S}\mathbf{t})^3},\\ b_{2,n}(\mathbf{t}) &= \frac{\frac{1}{n}\sum_{i=1}^n(\mathbf{t}'\mathbf{x}_i-\mathbf{t}'\bar{\mathbf{x}})^4}{(\mathbf{t}'\mathbf{S}\mathbf{t})^2}.\end{aligned}$$

Mardia's kurtosis test [Mardia (1970)] is a function of d_i^2 and his skewness test is a function of the scaled residuals

$$\mathbf{z}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i-\bar{\mathbf{x}}),\ i=1,\ldots,n.$$

Mardia's measures of multivariate skewness and kurtosis are

$$\begin{aligned}B_{1,n} &= \frac{1}{n^2}\sum_{i,j=1}^n\{(\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j-\bar{\mathbf{x}})\}^3,\\ B_{2,n} &= \frac{1}{n}\sum_{i=1}^n\{(\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\}^2,\end{aligned}$$

respectively. The tests of multivariate normality based on multivariate skewness, $\beta_{1,n}^M$ and $B_{1,n}$, are inconsistent against each fixed non-normal elliptical distribution [Baringhaus and Henze (1991)]. However, the tests based on multivariate kurtosis, $\beta_{2,n}^M$ and $B_{2,n}$, are consistent. An approximation formula of the power of the test $\beta_{2,n}^M$ against elliptically symmetric distributions was derived by Naito (1998). Cox and Small (1978) proposed tests based on linearity of regression rather than directly on normality. An omnibus test based on empirical characteristic function of the scaled residuals was also proposed [Henze and Zirkler (1990), v. also Henze and Wagner (1997)]. Goodness-of-fit tests for a general multivariate distribution by the empirical characteristic function was treated by Fan (1997). A characterization of multivariate normality by hermitian polynomials was recently proposed by Kariya et al. (1997) to build an omnibus test. A comparative study of goodness-of-fit tests for multivariate normality was carried out by Romeu and Ozturk (1993).

10.7 Problems

1. In morphometric studies, it is often the case that all variables are positively correlated. Prove that if $\mathbf{\Sigma}$ has all positive covariances, $\sigma_{ij} > 0$ for $i \neq j$, then all the coefficients in $\mathbf{h}_1$ of the first principal component may be taken non-negative.

2. For $\mathbf{\Sigma} \geq \mathbf{0}$ in $\mathbb{R}_p^p$ with spectral decomposition $\mathbf{\Sigma} = \mathbf{HDH}'$ as in Section 10.2, prove that $\mathbf{\Theta} = \sum_{i=1}^k \lambda_i \mathbf{h}_i \mathbf{h}_i'$ is the matrix of rank k such that

$$||\mathbf{\Sigma} - \mathbf{\Theta}||^2 = \sum_{i=1}^p \sum_{j=1}^p (\sigma_{ij} - \theta_{ij})^2$$

is minimum.
Hint: $||\mathbf{\Sigma} - \mathbf{\Theta}||^2 = \text{tr}(\mathbf{D} - \mathbf{E})(\mathbf{D} - \mathbf{E})'$, where $\mathbf{E} = \mathbf{H}'\mathbf{\Theta}\mathbf{H}$.

3. Assume $\mathbf{x} \in \mathbb{R}^p$ has density

$$f_{\mathbf{x}}(\mathbf{x}) = |\mathbf{\Lambda}|^{-1/2} g[(\mathbf{x} - \mu\mathbf{1})'\mathbf{\Lambda}^{-1}(\mathbf{x} - \mu\mathbf{1})],$$

where $\mathbf{\Lambda} = \sigma^2[(1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}']$. Prove there exists $\mathbf{H} \in \mathbf{O}_p$ such that $\mathbf{y} = \mathbf{Hx}$ has density

$$f_{\mathbf{y}}(\mathbf{y}) = \lambda_1^{-1/2} \lambda_2^{-(p-1)/2} g\left[\frac{(y_1 - p^{1/2}\mu)^2}{\lambda_1} + \frac{\sum_{i=2}^p y_i^2}{\lambda_2}\right].$$

4. **Parent-child interclass correlation** [Srivastava (1984)].
Assume $\mathbf{x} \in \mathbb{R}^{p+1}$ has density

$$f_{\mathbf{x}}(\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2} \exp[(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})],$$

where

$$\begin{aligned} \boldsymbol{\mu} &= \begin{pmatrix} \mu_m \\ \mu_s \mathbf{1} \end{pmatrix}, \\ \mathbf{\Sigma} &= \begin{pmatrix} \sigma_m^2 & \sigma_{ms}\mathbf{1}' \\ \sigma_{ms}\mathbf{1} & \sigma_s^2[(1-\rho_{ss})\mathbf{I} + \rho_{ss}\mathbf{1}\mathbf{1}'] \end{pmatrix}. \end{aligned}$$

Here, "m" stands for mother and "s" means siblings. Let $\mathbf{A}' = (\mathbf{1}/p, \mathbf{\Gamma}') \in \mathbb{R}^p$ for some $\mathbf{\Gamma}$ satisfying $\mathbf{\Gamma 1} = \mathbf{0}$ and $\mathbf{\Gamma\Gamma}' = \mathbf{I}_{p-1}$.

(i) Interpret the parameters $(\mu_m, \mu_s, \sigma_m^2, \sigma_{ms}, \rho_{ss})$.
(ii) Prove that if

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{A} \end{pmatrix},$$

then $\tilde{\mathbf{A}}\mathbf{\Sigma}\tilde{\mathbf{A}}' = \text{diag}(\mathbf{\Omega}, \gamma_s^2 \mathbf{I}_{p-1})$, where

$$\mathbf{\Omega} = \begin{pmatrix} \sigma_m^2 & \sigma_{ms} \\ \sigma_{ms} & \eta^2 \end{pmatrix},$$

$$\begin{aligned}\gamma_s^2 &= \sigma_s^2(1-\rho_{ss}),\\ \eta^2 &= [1+(p-1)\rho_{ss}]\sigma_s^2/p.\end{aligned}$$

(iii) Deduce that $\mathbf{y} = \tilde{\mathbf{A}}\mathbf{x}$ is such that $(y_1, y_2)' \sim N_2\left((\mu_m, \mu_s)', \mathbf{\Omega}\right)$, $y_i \sim N(0, \gamma_s^2)$ $(i = 3, \ldots, p+1)$, and $(y_1, y_2) \perp\!\!\!\perp (y_3, \ldots, y_{p+1})$.

(iv) What are the implications for maximum likelihood estimation of the unknown parameters in i)?

Remark: The y_i's are not the principal components but are closely related to the concept.

5. Let $\mathbf{f} = (f_1, \ldots, f_p)'$ be the eigenvalues of the sample correlation matrix $\mathbf{R}$. If the eigenvalues γ_α of the population correlation matrix $\boldsymbol{\rho}$ are all distinct, then find the limiting distribution of $\sum_{i=1}^k f_i/p$ for the ratio of total variance explained by the first k principal components.

11
Canonical correlations

11.1 Introduction

The objective of canonical correlation analysis is to get a simple description of the structure of correlation between subsets of variables. Assume that two subsets of variables $\mathbf{x}_1$ and $\mathbf{x}_2$ have a joint normal distribution,

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

The analysis searches for a pair of linear combinations $\mathbf{t}_1'\mathbf{x}_1$ and $\mathbf{t}_2'\mathbf{x}_2$ with maximum correlation. This is the first canonical correlation. Having found such a pair, the analysis is pursued one step further by searching for a second pair of linear combinations with maximum correlation among all those uncorrelated with the first pair. The correlation found is the second canonical correlation. The argument is repeated until all possible correlations are exhausted. This analysis is explained in detail in Section 11.2. In Section 11.3, tests of independence between $\mathbf{x}_1$ and $\mathbf{x}_2$ are derived. Not surprisingly, the tests proposed will be functions of the sample canonical correlations. Section 11.4 uses advantageously the context of testing independence to derive simple proofs of the properties of $U(p; m, n)$ distributions introduced earlier in Section 9.3.2. As a by-product we also obtain a method of constructing Q-Q plots of squared radii for a visual inspection of multivariate normality. Asymptotic distributions of sample canonical correlations is the subject of Section 11.5.

11.2 Definition and basic properties

Assume $\mathbf{\Sigma}_{jj} > \mathbf{0}$, $\mathbf{\Sigma}_{ij} \in \mathbb{R}^{p_i}_{p_j}$, $i,j = 1,2$. Without any loss of generality, suppose $p_1 \leq p_2$. Write $\mathbf{\Sigma}_{jj} = \mathbf{A}_j^2$, where $\mathbf{A}_j > \mathbf{0}$, $j = 1,2$. Now using the SVD (v. Proposition 1.11), we have

$$\mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1} = \mathbf{G}(\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0})\mathbf{H}',$$

where $\mathbf{D}_{\boldsymbol{\rho}} = \operatorname{diag}(\rho_1, \ldots, \rho_{p_1})$, $\rho_1 \geq \cdots \geq \rho_{p_1} \geq 0$,

$$\begin{aligned} \mathbf{G} &= (\mathbf{g}_1, \ldots, \mathbf{g}_{p_1}) \in \mathbf{O}_{p_1}, \\ \mathbf{H} &= (\mathbf{h}_1, \ldots, \mathbf{h}_{p_2}) \in \mathbf{O}_{p_2}. \end{aligned}$$

If we define

$$\begin{aligned} \mathbf{u} &= \mathbf{G}'\mathbf{A}_1^{-1}\mathbf{x}_1 = (u_1, \ldots, u_{p_1})', \\ \mathbf{v} &= \mathbf{H}'\mathbf{A}_2^{-1}\mathbf{x}_2 = (v_1, \ldots, v_{p_2})', \end{aligned}$$

then

$$\operatorname{var}\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{p_1} & (\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0}) \\ \begin{pmatrix} \mathbf{D}_{\boldsymbol{\rho}} \\ \mathbf{0} \end{pmatrix} & \mathbf{I}_{p_2} \end{pmatrix}.$$

Obviously, $\operatorname{var} u_i = \operatorname{var} v_j = 1$ and $\operatorname{cor}(u_i, v_j) = \rho_i \delta_{ij}$, $i = 1, \ldots, p_1$, $j = 1, \ldots, p_2$.

Definition 11.1 *The variables $u_1, \ldots, u_{p_1}$ and $v_1, \ldots, v_{p_2}$ are defined to be the* canonical variables. *The numbers ρ_i's, $1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_{p_1} \geq 0$, are the* canonical correlations.

Note that the number of nonzero canonical correlations is $\operatorname{rank} \mathbf{\Sigma}_{12} \equiv c$. In a similar manner as the principal components were interpreted, the canonical variables can also be derived sequentially.

First, we seek linear combinations $\mathbf{t}_1'\mathbf{x}_1$ and $\mathbf{t}_2'\mathbf{x}_2$ such that $\operatorname{cor}(\mathbf{t}_1'\mathbf{x}_1, \mathbf{t}_2'\mathbf{x}_2)$ is maximal. But, in general, since $\operatorname{cor}(x, y)$ is invariant with respect to linear transformations, $x \mapsto ax + b$, $y \mapsto cy + d$, $a, c > 0$, we may assume at the outset that $\operatorname{var} \mathbf{t}_j'\mathbf{x}_j = \mathbf{t}_j'\mathbf{\Sigma}_{jj}\mathbf{t}_j = 1$, $j = 1,2$. Introducing the ellipsoids

$$\mathcal{E}_j = \{\mathbf{t}_j : \ \mathbf{t}_j'\mathbf{\Sigma}_{jj}\mathbf{t}_j = 1\}, \ j = 1,2,$$

the problem is thus

$$\max_{\substack{\mathbf{t}_1 \in \mathcal{E}_1 \\ \mathbf{t}_2 \in \mathcal{E}_2}} \mathbf{t}_1'\mathbf{\Sigma}_{12}\mathbf{t}_2.$$

For $\mathbf{t}_j \in \mathcal{E}_j$, $|\mathbf{A}_j\mathbf{t}_j| = 1$, $j = 1,2$, the Cauchy-Schwarz inequality gives

$$\begin{aligned} (\mathbf{t}_1'\mathbf{\Sigma}_{12}\mathbf{t}_2)^2 &= \langle \mathbf{A}_1\mathbf{t}_1, \mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1}\mathbf{h} \rangle^2 \\ &\leq |\mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1}\mathbf{h}|^2, \end{aligned}$$

where $\mathbf{h} = \mathbf{A}_2\mathbf{t}_2$ has norm 1. Letting $\mathbf{B} = \mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1}$, then $|\mathbf{Bh}|^2 = \mathbf{h}'\mathbf{B}'\mathbf{Bh}$, where

$$\mathbf{B}'\mathbf{B} = (\mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1})'(\mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1}) = \mathbf{H}\begin{pmatrix}\mathbf{D}_{\boldsymbol{\rho}^2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{H}'.$$

Thus, from the method used for principal components, we find $\mathbf{h}'\mathbf{B}'\mathbf{Bh} \leq \rho_1^2$ with equality when $\mathbf{h} = \mathbf{h}_1$. This gives $\mathbf{t}_2'\mathbf{x}_2 = \mathbf{h}_1'\mathbf{A}_2^{-1}\mathbf{x}_2 = v_1$. Finally, the Cauchy-Schwarz inequality is, in fact, an equality iff $\mathbf{A}_1\mathbf{t}_1 \propto \mathbf{Bh}_1$ or, equivalently,

$$\begin{aligned}\mathbf{t}_1 &\propto \mathbf{A}_1^{-1}\mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1}\mathbf{h}_1\\ &= \mathbf{A}_1^{-1}\mathbf{G}(\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0})\mathbf{H}'\mathbf{h}_1\\ &= \mathbf{A}_1^{-1}\mathbf{G}(\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0})\mathbf{e}_1\\ &= \rho_1\mathbf{A}_1^{-1}\mathbf{g}_1,\end{aligned}$$

which, in turn, gives $\mathbf{t}_1'\mathbf{x}_1 = \mathbf{g}_1'\mathbf{A}_1^{-1}\mathbf{x}_1 = u_1$. We have proved that (u_1, v_1) is the pair of linear combinations with maximum correlation ρ_1.

Second, having found pairs of linear combinations

$$(u_i, v_i) = (\mathbf{g}_i'\mathbf{A}_1^{-1}\mathbf{x}_1, \mathbf{h}_i'\mathbf{A}_2^{-1}\mathbf{x}_2),\ i = 1, \ldots, k,\ k < \text{rank } \mathbf{\Sigma}_{12} \equiv c,$$

another pair $(\mathbf{t}_1'\mathbf{x}_1, \mathbf{t}_2'\mathbf{x}_2)$ is sought with maximum correlation among all those uncorrelated with the preceding pairs; i.e., the restriction

$$\text{cov}(\mathbf{t}_j'\mathbf{x}_j, u_i) = \text{cov}(\mathbf{t}_j'\mathbf{x}_j, v_i) = 0,\ i = 1, \ldots, k;\ j = 1, 2,$$

is imposed. This last restriction is characterized in terms of orthogonality:

$$\text{cov}(\mathbf{t}_1'\mathbf{x}_1, u_i) = \mathbf{t}_1'\mathbf{\Sigma}_{11}\mathbf{A}_1^{-1}\mathbf{g}_i = \mathbf{t}_1'\mathbf{A}_1\mathbf{g}_i = 0 \Longleftrightarrow \mathbf{t}_1 \perp \mathbf{A}_1\mathbf{g}_i.$$

Similarly, $\text{cov}(\mathbf{t}_2'\mathbf{x}_2, v_i) = 0$ iff $\mathbf{t}_2 \perp \mathbf{A}_2\mathbf{h}_i$. We note that when $\mathbf{t}_1 \perp \mathbf{A}_1\mathbf{g}_i$, the other condition, $\text{cov}(\mathbf{t}_1'\mathbf{x}_1, v_i) = 0$, is automatically satisfied:

$$\begin{aligned}\text{cov}(\mathbf{t}_1'\mathbf{x}_1, v_i) &= \mathbf{t}_1'\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1}\mathbf{h}_i\\ &= \mathbf{t}_1'\mathbf{A}_1(\mathbf{A}_1^{-1}\mathbf{\Sigma}_{12}\mathbf{A}_2^{-1})\mathbf{h}_i\\ &= \mathbf{t}_1'\mathbf{A}_1\mathbf{G}(\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0})\mathbf{H}'\mathbf{h}_i\\ &= \rho_i\mathbf{t}_1'\mathbf{A}_1\mathbf{g}_i = 0.\end{aligned}$$

Similarly, when $\mathbf{t}_2 \perp \mathbf{A}_2\mathbf{h}_i$ then $\text{cov}(\mathbf{t}_2'\mathbf{x}_2, u_i) = 0$ is automatically satisfied. So the problem becomes

$$\max_{\substack{\mathbf{t}_1 \in \mathcal{E}_1^{\perp}\\ \mathbf{t}_2 \in \mathcal{E}_2^{\perp}}} \mathbf{t}_1'\mathbf{\Sigma}_{12}\mathbf{t}_2,$$

where

$$\begin{aligned}\mathcal{E}_1^{\perp} &= \{\mathbf{t}_1 \in \mathcal{E}_1 :\ \mathbf{t}_1 \perp \mathbf{A}_1\mathbf{g}_1, \ldots, \mathbf{A}_1\mathbf{g}_k\},\\ \mathcal{E}_2^{\perp} &= \{\mathbf{t}_2 \in \mathcal{E}_2 :\ \mathbf{t}_2 \perp \mathbf{A}_2\mathbf{h}_1, \ldots, \mathbf{A}_2\mathbf{h}_k\}.\end{aligned}$$

The Cauchy-Schwarz inequality gives for $\mathbf{t}_j \in \mathcal{E}_j^{\perp}$,

$$(\mathbf{t}_1'\boldsymbol{\Sigma}_{12}\mathbf{t}_2)^2 = \langle \mathbf{A}_1\mathbf{t}_1, \mathbf{B}\mathbf{h}\rangle^2 \leq \mathbf{h}'\mathbf{B}'\mathbf{B}\mathbf{h},$$

where, as before, $\mathbf{B} = \mathbf{A}_1^{-1}\boldsymbol{\Sigma}_{12}\mathbf{A}_2^{-1}$ and $\mathbf{h} = \mathbf{A}_2\mathbf{t}_2$. But, using the orthogonality restrictions,

$$\begin{aligned}\mathbf{h}'\mathbf{B}'\mathbf{B}\mathbf{h} &= \mathbf{t}_2'\mathbf{A}_2\mathbf{H}\begin{pmatrix}\mathbf{D}_{\boldsymbol{\rho}^2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{H}'\mathbf{A}_2\mathbf{t}_2\\ &= \sum_{i=k+1}^{c}(\mathbf{t}_2'\mathbf{A}_2\mathbf{h}_i)^2\rho_i^2 \leq \rho_{k+1}^2,\end{aligned}$$

with equality when $\mathbf{h} = \mathbf{A}_2\mathbf{t}_2 = \mathbf{h}_{k+1}$, which yields $\mathbf{t}_2'\mathbf{x}_2 = \mathbf{h}_{k+1}'\mathbf{A}_2^{-1}\mathbf{x}_2 = v_{k+1}$. As before, the Cauchy-Schwarz inequality becomes an equality iff $\mathbf{A}_1\mathbf{t}_1 \propto \mathbf{B}\mathbf{h}_{k+1}$, which implies $\mathbf{t}_1 = \mathbf{A}_1^{-1}\mathbf{g}_{k+1}$ and $\mathbf{t}_1'\mathbf{x}_1 = \mathbf{g}_{k+1}'\mathbf{A}_1^{-1}\mathbf{x}_1 = u_{k+1}$. The solution is the pair of canonical variables (u_{k+1}, v_{k+1}).

Repeating the second stage for $k = 1, \ldots, c-1$, all the pairs of canonical variables (u_i, v_i), $i = 1, \ldots, c$, can be generated. Each pair of canonical variables is identified with the pair of linear combinations of $\mathbf{x}_1$ and $\mathbf{x}_2$ with maximum correlation among all those uncorrelated with the preceding pairs.

Finally, the canonical correlations can be characterized as solutions of a determinant equation. In fact, the nonzero squared canonical correlations ρ_i^2, $i = 1, \ldots, c$, are the nonzero eigenvalues of

$$\mathbf{B}'\mathbf{B} = (\mathbf{A}_1^{-1}\boldsymbol{\Sigma}_{12}\mathbf{A}_2^{-1})'(\mathbf{A}_1^{-1}\boldsymbol{\Sigma}_{12}\mathbf{A}_2^{-1}) = \mathbf{A}_2^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{A}_2^{-1}.$$

Hence, the nonzero ρ_i^2 are the nonzero solutions λ of the equation

$$|\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} - \lambda\mathbf{I}| = 0.$$

11.3 Tests of independence

Based on a random sample of size n from a $N_{p_1+p_2}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$$

with $\boldsymbol{\Sigma}_{ij} \in \mathbb{R}_{p_j}^{p_i}$, we construct a test of independence reflected by the hypothesis,

$$H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0} \Longleftrightarrow H_0 : \ \rho_1 = \cdots = \rho_{p_1} = 0,$$

against all alternatives. The unbiased estimator $\mathbf{S}$ of $\boldsymbol{\Sigma}$ is partitioned in conformity as

$$(n-1)\mathbf{S} \equiv \mathbf{V} = \begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}$$

and we know already that $\mathbf{V} \sim W_p(n-1, \mathbf{\Sigma})$. The MLE $\hat{\mathbf{\Sigma}} = \mathbf{V}/n$ is proportional to $\mathbf{S}$. Without any restriction, the MLE of $\mathbf{\Sigma}_{ij}$, $i, j = 1, 2$, is $\hat{\mathbf{\Sigma}}_{ij} = \mathbf{V}_{ij}/n$. However, under H_0, the restricted MLE's are given by

$$\hat{\hat{\mathbf{\Sigma}}}_{11} = \hat{\mathbf{\Sigma}}_{11},\ \hat{\hat{\mathbf{\Sigma}}}_{22} = \hat{\mathbf{\Sigma}}_{22},\ \hat{\hat{\mathbf{\Sigma}}}_{12} = \mathbf{0}.$$

The LRT takes the form

$$\begin{aligned} \Lambda &= \frac{L(\bar{\mathbf{x}}, \hat{\hat{\mathbf{\Sigma}}}_{11}, \hat{\hat{\mathbf{\Sigma}}}_{22}, \hat{\hat{\mathbf{\Sigma}}}_{12})}{L(\bar{\mathbf{x}}, \hat{\mathbf{\Sigma}}_{11}, \hat{\mathbf{\Sigma}}_{22}, \hat{\mathbf{\Sigma}}_{12})} \\ &= \frac{|\hat{\mathbf{\Sigma}}_{11}|^{-n/2}|\hat{\mathbf{\Sigma}}_{22}|^{-n/2}}{|\hat{\mathbf{\Sigma}}|^{-n/2}}. \end{aligned}$$

Thus, since $\hat{\mathbf{\Sigma}} \propto \mathbf{V}$ and using the relation $|\mathbf{V}| = |\mathbf{V}_{11}||\mathbf{V}_{22.1}|$,

$$\begin{aligned} \Lambda^{2/n} &= \frac{|\mathbf{V}_{22.1}|}{|\mathbf{V}_{22}|} = \frac{|\mathbf{V}_{11.2}|}{|\mathbf{V}_{11}|} \\ &= |\mathbf{I} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}| \\ &= \prod_{i=1}^{p_1}(1 - r_i^2) \end{aligned}$$

is a function of the sample canonical correlations r_i's, where r_i^2 is a solution λ of the equation

$$|\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1} - \lambda\mathbf{I}| = 0.$$

They satisfy w.p.1, $1 > r_1^2 > \cdots > r_{p_1}^2 > 0$.

Consider now the invariant tests. The group $\mathbf{G}_{p_1} \times \mathbf{G}_{p_2} \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ transforms the observations as

$$\begin{aligned} \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix} &\mapsto \begin{pmatrix} \mathbf{B}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix} + \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B}_1\mathbf{x}_{i1} + \mathbf{b}_1 \\ \mathbf{B}_2\mathbf{x}_{i2} + \mathbf{b}_2 \end{pmatrix},\ i = 1, \ldots, n, \end{aligned}$$

for any $(\mathbf{B}_1, \mathbf{B}_2, \mathbf{b}_1, \mathbf{b}_2) \in \mathbf{G}_{p_1} \times \mathbf{G}_{p_2} \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$. This induces the following transformations on the minimal sufficient statistic $(\bar{\mathbf{x}}, \mathbf{V})$:

$$\begin{aligned} \begin{pmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{pmatrix} &\mapsto \begin{pmatrix} \mathbf{B}_1\bar{\mathbf{x}}_1 + \mathbf{b}_1 \\ \mathbf{B}_2\bar{\mathbf{x}}_2 + \mathbf{b}_2 \end{pmatrix}, \\ \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} &\mapsto \begin{pmatrix} \mathbf{B}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{pmatrix} \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{B}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2' \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B}_1\mathbf{V}_{11}\mathbf{B}_1' & \mathbf{B}_1\mathbf{V}_{12}\mathbf{B}_2' \\ \mathbf{B}_2\mathbf{V}_{21}\mathbf{B}_1' & \mathbf{B}_2\mathbf{V}_{22}\mathbf{B}_2' \end{pmatrix}. \end{aligned}$$

A test function $f(\bar{\mathbf{x}}, \mathbf{V})$ is invariant iff

$$f(\mathbf{y}, \mathbf{W}) = f\left(\begin{pmatrix} \mathbf{B}_1\mathbf{y}_1 + \mathbf{b}_1 \\ \mathbf{B}_2\mathbf{y}_2 + \mathbf{b}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{B}_1\mathbf{W}_{11}\mathbf{B}_1' & \mathbf{B}_1\mathbf{W}_{12}\mathbf{B}_2' \\ \mathbf{B}_2\mathbf{W}_{21}\mathbf{B}_1' & \mathbf{B}_2\mathbf{W}_{22}\mathbf{B}_2' \end{pmatrix}\right),$$

$\forall(\mathbf{B}_1, \mathbf{B}_2, \mathbf{b}_1, \mathbf{b}_2) \in \mathbf{G}_{p_1} \times \mathbf{G}_{p_2} \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, $\forall(\mathbf{y}, \mathbf{W}) \in \mathbb{R}^p \times \mathcal{P}_p$. The choice $\mathbf{b}_i = -\mathbf{B}_i \bar{\mathbf{x}}_i$, $i = 1, 2$, immediately yields

$$f(\bar{\mathbf{x}}, \mathbf{V}) = f\left(\mathbf{0}, \begin{pmatrix} \mathbf{B}_1\mathbf{V}_{11}\mathbf{B}_1' & \mathbf{B}_1\mathbf{V}_{12}\mathbf{B}_2' \\ \mathbf{B}_2\mathbf{V}_{21}\mathbf{B}_1' & \mathbf{B}_2\mathbf{V}_{22}\mathbf{B}_2' \end{pmatrix}\right).$$

Using the same arguments as in the definition of canonical correlations, let $\mathbf{V}_{ii} = \mathbf{A}_i^2$, where $\mathbf{A}_i > \mathbf{0}$, $i = 1, 2$, and consider the SVD

$$\mathbf{A}_1^{-1}\mathbf{V}_{12}\mathbf{A}_2^{-1} = \mathbf{G}(\mathbf{D}_\mathbf{r}, \mathbf{0})\mathbf{H}',$$

where $\mathbf{D}_\mathbf{r} = \operatorname{diag}(r_1, \ldots, r_{p_1})$, $1 > r_1 > \cdots > r_{p_1} > 0$, and we still assume $p_1 \leq p_2$ without loss of generality. Then, the choice $\mathbf{B}_1 = \mathbf{G}'\mathbf{A}_1^{-1}$ and $\mathbf{B}_2 = \mathbf{H}'\mathbf{A}_2^{-1}$ finally gives

$$f(\bar{\mathbf{x}}, \mathbf{V}) = f\left(\mathbf{0}, \begin{pmatrix} \mathbf{I}_{p_1} & (\mathbf{D}_\mathbf{r}, \mathbf{0}) \\ \begin{pmatrix} \mathbf{D}_\mathbf{r} \\ \mathbf{0} \end{pmatrix} & \mathbf{I}_{p_2} \end{pmatrix}\right);$$

i.e., any invariant test is a function of the sample canonical correlations r_i's. A similar argument shows that the power function of any invariant test depends only on the population canonical correlations ρ_i's.

Proposition 11.1 *With respect to the block-diagonal group of transformations above, any invariant test depends on the minimal sufficient statistic* $(\bar{\mathbf{x}}, \mathbf{V})$ *only through the sample canonical correlations* r_i*'s. The power function of any invariant test depends on* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *only through the population canonical correlations* ρ_i*'s.*

We now derive the null distribution of the LRT test.

Proposition 11.2 *Under the hypothesis of independence,* $H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0}$ *and* $n - 1 > \min(p_1, p_2)$, $\Lambda^{2/n} \sim U(p_2; p_1, n - 1 - p_1)$.

Proof. By invariance, assume without loss of generality that $\boldsymbol{\Sigma} = \mathbf{I}$ and let $m = n - 1$. Write

$$\begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} (\mathbf{X}_1, \mathbf{X}_2),$$

where

$$(\mathbf{X}_1, \mathbf{X}_2) \sim N_p^m(\mathbf{0}, \mathbf{I}_m \otimes \mathbf{I}_p).$$

The conditional distribution of $\mathbf{X}_2$ given $\mathbf{X}_1$ is

$$\mathbf{X}_2 \mid \mathbf{X}_1 \sim N_{p_2}^m(\mathbf{0}, \mathbf{I}_m \otimes \mathbf{I}_{p_2}).$$

Now, for $\mathbf{X}_1 \in \mathbb{R}_{p_1}^m$, $\operatorname{rank} \mathbf{X}_1 \stackrel{\text{w.p.1}}{=} p_1$. Therefore, we have the SVD

$$\mathbf{G}'\mathbf{X}_1\mathbf{H} = \begin{pmatrix} \mathbf{D} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{G} \in \mathbf{O}_m$, $\mathbf{H} \in \mathbf{O}_{p_1}$, and $\mathbf{D} \in \mathbb{R}^{p_1}_{p_1}$ is diagonal and nonsingular. Thus,

$$\mathbf{G}'\mathbf{X}_1 = \begin{pmatrix} \mathbf{D}\mathbf{H}' \\ \mathbf{0} \end{pmatrix} \equiv \begin{pmatrix} \tilde{\mathbf{X}}_1 \\ \mathbf{0} \end{pmatrix},$$

where $\tilde{\mathbf{X}}_1 \in \mathbb{R}^{p_1}_{p_1}$ is nonsingular. Since $\mathbf{G}$ (a function of $\mathbf{X}_1$) is orthogonal,

$$\mathbf{G}'\mathbf{X}_2 \mid \mathbf{X}_1 \sim N^m_{p_2}(\mathbf{0}, \mathbf{I}_m \otimes \mathbf{I}_{p_2}),$$

which does not depend on $\mathbf{X}_1$ and so $\mathbf{G}'\mathbf{X}_2 \sim N^m_{p_2}(\mathbf{0}, \mathbf{I}_m \otimes \mathbf{I}_{p_2})$ unconditionally. Now, partition

$$\mathbf{G}'\mathbf{X}_2 = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix},$$

where $\mathbf{Y} \in \mathbb{R}^{p_1}_{p_2}$ and $\mathbf{Z} \in \mathbb{R}^{m-p_1}_{p_2}$. Then, $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$ and

$$\begin{aligned} \mathbf{V}_{22} &= \mathbf{X}_2'\mathbf{X}_2 = (\mathbf{G}'\mathbf{X}_2)'(\mathbf{G}'\mathbf{X}_2) = \mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z}, \\ \mathbf{V}_{11} &= \mathbf{X}_1'\mathbf{X}_1 = (\mathbf{G}'\mathbf{X}_1)'(\mathbf{G}'\mathbf{X}_1) = \tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1, \\ \mathbf{V}_{12} &= \mathbf{X}_1'\mathbf{X}_2 = (\mathbf{G}'\mathbf{X}_1)'(\mathbf{G}'\mathbf{X}_2) = \tilde{\mathbf{X}}_1'\mathbf{Y}. \end{aligned}$$

Finally,

$$\begin{aligned} \Lambda^{2/n} &= \frac{|\mathbf{V}_{22.1}|}{|\mathbf{V}_{22}|} \\ &= \frac{|\mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z} - \mathbf{Y}'\tilde{\mathbf{X}}_1(\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z}|} = \frac{|\mathbf{Z}'\mathbf{Z}|}{|\mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z}|}, \end{aligned}$$

where $\mathbf{Z}'\mathbf{Z} \sim W_{p_2}(m - p_1)$, and $\mathbf{Y}'\mathbf{Y} \sim W_{p_2}(p_1)$. By definition,

$$\Lambda^{2/n} \sim U(p_2; p_1, m - p_1).$$

□

Let $\mathbf{R} = \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}$. As for multivariate regression, other invariant tests can be constructed such as

$$\begin{aligned} \text{tr } \mathbf{R} &= \sum_{i=1}^{p_1} r_i^2, \\ \text{tr } \mathbf{R}(\mathbf{I} - \mathbf{R})^{-1} &= \sum_{i=1}^{p_1} \frac{r_i^2}{(1 - r_i^2)}, \\ r_1^2 &= \max\{r_1^2, \ldots, r_{p_1}^2\}. \end{aligned}$$

Again, none of these tests has a power function which uniformly dominates the others. It is shown in Example 14.10 how to perform a bootstrap test using the test statistics tr $\mathbf{R}$ or tr $\mathbf{R}(\mathbf{I} - \mathbf{R})^{-1}$.

11.4 Properties of U distributions

We end this chapter with some properties and characterizations useful for the tabulation and moments of U distributions. These simplified proofs are from Bilodeau (1996).

Assume $\mathbf{x}_1 \in \mathbb{R}^{p_1}$ is fixed and $\mathbf{x}_2 \sim N_{p_2}(\mathbf{0}, \mathbf{\Sigma}_{22})$, $p = p_1 + p_2$, $\mathbf{\Sigma}_{22} > \mathbf{0}$. Based on a random sample of size n, say $\mathbf{X} \in \mathbb{R}^n_p$, the matrix of sums of squares and cross-products $\mathbf{V}$ is partitioned in conformity as

$$\mathbf{V} = \mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} (\mathbf{X}_1, \mathbf{X}_2) = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}.$$

When $n > \min(p_1, p_2)$ and rank $\mathbf{X}_1 = p_1$, consider

$$\tilde{\Lambda} = \frac{|\mathbf{V}|}{|\mathbf{V}_{11}||\mathbf{V}_{22}|} = \frac{|\mathbf{V}_{22.1}|}{|\mathbf{V}_{22}|} = \frac{|\mathbf{V}_{11.2}|}{|\mathbf{V}_{11}|}.$$

Proposition 11.3 *If $n > \min(p_1, p_2)$ and* rank $\mathbf{X}_1 = p_1$, *then*

$$\tilde{\Lambda} \sim U(p_2; p_1, n - p_1).$$

Proof. Assume without loss of generality $\mathbf{\Sigma}_{22} = \mathbf{I}$ and thus $\mathbf{X}_2 \sim N^n_{p_2}(\mathbf{0}, \mathbf{I})$. Now, $\mathbf{X}_1 \in \mathbb{R}^n_{p_1}$ has rank $\mathbf{X}_1 = p_1$. Its singular value decomposition is

$$\mathbf{G}'\mathbf{X}_1\mathbf{H} = \begin{pmatrix} \mathbf{D} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{G} \in \mathbf{O}_n$, $\mathbf{H} \in \mathbf{O}_{p_1}$, and $\mathbf{D} \in \mathbb{R}^{p_1}_{p_1}$ is diagonal and nonsingular. Thus,

$$\mathbf{G}'\mathbf{X}_1 = \begin{pmatrix} \mathbf{D}\mathbf{H}' \\ \mathbf{0} \end{pmatrix} \equiv \begin{pmatrix} \tilde{\mathbf{X}}_1 \\ \mathbf{0} \end{pmatrix}$$

where $\tilde{\mathbf{X}}_1 \in \mathbb{R}^{p_1}_{p_1}$ is nonsingular. Since $\mathbf{G}$ (a function of $\mathbf{X}_1$) is orthogonal, $\mathbf{G}'\mathbf{X}_2 \sim N^n_{p_2}(\mathbf{0}, \mathbf{I})$. Partition

$$\mathbf{G}'\mathbf{X}_2 = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix},$$

where $\mathbf{Y} \in \mathbb{R}^{p_1}_{p_2}$ and $\mathbf{Z} \in \mathbb{R}^{n-p_1}_{p_2}$. Then, $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$ and $\mathbf{V}_{22} = \mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z}$, $\mathbf{V}_{11} = \tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1$, and $\mathbf{V}_{12} = \tilde{\mathbf{X}}_1'\mathbf{Y}$. Finally,

$$\tilde{\Lambda} = \frac{|\mathbf{V}_{22.1}|}{|\mathbf{V}_{22}|} = \frac{|\mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z} - \mathbf{Y}'\tilde{\mathbf{X}}_1(\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z}|} = \frac{|\mathbf{Z}'\mathbf{Z}|}{|\mathbf{Y}'\mathbf{Y} + \mathbf{Z}'\mathbf{Z}|},$$

where $\mathbf{Z}'\mathbf{Z} \sim W_{p_2}(n - p_1)$, $\mathbf{Y}'\mathbf{Y} \sim W_{p_2}(p_1)$. By definition, $\tilde{\Lambda} \sim U(p_2; p_1, n - p_1)$. □

Proposition 11.3 remains valid if $\mathbf{X}_1$ has any absolutely continuous distribution (and thus has rank $\mathbf{X}_1 = p_1$ w.p.1 (v. Lemma 7.1 and the remark on page 88)) and $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$. It suffices to notice the distribution of $\tilde{\Lambda}$ does not depend on $\mathbf{X}_1$.

Vice versa, writing $\tilde{\Lambda} = |\mathbf{V}_{11.2}|/|\mathbf{V}_{11}|$, if $\mathbf{X}_1$ is normal and $\mathbf{X}_2$ is fixed, rank $\mathbf{X}_2 = p_2$, it is clear the same proof yields $\tilde{\Lambda} \sim U(p_1; p_2, n - p_2)$. The duality property asserts that, in fact, $U(p_1; p_2, n - p_2) \stackrel{d}{=} U(p_2; p_1, n - p_1)$.

As a by-product, we show the "duality" property:

Corollary 11.1 *$U(p; m, n) \stackrel{d}{=} U(m; p, m + n - p)$ when $m + n > p$.*

Proof. Assume $\mathbf{X}_1$ and $\mathbf{X}_2$ are both normal and $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$. Since $\mathbf{X}_1$ is normal, $\tilde{\Lambda} \sim U(p_1; p_2, n-p_2)$, and since $\mathbf{X}_2$ is also normal, $\tilde{\Lambda} \sim U(p_2; p_1, n-p_1)$. The distribution of $\tilde{\Lambda}$ being unique, $U(p_1; p_2, n-p_2) \stackrel{d}{=} U(p_2; p_1, n-p_1)$. Substitute $(p, m, m + n)$ for (p_1, p_2, n). □

In order to obtain a characterization of U distributions as a product of independent beta variables, we prove the following lemma.

Lemma 11.1 *If $n \geq p$,*

$$U(p; 1, n) \stackrel{d}{=} \text{beta}\left(\tfrac{1}{2}(n - p + 1); \tfrac{1}{2}p\right).$$

Proof. When $m = 1$, recalling the identity $|\mathbf{I} + \mathbf{AB}| = |\mathbf{I} + \mathbf{BA}|$ (v. Problem 1.8.3),

$$U(p; 1, n) \stackrel{d}{=} \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{z}\mathbf{z}'|} = |\mathbf{I} + \mathbf{W}^{-1}\mathbf{z}\mathbf{z}'|^{-1} = (1 + \mathbf{z}'\mathbf{W}^{-1}\mathbf{z})^{-1},$$

where $\mathbf{z} \perp\!\!\!\perp \mathbf{W}$, $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$, and $\mathbf{W} \sim W_p(n)$. Using Proposition 8.2,

$$\mathbf{z}'\mathbf{W}^{-1}\mathbf{z} \sim F_c(p, n - p + 1).$$

Finally, using Problem 3.5.5,

$$(1 + \mathbf{z}'\mathbf{W}^{-1}\mathbf{z})^{-1} \sim \text{beta}\left(\tfrac{1}{2}(n - p + 1); \tfrac{1}{2}p\right).$$

□

Proposition 11.4 *A variable distributed as $U(p; m, n)$, $n \geq p$, has the two characterizations*

$$U(p; m, n) \stackrel{d}{=} \prod_{i=1}^{m} \text{beta}\left(\tfrac{1}{2}(n - p + i); \tfrac{1}{2}p\right)$$

and

$$U(p; m, n) \stackrel{d}{=} \prod_{i=1}^{p} \text{beta}\left(\tfrac{1}{2}(n - p + i); \tfrac{1}{2}m\right);$$

i.e., a $U(p; m, n)$ variable has the same distribution as a product of independent beta variables.

Proof. The second representation follows from the first and the duality property of U distributions. We need only show the first representation. Its proof proceeds by induction on m. From Lemma 11.1, the result is true for

$m = 1$. Assume the result is true for $m-1$ and show it holds for m. By definition,

$$\begin{aligned} U(p;m,n) &\stackrel{d}{=} \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{Z}'\mathbf{Z}|},\ \mathbf{W}\sim W_p(n),\ \mathbf{Z}\sim N_p^m(\mathbf{0},\mathbf{I}_m\otimes\mathbf{I}_p),\ \mathbf{Z}\perp\!\!\!\perp\mathbf{W} \\ &= \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{z}_1\mathbf{z}_1'|}\cdot\frac{|(\mathbf{W}+\mathbf{z}_1\mathbf{z}_1')|}{|(\mathbf{W}+\mathbf{z}_1\mathbf{z}_1')+\mathbf{Z}_2'\mathbf{Z}_2|}\equiv U_1\cdot U_2, \end{aligned}$$

where

$$\mathbf{Z}=\begin{pmatrix}\mathbf{z}_1'\\ \mathbf{Z}_2\end{pmatrix},$$

$\mathbf{z}_1 \sim N_p(\mathbf{0},\mathbf{I})$, $\mathbf{Z}_2 \sim N_p^{m-1}(\mathbf{0},\mathbf{I}_{m-1}\otimes\mathbf{I}_p)$, and $\mathbf{z}_1 \perp\!\!\!\perp \mathbf{Z}_2$. Consider now the distribution of U_2. Let $\mathbf{W}_1 = \mathbf{W}+\mathbf{z}_1\mathbf{z}_1'$ and $\mathbf{W}_2 = \mathbf{Z}_2'\mathbf{Z}_2$. Then, $\mathbf{W}_1 \sim W_p(n+1)$, $\mathbf{W}_2 \sim W_p(m-1)$, and $\mathbf{W}_1 \perp\!\!\!\perp \mathbf{W}_2$. Therefore, $U_2 \sim U(p;m-1,n+1)$ and the induction hypothesis gives $U_2 \sim \prod_{i=1}^{m-1}\text{beta}\left(\frac{1}{2}(n+1-p+i);\frac{1}{2}p\right)$. Translating $i\mapsto i+1$,

$$U_2\sim\prod_{i=2}^{m}\text{beta}\left(\tfrac{1}{2}(n-p+i);\tfrac{1}{2}p\right).$$

The factor missing for $i=1$ is U_1. The proof is complete if we prove $U_1 \perp\!\!\!\perp U_2$. First, note that if $U_1 \perp\!\!\!\perp \mathbf{W}_1$, then U_1, $\mathbf{W}_1$, and $\mathbf{Z}_2$ are mutually independent and, therefore, $U_1 \perp\!\!\!\perp U_2$. So, we prove $U_1 \perp\!\!\!\perp \mathbf{W}_1$. But, if $\mathbf{V}\sim W_p(n,\mathbf{\Sigma})$, $\mathbf{\Sigma}>\mathbf{0}$, $\mathbf{x}\sim N_p(\mathbf{0},\mathbf{\Sigma})$, and $\mathbf{V}\perp\!\!\!\perp\mathbf{x}$, then $(\mathbf{V},\mathbf{x})\stackrel{d}{=}(\mathbf{Y}'\mathbf{Y},\mathbf{x})$, where $\mathbf{Y}\sim N_p^n(\mathbf{0},\mathbf{I}_n\otimes\mathbf{\Sigma})$, $\mathbf{Y}\perp\!\!\!\perp\mathbf{x}$. In the model for $(\mathbf{Y},\mathbf{x})$, $\mathbf{Y}'\mathbf{Y}+\mathbf{x}\mathbf{x}'$ is complete and sufficient for $\mathbf{\Sigma}$. Therefore, $\mathbf{V}+\mathbf{x}\mathbf{x}'$ is complete and sufficient for $\mathbf{\Sigma}$. Using Basu's theorem in the footnote on page 118, $\mathbf{V}+\mathbf{x}\mathbf{x}'$ is independent of any ancillary statistic such as $|\mathbf{V}|/|\mathbf{V}+\mathbf{x}\mathbf{x}'|$. This proves $U_1 \perp\!\!\!\perp \mathbf{W}_1$. □

This representation is useful for finding the distribution function or quantiles of a $U(p;m,n)$ distribution since $\ln U(p;m,n)$ can be represented as a convolution of simple distributions. Of course, it is advantageous to use the representation with $\min(p,m)$ number of factors. This number of factors can be reduced further by $\frac{1}{2}$ by grouping adjacent factors by pairs [Anderson (1984), p. 304]. The following lemma allows the pairing.

Lemma 11.2 *For $n>1$,*

$$[\text{beta}(n-1;m)]^2\stackrel{d}{=}\text{beta}(\tfrac{1}{2}(n-1);\tfrac{1}{2}m)\cdot\text{beta}(\tfrac{1}{2}n;\tfrac{1}{2}m).$$

Proof. It is straightforward to check that all moments of order $h>0$ on the left and right sides of $\stackrel{d}{=}$ are the same (v. Problem 11.6.3). Since the domain is the bounded interval $[0,1]$, there is a unique distribution with these moments [Serfling (1980), p. 46]. □

The following representation has a reduced number of factors as p is even or odd.

Corollary 11.2 *For $n \geq p$, a $U(p; m, n)$ variable can be represented as*

$$\prod_{i=1}^{r} [\text{beta}(n+1-2i; m)]^2, \qquad \text{if } p = 2r$$

$$\text{beta}(\tfrac{1}{2}(n-p+1); \tfrac{1}{2}m) \cdot \prod_{i=1}^{r} [\text{beta}(n+1-2i; m)]^2, \qquad \text{if } p = 2r+1.$$

Proof. The proof for $p = 2r$ is as follows. From Proposition 11.4, we have

$$U(p; m, n) \stackrel{d}{=} \prod_{i=1}^{r} \text{beta}(\tfrac{1}{2}(n-p+2i-1); \tfrac{1}{2}m)\ \text{beta}(\tfrac{1}{2}(n-p+2i); \tfrac{1}{2}m),$$

and from Lemma 11.2,

$$U(p; m, n) \stackrel{d}{=} \prod_{i=1}^{r} [\text{beta}(n-p+2i-1; m)]^2.$$

The conclusion follows after reversing the index $i \mapsto r-i+1$. The proof for p odd is identical except for the first isolated factor. □

The asymptotic distribution as $n \to \infty$ of $U(p; m, n)$ should be clear from the asymptotic distribution of the likelihood ratio statistic in Proposition 11.2,

$$-n \ln U(p; m, n) \stackrel{d}{\to} \chi^2_{pm}. \tag{11.1}$$

The slight modification

$$-[n - \tfrac{1}{2}(p-m+1)] \ln U(p; m, n) \stackrel{d}{\to} \chi^2_{pm}$$

is often used as an improved approximation since it has a remainder of order $O(n^{-2})$, whereas the remainder in (11.1) is $O(n^{-1})$. The general asymptotic expansion of order $O(n^{-\alpha})$ [Box (1949)] is treated in Section 12.3. As an alternative to asymptotic expansion an S-plus program in Appendix C uses the fast Fourier transform [Press (1992)] to compute the density of $U(p; m, n)$ by convolution and thus calculates exact probabilities (up to a discretization of the beta variables) and quantiles. Srivastava and Yau (1989) presented the saddlepoint method for obtaining tail probabilities. An exact closed form solution without series representation was also recently derived [Coelho (1998)].

11.4.1 Q-Q plot of squared radii

The scaled residuals of n observations, $\mathbf{x}_i$, may be defined as

$$\mathbf{z}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}),\ i = 1, \ldots, n.$$

Then, the squared radii (or squared Mahalanobis distances of $\mathbf{x}_i$ to $\bar{\mathbf{x}}$) are

$$d_i^2 = |\mathbf{z}_i|^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}),\ i = 1, \ldots, n.$$

Note that if $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then d_i is an ancillary statistic; i.e., the distribution of d_i, say $F(\cdot)$, does not depend on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. One aspect of multivariate normality can thus be tested with a Q-Q plot of the ordered d_i^2 against the quantiles of the distribution $F(\cdot)$ [Small (1978)]. Gnanadesikan and Kettenring (1972) derived the following result.

Lemma 11.3 *If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then*

$$\frac{n}{(n-1)^2} d_i^2 \sim \text{beta}\left(\tfrac{1}{2}p; \tfrac{1}{2}(n-p-1)\right).$$

Proof.

$$\begin{aligned} 1 - \frac{n}{(n-1)^2} d_i^2 &= |\mathbf{V} - \frac{n}{(n-1)}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'|/|\mathbf{V}| \\ &= |\mathbf{W}_1|/|\mathbf{W}_1 + \mathbf{W}_2|, \end{aligned}$$

where $\mathbf{V} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, $\mathbf{W}_1 = \mathbf{V} - \mathbf{W}_2$, and $\mathbf{W}_2 = \frac{n}{(n-1)}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. Assume without loss of generality that $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$. Thus, with $\mathbf{Z} \sim N_p^n(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{I}_p)$,

$$(\mathbf{W}_1, \mathbf{W}_2) \stackrel{d}{=} (\mathbf{Z}'(\mathbf{Q} - \mathbf{H})\mathbf{Z}, \mathbf{Z}'\mathbf{H}\mathbf{Z}),$$

where

$$\begin{aligned} \mathbf{H} &= \frac{n}{(n-1)}(\mathbf{e}_i - n^{-1}\mathbf{1})(\mathbf{e}_i - n^{-1}\mathbf{1})', \\ \mathbf{Q} &= \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'. \end{aligned}$$

The following can be verified easily:

(i) $\mathbf{H}$ is idempotent of rank 1,

(ii) $\mathbf{Q}$ is idempotent of rank $n-1$,

(iii) $\mathbf{Q}(\mathbf{e}_i - n^{-1}\mathbf{1}) = (\mathbf{e}_i - n^{-1}\mathbf{1})$ and, thus, $\mathbf{Q}\mathbf{H} = \mathbf{H}$,

(iv) $\mathbf{Q} - \mathbf{H}$ is idempotent of rank $n-2$, $(\mathbf{Q} - \mathbf{H})\mathbf{H} = \mathbf{0}$.

Thus, $\mathbf{W}_1 \perp\!\!\!\perp \mathbf{W}_2$, $\mathbf{W}_1 \sim W_p(n-2)$, and $\mathbf{W}_2 \sim W_p(1)$ (v. Proposition 7.8 and Problem 6.4.3), which implies

$$|\mathbf{W}_1|/|\mathbf{W}_1 + \mathbf{W}_2| \sim U(p; 1, n-2) \stackrel{d}{=} \text{beta}\left(\tfrac{1}{2}(n-p-1); \tfrac{1}{2}p\right).$$

□

Consider the ordered d_i^2,

$$d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2.$$

Assuming d_i^2, $i = 1, \ldots, n$, are i.i.d. according to the distribution in Lemma 11.3, one could evaluate the expected order statistics, $E\, d_{(i)}^2$. Then, the Q-Q plot consists of a graph of the points

$$\left(d_{(i)}^2,\ E\, d_{(i)}^2\right),\ i = 1, \ldots, n.$$

To simplify matters, we can assign to $d^2_{(i)}$ a cumulative probability of i/n and approximate $E\ d^2_{(i)}$ by the quantile $\gamma_i = i/n$ of the distribution

$$[(n-1)^2/n]\ \text{beta}\left(\tfrac{1}{2}p; \tfrac{1}{2}(n-p-1)\right).$$

Blom (1958) has shown how to select α and β so that the expected order statistic $E\ d^2_{(i)}$ may be well approximated by the quantile

$$\gamma_i = (i-\alpha)/(n-\alpha-\beta+1). \tag{11.2}$$

For beta, the distribution at hand, the indicated choice is

$$\begin{aligned} \alpha &= \frac{(p-2)}{2p}, \\ \beta &= \frac{(n-p-2)}{2(n-p-1)}. \end{aligned} \tag{11.3}$$

Thus, the recommended Q-Q plot is the graph of the points

$$\left(d^2_{(i)},\ [(n-1)^2/n]\ \text{beta}_{\gamma_i}\left(\tfrac{1}{2}p; \tfrac{1}{2}(n-p-1)\right)\right),\ i = 1, \ldots, n,$$

where $\text{beta}_\alpha(a;b)$ denotes the quantile α of a beta$(a;b)$ distribution and γ_i is given by (11.2) and (11.3). The Splus function *qqbeta* in Appendix C produces the Q-Q plot. One should not forget, however, that the d^2_i are correlated, but from Wilks (1963),

$$\text{cor}(d^2_i, d^2_j) = -\frac{1}{(n-1)},\ i \neq j,$$

and the correlation, of the order $O(n^{-1})$, is negligible for moderate to large sample sizes. A Q-Q plot approaching a 45° straight line is consistent with multivariate normality. Figure 11.1 gives the Q-Q plot for 50 observations generated from a $N_3(\mathbf{0}, \mathbf{I})$ distribution and Figure 11.2 is the Q-Q plot for 50 observations generated from a trivariate Cauchy distribution. These are easily generated with Example 13.2. The deviations from the straight line are clearly more systematic in Figure 11.2 associated with a distribution with heavier "tails" than the multivariate normal.

For large n, the beta distribution can be approximated by a χ^2_p distribution. Gnanadesikan (1977, p. 172) remarked that in the bivariate case $n = 25$ may provide a sufficiently large sample for this chi-squared approximation to be adequate. However, $n = 100$ does not seem large enough for $p = 4$, for there is a marked deviation from linearity when the ordered d^2_i are plotted against expected order statistics of chi-squared, and this effect becomes more marked as p increases. We therefore recommend the use of the beta distribution.

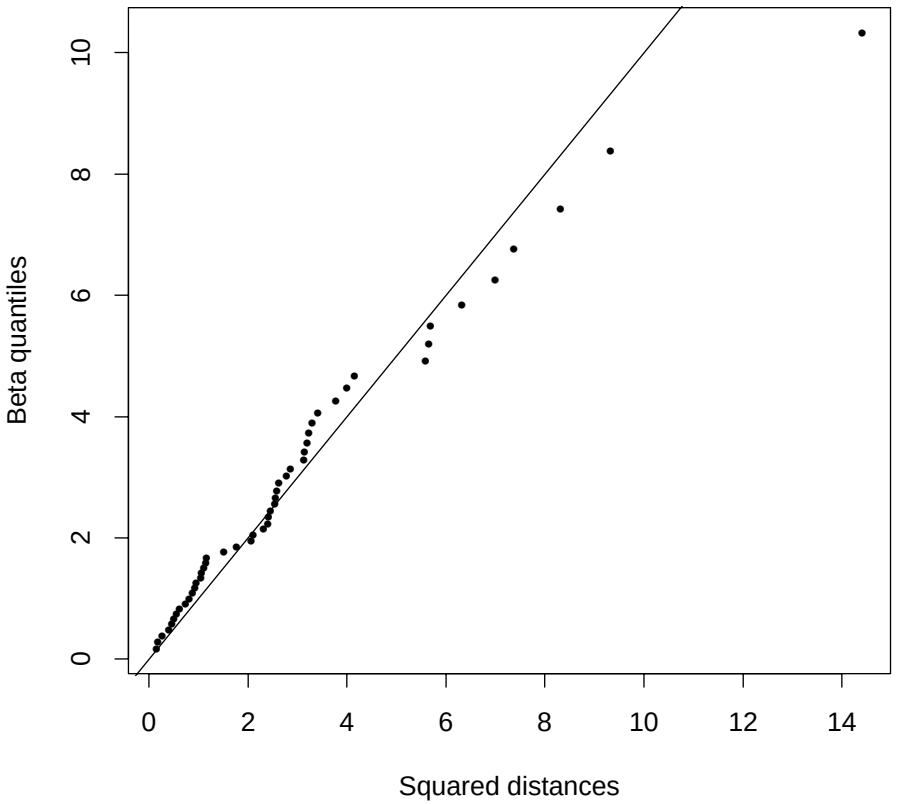

Figure 11.1. Q-Q plot for a sample of size $n = 50$ from a trivariate normal, $N_3(\mathbf{0}, \mathbf{I})$, distribution.

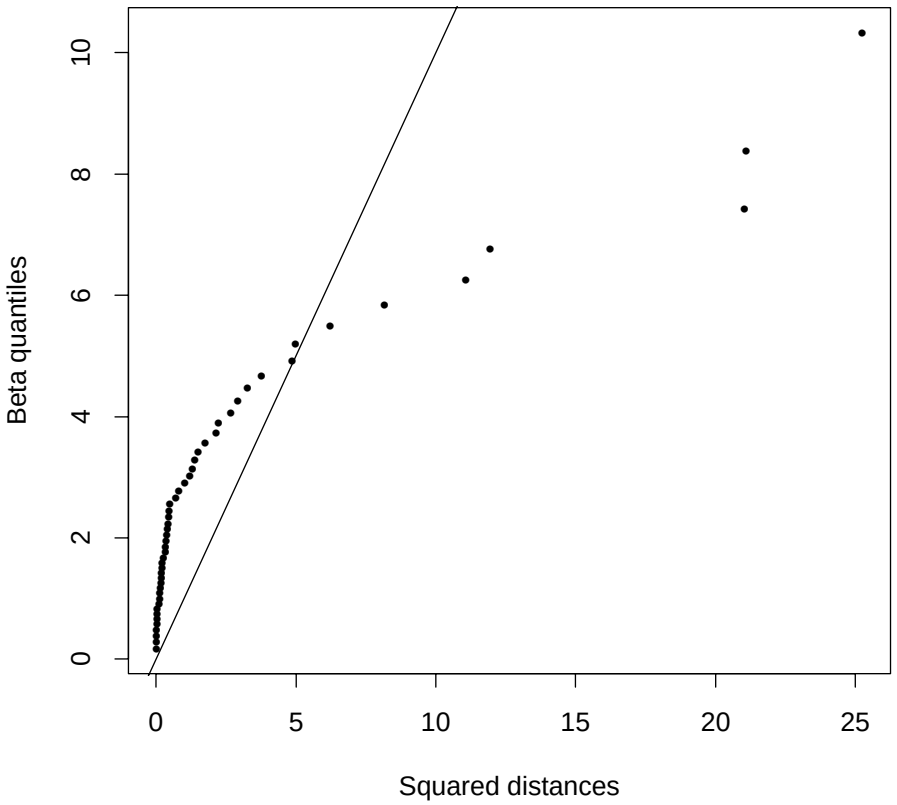

Figure 11.2. Q-Q plot for a sample of size $n = 50$ from a trivariate t on 1 degree of freedom, $t_{3,1}(\mathbf{0}, \mathbf{I}) \equiv Cauchy_3(\mathbf{0}, \mathbf{I})$, distribution.

11.5 Asymptotic distributions

Assuming $\mathbf{x}_i$, $i = 1, \ldots, n$, are i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we derive the asymptotic distribution of r_α^2, $\alpha = 1, \ldots, p_1$, when ρ_α is distinct from all other canonical correlations.

A squared sample canonical correlation, r_α^2, is a value of l for which there is a nonzero solution $\mathbf{c}$ to the equation

$$(\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} - l\ \mathbf{I})\mathbf{c} = \mathbf{0}. \tag{11.4}$$

Using the result $n^{1/2}(\mathbf{S} - \boldsymbol{\Sigma}) \xrightarrow{d} \mathbf{W}$ of Section 6.3, where

$$\mathbf{W} \sim N_p^p(\mathbf{0}, (\mathbf{I} + \mathbf{K})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})),$$

we write

$$\begin{aligned} \mathbf{S}_{11} &= \mathbf{I} + n^{-1/2}\mathbf{W}_{11}, \\ \mathbf{S}_{22} &= \mathbf{I} + n^{-1/2}\mathbf{W}_{22}, \\ \mathbf{S}_{12} &= (\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0}) + n^{-1/2}\mathbf{W}_{12}. \end{aligned}$$

Using Problem 1.8.15, $\mathbf{S}_{22}^{-1} = \mathbf{I} - n^{-1/2}\mathbf{W}_{22} + O_p(n^{-1})$ and similarly for $\mathbf{S}_{11}^{-1}$. Keeping terms up to order $n^{-1/2}$,

$$\begin{aligned} \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} &= \mathbf{D}_{\boldsymbol{\rho}}^2 + n^{-1/2}\Big[-\mathbf{D}_{\boldsymbol{\rho}}^2\mathbf{W}_{11} + (\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0})\mathbf{W}_{21} \\ &\quad -(\mathbf{D}_{\boldsymbol{\rho}}, \mathbf{0})\mathbf{W}_{22}\begin{pmatrix}\mathbf{D}_{\boldsymbol{\rho}} \\ \mathbf{0}\end{pmatrix} + \mathbf{W}_{12}\begin{pmatrix}\mathbf{D}_{\boldsymbol{\rho}} \\ \mathbf{0}\end{pmatrix}\Big] + O_p(n^{-1}). \end{aligned}$$

We now apply the perturbation method as in Section 8.8.1 and obtain from (8.11) the expansion

$$r_\alpha^2 = \rho_\alpha^2 + n^{-1/2}[-\rho_\alpha^2 w_{11}^{\alpha\alpha} - \rho_\alpha^2 w_{22}^{\alpha\alpha} + 2\rho_\alpha w_{21}^{\alpha\alpha}] + O_p(n^{-1}),$$

where $w_{ij}^{\alpha\alpha}$ is the element (α, α) of the matrix $\mathbf{W}_{ij}$. From (6.1), we have

$$(w_{11}^{\alpha\alpha}, w_{22}^{\alpha\alpha}, w_{21}^{\alpha\alpha})' \xrightarrow{d} N_3(\mathbf{0}, \boldsymbol{\Omega}),$$

where

$$\boldsymbol{\Omega} = \begin{pmatrix} 2 & 2\rho_\alpha^2 & 2\rho_\alpha \\ 2\rho_\alpha^2 & 2 & 2\rho_\alpha \\ 2\rho_\alpha & 2\rho_\alpha & 1 + \rho_\alpha^2 \end{pmatrix}.$$

Finally, defining the linear combination vector $\mathbf{a} = (-\rho_\alpha^2, -\rho_\alpha^2, 2\rho_\alpha)'$, we obtain $n^{1/2}(r_\alpha^2 - \rho_\alpha^2) \xrightarrow{d} N(0, \mathbf{a}'\boldsymbol{\Omega}\mathbf{a})$, whereby a direct calculation shows $\mathbf{a}'\boldsymbol{\Omega}\mathbf{a} = 4\rho_\alpha^2(1 - \rho_\alpha^2)^2$. We have shown:

Proposition 11.5 *The asymptotic distribution of the squared sample canonical correlation r_α^2, $\alpha = 1, \ldots, p_1$, assuming ρ_α is distinct from all other canonical correlations is $n^{1/2}(r_\alpha^2 - \rho_\alpha^2) \xrightarrow{d} N(0, 4\rho_\alpha^2(1 - \rho_\alpha^2)^2)$.*

Various extensions of Proposition 11.5 to the joint distribution of sample canonical correlations can be envisaged. The simplest extension is to the joint distribution of $r_1^2, \ldots, r_{p_1}^2$ when all canonical correlations are distinct, $\rho_1 > \cdots > \rho_{p_1}$.

Corollary 11.3 *The asymptotic joint distribution of $r_1^2, \ldots, r_{p_1}^2$ when all population canonical correlations are distinct, $\rho_1 > \cdots > \rho_{p_1}$, is*

$$n^{1/2}(r_1^2 - \rho_1^2, \ldots, r_{p_1}^2 - \rho_{p_1}^2) \xrightarrow{d} N_{p_1}\left(\mathbf{0}, 4 \text{ diag}\left(\rho_1^2(1-\rho_1^2)^2, \ldots, \rho_{p_1}^2(1-\rho_{p_1}^2)^2\right)\right).$$

Proof. It suffices to consider the asymptotic covariance of two squared sample canonical correlations, r_α^2 and r_β^2, when the population canonical correlations, ρ_α and ρ_β, are of multiplicity 1. But, it is immediate from the proof of Proposition 11.5 that

$$\text{cov}(-\rho_\alpha^2 w_{11}^{\alpha\alpha} - \rho_\alpha^2 w_{22}^{\alpha\alpha} + 2\rho_\alpha w_{21}^{\alpha\alpha}, -\rho_\beta^2 w_{11}^{\beta\beta} - \rho_\beta^2 w_{22}^{\beta\beta} + 2\rho_\beta w_{21}^{\beta\beta}) = 0$$

since from (6.1), all the covariances satisfy, $\text{cov}(w_{ij}^{\alpha\alpha}, w_{kl}^{\beta\beta}) = 0$, $i, j, k, l = 1, 2$. □

Hsu (1941) derived the asymptotic joint density when

$$1 > \rho_1 > \cdots > \rho_c > \rho_{c+1} = \cdots = \rho_{p_1} = 0.$$

Muirhead and Waternaux (1980) obtained the asymptotic joint distribution when all population canonical correlations are distinct, as in Proposition 11.3 but for any underlying distribution with finite fourth moments. Eaton and Tyler (1994), assuming an underlying elliptical distribution or, in fact, any other distribution with finite fourth moments, derived the asymptotic joint distribution in full generality,

$$\rho_1 \geq \cdots \geq \rho_c > \rho_{c+1} = \cdots = \rho_{p_1} = 0,$$

using an extension of Wielandt's inequality to singular values.

In canonical correlation analysis, the number of nonzero population correlations is called the dimensionality. Asymptotic distributions of the dimensionality estimated by Mallow's criterion and Akaike's criterion were derived [Gunderson and Muirhead (1997)] for non-normal multivariate populations with finite fourth moments.

11.6 Problems

1. Obtain the hth moment, $h > 0$, of $U \sim U(p; m, n)$, $n \geq p$,

$$E\, U^h = \prod_{i=1}^{m} \frac{\Gamma\left[\frac{1}{2}(n-p+i)+h\right]}{\Gamma\left[\frac{1}{2}(n-p+i)\right]} \cdot \frac{\Gamma\left[\frac{1}{2}(n+i)\right]}{\Gamma\left[\frac{1}{2}(n+i)+h\right]}$$

$$= \prod_{i=1}^{p} \frac{\Gamma\left[\frac{1}{2}(n-p+i)+h\right]}{\Gamma\left[\frac{1}{2}(n-p+i)\right]} \cdot \frac{\Gamma\left[\frac{1}{2}(m+n-p+i)\right]}{\Gamma\left[\frac{1}{2}(m+n-p+i)+h\right]}.$$

2. Establish the following exact results concerning U distributions:

$$\begin{aligned}
\frac{n[1-U(1;m,n)]}{m\,U(1;m,n)} &\sim F(m,n),\\
\frac{(n-p+1)[1-U(p;1,n)]}{p\,U(p;1,n)} &\sim F(p,n-p+1),\\
\frac{(n-1)[1-U(2;m,n)^{1/2}]}{m\,U(2;m,n)^{1/2}} &\sim F(2m,2(n-1)),\\
\frac{(n-p+1)[1-U(p;2,n)^{1/2}]}{p\,U(p;2,n)^{1/2}} &\sim F(2p,2(n-p+1)).
\end{aligned}$$

3. Prove that $[\text{beta}(n-1,m)]^2$ and $\text{beta}(\frac{1}{2}(n-1),\frac{1}{2}m)\cdot\text{beta}(\frac{1}{2}n,\frac{1}{2}m)$, a product of two independent betas, have the same moments of order $h>0$.

4. For

$$\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')' = (x_1,\ldots,x_{p_1};x_{p_1+1},\ldots,x_{p_1+p_2})'$$

establish that simple correlation and multiple correlation coefficients are bounded above as

(i) $|\rho_{x_i,x_j}| \le \rho_1,\ i=1,\ldots,p_1,\ j=p_1+1,\ldots,p_1+p_2,$
(ii) $R_{x_i,\mathbf{x}_2} \le \rho_1,\ i=1,\ldots,p_1,$

where ρ_1 is the largest canonical correlation.

5. Let $\boldsymbol{\Sigma}_{12} = \rho\mathbf{1}_{p_1}\mathbf{1}_{p_2}'$, $\boldsymbol{\Sigma}_{ii} = \rho\mathbf{1}_{p_i}\mathbf{1}_{p_i}' + (1-\rho)\mathbf{I}_{p_i}$, $i=1,2$, corresponding to the equicorrelated case. Determine the canonical variables corresponding to the nonzero canonical correlation.
Hint: $\boldsymbol{\Sigma}_{11}\mathbf{1}_{p_1} = [1+(p_1-1)\rho]\mathbf{1}_{p_1}$.

6. Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be i.i.d. $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

with $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{p_1}_{p_2}$, $p = p_1+p_2$. For testing $H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0}$ against $H_1 : \boldsymbol{\Sigma}_{12} \ne \mathbf{0}$, consider the test statistic [Escoufier (1973)]

$$E = \frac{\text{tr}(\mathbf{S}_{12}\mathbf{S}_{21})}{[\text{tr}(\mathbf{S}_{11}^2)]^{1/2}[\text{tr}(\mathbf{S}_{22}^2)]^{1/2}},$$

where $\mathbf{S}$ is the sample variance partitioned as $\boldsymbol{\Sigma}$. Prove:

(i) E is invariant under the group of transformations

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix} + \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}$$

for any $(\mathbf{H}_1, \mathbf{H}_2, \mathbf{b}_1, \mathbf{b}_2) \in \mathbf{O}_{p_1} \times \mathbf{O}_{p_2} \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$.

(ii) If H_0 holds, then the distribution of E is the same as when

$$\mathbf{\Sigma} = \begin{pmatrix} \text{diag}(\lambda_i) & \mathbf{0} \\ \mathbf{0} & \text{diag}(\gamma_j) \end{pmatrix},$$

where λ_i and γ_j are, respectively, the eigenvalues of $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$.

(iii) Under H_0, $n^{1/2}\mathbf{S}_{12} \xrightarrow{d} \mathbf{Z} = (z_{ij})$, where z_{ij} are independently distributed as $N(0, \lambda_i\gamma_j)$, $i = 1, \ldots, p_1$, $j = 1, \ldots, p_2$.

(iv) Conclude the null distribution

$$n\ E \xrightarrow{d} \left(\sum_{i=1}^{p_1} \lambda_i^2\right)^{-1/2} \left(\sum_{j=1}^{p_2} \gamma_j^2\right)^{-1/2} \sum_{i=1}^{p_1}\sum_{j=1}^{p_2} \lambda_i\gamma_j z_{ij}^2.$$

Remark: Unlike for canonical correlations, the asymptotic null distribution depends on unknown parameters because of the lack of invariance of E (the group $\mathbf{O}_{p_1} \times \mathbf{O}_{p_2} \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ is only a subgroup of $\mathbf{G}_{p_1} \times \mathbf{G}_{p_2} \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$). The asymptotic distribution of E for sampling from an elliptical distribution was derived by Cléroux and Ducharme (1989).

7. **Test of mutual independence of several subvectors.**
This problem given in the form of a project derives the exact null distribution of the likelihood ratio test for mutual independence. Consider a random sample of size $n \geq p + 1$ from $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1r} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{\Sigma}_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{r1} & \mathbf{\Sigma}_{r2} & \cdots & \mathbf{\Sigma}_{rr} \end{pmatrix}$$

with $\mathbf{\Sigma}_{ij} \in \mathbb{R}^{p_i}_{p_j}$, $p = \sum_{j=1}^r p_j$. We wish to test $H_0 : \mathbf{\Sigma}_{ij} = \mathbf{0},\ 1 \leq i < j \leq r$, versus all alternatives.

(i) Prove the likelihood ratio test Λ for H_0 can be written

$$\Lambda^{2/n} = \frac{|\mathbf{V}|}{\prod_{i=1}^r |\mathbf{V}_{ii}|},$$

where as usual $\mathbf{V} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \sim W_p(n-1, \mathbf{\Sigma})$.

(ii) Obtain the exact null moments of $\tilde{\Lambda} = \Lambda^{2/n}$,

$$E\ \tilde{\Lambda}^h = \frac{\Gamma_p(\frac{1}{2}m + h)}{\Gamma_p(\frac{1}{2}m)} \prod_{i=1}^r \frac{\Gamma_{p_i}(\frac{1}{2}m)}{\Gamma_{p_i}(\frac{1}{2}m + h)},$$

where $m = n - 1$.

Hint:

$$E\ \tilde{\Lambda}^h = \frac{c_{p,m}}{c_{p,m+2h}} \prod_{i=1}^{r} E\ |\mathbf{V}_{ii}|^{-h},$$

where $c_{p,m} = [2^{pm/2}\Gamma_p(\frac{1}{2}m)]^{-1}$ is the normalizing constant of a $W_p(m)$ density and where $\mathbf{V}_{ii}$ are mutually independent $W_{p_i}(m+2h)$.

(iii) Define the upper left corner of $\mathbf{V}$ to be

$$\tilde{\mathbf{V}}_{ii} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1i} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{i1} & \mathbf{V}_{i2} & \cdots & \mathbf{V}_{ii} \end{pmatrix} \in \mathbb{R}^{\bar{p}_i}_{\bar{p}_i},$$

where $\bar{p}_i = p_1 + \cdots + p_i$, and note that $\tilde{\mathbf{V}}_{rr} = \mathbf{V}$ and $\tilde{\mathbf{V}}_{11} = \mathbf{V}_{11}$. Derive the equivalent form $\Lambda^{2/n} = \prod_{i=2}^{r} U_i$, where

$$U_i = \frac{|\tilde{\mathbf{V}}_{ii}|}{|\mathbf{V}_{ii}||\tilde{\mathbf{V}}_{i-1,i-1}|},\ i = 2, \ldots, r.$$

(iv) Use Proposition 11.2 to obtain immediately under H_0

$$U_i \sim U\left(p_i; \bar{p}_{i-1}, n - 1 - \bar{p}_{i-1}\right),\ i = 2, \ldots, r.$$

(v) When

$$\mathbf{\Sigma} = \begin{pmatrix} \tilde{\mathbf{\Sigma}}_{r-1,r-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{rr} \end{pmatrix},$$

prove that $(\tilde{\mathbf{V}}_{r-1,r-1}, \mathbf{V}_{rr})$ is sufficient and complete and that U_r is ancillary. Conclude that, under H_0, $U_r \perp\!\!\!\perp \mathbf{V}_{r-1,r-1}, \mathbf{V}_{rr})$.

(vi) Using (iii), prove that $U_r \perp\!\!\!\perp U_2, \ldots, U_{r-1})$ under H_0.

(vii) Repeat this argument to prove $U_i \perp\!\!\!\perp U_2, \ldots, U_{i-1})$, $i = 3, \ldots, r$, whence, altogether, $U_2, \ldots, U_r$ are mutually independent under H_0.

(viii) Use Proposition 11.4 to obtain the exact null distribution

$$\tilde{\Lambda} \stackrel{d}{=} \prod_{i=2}^{r} \prod_{j=1}^{p_i} \text{beta}\left(\tfrac{1}{2}(n - \bar{p}_{i-1} - j); \tfrac{1}{2}\bar{p}_{i-1}\right).$$

Note that a further representation with a reduced number of factors as p_i is odd or even is immediate from Corollary 11.2.

(ix) Prove $U_i^{n/2}$ is the likelihood ratio test for $H_i : \mathbf{\Sigma}_{li} = \mathbf{0}$, $l = 1, \ldots, i-1$ when it is known that all the hypotheses H_{i+1}, $\cdots$, H_r are true. Note that $H_0 = \cap_{i=2}^{r} H_i$.

(x) Use (viii) to obtain the equivalent expression for the exact null moments of $\tilde{\Lambda}$:

$$E\ \tilde{\Lambda}^h = \prod_{i=2}^{r} \left\{ \prod_{j=1}^{p_i} \frac{\Gamma[\frac{1}{2}(n-\bar{p}_{i-1}-j)+h]\ \Gamma[\frac{1}{2}(n-j)]}{\Gamma[\frac{1}{2}(n-\bar{p}_{i-1}-j)]\ \Gamma[\frac{1}{2}(n-j)+h]} \right\}.$$

8. **Generalized squared interpoint distance** [Gnanadesikan and Kettenring (1972)]
Let $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)$ be the generalized squared interpoint distance (or squared Mahalanobis distance) between $\mathbf{x}_i$ and $\mathbf{x}_j$, $i \neq j$. Prove that if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\frac{1}{2(n-1)} d_{ij}^2 \sim \text{beta}\left(\tfrac{1}{2}p; \tfrac{1}{2}(n-p-1)\right).$$

12
Asymptotic expansions

12.1 Introduction

The exact distribution of likelihood ratio tests in multivariate analysis is often too complicated to be of any practical use. An asymptotic expansion due to Box (1949) is rather simple and easy to program on a computer to obtain the distribution function to any degree of accuracy. This approximation is applied on several of the testing situations previously encountered. In at least one situation where the exact distribution is known, an evaluation of the approximation is carried out for small to moderate sample sizes.

12.2 General expansions

The method can be used whenever the likelihood ratio criterion Λ (or a suitable power W) has moment of order h of the form

$$E\ W^h = K \left[\frac{\prod_{j=1}^{b} y_j^{y_j}}{\prod_{k=1}^{a} x_k^{x_k}} \right]^h \frac{\prod_{k=1}^{a} \Gamma[x_k(1+h)+\zeta_k]}{\prod_{j=1}^{b} \Gamma[y_j(1+h)+\eta_j]}, \tag{12.1}$$

where

$$\sum_{j=1}^{b} y_j = \sum_{k=1}^{a} x_k,$$

and K is just a constant (not depending on h) so that $E\ W^0 = 1$. Equation (12.1) is usually obtained for real h; it is, however, generally valid on the domain where the functions are analytic. This means if we let $M = -2\log W$, then we can write the characteristic function of ρM, for a constant $0 < \rho \leq 1$ to be determined later, as

$$\begin{aligned} c_{\rho M}(t) &= E\ W^{-2it\rho} \\ &= K\left[\frac{\prod_{j=1}^b y_j^{y_j}}{\prod_{k=1}^a x_k^{x_k}}\right]^{-2it\rho} \frac{\prod_{k=1}^a \Gamma[x_k(1-2it\rho)+\zeta_k]}{\prod_{j=1}^b \Gamma[y_j(1-2it\rho)+\eta_j]}. \end{aligned}$$

Taking logarithms and defining

$$\beta_k = (1-\rho)x_k, \qquad \epsilon_j = (1-\rho)y_j, \tag{12.2}$$

the cumulant generating function is

$$K_{\rho M}(t) = \log c_{\rho M}(t) = g(t) - g(0), \tag{12.3}$$

where

$$\begin{aligned} g(t) &= 2it\rho[\sum_{k=1}^a x_k \log x_k - \sum_{j=1}^b y_j \log y_j] \\ &\quad + \sum_{k=1}^a \log\Gamma[\rho x_k(1-2it) + \beta_k + \zeta_k] \\ &\quad - \sum_{j=1}^b \log\Gamma[\rho y_j(1-2it) + \epsilon_j + \eta_j], \end{aligned}$$

with

$$g(0) = -\log K = \sum_{k=1}^a \log\Gamma[\rho x_k + \beta_k + \zeta_k] - \sum_{j=1}^b \log\Gamma[\rho y_j + \epsilon_j + \eta_j].$$

We use the asymptotic expansion in z as $|z| \to \infty$ [Erdélyi et al. (1953), p. 48] for bounded h,

$$\begin{aligned} \log\Gamma(z+h) &= \log\sqrt{2\pi} + (z+h-\tfrac{1}{2})\log z - z \\ &\quad - \sum_{\alpha=1}^{l} (-1)^\alpha \frac{B_{\alpha+1}(h)}{\alpha(\alpha+1)} z^{-\alpha} + O(z^{-(l+1)}), \quad |\arg z| < \pi. \end{aligned} \tag{12.4}$$

The terms $B_r(h)$ are the Bernoulli polynomials defined to be the coefficients in the Taylor series

$$\frac{ze^{hz}}{e^z - 1} = \sum_{r=0}^{\infty} B_r(h)\frac{z^r}{r!}, \quad |z| < 2\pi.$$

The reader can verify the first few Bernoulli polynomials

$$B_0(h) = 1,$$

$$
\begin{aligned}
B_1(h) &= h - \tfrac{1}{2}, \\
B_2(h) &= h^2 - h + \tfrac{1}{6}, \\
B_3(h) &= h^3 - \tfrac{3}{2}h^2 + \tfrac{1}{2}h, \\
B_4(h) &= h^4 - 2h^3 + h^2 - \tfrac{1}{30}, \\
B_5(h) &= h^5 - \tfrac{5}{2}h^4 + \tfrac{5}{3}h^3 - \tfrac{1}{6}h, \\
B_6(h) &= h^6 - 3h^5 + \tfrac{5}{2}h^4 - \tfrac{1}{2}h^2 + \tfrac{1}{42}.
\end{aligned}
$$

Bernoulli polynomials can be generated at will with modern symbolic computations software such as the function *bernoulli(r,h);* in Maple [Redfern (1996)] or *BernoulliB[h,r]* in Mathematica [Wolfram (1996)]. Let $(z, h) = (\rho x_k(1 - 2it), \beta_k + \zeta_k)$, $(\rho y_j(1 - 2it), \epsilon_j + \eta_j)$, $(\rho x_k, \beta_k + \zeta_k)$, and $(\rho y_j, \epsilon_j + \eta_j)$ in turn in (12.4). We assume that x_k and y_j are terms behaving as $O(n)$, where n is the sample size. This will have to be checked in each application. When $\rho = 1$, then $\beta_k = \epsilon_j = 0$, and h is bounded in all cases. Later, ρ will be allowed to depend on the sample size n and we will need to check that β_k and ϵ_j are bounded.

Then substitute the four expansions for $\log\Gamma(z + h)$ in $g(t)$ and $g(0)$ of (12.3) to obtain, after long but straightforward simplifications,

$$
K_{\rho M}(t) = -\tfrac{1}{2} f \log(1 - 2it) + \sum_{\alpha=1}^{l} \omega_\alpha [(1 - 2it)^{-\alpha} - 1] + O(n^{-(l+1)}), \quad (12.5)
$$

where

$$
f = -2 \left[\sum_{k=1}^{a} \zeta_k - \sum_{j=1}^{b} \eta_j - \tfrac{1}{2}(a - b) \right], \quad (12.6)
$$

$$
\omega_\alpha = \frac{(-1)^{\alpha+1}}{\alpha(\alpha + 1)} \left[\sum_{k=1}^{a} \frac{B_{\alpha+1}(\beta_k + \zeta_k)}{(\rho x_k)^\alpha} - \sum_{j=1}^{b} \frac{B_{\alpha+1}(\epsilon_j + \eta_j)}{(\rho y_j)^\alpha} \right]. \quad (12.7)
$$

Note that $\omega_\alpha = O(n^{-\alpha})$ if x_k and y_j are $O(n)$, and β_k and ϵ_j are $O(1)$.

The next step consists of deriving the characteristic function $c_{\rho M}$ by exponentiation of $K_{\rho M}$ and then using the inversion formula (2.2) to derive the p.d.f.:

$$
\begin{aligned}
c_{\rho M}(t) &= e^{K_{\rho M}(t)} \\
&= (1 - 2it)^{-f/2} \prod_{\alpha=1}^{l} \exp[\omega_\alpha (1 - 2it)^{-\alpha}] \prod_{\alpha=1}^{l} \exp(-\omega_\alpha) \\
&\quad \cdot [1 + O(n^{-(l+1)})] \\
&= (1 - 2it)^{-f/2} \prod_{\alpha=1}^{l} \sum_{k=0}^{\infty} \frac{\omega_\alpha^k}{k!} (1 - 2it)^{-\alpha k} \prod_{\alpha=1}^{l} \sum_{k=0}^{\infty} (-1)^k \frac{\omega_\alpha^k}{k!}
\end{aligned}
$$

$$\cdot[1 + O(n^{-(l+1)})].$$

The expansion of order $O(n^{-(l+1)})$ is then obtained by keeping and collecting terms of order up to $O(n^{-l})$. Let us illustrate the procedure for an expansion of order $O(n^{-2})$ and leave the higher-order expansion to the symbolic calculators Mathematica [Wolfram (1996)] or Maple [Redfern (1996)]. For the expansion of order $O(n^{-2})$, set $l = 1$ and note that

$$\begin{aligned} c_{\rho M}(t) &= (1-2it)^{-f/2}[1+\omega_1(1-2it)^{-1}](1-\omega_1) + O(n^{-2}) \\ &= (1-2it)^{-f/2}\{1+\omega_1[(1-2it)^{-1}-1]\} + O(n^{-2}) \\ &= c_f(t) + \omega_1[c_{f+2}(t) - c_f(t)] + O(n^{-2}), \end{aligned} \tag{12.8}$$

where $c_f(t) = (1-2it)^{-f/2}$ denotes the characteristic function of χ^2_f on f degrees of freedom. Then, by the inversion formula (2.2),

$$\begin{aligned} f_{\rho M}(s) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} c_{\rho M}(t)e^{-its}dt \\ &= g_f(s) + \omega_1[g_{f+2}(s) - g_f(s)] + O(n^{-2}), \end{aligned} \tag{12.9}$$

where $g_f(t)$ is the p.d.f. of χ^2_f. Finally, by integration on $(-\infty, x]$, the d.f. takes the form

$$P(\rho M \le x) = G_f(x) + \omega_1[G_{f+2}(x) - G_f(x)] + O(n^{-2}),$$

where $G_f(t)$ is the d.f. of χ^2_f. A full justification of the last two integrations would require one to show that the remainders in (12.8) and (12.9) are $O(n^{-2})$ uniformly in t and s, respectively; v. Box (1949) for details.

The whole purpose of introducing ρ in the expansion is to reduce the number of terms. In the above example, one can choose ρ to annihilate the term of order $O(n^{-1})$, i.e., to make $\omega_1 = 0$. Recalling (12.2) and $B_2(h) = h^2 - h + \frac{1}{6}$, we have

$$\begin{aligned} \omega_1 &= \frac{1}{2\rho}\left[\sum_{k=1}^{a} x_k^{-1}B_2(\beta_k+\zeta_k) - \sum_{j=1}^{b} y_j^{-1}B_2(\epsilon_j+\eta_j)\right] \\ &= \frac{1}{2\rho}\left[-(1-\rho)f + \sum_{k=1}^{a} x_k^{-1}(\zeta_k^2 - \zeta_k + \tfrac{1}{6}) - \sum_{j=1}^{b} y_j^{-1}(\eta_j^2 - \eta_j + \tfrac{1}{6})\right]. \end{aligned}$$

Thus, ω_1 vanishes by choosing

$$\rho = 1 - f^{-1}\left[\sum_{k=1}^{a} x_k^{-1}(\zeta_k^2 - \zeta_k + \tfrac{1}{6}) - \sum_{j=1}^{b} y_j^{-1}(\eta_j^2 - \eta_j + \tfrac{1}{6})\right]. \tag{12.10}$$

Even though ρ now depends on x_k and y_j, assumed to be of order $O(n)$, the asymptotic expansion is still valid since for this choice of ρ, $\beta_k = (1-\rho)x_k$ and $\epsilon_j = (1-\rho)y_j$ are terms of order $O(1)$ and, thus, ω_α is still $O(n^{-\alpha})$.

Proposition 12.1 *If W has moments* (12.1)*, where x_k and y_j are terms $O(n)$, then with the choice of ρ in* (12.10)*,*

$$P(\rho M \le x) = G_f(x) + O(n^{-2}). \tag{12.11}$$

The asymptotic expansions of order $O(n^{-2})$ were proposed by Bartlett (1938). The factor ρ which annihilates the term ω_1 of order $O(n^{-1})$ is referred to as the Bartlett correction factor.

We now give the general result for an expansion of order $O(n^{-6})$ ($l = 5$) calculated with the help of Maple [Redfern (1996)]:

$$\begin{aligned}
P(\rho M \le x) &= G_f(x) + \omega_1[G_{f+2}(x) - G_f(x)] + \omega_2[G_{f+4}(x) - G_f(x)] \\
&\quad + \tfrac{1}{2}\omega_1^2[G_{f+4}(x) - 2G_{f+2}(x) + G_f(x)] + \omega_3[G_{f+6}(x) - G_f(x)] \\
&\quad + \omega_2\omega_1[G_{f+6}(x) - G_{f+4}(x) - G_{f+2}(x) + G_f(x)] \\
&\quad + \tfrac{1}{6}\omega_1^3[G_{f+6}(x) - 3G_{f+4}(x) + 3G_{f+2}(x) - G_f(x)] \\
&\quad + \omega_4[G_{f+8}(x) - G_f(x)] + \omega_3\omega_1[G_{f+8}(x) - G_{f+6}(x) - G_{f+2}(x) + G_f(x)] \\
&\quad + \tfrac{1}{2}\omega_2\omega_1^2[G_{f+8}(x) - 2G_{f+6}(x) + 2G_{f+2}(x) - G_f(x)] \\
&\quad + \tfrac{1}{24}\omega_1^4[G_{f+8}(x) - 4G_{f+6}(x) + 6G_{f+4}(x) - 4G_{f+2}(x)] + G_f(x)] \\
&\quad + \tfrac{1}{2}\omega_2^2[G_{f+8}(x) - 2G_{f+4}(x) + G_f(x)] + \omega_5[G_{f+10}(x) - G_f(x)] \\
&\quad + \omega_4\omega_1[G_{f+10}(x) - G_{f+8}(x) - G_{f+2}(x) + G_f(x)] \\
&\quad + \omega_3\omega_2[G_{f+10}(x) - G_{f+6}(x) - G_{f+4}(x) + G_f(x)] \\
&\quad + \tfrac{1}{2}\omega_3\omega_1^2[G_{f+10}(x) - 2G_{f+8}(x) + G_{f+6}(x) \\
&\qquad - G_{f+4}(x) + 2G_{f+2}(x) - G_f(x)] \\
&\quad + \tfrac{1}{6}\omega_2\omega_1^3[G_{f+10}(x) - 3G_{f+8}(x) + 2G_{f+6}(x) \\
&\qquad + 2G_{f+4}(x) - 3G_{f+2}(x) + G_f(x)] \\
&\quad + \tfrac{1}{120}\omega_1^5[G_{f+10}(x) - 5G_{f+8}(x) + 10G_{f+6}(x) \\
&\qquad - 10G_{f+4}(x) + 5G_{f+2}(x) - G_f(x)] \\
&\quad + \tfrac{1}{2}\omega_2^2\omega_1[G_{f+10}(x) - G_{f+8}(x) - 2G_{f+6}(x) \\
&\qquad + 2G_{f+4}(x) + G_{f+2}(x) - G_f(x)] \\
&\quad + O(n^{-6}).
\end{aligned}$$

When ρ is chosen as in (12.10) so that $\omega_1 = 0$, then things reduce considerably.

Proposition 12.2 *If W has moments* (12.1) *where x_k and y_j are terms $O(n)$, then with the choice of ρ in* (12.10)*,*

$$\begin{aligned}
P(\rho M \le x) &= G_f(x) + \omega_2[G_{f+4}(x) - G_f(x)] + \omega_3[G_{f+6}(x) - G_f(x)] \\
&\quad + \omega_4[G_{f+8}(x) - G_f(x)] + \tfrac{1}{2}\omega_2^2[G_{f+8}(x) - 2G_{f+4}(x) + G_f(x)] \\
&\quad + \omega_5[G_{f+10}(x) - G_f(x)] + \omega_3\omega_2[G_{f+10}(x) - G_{f+6}(x) - G_{f+4}(x) + G_f(x)] \\
&\quad + O(n^{-6}).
\end{aligned}$$

Automatic correction of coverage probability of confidence intervals [Martin (1990) or of error in rejection probability of tests [Beran (1987, 1988)] is now made possible by the resampling or "bootstrap" technology (v. Chapter 14). Bootstrap of a Bartlett corrected likelihood ratio test reduces level error in Proposition 12.1 from $O(n^{-2})$ to $O(n^{-3})$ automatically without further analytical expansions.

12.3 Examples

We now present some examples.

Example 12.1 Test of sphericity.
The likelihood ratio test (LRT) of sphericity was derived in Proposition 8.11 *and its moments are given in Proposition* 8.12. *Now, it is simply a matter of rewriting things in the form* (12.1) *to obtain the asymptotic expansion. Hence, for* $W = \Lambda^{m/n}$, $m = n - 1$,

$$\begin{aligned} E\ W^h &= Kp^{pmh/2} \frac{\Gamma_p[\frac{1}{2}m + \frac{1}{2}mh]}{\Gamma[\frac{1}{2}mp + \frac{1}{2}pmh]} \\ &= Kp^{pmh/2} \frac{\prod_{k=1}^{p} \Gamma[\frac{1}{2}m + \frac{1}{2}mh - \frac{1}{2}(k-1)]}{\Gamma[\frac{1}{2}mp + \frac{1}{2}pmh]} \\ &= Kp^{pmh/2} \frac{\prod_{k=1}^{p} \Gamma[\frac{1}{2}m(1+h) - \frac{1}{2}(k-1)]}{\Gamma[\frac{1}{2}mp(1+h)]}, \end{aligned}$$

so that we have the form (12.1) *with*

$$a = p, \quad x_k = \tfrac{1}{2}m, \quad \zeta_k = -\tfrac{1}{2}(k-1),$$

$$b = 1, \quad y_1 = \tfrac{1}{2}mp, \quad \eta_1 = 0.$$

Observe that $\sum_{k=1}^{p} x_k = y_1$ *is satisfied and* x_k *and* y_1 *are terms behaving as* $O(n)$. *The asymptotic expansion with remainder* $O(n^{-6})$ *as in Proposition* 12.2 *is now a simple matter of calculating with* (12.6) *and* (12.10),

$$\begin{aligned} f &= \tfrac{1}{2}(p+2)(p-1), \\ \rho &= 1 - \frac{2p^2 + p + 2}{6pm}, \end{aligned}$$

β_k *and* ϵ_1 *in* (12.2), *and, finally,* ω_α, $\alpha = 2, 3, 4, 5$, *in* (12.7). *Of course, one could go to great lengths to obtain the most simplified algebraic expressions in terms of* p *and* n. *For example, Davis* (1971) *using properties of Bernoulli polynomials showed*

$$\omega_\alpha = \frac{2(-1)^\alpha}{\alpha(\alpha+1)(\alpha+2)\rho^\alpha}$$

s	B_s	δ_s
0	1	$-\frac{1}{2}p$
1	$-\frac{1}{2}$	$\frac{1}{4}p(p+1)$
2	$\frac{1}{6}$	$-\frac{1}{16}p(2p^2+3p-1)$
3	0	$\frac{1}{16}p(p-1)(p+1)(p+2)$
4	$-\frac{1}{30}$	$-\frac{1}{192}p(6p^4+15p^3-10p^2-30p+3)$
5	0	$\frac{1}{128}p(p-1)(p+1)(p+2)(2p^2+2p-7)$
6	$\frac{1}{42}$	$-\frac{1}{768}p(6p^6+21p^5-21p^4-105p^3+21p^2+147p-5)$
7	0	$\frac{1}{768}p(p-1)(p+1)(p+2)(3p^4+6p^3-23p^2-26p+62)$

Table 12.1. Polynomials δ_s and Bernoulli numbers B_s for asymptotic expansions.

$$\cdot \sum_{s=1}^{\alpha+1} \binom{\alpha+2}{s+1} (1-\rho)^{\alpha+1-s} \left[\delta_s + \tfrac{1}{2}(s+1)\frac{B_s}{p^{s-1}}\right] (\tfrac{1}{2}m)^{1-s},$$

where $B_s \equiv B_s(0)$ *are the Bernoulli numbers and the* δ_s *are certain polynomials in* p *defined by Box* (1949) *(v. Table* 12.1*).*

Example 12.2 Asymptotics for $U(p;m,n)$ distributions.
The LRT for the general linear hypothesis in multivariate regression was described in Proposition 9.3 *as* $\Lambda^{2/n} \sim U(p;r,n-k)$*. Another example is the LRT for independence between two subvectors in Proposition* 11.2 *where* $\Lambda^{2/n} \sim U(p_2;p_1,n-1-p_1)$*.*

Thus, we derive the asymptotic expansion for $W \sim [U(p;m,n-c)]^{n/2}$*, where* $n-c \geq p$*, which includes both cases. Now, the moments of* U *distributions were given in Problem* 11.6.1*. Hence,*

$$\begin{aligned} E\ W^h &= E\ [U(p;m,n-c)]^{nh/2} \\ &= K\frac{\prod_{k=1}^{p}\Gamma[\frac{1}{2}(n-c-p+k)+\frac{1}{2}nh]}{\prod_{j=1}^{p}\Gamma[\frac{1}{2}(m+n-c-p+j)+\frac{1}{2}nh]} \\ &= K\frac{\prod_{k=1}^{p}\Gamma[\frac{1}{2}n(1+h)+\frac{1}{2}(-c-p+k)]}{\prod_{j=1}^{p}\Gamma[\frac{1}{2}n(1+h)+\frac{1}{2}(m-c-p+j)]}, \end{aligned}$$

which is of the form (12.1) *with*

$$a = p, \quad x_k = \tfrac{1}{2}n, \quad \zeta_k = \tfrac{1}{2}(-c - p + k),$$

$$b = p, \quad y_j = \tfrac{1}{2}n, \quad \eta_j = \tfrac{1}{2}(m - c - p + j).$$

Note, again, that x_k *and* y_j *are* $O(n)$ *and* $\sum_{k=1}^{p} x_k = \sum_{j=1}^{p} y_j$ *is satisfied. Using* (12.6) *and* (12.10)*, we have*

$$\begin{aligned} f &= pm, \\ \rho &= 1 - n^{-1}[c - \tfrac{1}{2}(m - p - 1)]. \end{aligned}$$

An interesting peculiarity in this case which derives from a symmetry property of Bernoulli polynomials [Erdélyi et al. (1953)*, p.* 37*], namely* $B_\alpha(1-h) = (-1)^\alpha B_\alpha(h)$*, is that* $\omega_{2\alpha-1} = 0$*,* $\alpha = 1, 2, \ldots$*, which means the series involves only terms of even powers of* n^{-1} *[Lee* (1972)*]. To see this, first note*

$$\begin{aligned} \beta_k = (1-\rho)x_k &= \tfrac{1}{2}[c - \tfrac{1}{2}(m - p - 1)], \\ \beta_k + \zeta_k &= -\tfrac{1}{2}[\tfrac{1}{2}(m + p - 1) - k], \end{aligned}$$

and, similarly,

$$\epsilon_k + \eta_k = \tfrac{1}{2}[\tfrac{1}{2}(m - p + 1) + k].$$

Therefore, we find (note that $k \mapsto p - k + 1$ *reverses the order of terms in the following sums)*

$$\begin{aligned} \sum_{k=1}^{p} B_{2\alpha}(\beta_k + \zeta_k) &= \sum_{k=1}^{p} B_{2\alpha}\left(-\tfrac{1}{2}[\tfrac{1}{2}(m + p - 1) - k]\right) \\ &= \sum_{k=1}^{p} B_{2\alpha}\left(-\tfrac{1}{2}[\tfrac{1}{2}(m + p - 1) - (p - k + 1)]\right) \\ &= \sum_{k=1}^{p} B_{2\alpha}\left(1 + \tfrac{1}{2}[\tfrac{1}{2}(m + p - 1) - (p - k + 1)]\right) \\ &= \sum_{k=1}^{p} B_{2\alpha}(\epsilon_k + \eta_k), \end{aligned}$$

and from (12.7)*,* $\omega_{2\alpha-1} = 0$*,* $\alpha = 1, 2, \ldots$*. Thus, the expansion in Proposition* 12.2 *further reduces to*

$$P(\rho M \le x) = G_f(x) + \omega_2[G_{f+4}(x) - G_f(x)] + \omega_4[G_{f+8}(x) - G_f(x)]$$

$$+\tfrac{1}{2}\omega_2^2[G_{f+8}(x) - 2G_{f+4}(x) + G_f(x)] + O(n^{-6}), \tag{12.12}$$

where from the same symmetry property of Bernoulli polynomials, one can easily establish that

$$\omega_{2\alpha} = \frac{2^{2\alpha}}{\alpha(2\alpha+1)(\rho n)^{2\alpha}} \sum_{k=1}^{p} B_{2\alpha+1}\left(\tfrac{1}{2}[\tfrac{1}{2}(m - p + 1) + k]\right), \ \alpha = 1, 2, \ldots.$$

	$n=2$	$n=5$	$n=10$	$n=15$	$n=20$	$n=30$
$O(n^{-2})$	.9500	.9500	.9500	.9500	.9500	.9500
$O(n^{-4})$	.8107	.8848	.9220	.9345	.9402	.9451
$O(n^{-6})$	.7168	.8642	.9182	.9334	.9397	.9449
exact	.4714	.8315	.9139	.9322	.9393	.9448

Table 12.2. Asymptotic expansions for $U(2;12,n)$ distributions.

A small-scale numerical evaluation of (12.12) would help to determine how large n should be for the asymptotics of $U(p;m,n)$ distributions to be accurate. Fix $p=2$ and $m=12$, and vary n =2, 5, 10, 15, 20, and 30. The asymptotic distribution of $-n\log U(2;12,n)$ is χ^2_{24}. So, we choose the critical point $x=\chi^2_{.95,24}=36.41502$. The evaluation of

$$P\left[-\left(n+\tfrac{1}{2}(12-2-1)\right)\log U(2;12,n)\le 36.41502\right]$$

using (12.12) led to Table 12.2.

The exact values were obtained for $p=2$ with the transformation in Problem 11.6.2:

$$\begin{aligned} P\left(-\rho n\log U(2;12,n)\le x\right) &= P\left(U(2;12,n)\ge e^{-x/\rho n}\right) \\ &= P\left(F\left(24,2(n-1)\right)\le \frac{(n-1)(1-y^{1/2})}{12y^{1/2}}\right), \end{aligned}$$

with $y=e^{-x/\rho n}$. The approximations of order $O(n^{-6})$ can thus be used in practice for n as small as 10 in this case. They are nearly exact to four decimal places for $n=30$.

Example 12.3 Test of mutual independence between subvectors. *This is a continuation of Problem* 11.6.7*, where in item (ii), we found the moments of* $\tilde{\Lambda}=\Lambda^{2/n}$*,*

$$E\ \tilde{\Lambda}^h=\frac{\Gamma_p(\frac{1}{2}m+h)}{\Gamma_p(\frac{1}{2}m)}\prod_{j=1}^{r}\frac{\Gamma_{p_j}(\frac{1}{2}m)}{\Gamma_{p_j}(\frac{1}{2}m+h)},$$

with $m=n-1$*. This can be written in the form of* (12.1) *for* $W=\Lambda$ *as*

$$E\ W^h=E\ \tilde{\Lambda}^{nh/2}=K\frac{\prod_{k=1}^{p}\Gamma[\frac{1}{2}n(1+h)-\frac{1}{2}k]}{\prod_{j=1}^{r}\prod_{l=1}^{p_j}\Gamma[\frac{1}{2}n(1+h)-\frac{1}{2}l]},$$

with the identification

$$a=p, \qquad x_k=\tfrac{1}{2}n, \quad \zeta_k=-\tfrac{1}{2}k,\ k=1,\ldots,a,$$

$$b=\textstyle\sum_{j=1}^{r}p_j=p, \quad y_{j_l}=\tfrac{1}{2}n, \quad \eta_{j_l}=-\tfrac{1}{2}l,\ j=1,\ldots,r,\ l=1,\ldots,p_j.$$

The constants f and ρ can be verified with (12.6) *and* (12.10)*:*

$$\begin{aligned} f &= -2\left[\sum_{k=1}^{p} -\tfrac{1}{2}k - \sum_{j=1}^{r}\sum_{l=1}^{p_j} -\tfrac{1}{2}l\right] = \tfrac{1}{2}\Sigma_2, \\ \rho &= 1 - \frac{2\Sigma_3 + 9\Sigma_2}{6n\Sigma_2}, \end{aligned}$$

where $\Sigma_s \equiv p^s - \sum_{j=1}^{r} p_j^s$. For simplified algebraic expressions of ω_2 through ω_6, the reader is referred to Box (1949)*.*

Example 12.4 Test of equality of variances.
The null moments of the modified likelihood ratio test Λ^* for the hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a$ were obtained in Proposition 8.17*. Thus, for $W = \Lambda^*$, we can write*

$$\begin{aligned} E\ W^h &= \frac{m^{pmh/2}}{\prod_{i=1}^{a} m_i^{pm_ih/2}} \frac{\Gamma_p(\frac{1}{2}m)}{\Gamma_p[\frac{1}{2}m(1+h)]} \prod_{i=1}^{a} \frac{\Gamma_p[\frac{1}{2}m_i(1+h)]}{\Gamma_p(\frac{1}{2}m_i)} \\ &= K\left[\frac{\prod_{i=1}^{a}(\frac{1}{2}m)^{m/2}}{\prod_{i=1}^{a}\prod_{l=1}^{p}(\frac{1}{2}m_i)^{m_i/2}}\right]^h \frac{\prod_{i=1}^{a}\prod_{l=1}^{p}\Gamma[\frac{1}{2}m_i(1+h) - \frac{1}{2}(l-1)]}{\prod_{j=1}^{p}\Gamma[\frac{1}{2}m(1+h) - \frac{1}{2}(j-1)]} \end{aligned}$$

which is of the form (12.1) *with the identification*

$$a = pa, \quad x_{k_l} = \tfrac{1}{2}m_k, \quad \zeta_{k_l} = -\tfrac{1}{2}(l-1),\ k = 1, \ldots, a,\ l = 1, \ldots, p,$$

$$b = p, \qquad y_j = \tfrac{1}{2}m, \qquad \eta_j = -\tfrac{1}{2}(j-1),\ j = 1, \ldots, p.$$

The degrees of freedom f and ρ in (12.6) *and* (12.10) *are*

$$\begin{aligned} f &= -2\left[\sum_{k=1}^{a}\sum_{l=1}^{p}\zeta_{k_l} - \sum_{j=1}^{p}\eta_j - \tfrac{1}{2}(pa-p)\right] \\ &= \tfrac{1}{2}p(p+1)(a-1), \\ \rho &= 1 - f^{-1}\left[\sum_{k=1}^{a}\sum_{l=1}^{p} x_{k_l}^{-1}(\zeta_{k_l}^2 - \zeta_{k_l} + \tfrac{1}{6}) - \sum_{j=1}^{p} y_j^{-1}(\eta_j^2 - \eta_j + \tfrac{1}{6})\right] \\ &= 1 - \frac{(2p^2+3p-1)}{6(p+1)(a-1)}\left(\sum_{k=1}^{a}\frac{1}{m_k} - \frac{1}{m}\right). \end{aligned}$$

Values of ω_α can be calculated from (12.7) *in simplified algebraic form but this is unnecessary since they can be easily programmed for the computer to evaluate the expansion in Proposition* 12.2*. Note finally that since we require $(1-\rho)x_{k_l}$ and $(1-\rho)y_j$ to remain bounded, the expansion is asymptotic as $m \to \infty$ while $m_k/m \to \alpha_k$ for some proportions $0 < \alpha_k < 1$ such that $\sum_{k=1}^{a}\alpha_k = 1$.*

The basic idea in asymptotic expansions was to represent the cumulant generating function in the form (12.5). This is often possible even though the moments may not be of the form (12.1).

Example 12.5 *An example is provided by the modified likelihood ratio test for a given variance in Problem* 8.9.8*, where*

$$E\ \Lambda^{*h} = \left(\frac{2e}{m}\right)^{mph/2} (1+h)^{-mp(1+h)/2} \frac{\Gamma_p[\frac{1}{2}m(1+h)]}{\Gamma_p(\frac{1}{2}m)},\ m = n-1.$$

For $W = \Lambda^*$*, Davis* (1971) *showed that Proposition* 12.2 *holds with*

$$\begin{aligned} f &= \tfrac{1}{2}p(p+1), \\ \rho &= 1 - \frac{2p^2+3p-1}{6m(p+1)}, \\ \omega_\alpha &= \frac{2(-1)^\alpha}{\alpha(\alpha+1)(\alpha+2)\rho^\alpha} \sum_{s=1}^{\alpha+1} \binom{\alpha+2}{s+1} (1-\rho)^{\alpha+1-s} \delta_s (\tfrac{1}{2}m)^{1-s},\ (\omega_1 = 0). \end{aligned}$$

The δ_s *are the same as those of Table* 12.1*.*

For asymptotic expansions of the null distribution of Lawley-Hotelling and Pillai trace tests, the reader is referred to Muirhead (1970) and Fujikoshi (1970).

12.4 Problem

1. This problem develops the asymptotic expansion of the LRT for the equality of means and variances

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_a;\ \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_a$$

between a multivariate normal populations. The LRT Λ, together with its moments, are given in Problem 8.9.14. Using the same notation and $W = \Lambda$, establish the following:

(i) The moments of W have the equivalent form

$$E\, W^h = K \left[\frac{\prod_{j=1}^p (\frac{1}{2}n)^{n/2}}{\prod_{k=1}^p \prod_{l=1}^a (\frac{1}{2}n_l)^{n_l/2}}\right]^h \frac{\prod_{k=1}^p \prod_{l=1}^a \Gamma[\frac{1}{2}n_l(1+h) - \frac{1}{2}k]}{\prod_{j=1}^p \Gamma[\frac{1}{2}n(1+h) - \frac{1}{2}j]}$$

(ii) Perform the usual identification to conclude the validity of Proposition 12.2 with

$$\begin{aligned} f &= \tfrac{1}{2}(a-1)p(p+3), \\ \rho &= 1 - \frac{(2p^2+9p+11)}{6(a-1)(p+3)} \left(\sum_{i=1}^a \frac{1}{n_i} - \frac{1}{n}\right). \end{aligned}$$

13
Robustness

13.1 Introduction

Many inference methods were presented in previous chapters for multivariate normal populations. A question of theoretical and utmost practical importance is the effect of non-normality on the inference. For example, what happens if the likelihood ratio test of sphericity, derived assuming normality, is performed, but, in fact, the population follows a multivariate student distribution on 10 degrees of freedom? Is the significance level of $\alpha = 5\%$, say, still close to 5%? The theory of robustness gives answers as to how sensitive multivariate normal inferences are to departures from normality. Most importantly, it proposes some remedies, i.e., more robust procedures. In Section 13.2, we present some non-normal models often used in robustness, the so-called elliptical distributions. The rest of the chapter is devoted to robust estimation and adjusted likelihood ratio tests.

A robust analysis of data is useful in several ways. It can validate or rebuff data analysis done on classical assumptions of multivariate normality. It also comes into play in the identification of outliers, which is a challenging task for data sets with more than two variables. Robust estimates of location vector and scale matrix serve this role admirably. They can be used to evaluate robust Mahalanobis distances from an observation vector $\mathbf{x}_i$ to the location vector. Points with large Mahalanobis distances can then be singled out and scrutinized.

13.2 Elliptical distributions

Suppose that $\mathbf{x} \in \mathbb{R}^p$ has a density

$$f_{\mathbf{x}}(\mathbf{x}) = |\mathbf{\Lambda}|^{-1/2} g[(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})],$$

where $g : [0,\infty) \to [0,\infty)$ is a fixed function independent of $\boldsymbol{\mu}$ and $\mathbf{\Lambda} = (\Lambda_{ij})$ and depends on $\mathbf{x}$ only through $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. Denote this *elliptical distribution* by $\mathbf{x} \sim E_p(\boldsymbol{\mu}, \mathbf{\Lambda})$. The main reference for elliptical distributions is Kelker (1970). The affine linear transformation $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{b}$ with $\mathbf{B} \in \mathbf{G}_p$ and $\mathbf{b} \in \mathbb{R}^p$ has density

$$f_{\mathbf{y}}(\mathbf{y}) = |\mathbf{B}\mathbf{\Lambda}\mathbf{B}'|^{-1/2} g[(\mathbf{y}-\mathbf{B}\boldsymbol{\mu}-\mathbf{b})'(\mathbf{B}\mathbf{\Lambda}\mathbf{B}')^{-1}(\mathbf{y}-\mathbf{B}\boldsymbol{\mu}-\mathbf{b})].$$

Thus, $\mathbf{y} \sim E_p(\mathbf{B}\boldsymbol{\mu}+\mathbf{b}, \mathbf{B}\mathbf{\Lambda}\mathbf{B}')$; i.e., the transformation $\mathbf{x} \mapsto \mathbf{B}\mathbf{x}+\mathbf{b}$ induces the parameter transformation $\boldsymbol{\mu} \mapsto \mathbf{B}\boldsymbol{\mu}+\mathbf{b}$ and $\mathbf{\Lambda} \mapsto \mathbf{B}\mathbf{\Lambda}\mathbf{B}'$. In particular, $\mathbf{z} = \mathbf{\Lambda}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}) \sim E_p(\mathbf{0}, \mathbf{I})$ has a spherical or rotationally invariant distribution. Elliptical distributions are a location scale generalization of spherical distributions. Thus, for example, if $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$ with characteristic function necessarily of the form $c_{\mathbf{z}}(\mathbf{t}) = \phi(\mathbf{t}'\mathbf{t})$ (v. Problem 4.6.6), then $\mathbf{x} = \mathbf{\Lambda}^{1/2}\mathbf{z} + \boldsymbol{\mu} \sim E_p(\boldsymbol{\mu}, \mathbf{\Lambda})$ has characteristic function $c_{\mathbf{x}}(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu})\phi(\mathbf{t}'\mathbf{\Lambda}\mathbf{t})$. Moreover, if $\mathbf{z}$ has a finite second moment, $E\ \mathbf{z} = \mathbf{0}$ and $\text{var}\ \mathbf{z} = \alpha\mathbf{I}$, for some constant α, implies $E\ \mathbf{x} = \boldsymbol{\mu}$ and $\text{var}\ \mathbf{x} \equiv \mathbf{\Sigma} = \alpha\mathbf{\Lambda}$. An important implication is that all elliptical distributions with finite second moments have the same correlation matrix. The constant $\alpha = -2\phi'(0)$ (v. Problem 4.6.15) is easily found by differentiation of $c_{\mathbf{z}}(\mathbf{t})$.

Examples of spherical distributions commonly used in robustness are members of the normal mixture family with density

$$f_{\mathbf{x}}(\mathbf{x}) = \int_0^\infty (2\pi w)^{-p/2} \exp(-\tfrac{1}{2}w^{-1}\mathbf{x}'\mathbf{x}) dF(w),$$

where $F(\cdot)$ is the "mixing" distribution function on $[0,\infty)$. These can be simulated easily using the representation $\mathbf{x} \stackrel{d}{=} w^{1/2}\mathbf{z}$, where $w \sim F$, $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$, $w \perp\!\!\!\perp \mathbf{z}$ (v. Problem 13.6.1).

Example 13.1 *Obviously,* $P(w = \sigma^2) = 1$ *yields the* $N_p(\mathbf{0}, \sigma^2\mathbf{I})$ *distribution.*

Example 13.2 *The two-point distribution,*

$$\begin{aligned} P(w=1) &= 1-\epsilon, \\ P(w=\sigma^2) &= \epsilon \end{aligned}$$

for some "contamination" proportion $0 < \epsilon < 1$*, yields the symmetric contaminated normal distribution.*

Example 13.3 *The multivariate* t *on* ν *degrees of freedom denoted* $t_{p,\nu}$ *is obtained with* $\nu w^{-1} \sim \chi^2_\nu$*. The reader is asked to show in Problem* 13.6.1

that $\mathbf{x}$ *has density*

$$f_{\mathbf{x}}(\mathbf{x}) = c_{p,\nu}(1 + \mathbf{x}'\mathbf{x}/\nu)^{-(\nu+p)/2},\ \mathbf{x} \in \mathbb{R}^p,$$

where $c_{p,\nu} = (\nu\pi)^{-p/2}\Gamma\left[\frac{1}{2}(\nu+p)\right]/\Gamma\left(\frac{1}{2}\nu\right)$. *The general multivariate* $t_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ *is obtained by relocating and rescaling,* $\mathbf{y} = \boldsymbol{\Lambda}^{1/2}\mathbf{x} + \boldsymbol{\mu}$, *and has density*

$$f_{\mathbf{y}}(\mathbf{y}) = c_{p,\nu}|\boldsymbol{\Lambda}|^{-1/2}\left[1 + (\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\mathbf{y}-\boldsymbol{\mu})/\nu\right]^{-(\nu+p)/2},\ \mathbf{y} \in \mathbb{R}^p.$$

The multivariate t on 1 *degree of freedom is also known as the multivariate Cauchy distribution.*

The Kotz-type distributions form another important class of elliptical distributions [Fang et al. (1991), p. 76]. Their characteristic function was obtained recently by Kotz and Ostrovskii (1994). Elliptical distributions that can be expanded as a power series are defined in Steyn (1993) and used to define other nonelliptical distributions with heterogeneous kurtosis.

The following result gives the marginal and conditional distributions for an $E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ distribution. Let $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ with $\mathbf{x}_i \in \mathbb{R}^{p_i}$, $i = 1, 2$, $p = p_1 + p_2$, and partition $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ in conformity as

$$\begin{aligned} \boldsymbol{\mu} &= (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')', \\ \boldsymbol{\Lambda} &= \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix}. \end{aligned}$$

Proposition 13.1 *The marginal and conditional distributions of an* $E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ *distribution are elliptical:*

(i) $\mathbf{x}_2 \sim E_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Lambda}_{22})$,

(ii) $\mathbf{x}_1|\mathbf{x}_2 \sim E_{p_1}(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Lambda}_{11.2})$, *where*

$$\begin{aligned} \boldsymbol{\mu}_{1.2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \\ \boldsymbol{\Lambda}_{11.2} &= \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}. \end{aligned}$$

The conditional variance is of the form $\mathrm{var}(\mathbf{x}_1|\mathbf{x}_2) = w(\mathbf{x}_2)\boldsymbol{\Lambda}_{11.2}$, *for some function* $w(\mathbf{x}_2) \in \mathbb{R}$ *which depends on* $\mathbf{x}_2$ *only through the quadratic form*

$$(\mathbf{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2).$$

Proof. Letting $\mathbf{t} = (\mathbf{0}', \mathbf{t}_2')'$ in $c_{\mathbf{x}}(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu})\phi(\mathbf{t}'\boldsymbol{\Lambda}\mathbf{t})$, we find $c_{\mathbf{x}_2}(\mathbf{t}_2) = \exp(i\mathbf{t}_2'\boldsymbol{\mu}_2)\phi(\mathbf{t}_2'\boldsymbol{\Lambda}_{22}\mathbf{t}_2)$ and, thus, $\mathbf{x}_2 \sim E_{p_2}(\boldsymbol{\mu}_2, \boldsymbol{\Lambda}_{22})$. For the conditional distribution, let

$$\mathbf{z} = \mathbf{x}_1 - [\boldsymbol{\mu}_1 + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)]$$

with jacobian $J(\mathbf{x} \to \mathbf{z}, \mathbf{x}_2) = 1$. Upon using Problem 1.8.2, the conditional density $\mathbf{z}|\mathbf{x}_2$ is

$$\frac{|\boldsymbol{\Lambda}|^{-1/2}g[\mathbf{z}'\boldsymbol{\Lambda}_{11.2}^{-1}\mathbf{z} + (\mathbf{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)]}{|\boldsymbol{\Lambda}_{22}|^{-1/2}f_{\mathbf{x}_2}(\mathbf{x}_2)},$$

where $f_{\mathbf{x}_2}(\mathbf{x}_2)$ depends only on $(\mathbf{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$. Thus, we have $\mathbf{z}|\mathbf{x}_2 \sim E_{p_1}(\mathbf{0}, \boldsymbol{\Lambda}_{11.2})$ and, in turn, $\mathbf{x}_1|\mathbf{x}_2 \sim E_{p_1}(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Lambda}_{11.2})$. □

Example 13.4 *The univariate power exponential distribution has p.d.f.*

$$f_x(x) = c_{1,\alpha}\Lambda^{-1/2} \exp\left(-\tfrac{1}{2}\left|\frac{x-\mu}{\Lambda^{1/2}}\right|^{2\alpha}\right), \ \alpha > 0. \tag{13.1}$$

A multivariate extension seems to be

$$f_{\mathbf{x}}(\mathbf{x}) = c_{p,\alpha}|\boldsymbol{\Lambda}|^{-1/2} \exp\left\{-\tfrac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{\alpha}\right\}. \tag{13.2}$$

This elliptical distribution has an advantage of generating distributions with heavier and lighter tails than the multivariate normal by taking $\alpha < 1$ or $\alpha > 1$, whereas many other elliptical distributions including the multivariate t cannot generate lighter-tail distributions. Kuwana and Kariya (1991) *used this property to derive a locally best invariant test of multivariate normality ($\alpha = 1$). Taking $\alpha = 0.5$ simply, in* (13.2),

$$E\ (x_1 - \mu_1)^2 = 4(p+1)\Lambda_{11},$$

which depends on p (v. Problem 13.6.5*); the corresponding moment in* (13.1) *is*

$$E\ (x-\mu)^2 = 8\Lambda_{11}, \ \textit{with}\ \Lambda = \Lambda_{11}.$$

So, the marginal distribution of x_1 in (13.2) *is not that of x in* (13.1). *The inconsistency takes place for many other elliptical distributions. Kano* (1994) *characterized the consistency property of elliptical distributions: An elliptical family is consistent iff it is a normal mixture family. In particular, the multivariate normal and multivariate t families are consistent. In Proposition* 13.1 *the marginal is elliptical but possibly of a different functional form since the characteristic function ϕ may be related to p.*

For the estimation of $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, it seems natural to ask that location and scatter estimates transform in exactly the same manner as the parameters; i.e., that they be "affine equivariant" as described in the following definition. Formally, let

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \in \mathbb{R}_p^n$$

be the sample matrix.

Definition 13.1 *The location and scatter estimates $\hat{\boldsymbol{\mu}}(\mathbf{X})$ and $\hat{\boldsymbol{\Lambda}}(\mathbf{X})$ are* affine equivariant *iff for all $\mathbf{B} \in \mathbf{G}_p$ and $\mathbf{b} \in \mathbb{R}^p$,*

$$\begin{aligned} \hat{\boldsymbol{\mu}}(\mathbf{X}\mathbf{B}' + \mathbf{1}\mathbf{b}') &= \mathbf{B}\hat{\boldsymbol{\mu}}(\mathbf{X}) + \mathbf{b}, \\ \hat{\boldsymbol{\Lambda}}(\mathbf{X}\mathbf{B}' + \mathbf{1}\mathbf{b}') &= \mathbf{B}\hat{\boldsymbol{\Lambda}}(\mathbf{X})\mathbf{B}'. \end{aligned}$$

When the underlying distribution belongs to an elliptical family, the distribution of affine equivariant estimates has a special structure. In particular, the general form of the mean and variance estimates can be characterized for finite samples. To establish this general form, we need to extend the notion of rotational invariance of random vectors in Section 4.4 to symmetric random matrices.

Definition 13.2 *A random symmetric matrix* $\mathbf{W}$ *is* rotationally invariant *iff* $\mathbf{W} \stackrel{d}{=} \mathbf{HWH}'$, $\forall \mathbf{H} \in \mathbf{O}_p$.

The following lemma [Tyler (1982)] characterizes the general form of the mean and variance of any rotationally invariant random matrix.

Proposition 13.2 *Let* $\mathbf{W} \in \mathbb{R}_p^p$ *symmetric be rotationally invariant with finite second moments. Then, there exist constants* η, $\sigma_1 \geq 0$, *and* $\sigma_2 \geq -2\sigma_1/p$ *such that*

$$\begin{aligned} E\ \mathbf{W} &= \eta \mathbf{I}, \\ \text{var}\ \mathbf{W} &= \sigma_1(\mathbf{I} + \mathbf{K}_p) + \sigma_2\ \text{vec}(\mathbf{I})[\text{vec}(\mathbf{I})]'. \end{aligned}$$

Proof. For the mean, let $E\ \mathbf{W} \equiv \mathbf{A}$. By rotational invariance, $\mathbf{A} = \mathbf{HAH}'$, $\forall \mathbf{H} \in \mathbf{O}_p$. Hence,

$$\mathbf{x}'\mathbf{Ax} = \mathbf{x}'\mathbf{HAH}'\mathbf{x} = \mathbf{y}'\mathbf{Ay},\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p,\ |\mathbf{x}| = |\mathbf{y}| = 1.$$

Choosing $\mathbf{x} = \mathbf{h}_i$ and $\mathbf{y} = \mathbf{h}_j$, the ith and jth eigenvectors of $\mathbf{A}$ corresponding to eigenvalues λ_i and λ_j, respectively, we get $\lambda_i = \lambda_j \equiv \eta$ (say). This means $\mathbf{A} = \eta\mathbf{I}$. For the variance, let

$$\mathbf{\Omega} \equiv \text{var}\ \mathbf{W} = \sum \Omega_{ijkl}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{e}_k\mathbf{e}_l',$$

where $\text{cov}(w_{ki}, w_{lj}) = \Omega_{ijkl}$. Note that $\{\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{e}_k\mathbf{e}_l', i, j, k, l = 1, \ldots, p\}$ forms a basis for $\mathbb{R}_{p^2}^{p^2}$. Since $\mathbf{W} \stackrel{d}{=} \mathbf{HWH}'$, $\forall \mathbf{H} \in \mathbf{O}_p$, then $\text{vec}(\mathbf{W}) \stackrel{d}{=} (\mathbf{H} \otimes \mathbf{H})\text{vec}(\mathbf{W})$ and, thus, $\mathbf{\Omega} = (\mathbf{H} \otimes \mathbf{H})\mathbf{\Omega}(\mathbf{H}' \otimes \mathbf{H}')$, or

$$\sum \Omega_{ijkl}\mathbf{h}_i\mathbf{h}_j' \otimes \mathbf{h}_k\mathbf{h}_l' = \sum \Omega_{ijkl}\mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{e}_k\mathbf{e}_l',$$

where $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_p)$. By choosing for some m, $\mathbf{h}_m = -\mathbf{e}_m$ and $\mathbf{h}_r = \mathbf{e}_r$, $r \neq m$, we obtain $\Omega_{ijkl} = 0$ unless $i = j = k = l$, $i = j$ and $k = l$, $i = k$ and $j = l$, or $i = l$ and $j = k$. By choosing $\mathbf{H}$ to give a permutation of the rows, we obtain $\Omega_{iiii} = \sigma_0$, $\forall i = 1, \ldots, p$, $\Omega_{iikk} = \sigma_1$ for $i \neq k$, $\Omega_{ijij} = \sigma_2$ for $i \neq j$, and $\Omega_{ijji} = \sigma_3$ for $i \neq j$. Thus,

$$\begin{aligned} \mathbf{\Omega} &= \sigma_0 \left(\sum_i \mathbf{e}_i\mathbf{e}_i' \otimes \mathbf{e}_i\mathbf{e}_i' \right) + \sigma_1 \left(\sum_{i \neq k} \mathbf{e}_i\mathbf{e}_i' \otimes \mathbf{e}_k\mathbf{e}_k' \right) \\ &\quad + \sigma_2 \left(\sum_{i \neq j} \mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{e}_i\mathbf{e}_j' \right) + \sigma_3 \left(\sum_{i \neq j} \mathbf{e}_i\mathbf{e}_j' \otimes \mathbf{e}_j\mathbf{e}_i' \right) \end{aligned}$$

$$\begin{aligned} &= \sigma_1 \mathbf{I} + \sigma_2 \text{ vec}(\mathbf{I})[\text{vec}(\mathbf{I})]' + \sigma_3 \mathbf{K}_p \\ &\quad + (\sigma_0 - \sigma_1 - \sigma_2 - \sigma_3) \left(\sum_i \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{e}_i \mathbf{e}_i' \right). \end{aligned}$$

Since $\forall \mathbf{H} \in \mathbf{O}_p$,

$$\begin{aligned} (\mathbf{H} \otimes \mathbf{H}) \mathbf{I} (\mathbf{H}' \otimes \mathbf{H}') &= \mathbf{I}, \\ (\mathbf{H} \otimes \mathbf{H}) \text{vec}(\mathbf{I})[\text{vec}(\mathbf{I})]' (\mathbf{H}' \otimes \mathbf{H}') &= \text{vec}(\mathbf{I})[\text{vec}(\mathbf{I})]', \\ (\mathbf{H} \otimes \mathbf{H}) \mathbf{K}_p (\mathbf{H}' \otimes \mathbf{H}') &= \mathbf{K}_p, \end{aligned}$$

and

$$(\mathbf{H} \otimes \mathbf{H}) \left(\sum_i \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{e}_i \mathbf{e}_i' \right) (\mathbf{H}' \otimes \mathbf{H}') \neq \left(\sum_i \mathbf{e}_i \mathbf{e}_i' \otimes \mathbf{e}_i \mathbf{e}_i' \right),$$

for some $\mathbf{H} \in \mathbf{O}_p$, it follows that $\sigma_0 - \sigma_1 - \sigma_2 - \sigma_3 = 0$. Also, since $\mathbf{W}$ is symmetric, $\text{cov}(w_{ij}, w_{ji}) = \text{var } w_{ij}$, which implies $\sigma_1 = \sigma_3$. Therefore,

$$\mathbf{\Omega} = \sigma_1 (\mathbf{I} + \mathbf{K}_p) + \sigma_2 \text{vec}(\mathbf{I})[\text{vec}(\mathbf{I})]'.$$

The conditions on σ_1 and σ_2 follow since $\mathbf{\Omega}$ is positive semidefinite. □

The variance of $\mathbf{W} = (w_{ij})$ can be written componentwise with the Kronecker delta

$$\text{cov}(w_{ki}, w_{lj}) = \sigma_1 (\delta_{ij} \delta_{kl} + \delta_{kj} \delta_{il}) + \sigma_2 \delta_{ki} \delta_{lj}.$$

The form of var $\mathbf{W}$ states that the off-diagonal elements of $\mathbf{W}$ are uncorrelated with each other and uncorrelated with the diagonal elements. Each off-diagonal element has variance σ_1. The diagonal elements all have variance $2\sigma_1 + \sigma_2$ with the covariance between any two diagonal elements being σ_2.

Example 13.5 *A simple example is* $\mathbf{W} \sim W_p(m)$ *which is rotationally invariant with* var $\mathbf{W} = m(\mathbf{I} + \mathbf{K_p})$.

Example 13.6 *Assume* $\mathbf{x} \sim E_p(\mathbf{0}, \mathbf{\Lambda})$ *and let* $\mathbf{W} = \mathbf{x}\mathbf{x}'$. *Then*

$$\text{var } \mathbf{W} = (\mathbf{\Lambda}^{1/2} \otimes \mathbf{\Lambda}^{1/2}) \text{var}(\mathbf{z}\mathbf{z}') (\mathbf{\Lambda}^{1/2} \otimes \mathbf{\Lambda}^{1/2}),$$

where $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$. *Using Proposition* 13.2 var$(\mathbf{z}\mathbf{z}')$ *is evaluated with*

$$\begin{aligned} \sigma_1 &= \text{var}(z_1 z_2) = E(z_1^2 z_2^2) = \mu_{22}, \\ \sigma_2 &= \text{cov}(z_1^2, z_2^2) = E(z_1^2 z_2^2) - E(z_1^2) E(z_2^2) = \mu_{22} - \mu_2^2. \end{aligned}$$

In terms of cumulants we have $\sigma_1 = k_{22} + k_2^2$ *and* $\sigma_2 = k_{22}$. *These cumulants are easily found with the Taylor series*

$$\begin{aligned} \ln \phi(t_1^2 + t_2^2) &= k_2 \frac{(it_1)^2}{2!} + k_2 \frac{(it_2)^2}{2!} + k_4 \frac{(it_1)^4}{4!} + k_4 \frac{(it_2)^4}{4!} \\ &\quad + k_{22} \frac{(it_1)^2}{2!} \frac{(it_2)^2}{2!} + o(|\mathbf{t}|^4). \end{aligned}$$

The reader can verify by differentiation (v. Problem 13.6.3*)*

$$\begin{aligned} k_2 &= -2\phi'(0) = \alpha, \\ k_4 &= 12(\phi''(0) - \phi'(0)^2), \\ k_{22} &= 4(\phi''(0) - \phi'(0)^2). \end{aligned}$$

The kurtosis of z_1 *is*

$$\frac{k_4}{k_2^2} = 3\frac{(\phi''(0) - \phi'(0)^2)}{\phi'(0)^2} \equiv 3k,$$

where k *represents a kurtosis parameter. Thus,* $k_4 = 3k\alpha^2$ *and* $k_{22} = k\alpha^2$. *Finally, we obtain* $\sigma_1 = (1+k)\alpha^2$ *and* $\sigma_2 = k\alpha^2$ *from which*

$$\text{var}(\mathbf{z}\mathbf{z}') = \alpha^2(1+k)(\mathbf{I} + \mathbf{K}_p) + \alpha^2 k \text{ vec}(\mathbf{I})[\text{vec}(\mathbf{I})]'$$

and

$$\text{var } \mathbf{W} = (1+k)(\mathbf{I} + \mathbf{K}_p)(\mathbf{\Sigma} \otimes \mathbf{\Sigma}) + k \text{ vec}(\mathbf{\Sigma})[\text{vec}(\mathbf{\Sigma})]',$$

where $\mathbf{\Sigma} = \alpha\mathbf{\Lambda}$ *is the variance of* $\mathbf{x}$.

Corollary 13.1 *If* $\hat{\boldsymbol{\mu}}(\mathbf{X})$ *and* $\hat{\mathbf{\Lambda}}(\mathbf{X})$ *are affine equivariant with finite second moment and* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *are i.i.d.* $E_p(\boldsymbol{\mu}, \mathbf{\Lambda})$, *then there exist constants* η, $\beta \geq 0$, $\sigma_1 \geq 0$ *and* $\sigma_2 \geq -2\sigma_1/p$ *such that*

$$\begin{aligned} E\ \hat{\boldsymbol{\mu}}(\mathbf{X}) &= \boldsymbol{\mu}, \\ \text{var } \hat{\boldsymbol{\mu}}(\mathbf{X}) &= \beta\mathbf{\Lambda}, \\ E\ \hat{\mathbf{\Lambda}}(\mathbf{X}) &= \eta\mathbf{\Lambda}, \\ \text{var } \hat{\mathbf{\Lambda}}(\mathbf{X}) &= \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{\Lambda} \otimes \mathbf{\Lambda}) + \sigma_2 \text{ vec}(\mathbf{\Lambda})[\text{vec}(\mathbf{\Lambda})]'. \end{aligned}$$

Proof. First, $\mathbf{X} \stackrel{d}{=} \mathbf{Z}\mathbf{\Lambda}^{1/2} + \mathbf{1}\boldsymbol{\mu}'$, where

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_n' \end{pmatrix}$$

and $\mathbf{z}_i$'s are i.i.d. $E_p(\mathbf{0}, \mathbf{I})$. Hence, $\hat{\boldsymbol{\mu}}(\mathbf{X}) \stackrel{d}{=} \hat{\boldsymbol{\mu}}(\mathbf{Z}\mathbf{\Lambda}^{1/2} + \mathbf{1}\boldsymbol{\mu}') = \mathbf{\Lambda}^{1/2}\hat{\boldsymbol{\mu}}(\mathbf{Z}) + \boldsymbol{\mu}$. Obviously, $\hat{\boldsymbol{\mu}}(\mathbf{Z})$ is a rotationally invariant random vector. Using the result of Section 4.5, $E\ \hat{\boldsymbol{\mu}}(\mathbf{Z}) = 0$ and var $\hat{\boldsymbol{\mu}}(\mathbf{Z}) = \beta\mathbf{I}$, for some $\beta \geq 0$. Therefore, $E\ \hat{\boldsymbol{\mu}}(\mathbf{X}) = \boldsymbol{\mu}$ and var $\hat{\boldsymbol{\mu}}(\mathbf{X}) = \beta\mathbf{\Lambda}$. Similarly, $\hat{\mathbf{\Lambda}}(\mathbf{X}) \stackrel{d}{=} \mathbf{\Lambda}^{1/2}\hat{\mathbf{\Lambda}}(\mathbf{Z})\mathbf{\Lambda}^{1/2}$, where $\hat{\mathbf{\Lambda}}(\mathbf{Z})$ is a rotationally invariant matrix whose mean and variance have the general form in Proposition 13.2. Hence, $E\ \hat{\mathbf{\Lambda}}(\mathbf{X}) = \eta\mathbf{\Lambda}$, for some η, and

$$\begin{aligned} \text{var } \hat{\mathbf{\Lambda}}(\mathbf{X}) &= \text{var}\left[(\mathbf{\Lambda}^{1/2} \otimes \mathbf{\Lambda}^{1/2})\text{vec}(\hat{\mathbf{\Lambda}}(\mathbf{Z}))\right] \\ &= (\mathbf{\Lambda}^{1/2} \otimes \mathbf{\Lambda}^{1/2})\left[\text{var } \hat{\mathbf{\Lambda}}(\mathbf{Z})\right](\mathbf{\Lambda}^{1/2} \otimes \mathbf{\Lambda}^{1/2}) \\ &= \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{\Lambda} \otimes \mathbf{\Lambda}) + \sigma_2 \text{ vec}(\mathbf{\Lambda})[\text{vec}(\mathbf{\Lambda})]' \end{aligned}$$

for some $\sigma_1 \geq 0$, $\sigma_2 \geq -2\sigma_1/p$. □

Complicated expressions using tensor methods for third-order and fourth-order cumulants of affine equivariant estimates in elliptical families were obtained by Grübel and Rocke (1990).

Another way of writing var $\hat{\mathbf{\Lambda}}(\mathbf{X})$ is to give the covariances between any two elements of $\hat{\mathbf{\Lambda}}(\mathbf{X}) = (\hat{\Lambda}_{ij})$:

$$\operatorname{cov}(\hat{\Lambda}_{ki}, \hat{\Lambda}_{lj}) = \sigma_1(\Lambda_{ij}\Lambda_{kl} + \Lambda_{kj}\Lambda_{il}) + \sigma_2\Lambda_{ki}\Lambda_{lj}.$$

One should note that a reasonable estimate of $\mathbf{\Lambda}$ assumed positive definite should satisfy $\hat{\mathbf{\Lambda}}(\mathbf{X}) > \mathbf{0}$ w.p.1, and in that case, $\eta > 0$.

13.3 Maximum likelihood estimates

Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x} \sim E_p(\boldsymbol{\mu}, \mathbf{\Lambda})$ with var $\mathbf{x} = \alpha\mathbf{\Lambda} = \mathbf{\Sigma}$. The simplest but inefficient method to estimate $(\boldsymbol{\mu}, \mathbf{\Lambda})$ would be to use the MLE under a $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution, $\bar{\mathbf{x}}$ and $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. A more efficient procedure would be the MLE under the "true" $E_p(\boldsymbol{\mu}, \mathbf{\Lambda})$ model. These two possibilities are now investigated.

13.3.1 Normal MLE

When $\mathbf{x}$ has finite fourth-order moments, the general discussion of Section 6.3 showed that

$$n^{1/2}\begin{pmatrix} \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \\ \bar{\mathbf{x}}' - \boldsymbol{\mu}' \end{pmatrix} \xrightarrow{d} N_p^{p+1}\left(\mathbf{0}, \begin{pmatrix} \operatorname{var}\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} \end{pmatrix}\right),$$

where $\mathbf{W} = \mathbf{x}\mathbf{x}'$. From the calculation of var $\mathbf{W}$ in Example 13.6, it follows that

$$n^{1/2}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}, \bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{d} (\mathbf{N}, \mathbf{n}),$$

where $\mathbf{n} \perp\!\!\!\perp \mathbf{N}$,

$$\begin{aligned} \mathbf{n} &\sim N_p(\mathbf{0}, \mathbf{\Sigma}) \\ \mathbf{N} &\sim N_p^p(\mathbf{0}, (1+k)(\mathbf{I} + \mathbf{K}_p)(\mathbf{\Sigma} \otimes \mathbf{\Sigma}) + k\ \operatorname{vec}(\mathbf{\Sigma})[\operatorname{vec}(\mathbf{\Sigma})]'). \end{aligned}$$

13.3.2 Elliptical MLE

For $\mathbf{x} \sim E_p(\boldsymbol{\mu}, \mathbf{\Lambda})$ defined with a known function $g(\cdot)$ the *log-likelihood* for $(\boldsymbol{\mu}, \mathbf{\Lambda})$ is simply

$$l_n(\boldsymbol{\mu}, \mathbf{\Lambda}) = cte + \sum_{i=1}^n \ln g[(\mathbf{x}_i - \boldsymbol{\mu})'\mathbf{\Lambda}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})] - \tfrac{1}{2}n \ln|\mathbf{\Lambda}|. \qquad (13.3)$$

Differention with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ (v. Problems 1.8.9 and 1.8.10) leads to the equations

$$\begin{aligned} \sum_{i=1}^{n} u(s_i)\hat{\boldsymbol{\Lambda}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) &= \mathbf{0} \\ \tfrac{1}{2}\sum_{i=1}^{n} u(s_i)\hat{\boldsymbol{\Lambda}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Lambda}}^{-1} - \tfrac{1}{2}n\hat{\boldsymbol{\Lambda}}^{-1} &= \mathbf{0}, \end{aligned}$$

where $u(s) = -2g'(s)/g(s)$ and $s_i = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Lambda}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$. Thus, the MLE satisfies the implicit (because s_i depends on $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}})$) estimating equations

$$\hat{\boldsymbol{\mu}} = \text{ave}\left[u(s_i)\mathbf{x}_i\right] / \text{ave}\left[u(s_i)\right], \tag{13.4}$$

$$\hat{\boldsymbol{\Lambda}} = \text{ave}\left[u(s_i)(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\right]. \tag{13.5}$$

The notation "ave" means arithmetic average over $i = 1, \ldots, n$.

Example 13.7 *The multivariate Student's $t_{p,\nu}$ has $g(s) \propto (1+s/\nu)^{-(\nu+p)/2}$ and $u(s) = (\nu+p)/(\nu+s)$. Note that $u(s) \geq 0$ and is strictly decreasing. It acts as a weight function, giving more weight to data points with small squared Mahalanobis distances.*

The existence and unicity of a solution to the estimating equations is a difficult problem. For the location-only problem, it is known in the univariate case [Reeds (1985)] that the estimating equation is susceptible to multiple solutions. Uniqueness of the solution in the univariate location-scale Cauchy ($\nu = 1$) problem was established by Copas (1975) and for $\nu > 1$ by Märkeläinen et al. (1981). The approach presented here is that of Kent and Tyler (1991), which works equally well in the multivariate case. The location-scale problem is very tricky, but the scale-only problem is quite simple. We will thus concentrate on the latter problem.

Scale-only problem

For the scale-only problem, we assume without any loss of generality that $\boldsymbol{\mu} = \mathbf{0}$. The log-likelihood reduces to

$$l(\mathbf{A}) = cte + \sum_{i=1}^{n} \ln g(\mathbf{x}_i'\mathbf{A}^{-1}\mathbf{x}_i) - \tfrac{1}{2}n \ln |\mathbf{A}|$$

and the estimating equation simplifies to

$$\hat{\mathbf{A}} = \text{ave}\left[u(s_i)\mathbf{x}_i\mathbf{x}_i'\right], \tag{13.6}$$

where $u(\cdot)$ is as before and $s_i = \mathbf{x}_i'\hat{\mathbf{A}}^{-1}\mathbf{x}_i$. Let $\psi(s) = su(s)$ and assume that

$$\lim_{s\to\infty} \psi(s) = a_0 > 0.$$

This condition is satisfied for the $t_{p,\nu}$ distribution, as $\lim_{s\to\infty}\psi(s) = \nu + p$. The following condition on the data is to ensure the existence of a solution to (13.6). It specifies that the data points should not be too concentrated in low-dimensional linear subspaces of $\mathbb{R}^p$. Let $P_n(\cdot)$ denote the empirical distribution of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, i.e., for any borel set $B \subset \mathbb{R}^p$

$$P_n(B) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{x}_i \in B).$$

Condition D. For all linear subspaces $\mathcal{V} \subset \mathbb{R}^p$ with $\dim \mathcal{V} \le p-1$,

$$P_n(\mathcal{V}) < 1 - [p - \dim\ \mathcal{V}]/a_0.$$

The existence of a solution under condition D is proved in Kent and Tyler (1991). Proving existence is the most difficult part, but uniqueness of the solution and convergence of a numerical algorithm is much simpler.

Proposition 13.3 *Under condition D, there exists* $\hat{\mathbf{A}} > \mathbf{0}$ *such that* $l(\hat{\mathbf{A}}) \le l(\mathbf{A})$, $\forall \mathbf{A} > \mathbf{0}$.

Note that when sampling from an absolutely continuous distribution condition D is satisfied w.p.1 for $a_0 > p$ and sample sizes $n \ge p$ since for any subspace $\mathcal{V}$, $k = \dim\ \mathcal{V} \le p-1$,

$$P_n(\mathcal{V}) \overset{\text{w.p.1}}{\le} \frac{k}{n} \le \frac{k}{p} = 1 - \frac{(p-k)}{p} < 1 - \frac{(p-k)}{a_0}.$$

The following condition of monotonicity is for unicity of the solution.

Condition M.

(i) For $s \ge 0$, $u(s) \ge 0$ and $u(s)$ is continuous and nonincreasing.

(ii) For $s \ge 0$, $\psi(s) = su(s)$ is strictly increasing.

Proposition 13.4 *If conditions D and M hold, then there exists a unique solution* $\hat{\mathbf{A}}$ *to* (13.6).

Proof. Existence is ensured by condition D. Now, assume there are two solutions $\hat{\mathbf{A}} = \mathbf{I}$ and $\hat{\mathbf{A}} = \mathbf{A}$. Let $\mathbf{A}$ have eigenvalues $\lambda_1 \ge \cdots \ge \lambda_p$ and assume, if possible, $\lambda_1 > 1$. Since $su(s)$ is strictly increasing and $u(s)$ is nonincreasing, it follows that for $\mathbf{x} \ne \mathbf{0}$,

$$u(\mathbf{x}'\mathbf{A}^{-1}\mathbf{x}) \le u(\lambda_1^{-1}\mathbf{x}'\mathbf{x}) \le \frac{\lambda_1^{-1}\mathbf{x}'\mathbf{x}}{\lambda_1^{-1}\mathbf{x}'\mathbf{x}} u(\lambda_1^{-1}\mathbf{x}'\mathbf{x}) < \lambda_1 u(\mathbf{x}'\mathbf{x}),$$

where the first inequality used Rayleigh's quotient. This implies

$$\mathbf{A} = \text{ave}\left[u(\mathbf{x}_i'\mathbf{A}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right] < \lambda_1\ \text{ave}\left[u(\mathbf{x}_i'\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right] = \lambda_1\mathbf{I}.$$

This gives the contradiction $\lambda_1 < \lambda_1$, and so $\lambda_1 \leq 1$. A similar argument shows $\lambda_p \geq 1$. Thus, $\mathbf{A} = \mathbf{I}$. □

Under conditions D and M, the unique solution can be found by regarding the estimating equation as a fixed-point equation. Given a starting value $\mathbf{A}_0 > \mathbf{0}$, define the iterative numerical algorithm

$$\mathbf{A}_{m+1} = \text{ave}\left[u(\mathbf{x}_i'\mathbf{A}_m^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right], \ m = 0, 1, \ldots.$$

Proposition 13.5 *Under conditions D and M, for any* $\mathbf{A}_0 > \mathbf{0}$, $\mathbf{A}_m$ *converges as* $m \to \infty$ *to the unique solution (the MLE) of* (13.6).

Proof. Conditions D and M ensure existence and uniqueness of a solution $\hat{\mathbf{A}}$. Since $\mathbf{x}_i \mapsto \mathbf{B}\mathbf{x}_i$, $\mathbf{B} \in \mathbf{G}_p$, induces the new solution $\hat{\mathbf{A}} \mapsto \mathbf{B}\hat{\mathbf{A}}\mathbf{B}'$, one can assume without loss of generality that $\hat{\mathbf{A}} = \mathbf{I}$ is the solution. Let $\lambda_{1,m} \geq \cdots \geq \lambda_{p,m}$ be the eigenvalues of $\mathbf{A}_m$, $m = 1, 2, \ldots$.
Step 1: The following results are established:

(i) $\lambda_{1,m} \leq 1 \Longrightarrow \lambda_{1,m+1} \leq 1$,

(ii) $\lambda_{1,m} > 1 \Longrightarrow \lambda_{1,m+1} < \lambda_{1,m}$,

(iii) $\lambda_{p,m} \geq 1 \Longrightarrow \lambda_{p,m+1} \geq 1$, and

(iv) $\lambda_{p,m} < 1 \Longrightarrow \lambda_{p,m+1} > \lambda_{p,m}$.

Note that (iii) and (iv) imply $\mathbf{A}_{m+1} > \mathbf{0}$ whenever $\mathbf{A}_m > \mathbf{0}$. To prove (i), if $\lambda_{1,m} \leq 1$, then

$$\mathbf{x}'\mathbf{A}_m^{-1}\mathbf{x} \geq \lambda_{1,m}^{-1}\mathbf{x}'\mathbf{x} \geq \mathbf{x}'\mathbf{x},$$

and since $u(s)$ is nonincreasing, $u(\mathbf{x}'\mathbf{A}_m^{-1}\mathbf{x}) \leq u(\mathbf{x}'\mathbf{x})$. Given that $\hat{\mathbf{A}} = \mathbf{I}$ is the solution, this implies $\mathbf{A}_{m+1} \leq \text{ave}\left[u(\mathbf{x}_i'\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right] = \mathbf{I}$. Thus, $\lambda_{1,m+1} \leq 1$. The proof of (iii) is similar. To prove (ii), since $\mathbf{x}'\mathbf{A}_m^{-1}\mathbf{x} \geq \lambda_{1,m}^{-1}\mathbf{x}'\mathbf{x}$, $u(s)$ is nonincreasing and $su(s)$ is strictly increasing, it follows that if $\lambda_{1,m} > 1$, then

$$u(\mathbf{x}'\mathbf{A}_m^{-1}\mathbf{x}) \leq u(\lambda_{1,m}^{-1}\mathbf{x}'\mathbf{x}) \leq \lambda_{1,m}u(\mathbf{x}'\mathbf{x}),$$

with the second inequality strict for $\mathbf{x} \neq \mathbf{0}$. This implies

$$\mathbf{A}_{m+1} < \lambda_{1,m} \text{ ave}\left[u(\mathbf{x}_i'\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right] = \lambda_{1,m}\mathbf{I}.$$

Thus, $\lambda_{1,m+1} < \lambda_{1,m}$. The proof of (iv) is similar.
Step 2: We shall now show that

(v) $\limsup \lambda_{1,m} \leq 1$,

(vi) $\liminf \lambda_{p,m} \geq 1$,

from which it follows that $\lambda_{1,m} \to 1$ and $\lambda_{p,m} \to 1$, so that $\mathbf{A}_m \to \mathbf{I}$ (v. Problem 1.8.13). Given $\mathbf{A} > \mathbf{0}$, let $\lambda_1(\mathbf{A})$ denote the largest eigenvalue and define

$$\phi(\mathbf{A}) = \text{ave}\left[u(\mathbf{x}_i'\mathbf{A}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right].$$

Step 1 implies that if $\lambda_1(\mathbf{A}) > 1$ and $\mathbf{B} = \phi(\mathbf{A})$, then $\lambda_1(\mathbf{B}) < \lambda_1(\mathbf{A})$. In view of step 1, statement (v) requires proof only in the case in which $\lambda_{1,m} = \lambda_1(\mathbf{A}_m) > 1$, $\forall m$. Note that $\lambda_{1,m}$ is a decreasing sequence in this case. Let $\lambda^* = \lim \lambda_{1,m} \geq 1$ and suppose, if possible, that $\lambda^* > 1$. From step 1, the eigenvalues of the sequence $\mathbf{A}_m$ are bounded away from 0 and ∞. Thus, we can find a convergent subsequence $\mathbf{A}_{m_j} \to \mathbf{B}_0$ say, where $\mathbf{B}_0 > \mathbf{0}$. Further, $\mathbf{A}_{m_j+1} = \phi(\mathbf{A}_{m_j}) \to \phi(\mathbf{B}_0) = \mathbf{B}_1$, say. Since $\lambda_{1,m}$ is decreasing, $\lambda_1(\mathbf{B}_0) = \lim \lambda_{1,m_j} = \lambda^*$ and $\lambda_1(\mathbf{B}_1) = \lim \lambda_{1,m_j+1} = \lambda^*$. However, step 1 implies that $\lambda_1(\mathbf{B}_1) < \lambda_1(\mathbf{B}_0)$, giving a contradiction. Hence, (v) follows. Item (vi) is proved similarly. □

Location-scale problem

Results for location scale are derived by embedding the p-dimensional location-scale problem into a $(p+1)$-dimensional scale-only problem. For given $\mathbf{\Lambda} \in \mathcal{P}_p$, $\boldsymbol{\mu} \in \mathbb{R}^p$, and $\gamma > 0$, let

$$\mathbf{A} = \begin{pmatrix} \mathbf{\Lambda} + \gamma^{-1}\boldsymbol{\mu}\boldsymbol{\mu}' & \gamma^{-1}\boldsymbol{\mu} \\ \gamma^{-1}\boldsymbol{\mu}' & \gamma^{-1} \end{pmatrix} \in \mathbb{R}^{p+1}_{p+1} \tag{13.7}$$

and observe that any $\mathbf{A} \in \mathcal{P}_{p+1}$ can be written in this form. On using the inverse of a partitioned matrix (v. Problem 1.8.1), one finds

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{\Lambda}^{-1} & -\mathbf{\Lambda}^{-1}\boldsymbol{\mu} \\ -\boldsymbol{\mu}'\mathbf{\Lambda}^{-1} & \gamma + \boldsymbol{\mu}'\mathbf{\Lambda}^{-1}\boldsymbol{\mu} \end{pmatrix}.$$

Now define the artificial vectors $\mathbf{y}_i = (\mathbf{x}_i', 1)' \in \mathbb{R}^{p+1}$ and note that

$$\mathbf{y}_i'\mathbf{A}^{-1}\mathbf{y}_i = (\mathbf{x}_i - \boldsymbol{\mu})'\mathbf{\Lambda}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + \gamma. \tag{13.8}$$

Let $\mathbf{A}_{(1)}$ be defined as in (13.7) but with $\gamma = 1$. Upon using (13.8) and $|\mathbf{A}_{(1)}| = |\mathbf{\Lambda}|$, the objective function (13.3) can be expressed as

$$l_n(\boldsymbol{\mu}, \mathbf{\Lambda}) = l(\mathbf{A}_{(1)}) = cte + \sum_{i=1}^{n} \ln g(\mathbf{y}_i'\mathbf{A}_{(1)}^{-1}\mathbf{y}_i - 1) - \tfrac{1}{2}n \ln |\mathbf{A}_{(1)}|. \tag{13.9}$$

Thus, the problem of maximizing (13.3) over $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\mathbf{\Lambda} \in \mathcal{P}_p$ is equivalent to maximizing $l(\mathbf{A}_{(1)})$ over $\mathbf{A}_{(1)} \in \mathcal{P}_{p+1}$ with the restriction that the $(p+1, p+1)$ element of $\mathbf{A}_{(1)}$ be 1. Moreover, the estimating equations (13.4) and (13.5) can be rewritten in a single estimating equation as

$$\hat{\mathbf{A}} = \begin{pmatrix} \hat{\mathbf{\Lambda}} + \hat{\gamma}^{-1}\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}' & \hat{\gamma}^{-1}\hat{\boldsymbol{\mu}} \\ \hat{\gamma}^{-1}\hat{\boldsymbol{\mu}}' & \hat{\gamma}^{-1} \end{pmatrix} = \text{ave}[u(s_i)\mathbf{y}_i\mathbf{y}_i'], \tag{13.10}$$

where $\hat{\gamma}^{-1} = \text{ave}[u(s_i)]$ with $s_i = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\hat{\mathbf{\Lambda}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$, as in the original location-scale formulation. Using (13.8), the single estimating equation can be reexpressed as

$$\hat{\mathbf{A}} = \text{ave}\left[u^*(\mathbf{y}_i'\hat{\mathbf{A}}^{-1}\mathbf{y}_i; \hat{\gamma})\mathbf{y}_i\mathbf{y}_i'\right], \tag{13.11}$$

where $u^*(s;\gamma) = u(s-\gamma)$, for $s \geq \gamma$. This looks very similar to the estimating equation of a scale-only problem, the difference being that the function $u^*(\cdot;\hat{\gamma})$ depends on the data through $\hat{\gamma}$. The next condition for existence of a solution is just the previous condition D on $\mathbf{y}_i$'s recast in terms of $\mathbf{x}_i$'s.

Condition D1. For all translated linear subspaces (hyperplanes) $\mathcal{H} \subset \mathbb{R}^p$ with dim $\mathcal{H} \leq p-1$,

$$P_n(\mathcal{H}) < 1 - (p - \dim\ \mathcal{H})/a_0.$$

This time if $a_0 > p+1$, $n \geq p+1$, then condition D1 is satisfied w.p.1 when sampling from an absolutely continuous distribution.

Proposition 13.6 *If conditions D1 and M hold, then there exists a solution* $\hat{\boldsymbol{\mu}} \in \mathbb{R}^p$ *and* $\hat{\boldsymbol{\Lambda}} > \mathbf{0}$ *to* (13.11). *This solution is unique if* $(s+\hat{\gamma})u(s)$ *is strictly increasing in* $s \geq 0$ *for* $\hat{\gamma}^{-1} = \text{ave}[u(s_i)]$ *defined above.*

A difficulty in applying Proposition 13.6 is the strictly increasing condition which depends on the unknown $\hat{\gamma}$. However, given a solution $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}})$ the condition guarantees that no other solutions exist. For the $t_{p,\nu}$ distribution, $\nu \geq 1$, we prove that $\hat{\gamma}$ is independent of the data ($\hat{\gamma} = 1$) and the condition is thus automatically satisfied for $\nu > 1$ since

$$(s+\hat{\gamma})u(s) = (\nu+p)(s+1)/(s+\nu)$$

is strictly increasing.

Lemma 13.1 *For the* $t_{p,\nu}$ *distribution* $\nu \geq 1$, $\hat{\gamma} = 1$.

Proof. If $\gamma_u \geq \gamma$ and $(s+\gamma_u)u(s)$ is strictly increasing and condition M holds, then $(s+\gamma)u(s)$ is also strictly increasing. Multiplying by $\hat{\boldsymbol{\Lambda}}^{-1}$ and taking the trace of (13.5), we get $\text{ave}[s_i u(s_i)] = p$. Thus, $\forall b > 0$,

$$p = \text{ave}[(s_i+b)u(s_i)] - b\hat{\gamma}^{-1},$$

which implies $\gamma_l \leq \hat{\gamma} \leq \gamma_u$, where

$$\begin{aligned} \gamma_u^{-1} &= \sup_{b>0} \inf_{s>0} [(s+b)u(s) - p]/b, \\ \gamma_l^{-1} &= \inf_{b>0} \sup_{s>0} [(s+b)u(s) - p]/b. \end{aligned}$$

Letting $b = \nu$, we obtain $1 \leq \gamma_l \leq \hat{\gamma} \leq \gamma_u \leq 1$. □

The Cauchy case, $\nu = 1$, has $(s+1)u(s) = p+1$, which is not strictly increasing. It requires a special treatment, but the MLE is also unique under condition D1 [Kent and Tyler (1991)]. For the $t_{p,\nu}$ case, since $\hat{\gamma}$ is independent of the data, this means that when condition D1 is satisfied, the fixed-point algorithm still converges to the MLE. So, for any starting

values $\boldsymbol{\mu}_0$ and $\boldsymbol{\Lambda}_0 > \mathbf{0}$, the iterative equations

$$\begin{aligned}\boldsymbol{\mu}_{m+1} &= \frac{\text{ave}\left\{u\left[(\mathbf{x}_i-\boldsymbol{\mu}_m)'\boldsymbol{\Lambda}_m^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_m)\right]\mathbf{x}_i\right\}}{\text{ave}\left\{u\left[(\mathbf{x}_i-\boldsymbol{\mu}_m)'\boldsymbol{\Lambda}_m^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_m)\right]\right\}},\\ \boldsymbol{\Lambda}_{m+1} &= \text{ave}\left\{u\left[(\mathbf{x}_i-\boldsymbol{\mu}_m)'\boldsymbol{\Lambda}_m^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_m)\right](\mathbf{x}_i-\boldsymbol{\mu}_m)(\mathbf{x}_i-\boldsymbol{\mu}_m)'\right\}\end{aligned}$$

converge to the MLE.

Asymptotics for the MLE

The general theory of maximum likelihood coupled with the fact that the MLE is affine equivariant tells us that for some constants β, σ_1, and σ_2,

$$n^{1/2}(\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}, \hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \xrightarrow{d} (\mathbf{N}, \mathbf{n}),$$

where

$$\begin{aligned}\mathbf{n} &\sim N_p(\mathbf{0}, \beta\boldsymbol{\Lambda})\\ \mathbf{N} &\sim N_p^p(\mathbf{0}, \sigma_1(\mathbf{I}+\mathbf{K}_p)(\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda}) + \sigma_2\ \text{vec}(\boldsymbol{\Lambda})[\text{vec}(\boldsymbol{\Lambda})]').\end{aligned}$$

Using Fisher's information, these constants can now be evaluated and we can also show that $\mathbf{N} \perp\!\!\!\perp \mathbf{n}$; i.e., they are asymptotically independent. The *score function* is the derivative of

$$l(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = cte + \ln g[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})] - \tfrac{1}{2}\ln|\boldsymbol{\Lambda}|$$

with respect to $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ and its variance is called *Fisher's information* and is denoted by $\mathcal{I}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. It is also well known that the asymptotic variance is the inverse of Fisher's information. Let us show that $\mathcal{I}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is block-diagonal and thus $\mathbf{N} \perp\!\!\!\perp \mathbf{n}$. We have

$$\begin{aligned}\partial l/\partial\boldsymbol{\mu} &= u(s)\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\\ \partial l/\partial\boldsymbol{\Lambda} &= -\tfrac{1}{2}\boldsymbol{\Lambda}^{-1} + \tfrac{1}{2}u(s)\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1},\end{aligned}$$

where $s = (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. The constants β, σ_1, and σ_2 being independent of $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, it suffices to evaluate the variance of the score while assuming $(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = (\mathbf{0}, \mathbf{I})$ and $\mathbf{x} \overset{d}{=} \mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$. The expectation

$$E\{\partial l/\partial\mu_i \cdot \partial l/\partial\Lambda_{jk}\}$$

involves only first-order and third-order product moments of $\mathbf{z}$, which is spherical. Since these moments are all null, it follows that $\mathcal{I}(\mathbf{0}, \mathbf{I})$ is block-diagonal with blocks $\mathcal{I}_1$ and $\mathcal{I}_2$, say. We then calculate β from $\mathcal{I}_1^{-1}$. Now,

$$\begin{aligned}\mathcal{I}_1 = E\left\{(\partial l/\partial\boldsymbol{\mu})(\partial l/\partial\boldsymbol{\mu})'\right\} &= E\ u^2(s)\mathbf{z}\mathbf{z}'\\ &= E[su^2(s)]E[\mathbf{u}\mathbf{u}'],\end{aligned}$$

where we have let $\mathbf{z} = s^{1/2}\mathbf{u}$, where $s \overset{d}{=} |\mathbf{z}|^2$, $\mathbf{u} \sim \text{unif}(S^{p-1})$, and $s \perp\!\!\!\perp \mathbf{u}$. Then, $\mathcal{I}_1 = p^{-1}E[su^2(s)]\mathbf{I}$ if we note that $E\ \mathbf{u}\mathbf{u}' = p^{-1}\mathbf{I}$ (v. Problem 13.6.4). Thus, we have shown that $\beta = p/E[su^2(s)]$. We now evaluate

σ_1 and σ_2 from $\mathcal{I}_2$. We would like to identify var $\mathbf{N}$ with $\mathcal{I}_2^{-1}$, but we must first eliminate the redundant elements of the symmetric $\mathbf{N}$ for var $\mathbf{N}$ to become nonsingular. For this reason, define

$$\begin{aligned}\mathbf{A}_j &= (\mathbf{0}, \mathbf{I}_j)' : p \times j,\ j = 1, \ldots, p,\\ \mathbf{M}_p &= \operatorname{diag}(\mathbf{A}_p, \ldots, \mathbf{A}_1) : p^2 \times \tfrac{1}{2}p(p+1),\end{aligned}$$

and verify that for any symmetric $\mathbf{A} \in \mathbb{R}_p^p$, $\mathbf{M}_p'\text{vec}(\mathbf{A})$ is the $\frac{1}{2}p(p+1)$-dimensional vector formed by stacking the columns of $\mathbf{A}$ after deleting the upper triangular part of $\mathbf{A}$. Now, $\text{var}(\mathbf{M}_p'\ \text{vec}(\mathbf{N})) = \mathbf{M}_p'\ \text{var}(\mathbf{N})\mathbf{M}_p$. It is easy to check that

$$\mathbf{M}_p'\mathbf{M}_p = \mathbf{I},\quad \mathbf{M}_p'\mathbf{K}_p\mathbf{M}_p = \mathbf{D}_p,\quad \mathbf{M}_p'\ \text{vec}(\mathbf{I}_p) = \mathbf{a}_p,$$

where

$$\begin{aligned}\boldsymbol{\alpha}_j &= (1, 0, \ldots, 0)' \in \mathbb{R}^j,\ j = 1, \ldots, p,\\ \mathbf{a}_p &= (\boldsymbol{\alpha}_p', \ldots, \boldsymbol{\alpha}_1')' : \tfrac{1}{2}p(p+1) \times 1,\\ \mathbf{D}_p &= \operatorname{diag}(\mathbf{a}_p).\end{aligned}$$

Then, we can identify

$$\mathcal{I}_2^{-1} = \sigma_1(\mathbf{I} + \mathbf{D}_p) + \boldsymbol{\sigma}_2\mathbf{a}_p\mathbf{a}_p'.$$

Using the inverse of a perturbed matrix (v. Problem 1.8.8), we have with the relations $(\mathbf{I} + \mathbf{D}_p)^{-1}\mathbf{a}_p = \frac{1}{2}\mathbf{a}_p$ and $\mathbf{a}_p'\mathbf{a}_p = p$,

$$\begin{aligned}\mathcal{I}_2 &= \sigma_1^{-1}(\mathbf{I} + \mathbf{D}_p)^{-1} - \sigma_2[4\sigma_1^2(1 + \tfrac{1}{2}p\sigma_2\sigma_1^{-1})]^{-1}\mathbf{a}_p\mathbf{a}_p'\\ &= i_1(\mathbf{I} + \mathbf{D}_p)^{-1} + i_2\mathbf{a}_p\mathbf{a}_p',\end{aligned}$$

where

$$i_1 = \sigma_1^{-1}, \qquad i_2 = -\sigma_2[4\sigma_1^2(1 + \tfrac{1}{2}p\sigma_2\sigma_1^{-1})]^{-1}. \tag{13.12}$$

As an example for $p = 2$, we thus have the identification

$$\begin{aligned}\mathcal{I}_2 &= i_1(\mathbf{I} + \mathbf{D}_p)^{-1} + i_2\mathbf{a}_p\mathbf{a}_p'\\ &= \begin{pmatrix} \frac{1}{2}i_1 + i_2 & 0 & i_2\\ 0 & i_1 & 0\\ i_2 & 0 & \frac{1}{2}i_1 + i_2 \end{pmatrix}\\ &= E\begin{pmatrix} \left(\frac{\partial l}{\partial \Lambda_{11}}\right)^2 & \left(\frac{\partial l}{\partial \Lambda_{11}}\right)\left(\frac{\partial l}{\partial \Lambda_{21}}\right) & \left(\frac{\partial l}{\partial \Lambda_{11}}\right)\left(\frac{\partial l}{\partial \Lambda_{22}}\right)\\ \left(\frac{\partial l}{\partial \Lambda_{11}}\right)\left(\frac{\partial l}{\partial \Lambda_{21}}\right) & \left(\frac{\partial l}{\partial \Lambda_{21}}\right)^2 & \left(\frac{\partial l}{\partial \Lambda_{21}}\right)\left(\frac{\partial l}{\partial \Lambda_{22}}\right)\\ \left(\frac{\partial l}{\partial \Lambda_{11}}\right)\left(\frac{\partial l}{\partial \Lambda_{22}}\right) & \left(\frac{\partial l}{\partial \Lambda_{21}}\right)\left(\frac{\partial l}{\partial \Lambda_{22}}\right) & \left(\frac{\partial l}{\partial \Lambda_{22}}\right)^2 \end{pmatrix}.\end{aligned}$$

Thus, in general for $i \neq j$,

$$\begin{aligned}i_1 &= E\left\{(\partial l/\partial \Lambda_{ij})^2\right\},\\ i_2 &= E\left\{(\partial l/\partial \Lambda_{ii})(\partial l/\partial \Lambda_{jj})\right\}.\end{aligned}$$

These are evaluated with

$$\begin{aligned} \partial l/\partial \Lambda_{ij} &= u(s) z_i z_j \stackrel{d}{=} \psi(s) u_i u_j, \\ \partial l/\partial \Lambda_{ii} &= -\tfrac{1}{2} + \tfrac{1}{2} u(s) z_i^2 \stackrel{d}{=} -\tfrac{1}{2} + \tfrac{1}{2} \psi(s) u_i^2. \end{aligned}$$

Using Problem 13.6.4, $E\ u_i^2 = p^{-1}$ and $E\ u_i^2 u_j^2 = [p(p+2)]^{-1}$, $i \neq j$, the final result is thus

$$\begin{aligned} i_1 &= [p(p+2)]^{-1} E\ \psi^2(s), \\ i_2 &= -\tfrac{1}{4} + [p(p+2)]^{-1} E\ \psi^2(s) \end{aligned}$$

if we note that $E\ \psi(s) = p$. The constants σ_1 and σ_2 are obtained by solving equation (13.12). The density of s was given in Problem 4.5.13. We have proved that under regularity conditions for the MLE [Lehmann (1983), pp. 429-430]

Proposition 13.7

$$n^{1/2}(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}, \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \stackrel{d}{\to} (\mathbf{N}, \mathbf{n}),$$

where $\mathbf{N} \perp\!\!\!\perp \mathbf{n}$ *and*

$$\begin{aligned} \mathbf{n} &\sim N_p(\mathbf{0}, \beta \mathbf{\Lambda}), \\ \mathbf{N} &\sim N_p^p(\mathbf{0}, \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{\Lambda} \otimes \mathbf{\Lambda}) + \sigma_2\ \mathrm{vec}(\mathbf{\Lambda})[\mathrm{vec}(\mathbf{\Lambda})]'), \end{aligned}$$

with

$$\begin{aligned} \beta &= p/E[su^2(s)], \\ \sigma_1 &= p(p+2)/E[\psi^2(s)], \\ \sigma_2 &= -2\sigma_1(1-\sigma_1)/[2 + p(1-\sigma_1)] \end{aligned}$$

and s *has density*

$$\frac{\pi^{p/2}}{\Gamma(\frac{1}{2}p)} s^{\frac{1}{2}p-1} g(s),\ \ s > 0.$$

The parameter σ_1 of the asymptotic variance will play a major role as an index of relative efficiency for robust tests.

Example 13.8 *For the* $t_{p,\nu}$ *distribution, the reader can check* $\sigma_1 = 1 + 2/(p+\nu)$.

The maximum likelihood estimation of the multivariate $t_{p,\nu}$ distribution with possibly missing data and unknown degrees of freedom was treated by Liu (1997). Missing data imputation using the multivariate $t_{p,\nu}$ distribution was also the subject of Liu (1995).

13.4 Robust estimates

An alternative approach to MLE consists of robust location and scatter estimates such as the M estimate [Maronna (1976), Huber (1981)] or the S estimate [Davies (1987), Lopuhaä (1989)]. The theoretical proofs for existence, unicity, consistency, and asymptotic normality of these estimates go beyond the scope of this book. Of importance to us, however, is to show how easily these affine equivariant and $\sqrt{n}$-asymptotically normal estimates can serve as the building block to robust tests on location and scatter. They are succintly introduced now and invoked later to construct robust tests.

13.4.1 M estimate

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ and $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$. The idea behind M estimate is to modify the MLE estimating equations to gain robustness. The M estimate of location and scatter are defined as solution to the equations

$$\boldsymbol{\mu}_n = \text{ave}\left[u_1(t_i)\mathbf{x}_i\right] / \text{ave}\left[u_1(t_i)\right], \tag{13.13}$$

$$\mathbf{V}_n = \text{ave}\left[u_2(t_i^2)(\mathbf{x}_i - \boldsymbol{\mu}_n)(\mathbf{x}_i - \boldsymbol{\mu}_n)'\right], \tag{13.14}$$

where $t_i = \left[(\mathbf{x}_i - \boldsymbol{\mu}_n)'\mathbf{V}_n^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_n)\right]^{1/2}$.

The M estimates are obviously affine equivariant. Interestingly, they include, as a particular case, the MLE estimate with the functions $u_1(t) = -2g'(t^2)/g(t^2)$ and $u_2(t^2) = u_1(t)$. Define $\psi_i(s) = su_i(s)$, $i = 1, 2$. The following conditions on the functions are needed and will always be assumed:

M1. u_1 and u_2 are non-negative, nonincreasing, and continuous on $[0, \infty)$.

M2. ψ_1 and ψ_2 are bounded. Let $K_i = \sup_{s \geq 0} \psi_i(s)$.

M3. ψ_2 is nondecreasing and is strictly increasing in the interval where $\psi_2 < K_2$.

M4. There exists s_0 such that $\psi_2(s_0^2) > p$ and that $u_1(s) > 0$ for $s \leq s_0$ (and, hence, $K_2 > p$).

Example 13.9 *The $t_{p,\nu}$ MLE has $\psi_1(s) = (\nu + p)s/(\nu + s^2)$ and $\psi_2(s) = (\nu + p)s/(\nu + s)$. It is easy to verify M1 through M4.*

Example 13.10 *Huber's ψ function is defined as*

$$\psi(s, k) = \max[-k, \min(s, k)].$$

Let $k > 0$ be a constant and take $\psi_1(s) = \psi(s, k)$ and $\psi_2(s) = \psi(s, k^2)$.

A further condition on the data is needed for existence of the M estimate.

Condition D2. There exists $a > 0$ such that for every hyperplane $\mathcal{H}$, dim $\mathcal{H} \leq p-1$,

$$P_n(\mathcal{H}) \leq 1 - \frac{p}{K_2} - a.$$

When sampling from an absolutely continuous distribution, condition D2 is satisfied w.p.1 for n sufficiently large.

Proposition 13.8 *If condition D2 is satisfied, there exists a solution* $(\boldsymbol{\mu}_n, \mathbf{V}_n)$ *to* (13.13) *and* (13.14). *Moreover,* $\boldsymbol{\mu}_n$ *belongs to the convex hull of* $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$.

Proposition 13.9 *Assume condition D2 and* g *is decreasing. Let* $(\boldsymbol{\mu}_n, \mathbf{V}_n)$ *be a solution to* (13.13) *and* (13.14), *then* $(\boldsymbol{\mu}_n, \mathbf{V}_n) \to (\boldsymbol{\mu}, \mathbf{V})$ *almost surely, where* $\mathbf{V} = \sigma^{-1}\boldsymbol{\Lambda}$ *with* σ *being the solution to* $E\ \psi_2(\sigma t^2) = p$ *and* $t = |\mathbf{z}|$.

The reason for the presence of σ is that $\mathbf{V}_n$ is consistent for a certain multiple of $\boldsymbol{\Lambda}$, $\sigma^{-1}\boldsymbol{\Lambda}$ say, defined by the implicit equation

$$\mathbf{V} = E\ u_2[(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})](\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'.$$

Multiplying by $\mathbf{V}^{-1}$ and taking trace yields $E\ \psi_2(\sigma|\mathbf{z}|^2) = p$. This expectation can be evaluated as a simple integral if one recalls the density of $t = |\mathbf{z}|$ (v. Problem 4.6.13):

$$f(t) = \frac{2\pi^{p/2}}{\Gamma(\frac{1}{2}p)} t^{p-1} g(t^2),\ t \geq 0.$$

Proposition 13.10 *Assume* $s\psi_i'(s)$ *are bounded* $(i = 1, 2)$ *and* g *is decreasing such that* $E\ \psi_1'(\sigma^{1/2}t) > 0$. *Then,*

$$n^{1/2}(\mathbf{V}_n - \mathbf{V}, \boldsymbol{\mu}_n - \boldsymbol{\mu}) \xrightarrow{d} (\mathbf{N}, \mathbf{n}),$$

where $\mathbf{n} \perp\!\!\!\perp \mathbf{N}$ *and*

$$\begin{aligned} \mathbf{n} &\sim N_p\left(\mathbf{0}, (\alpha/\beta^2)\mathbf{V}\right), \\ \mathbf{N} &\sim N_p^p\left(\mathbf{0}, \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{V} \otimes \mathbf{V}) + \sigma_2\ \mathrm{vec}(\mathbf{V})[\mathrm{vec}(\mathbf{V})]'\right), \end{aligned}$$

with σ *being the solution to* $E\ \psi_2(\sigma t^2) = p$, *where*

$$\begin{aligned} \alpha &= p^{-1}E\ \psi_1^2(\sigma^{1/2}t), \\ \beta &= E\left[\left(1 - p^{-1}\right)\ u_1(\sigma^{1/2}t) + p^{-1}\psi_1'(\sigma^{1/2}t)\right], \\ \sigma_1 &= a_1(p+2)^2(2a_2+p)^{-2}, \\ \sigma_2 &= a_2^{-2}\left\{(a_1-1) - 2a_1(a_2-1)[p + (p+4)a_2](2a_2+p)^{-2}\right\}, \end{aligned}$$

and

$$\begin{aligned} a_1 &= [p(p+2)]^{-1}E\ \psi_2^2(\sigma t^2), \\ a_2 &= p^{-1}E\ [\sigma t^2 \psi_2'(\sigma t^2)]. \end{aligned}$$

Those results are due to Maronna (1976), but Tyler (1982) found the asymptotic variance parameters σ_1 and σ_2 in Proposition 13.10. Asymptotic theory for robust principal components was developed by Tyler (1983b) and Boente (1987).

13.4.2 S estimate

Recently, Davies (1987) and Lopuhaä (1989) investigated properties of the S estimate for multivariate location and scatter. As before, consider a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ and $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$. Again, let $t = |\mathbf{z}|$. In the context of regression, Rousseeuw and Yohai (1984) obtained an asymptotically normal and robust estimate from a function ρ assumed to satisfy the following:

S1: ρ is symmetric, has a continuous derivative ψ, and $\rho(0) = 0$.

S2: There exists a finite constant $c_0 > 0$ such that ρ is strictly increasing on $[0, c_0]$ and constant on $[c_0, \infty)$. Let $a_0 = \sup \rho$.

A typical ρ function is Tukey's biweight

$$\rho(t) = \begin{cases} t^2/2 - t^4/(2c_0^2) + t^6/(6c_0^4) & \text{if } |t| \le c_0 \\ c_0^2/6 & \text{if } |t| \ge c_0. \end{cases}$$

The S estimate $(\boldsymbol{\mu}_n, \mathbf{V}_n)$ is defined as the solution of the optimization problem where $t_i = \left[(\mathbf{x}_i - \boldsymbol{\mu}_n)' \mathbf{V}_n^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_n)\right]^{1/2}$:

$$\min |\mathbf{V}_n| \text{ subject to } \frac{1}{n} \sum_{i=1}^{n} \rho(t_i) = b_0$$

over all $\boldsymbol{\mu}_n \in \mathbb{R}^p$ and $\mathbf{V}_n > \mathbf{0}$. The constant b_0, $0 < b_0 < a_0$, chosen so that $0 < b_0/a_0 \equiv r \le (n-p)/2n$, leads to a finite-sample breakdown point [Lopuhaä and Rousseeuw (1991)] of $\epsilon_n^* = \lceil nr \rceil / n$. The choice $r = (n-p)/2n$ results in the maximal breakdown point $\lfloor (n-p+1)/2 \rfloor / n$ (asymptotically 50%). Roughly speaking, the *breakdown point* is the minimum percentage of contaminated data necessary to bring the estimate beyond any given bound. The sample mean requires only one point and thus has a breakdown point $1/n$, or asymptotically 0%. To obtain simultaneously a breakdown point of $\epsilon_n^* = \lceil nr \rceil / n$ and a consistent estimate of scale, i.e., $\mathbf{V}_n \to \boldsymbol{\Lambda}$ w.p.1, for a given $E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ distribution the constant c_0 is chosen so that $E\ \rho(t)/a_0 = r$ and then b_0 is set to $E\ \rho(t)$.

A geometrical interpretation of S estimate can be given with the ellipsoidal contours of an $E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. First, the volume of a p-dimensional ellipsoid $\mathbf{z}' \boldsymbol{\Lambda}^{-1} \mathbf{z} \le 1$ is $|\boldsymbol{\Lambda}|^{1/2} 2\pi^{p/2} / [p\Gamma(\frac{1}{2}p)]$; thus, minimizing $|\boldsymbol{\Lambda}|$ corresponds to finding a minimum volume ellipsoid [Rousseeuw (1985)]. Second, if we could allow discontinuous ρ, then $\rho(t) = 1 - I_{[-c_0, c_0]}(t)$ would count the points outside the ellipsoid. So, for $r = 25\%$, the optimization would

find the minimum volume ellipsoid containing 75% of the data. An S estimate is thus a smoothed version of a minimum volume ellipsoid. The smoothing is done to get $\sqrt{n}$-asymptotically normal estimates. Assuming further

S3: ρ has a second derivative ψ', both $\psi'(t)$ and $u(t) = \psi(t)/t$ are bounded and continuous,

the asymptotic normality of the S estimate holds.

Proposition 13.11 *Assume S1 through S3. Let* $\mathbf{V} = \mathbf{\Lambda}$ *and assume*

$$E\ \psi'(t) > 0,$$
$$E\left[\psi'(t)t^2 + (p+1)\psi(t)t\right] > 0.$$

Let

$$\begin{aligned}
\alpha &= p^{-1}E\ \psi^2(t), \\
\beta &= E\left[\left(1-p^{-1}\right)\ u(t) + p^{-1}\psi'(t)\right], \\
\sigma_1 &= \frac{p(p+2)E[\psi^2(t)t^2]}{E^2[\psi'(t)t^2 + (p+1)\psi(t)t]}, \\
\sigma_2 &= -2p^{-1}\sigma_1 + 4\frac{E\left[\rho(t) - b_0\right]^2}{E^2[\psi(t)t]},
\end{aligned}$$

then

$$n^{1/2}(\mathbf{V}_n - \mathbf{V}, \boldsymbol{\mu}_n - \boldsymbol{\mu}) \xrightarrow{d} (\mathbf{N}, \mathbf{n}),$$

where $\mathbf{n} \perp\!\!\!\perp \mathbf{N}$ *and*

$$\begin{aligned}
\mathbf{n} &\sim N_p\left(\mathbf{0}, (\alpha/\beta^2)\mathbf{V}\right) \\
\mathbf{N} &\sim N_p^p\left(\mathbf{0}, \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{V}\otimes\mathbf{V}) + \sigma_2\ \mathrm{vec}(\mathbf{V})[\mathrm{vec}(\mathbf{V})]'\right).
\end{aligned}$$

	$p=1$	$p=2$	$p=10$
$r=.5$	26.9%	37.7%	91.5%
$r=.3$	40.5%	77.0%	98.0%
$r=.1$	49.1%	98.9%	99.9%

Table 13.1. Asymptotic efficiency of S estimate of scatter at the normal distribution.

According to Lopuhaä (1989) the asymptotic efficiency for the estimation of the scatter as measured by the index σ_1 (or $2\sigma_1 + \sigma_2$ for $p = 1$) are as in Table 13.1 at the normal distribution. The asymptotic efficiency of the location estimate are even higher.

For the S estimate, a high breakdown point corresponds to a low efficiency and vice versa. Let us mention that S estimates are able to achieve the asymptotic variance of M estimates. However, S estimates can have a

high breakdown point in any dimension, whereas the asymptotic breakdown point of an M estimate is at most $1/(p+1)$ [Tyler (1986)]. Lopuhaä (1991) defines τ estimates which can have the same high breakdown point as S estimates but can attain simultaneously high efficiency. The τ estimates are also $\sqrt{n}$-asymptotically normal.

An S-plus [Statistical Sciences, (1995)] function, *s.estimate*, to evaluate S estimate is described in Appendix C. The implementation follows the recommendations of Ruppert (1992) to increase the speed of numerical convergence of this numerically intensive problem. The S-plus function *asymp* evaluates the asymptotic variance constants $\lambda = \alpha/\beta^2$, σ_1, and σ_2, at the normal distribution.

13.4.3 Robust Hotelling-T^2

Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. $\mathbf{x} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. Consider a test of hypothesis on the mean, $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, using a robust version of the classical Hotelling-T^2. Assume $(\mathbf{V}_n, \boldsymbol{\mu}_n)$ is a robust affine equivariant and asymptotically normal estimate (M or S estimate for example),

$$n^{1/2}(\mathbf{V}_n - \mathbf{V}, \boldsymbol{\mu}_n - \boldsymbol{\mu}) \xrightarrow{d} (\mathbf{N}, \mathbf{n}),$$

where $\mathbf{n} \perp\!\!\!\perp \mathbf{N}$ and

$$\begin{aligned} \mathbf{n} &\sim N_p\left(\mathbf{0}, (\alpha/\beta^2)\mathbf{V}\right), \\ \mathbf{N} &\sim N_p^p\left(\mathbf{0}, \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{V} \otimes \mathbf{V}) + \sigma_2 \operatorname{vec}(\mathbf{V})[\operatorname{vec}(\mathbf{V})]'\right). \end{aligned}$$

Proposition 13.12 *Under the sequence of contiguous alternatives* $H_{1,n}$: $\boldsymbol{\mu} = \boldsymbol{\mu}_0 + n^{-1/2}\boldsymbol{\gamma}$,

$$T_R^2 = n(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0)'\mathbf{V}_n^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_0),$$

where $(\mathbf{V}_n, \boldsymbol{\mu}_n)$ *is asymptotically normal as above, satisfies*

$$T_R^2 \xrightarrow{d} \frac{\alpha}{\beta^2} \cdot \chi_p^2\left(\frac{\beta^2}{2\alpha}\boldsymbol{\gamma}'\mathbf{V}^{-1}\boldsymbol{\gamma}\right).$$

In particular, $T_R^2 \xrightarrow{d} \frac{\alpha}{\beta^2}\chi_p^2$ *under* H_0.

Proof. Let $\mathbf{X}$ and $\mathbf{Y}$ be the sample matrices under H_0 and $H_{1,n}$, respectively. Then, we can write

$$\mathbf{Y} \stackrel{d}{=} \mathbf{X} + n^{-1/2}\mathbf{1}\boldsymbol{\gamma}'.$$

Affine equivariance of the estimate immediately gives

$$\begin{aligned} \boldsymbol{\mu}_n(\mathbf{Y}) &\stackrel{d}{=} \boldsymbol{\mu}_n(\mathbf{X}) + n^{-1/2}\boldsymbol{\gamma}, \\ \mathbf{V}_n(\mathbf{Y}) &\stackrel{d}{=} \mathbf{V}_n(\mathbf{X}) \to \mathbf{V} \text{ w.p.1.} \end{aligned}$$

Since

$$n^{1/2}(\boldsymbol{\mu}_n(\mathbf{Y}) - \boldsymbol{\mu}_0) \stackrel{d}{=} n^{1/2}(\boldsymbol{\mu}_n(\mathbf{X}) - \boldsymbol{\mu}_0) + \boldsymbol{\gamma} \xrightarrow{d} N_p\left(\boldsymbol{\gamma}, (\alpha/\beta^2)\mathbf{V}\right),$$

it follows from Corollary 5.1 on quadratic forms (with $\mathbf{A} = (\beta^2/\alpha)\mathbf{V}^{-1}$) that

$$T_R^2 = n(\boldsymbol{\mu}_n(\mathbf{Y}) - \boldsymbol{\mu}_0)'\mathbf{V}_n^{-1}(\mathbf{Y})(\boldsymbol{\mu}_n(\mathbf{Y}) - \boldsymbol{\mu}_0) \xrightarrow{d} \frac{\alpha}{\beta^2}\chi_p^2\left(\frac{\beta^2}{2\alpha}\boldsymbol{\gamma}'\mathbf{V}^{-1}\boldsymbol{\gamma}\right).$$

□

Another type of robustness found in the literature assumes an elliptical distribution on the whole data matrix $\mathbf{X} \in \mathbb{R}_p^n$ with mean $\mathbf{1}\boldsymbol{\mu}'$ and variance of the form $\mathbf{I} \otimes \boldsymbol{\Sigma}$. Under weak assumptions on the p.d.f., the classical Hotelling-T^2 (8.1) remains UMPI and the null distribution of T^2 is the same as if $\mathbf{x}_i$ had been i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e., $T^2 \sim F_c(p, n-p)$ [v. Corollary 8.1]. The main difference in the two approaches resides in that the observations under an elliptical distribution on $\mathbf{X}$ cannot be independent, although they are uncorrelated, unless the elliptical distribution is normal. Independence and spherical symmetry do not go together, except in the normal case, by virtue of the Maxwell-Hershell theorem [v. Proposition 4.11]. One may consult the book by Kariya and Sinha (1989) on this type of robustness for many statistical tests.

Having found the asymptotic null distribution of Hotelling-T^2, it is now a simple matter to extend the results of Section 8.3 to construct robust simultaneous confidence intervals on means. For example, asymptotically, we are at least $(1-\gamma)\times 100\%$ confident in simultaneously presenting all of the observed "Scheffé" intervals:

$$\mathbf{a}'\boldsymbol{\mu}_n - \left(\frac{\alpha}{\beta^2}\frac{\chi_{\gamma,p}^2}{n}\right)^{1/2}(\mathbf{a}'\mathbf{V}_n\mathbf{a})^{1/2} \le \mathbf{a}'\boldsymbol{\mu} \le \mathbf{a}'\boldsymbol{\mu}_n + \left(\frac{\alpha}{\beta^2}\frac{\chi_{\gamma,p}^2}{n}\right)^{1/2}(\mathbf{a}'\mathbf{V}_n\mathbf{a})^{1/2},$$

$\forall \mathbf{a} \in \mathbb{R}^p$. Realistically, the parametric family $E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is unknown. Thus, α and β will have to be replaced by consistent estimates.

13.5 Robust tests on scale matrices

Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. $\mathbf{x} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. Consider a test of hypothesis on $\boldsymbol{\Lambda}$ which is of the general form $\mathbf{h}(\boldsymbol{\Lambda}) = \mathbf{0}$, where $\mathbf{h}(\boldsymbol{\Lambda}) \in \mathbb{R}^q$ is a continuously differentiable function. We will assume $\boldsymbol{\mu} = \mathbf{0}$. Under a $N_p(\mathbf{0}, \boldsymbol{\Lambda})$ distribution, recall that a likelihood ratio test on $\boldsymbol{\Lambda}$ is based uniquely on the likelihood statistic $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$. We know that $n\mathbf{S}_n \sim W_p(n, \boldsymbol{\Lambda})$. Thus, $\mathbf{S}_n$ has density (up to a multiplicative constant) $|\boldsymbol{\Lambda}|^{-n/2}\ \mathrm{etr}(-\frac{1}{2}n\boldsymbol{\Lambda}^{-1}\mathbf{S}_n)$. So, we define

$$\begin{aligned} f(\mathbf{A}, \boldsymbol{\Lambda}) &= |\boldsymbol{\Lambda}|^{-n/2}\ \mathrm{etr}(-\tfrac{1}{2}\boldsymbol{\Lambda}^{-1}\mathbf{A}), \\ f_{\mathbf{h}}(\mathbf{A}) &= \sup_{\mathbf{h}(\boldsymbol{\Lambda})=\mathbf{0}} f(\mathbf{A}, \boldsymbol{\Lambda}), \\ L_{\mathbf{h}}(\mathbf{A}) &= \frac{f_{\mathbf{h}}(\mathbf{A})}{f(\mathbf{A}, \mathbf{A})}. \end{aligned}$$

Note that $L_{\mathbf{h}}(\mathbf{S}_n)$ is the likelihood ratio test for $H_0 : \mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}$ when $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{\Lambda})$. The idea to build a robust test when $\mathbf{x} \sim E_p(\mathbf{0}, \mathbf{\Lambda})$ is to use the test statistic $L_{\mathbf{h}}(\hat{\mathbf{\Lambda}}_n)$ where $\hat{\mathbf{\Lambda}}_n$ could be $\mathbf{S}_n$ [Muirhead and Waternaux, (1980)] or, preferably, a more robust estimate [Tyler (1983a)]. Other approaches which will not be considered here include those based on minimum discrepancy test statistics [Browne and Shapiro (1987), Shapiro and Browne (1987)].

13.5.1 Adjusted likelihood ratio tests

A general method of making a simple correction to the likelihood ratio test is possible for hypotheses satisfying the following condition H on the function $\mathbf{h}$.

Condition H. $\mathbf{h}(\mathbf{\Gamma}) = \mathbf{h}(\gamma\mathbf{\Gamma})$, $\forall\gamma > 0$, $\forall\mathbf{\Gamma} > \mathbf{0}$.

Examples of hypothesis satisfying condition H are the test of sphericity and the test of covariance.

Example 13.11 *The test of sphericity* $H_0 : \mathbf{\Lambda} = \gamma\mathbf{I}$ *for some unknown* γ *can be written as* $H_0 : \mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}$ *with* $h_{ij}(\mathbf{\Lambda}) = \Lambda_{ij}/\Lambda_{pp}$, $1 \le i < j \le p$, *and* $h_{ii}(\mathbf{\Lambda}) = \Lambda_{ii}/\Lambda_{pp} - 1$, $i = 1, \ldots, p-1$. *Here, we have* $q = \frac{1}{2}(p-1)p + p - 1$.

Example 13.12 *The test of covariance between two subvectors* $H_0 : \mathbf{\Lambda}_{12} = \mathbf{0}$, *where* $\mathbf{\Lambda}_{12} \in \mathbb{R}^{p_1}_{p_2}$ *can be written as* $H_0 : \mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}$ *by choosing* $\mathbf{h}(\mathbf{\Lambda}) = \mathrm{vec}(\mathbf{\Lambda}_{11}^{-1/2}\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1/2})$. *Obviously,* $q = p_1 p_2$.

Condition H is not dependent on the location or the spread of the elliptical contours, but concerns only the direction and relative lengths of the axes of the contours. A condition E on the estimate $\hat{\mathbf{\Lambda}}_n$ is also necessary. However, as we encountered in M and S estimation, we usually have an estimate $\mathbf{V}_n$ of a multiple $\mathbf{V}$ of $\mathbf{\Lambda}$. Note that hypothesis $H_0 : \mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}$ is equivalent to $H_0 : \mathbf{h}(\mathbf{V}) = \mathbf{0}$ under condition H.

Condition E. $\mathbf{V}_n$ is affine equivariant and $n^{1/2}(\mathbf{V}_n - \mathbf{V}) \xrightarrow{d} \mathbf{Z}$, where

$$\mathbf{Z} \sim N_p^p(\mathbf{0}, \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{V} \otimes \mathbf{V}) + \sigma_2 \mathrm{vec}(\mathbf{V})[\mathrm{vec}(\mathbf{V})]').$$

Normal and elliptical MLE, the M estimate, and the S estimate satisfy condition E under regularity conditions. An estimate of $\mathbf{h}(\mathbf{V})$ is $\mathbf{h}(\mathbf{V}_n)$, whose asymptotic distribution follows from the delta method (v. Proposition 6.2). A difficulty is the redundance of variables due to the symmetry of $\mathbf{V}$. For this reason, the following derivative will be very useful. Define $d\mathbf{a}/d\mathbf{b} = (da_i/db_j)$, where i varies over rows and j runs over columns. The derivative of $\mathbf{h}(\mathbf{V})$ with respect to $\mathbf{V}$ is defined as

$$\mathbf{h}'(\mathbf{V}) = \tfrac{1}{2}[d\ \mathbf{h}(\mathbf{V})/d\ \mathrm{vec}(\mathbf{V})](\mathbf{I} + \mathbf{J}_p) \in \mathbb{R}^q_{p^2},$$

where $\mathbf{J}_p = \sum_{i=1}^{p} \mathbf{e}_i\mathbf{e}_i' \otimes \mathbf{e}_i\mathbf{e}_i'$ and $\mathbf{e}_i \in \mathbb{R}^p$ is a vector of zero but a 1 in position i. An example when $q = 1$ and $p = 2$ is enlightening:

$$\tfrac{1}{2}\left(dh/ds_{11}, dh/ds_{21}, dh/ds_{12}, dh/ds_{22}\right)\begin{pmatrix} 2 & 0 & \vdots & 0 & 0 \\ 0 & 1 & \vdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \vdots & 1 & 0 \\ 0 & 0 & \vdots & 0 & 2 \end{pmatrix}$$

$$= \left(dh/ds_{11}, \tfrac{1}{2}dh/ds_{21}, \tfrac{1}{2}dh/ds_{12}, dh/ds_{22}\right)$$

is the usual gradiant of h taking into account the symmetry. Before stating the result, we need a lemma on gradiants.

Lemma 13.2 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be continuously differentiable. Then,*

$$\langle df(\mathbf{x})/d\mathbf{x}, \mathbf{x}\rangle = 0$$

for all $\mathbf{x}$ in a neighborhood of $\mathbf{x}_0$ iff $f(\mathbf{x}) = f(\alpha\mathbf{x})$ for all $\mathbf{x}$ and $\alpha\mathbf{x}$ in a neighborhood of $\mathbf{x}_0$.

Proof. It suffices to notice that the contours of f are rays coming out of the origin and that the gradiant is a perpendicular vector to the contour.□

Proposition 13.13 *Under conditions H and E, $n^{1/2}[\mathbf{h}(\mathbf{V}_n) - \mathbf{h}(\mathbf{V})] \xrightarrow{d} \mathbf{Z}_\mathbf{h}$, where*

$$\mathbf{Z}_\mathbf{h} \sim N_q\left(\mathbf{0}, 2\sigma_1[\mathbf{h}'(\mathbf{V})](\mathbf{V}\otimes\mathbf{V})[\mathbf{h}'(\mathbf{V})]'\right).$$

Proof. As in Proposition 6.2, we can write

$$n^{1/2}[\mathbf{h}(\mathbf{V}_n) - \mathbf{h}(\mathbf{V})] = \mathbf{h}'(\mathbf{V})\; n^{1/2}\;\mathrm{vec}(\mathbf{V}_n - \mathbf{V}) + o_p(1).$$

Therefore,

$$n^{1/2}[\mathbf{h}(\mathbf{V}_n) - \mathbf{h}(\mathbf{V})] \xrightarrow{d} \mathbf{h}'(\mathbf{V})\;\mathrm{vec}(\mathbf{Z}).$$

From Lemma 13.2 and condition H, we have $\mathbf{h}'(\mathbf{V})\;\mathrm{vec}(\mathbf{V}) = \mathbf{0}$. Thus, the asymptotic variance is

$$\begin{aligned} \mathrm{var}\;\mathbf{h}'(\mathbf{V})\;\mathrm{vec}(\mathbf{Z}) &= \sigma_1[\mathbf{h}'(\mathbf{V})](\mathbf{I}+\mathbf{K}_p)(\mathbf{V}\otimes\mathbf{V})[\mathbf{h}'(\mathbf{V})]', \\ &= \sigma_1[\mathbf{h}'(\mathbf{V})](\mathbf{V}\otimes\mathbf{V})(\mathbf{I}+\mathbf{K}_p)[\mathbf{h}'(\mathbf{V})]'. \end{aligned}$$

Applying the identity $\mathbf{K}_p\;\mathrm{vec}(\mathbf{A}) = \mathrm{vec}(\mathbf{A}')$ to the columns of $[\mathbf{h}'(\mathbf{V})]'$ gives $(\mathbf{I}+\mathbf{K}_p)[\mathbf{h}'(\mathbf{V})]' = 2[\mathbf{h}'(\mathbf{V})]'$ and the conclusion follows. □

An important consequence of condition H is the asymptotic variance which becomes independent of σ_2. This means that for $\mathbf{V}_n$ satisfying condition E, all the asymptotic distributions of $\mathbf{h}(\mathbf{V}_n)$, under condition H, such as a simple correlation coefficient, a multiple correlation coefficient, a ratio of eigenvalues, etc., are the same as those for the sample variance $\mathbf{S}$, when

sampling from a multivariate normal distribution, except for the factor σ_1 in the asymptotic variance. For completeness, the results for correlations are now given.

Denote by r_{ij} the simple correlation defined from the scale estimate $\mathbf{V}_n = (v_{n,ij})$, i.e.,

$$r_{ij} = \frac{v_{n,ij}}{v_{n,ii}^{1/2} v_{n,jj}^{1/2}},$$

and let

$$\rho_{ij} = \frac{\Lambda_{ij}}{\Lambda_{ii}^{1/2} \Lambda_{jj}^{1/2}}$$

be the correlation for the $E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ distribution.

Proposition 13.14 *Assume condition E holds on* $\mathbf{V}_n$. *Then,*

$$n^{1/2}(r_{ij} - \rho_{ij}) \xrightarrow{d} \sigma_1^{1/2} \cdot N(0, (1 - \rho_{ij}^2)^2).$$

From the delta method it is also clear that an arbitrary number of correlation coefficients is jointly asymptotically normal. Thus, it suffices to consider the case of two correlation coefficients r_{ij} and r_{kl}.

Proposition 13.15 *Assume condition E holds on* $\mathbf{V}_n$. *Then,*

$$n^{1/2} \begin{pmatrix} r_{ij} - \rho_{ij} \\ r_{kl} - \rho_{kl} \end{pmatrix} \xrightarrow{d} \sigma_1^{1/2} \cdot N_2 \left(\mathbf{0}, \begin{pmatrix} (1 - \rho_{ij}^2)^2 & \omega \\ \omega & (1 - \rho_{kl}^2)^2 \end{pmatrix} \right),$$

where the asymptotic covariance ω *is given by*

$$\begin{aligned} \omega = {} & \rho_{ij}\rho_{kl} + \rho_{kj}\rho_{il} - \rho_{lj}(\rho_{ij}\rho_{kj} + \rho_{il}\rho_{kl}) - \rho_{ki}(\rho_{ij}\rho_{il} + \rho_{kj}\rho_{kl}) \\ & + \tfrac{1}{2}\rho_{ki}\rho_{lj}(\rho_{ij}^2 + \rho_{il}^2 + \rho_{kj}^2 + \rho_{kl}^2). \end{aligned}$$

Proof. Assume $\mathbf{V} = (\rho_{ij})$ without loss of generality. Write down the asymptotic distribution

$$n^{1/2} \left[\begin{pmatrix} v_{n,ij} \\ v_{n,ii} \\ v_{n,jj} \\ v_{n,kl} \\ v_{n,kk} \\ v_{n,ll} \end{pmatrix} - \begin{pmatrix} \rho_{ij} \\ 1 \\ 1 \\ \rho_{kl} \\ 1 \\ 1 \end{pmatrix} \right] \xrightarrow{d} N_6(\mathbf{0}, \boldsymbol{\Omega})$$

for a certain $\boldsymbol{\Omega}$ and apply the delta method. □

Similarly, for the multiple correlation coefficient $\hat{R} \equiv \hat{R}(\mathbf{V}_n)$ and partial correlation coefficient $r_{ij|\mathbf{x}_2} \equiv r_{ij|\mathbf{x}_2}(\mathbf{V}_n)$, obtained from $\mathbf{V}_n$ satisfying condition E, we can write the asymptotic distributions:

$$\begin{aligned} n^{1/2}(\hat{R}^2 - R^2) &\xrightarrow{d} \sigma_1^{1/2} \cdot N\left(0, 4R^2(1 - R^2)^2\right), \\ n^{1/2}\left(r_{ij|\mathbf{x}_2} - \rho_{ij|\mathbf{x}_2}\right) &\xrightarrow{d} \sigma_1^{1/2} \cdot N\left(0, (1 - \rho_{ij|\mathbf{x}_2}^2)^2\right). \end{aligned}$$

Higher-order asymptotic distributions for functions of the sample variance $\mathbf{S}$ can also be derived with the use of zonal polynomials [Iwashita and Siotani (1994)].

For the same reason, adjustment to the likelihood ratio test will take a rather simple form. The asymptotic distribution of the modified likelihood ratio test $L_{\mathbf{h}}(\mathbf{V}_n)$, where $\mathbf{V}_n$ may be a robust estimate, is obtained with the equivalent form of Wald's test for the same hypothesis. Let $u_n \sim v_n$ mean $u_n - v_n \xrightarrow{p} 0$. The following result on Wald's formulation holds regardless of condition H.

Proposition 13.16 *Let* $\mathbf{A}_n > \mathbf{0}$, $n = 1, 2, \ldots$, *be such that* $n^{1/2}(\mathbf{A}_n - \mathbf{A}) \xrightarrow{d} (\cdot)$ *for a fixed* $\mathbf{A} > \mathbf{0}$ *satisfying* $\mathbf{h}(\mathbf{A}) = \mathbf{0}$. *If* rank $\mathbf{h}'(\mathbf{\Gamma}) = q$, $\forall \mathbf{\Gamma}$ *in a neighborhood of* $\mathbf{A}$, *then*

$$-2 \ln L_{\mathbf{h}}(\mathbf{A}_n) \sim n[\mathbf{h}(\mathbf{A}_n)]'[C_{\mathbf{h}}(\mathbf{A}_n)]^{-1}\mathbf{h}(\mathbf{A}_n),$$

where $C_{\mathbf{h}}(\mathbf{\Gamma}) = 2[\mathbf{h}'(\mathbf{\Gamma})](\mathbf{\Gamma} \otimes \mathbf{\Gamma})[\mathbf{h}'(\mathbf{\Gamma})]'$.

Proof. This is a generalization of Wald's formulation for the asymptotic behavior of the likelihood ratio statistic. Refer to Tyler (1983a) for details. □

Corollary 13.2 *Assume conditions H and E. Then:*

(i) *under* H_0, $-2 \ln L_{\mathbf{h}}(\mathbf{V}_n) \xrightarrow{d} \sigma_1 \chi^2_q$,

(ii) *under the sequence of contiguous alternatives* $\mathbf{\Lambda}_n = \mathbf{\Lambda} + n^{-1/2}\mathbf{B}$, *where* $\mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}$ *and* $\mathbf{B}$ *is a fixed symmetric matrix,*

$$-2 \ln L_{\mathbf{h}}(\mathbf{V}_n) \xrightarrow{d} \sigma_1 \chi^2_q \left(\delta_{\mathbf{h}}(\mathbf{\Lambda}, \mathbf{B})/2\sigma_1\right),$$

where

$$\delta_{\mathbf{h}}(\mathbf{\Lambda}, \mathbf{B}) = [\mathrm{vec}(\mathbf{B})]'[\mathbf{h}'(\mathbf{\Lambda})]'[C_{\mathbf{h}}(\mathbf{\Lambda})]^{-1}\mathbf{h}'(\mathbf{\Lambda})\mathrm{vec}(\mathbf{B}).$$

Proof. From conditions H and E, $n^{1/2}(\mathbf{V}_n - \mathbf{V}) \xrightarrow{d} \mathbf{Z}$ and $n^{1/2}[\mathbf{h}(\mathbf{V}_n) - \mathbf{h}(\mathbf{V})] \xrightarrow{d} \mathbf{Z}_{\mathbf{h}}$, where $\mathbf{Z}_{\mathbf{h}} \sim N_q(\mathbf{0}, \sigma_1 C_{\mathbf{h}}(\mathbf{V}))$. Under $H_0 : \mathbf{h}(\mathbf{V}) = \mathbf{0}$, $n^{1/2}\mathbf{h}(\mathbf{V}_n) \xrightarrow{d} \mathbf{Z}_{\mathbf{h}}$, and since $\mathbf{h}'(\cdot)$ is continuous, $C_{\mathbf{h}}(\mathbf{V}_n) \xrightarrow{p} C_{\mathbf{h}}(\mathbf{V})$. Hence, we have

$$[n^{1/2}\mathbf{h}(\mathbf{V}_n)]'[C_{\mathbf{h}}(\mathbf{V}_n)]^{-1}[n^{1/2}\mathbf{h}(\mathbf{V}_n)] \xrightarrow{d} \mathbf{Z}'_{\mathbf{h}}[C_{\mathbf{h}}(\mathbf{V})]^{-1}\mathbf{Z}_{\mathbf{h}} \stackrel{d}{=} \sigma_1 \chi^2_q.$$

For contiguous alternatives, under condition H, the noncentrality parameter is invariant with respect to scalar multiplication $\delta_{\mathbf{h}}(\mathbf{\Lambda}, \mathbf{B}) = \delta_{\mathbf{h}}(\alpha\mathbf{\Lambda}, \alpha\mathbf{B})$, $\forall \alpha > 0$. □

As a particular case for $\mathbf{S}_n$ which has $\sigma_1 = 1 + k$, we have, under H_0,

$$-2 \ln L_{\mathbf{h}}(\mathbf{S}_n)/(1 + \hat{k}) \xrightarrow{d} \chi^2_q$$

for some consistent estimate $\hat{k}$ of the kurtosis parameter. So, in the class of $E_p(\mathbf{0}, \mathbf{\Lambda})$ with finite fourth-order moments, this adjusted LRT is robust in the sense that the asymtotic distribution is the same as if $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{\Lambda})$. Note that a consistent estimate $\hat{k}$ can be obtained by the method of moment with the identity

$$1 + k = pE(s^2)/[(p+2)E^2(s)], \tag{13.15}$$

where $s = |\mathbf{z}|^2$ and $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$ has fourth-order moments (v. Problem 13.6.12). More generally, the test statistic $-2 \ln L_{\mathbf{h}}(\mathbf{V}_n)/\hat{\sigma}_1$ will be referred to as an adjusted LRT.

The test of sphericity can serve as an example to illustrate Proposition 13.16 and Corollary 13.2. Wald's formulation is generally obtained by a Taylor series of $-2 \ln L_{\mathbf{h}}(\mathbf{V}_n)$ around $\mathbf{V}$, satisfying $H_0 : \mathbf{h}(\mathbf{V}) = \mathbf{0}$. For the test of sphericity, we have

$$-2 \ln L_{\mathbf{h}}(\mathbf{V}_n) = -n \ln |\mathbf{V}_n| + pn \ln(p^{-1} \text{tr } \mathbf{V}_n).$$

Under $H_0 : \mathbf{V} = \gamma \mathbf{I}$ and condition E, we can write $\mathbf{V}_n = \gamma \mathbf{I} + n^{-1/2} \mathbf{Z}_n$, where $\mathbf{Z}_n$ is bounded in probability. Since $\ln(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} x^i / i$, $-1 < x < 1$, it follows that for a fixed symmetric $\mathbf{A}$,

$$\ln |\mathbf{I} + t\mathbf{A}| = \sum_{i=1}^{\infty} (-1)^{i+1} \text{tr}(\mathbf{A}^i) t^i / i$$

for all t sufficiently small. Hence, we get the expansion

$$\begin{aligned} -2 \ln L_{\mathbf{h}}(\mathbf{V}_n) &= \tfrac{1}{2} \gamma^{-2} [\text{tr}(\mathbf{Z}_n^2) - p^{-1} (\text{tr } \mathbf{Z}_n)^2] + O_p(n^{-1/2}) \\ &\xrightarrow{d} \tfrac{1}{2} \gamma^{-2} [\text{tr}(\mathbf{Z}^2) - p^{-1} (\text{tr } \mathbf{Z})^2], \end{aligned}$$

where $\mathbf{Z} \sim \gamma N_p^p(\mathbf{0}, \sigma_1(\mathbf{I} + \mathbf{K}_p) + \sigma_2 \text{ vec}(\mathbf{I})[\text{vec}(\mathbf{I})]')$. From the relations (v. Problem 6.4.2) $\text{tr } \mathbf{Z}^2 = [\text{vec}(\mathbf{Z})]' \frac{1}{2}(\mathbf{I} + \mathbf{K}_p) \text{ vec}(\mathbf{Z})$ and $\text{tr } \mathbf{Z} = [\text{vec}(\mathbf{I})]' \text{ vec}(\mathbf{Z})$, it follows that

$$\tfrac{1}{2} \gamma^{-2} [\text{tr}(\mathbf{Z}^2) - p^{-1} (\text{tr } \mathbf{Z})^2] = [\text{vec}(\mathbf{Z})]' \mathbf{A} \text{ vec}(\mathbf{Z}),$$

where $\mathbf{A} = \frac{1}{2} \gamma^{-2} \{ \frac{1}{2} (\mathbf{I} + \mathbf{K}_p) - p^{-1} \text{ vec}(\mathbf{I})[\text{vec}(\mathbf{I})]' \}$ is a quadratic form. This is Wald's equivalent formulation for this test. The asymptotic result

$$-2 \ln L_{\mathbf{h}}(\mathbf{V}_n)/\sigma_1 \xrightarrow{d} \chi_q^2, \quad q = \tfrac{1}{2}(p-1)(p+2)$$

follows from Corollary 5.1 on quadratic forms.

When condition H is not satisfied, simple adjustments to the LRT is generally not possible, as the following corollary shows.

Corollary 13.3 *Under* $H_0 : \mathbf{h}(\mathbf{V}) = \mathbf{0}$ *and condition E,*

$$-2 \ln L_{\mathbf{h}}(\mathbf{V}_n) \xrightarrow{d} \sigma_1 \chi_{q-1}^2 + [\sigma_1 + \sigma_2 \delta_{\mathbf{h}}(\mathbf{V}, \mathbf{V})] \chi_1^2,$$

with $\chi_{q-1}^2 \perp\!\!\!\perp \chi_1^2$. *The term* $\delta_{\mathbf{h}}(\mathbf{V}, \mathbf{V}) = 0$ *iff for some neighborhood of* $\mathbf{V}$, $\mathbf{h}(\mathbf{\Gamma}) = \mathbf{h}(\gamma \mathbf{\Gamma})$ *for all* $\mathbf{\Gamma}$ *and* $\gamma \mathbf{\Gamma}$ *in this neighborhood.*

Proof. Take a closer look at the distribution of $\mathbf{Z_h}$ in Proposition 13.13 when condition H is not satisfied. Under H_0, we still have $n^{1/2}\mathbf{h}(\mathbf{V}_n) \xrightarrow{d} \mathbf{h}'(\mathbf{V})\ \mathrm{vec}(\mathbf{Z})$ but with an added term in the variance:

$$\begin{aligned}\text{var } \mathbf{h}'(\mathbf{V})\ \mathrm{vec}(\mathbf{Z}) &= \sigma_1[\mathbf{h}'(\mathbf{V})](\mathbf{V}\otimes\mathbf{V})(\mathbf{I}+\mathbf{K}_p)[\mathbf{h}'(\mathbf{V})]' \\ &\quad +\sigma_2\mathbf{h}'(\mathbf{V})\ \mathrm{vec}(\mathbf{V})[\mathrm{vec}(\mathbf{V})]'[\mathbf{h}'(\mathbf{V})]' \\ &\equiv D_{\mathbf{h}}(\mathbf{V}).\end{aligned}$$

Using Proposition 13.16, the equivalent Wald's formulation is

$$-2\ln L_{\mathbf{h}}(\mathbf{V}_n) \sim [n^{1/2}\mathbf{h}(\mathbf{V}_n)]'[C_{\mathbf{h}}(\mathbf{V}_n)]^{-1}[n^{1/2}\mathbf{h}(\mathbf{V}_n)],$$

and, thus,

$$-2\ln L_{\mathbf{h}}(\mathbf{V}_n) \xrightarrow{d} \mathbf{Z}'_{\mathbf{h}}[C_{\mathbf{h}}(\mathbf{V})]^{-1}\mathbf{Z}_{\mathbf{h}},$$

where $\mathbf{Z_h} \sim N_q(\mathbf{0}, D_{\mathbf{h}}(\mathbf{V}))$. The result follows since $[C_{\mathbf{h}}(\mathbf{V})]^{-1}D_{\mathbf{h}}(\mathbf{V})$ has eigenvalues σ_1 of multiplicity $(q-1)$ and $\sigma_1+\sigma_2\delta_{\mathbf{h}}(\mathbf{V},\mathbf{V})$. The second statement follows since $\mathbf{h}'(\mathbf{V})\ \mathrm{vec}(\mathbf{V}) = \mathbf{0}$ iff the stated condition holds. □

13.5.2 Weighted Nagao's test for a given variance

In this section, we consider an example where the condition H on the hypothesis is not satisfied, but a simple test, robust to large kurtosis, can still be built. For testing the hypothesis, $H_0 : \mathbf{\Sigma} = \mathbf{I}$, against $H_1 : \mathbf{\Sigma} \neq \mathbf{I}$, the modified likelihood ratio test based on n i.i.d. vectors from a $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution is (v. Problem 8.9.8)

$$\Lambda^* = e^{pm/2}|\mathbf{S}_n|^{m/2}\ \mathrm{etr}(-\tfrac{1}{2}m\mathbf{S}_n),\ \ m = n-1,$$

where $\mathbf{S}_n = \sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'/m$. It is invariant to orthogonal transformations, unbiased, and $-2\ln\Lambda^*$ is asymptotically distributed as a noncentral chi-square [Khatri and Srivastava (1974)], $\chi^2_f(\delta)$, with $f = \frac{1}{2}p(p+1)$ and $\delta = \sum_{i=1}^p d_i^2/4$, under the sequence of local alternatives

$$\mathbf{\Sigma}_n = \mathbf{I} + n^{-1/2}\mathbf{D},\ \mathbf{D} = \mathrm{diag}(d_1,\ldots,d_p). \tag{13.16}$$

However, Muirhead (1982, p. 365) showed that if the sample came from an elliptical distribution, $E_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, with kurtosis $3k$, then the asymptotic null distribution is

$$-2\ln\Lambda^*/(1+\hat{k}) \xrightarrow{d} \left[1+\frac{kp}{2(1+k)}\right]\chi^2_1 + \chi^2_{f-1},$$

where χ^2_1 and χ^2_{f-1} are independently distributed and $\hat{k}$ is a consistent estimate of k. A generalization to robust estimates of scale is proposed in Problem 13.6.16. Therefore, even the adjusted test statistic $-2\ln\Lambda^*/(1+\hat{k})$ is not robust to non-normality of the data, especially for large values of k or long-tailed distribution. Moreover, the procedure of estimating k

in the asymptotic distribution and calculating the critical points of the convolution of

$$\left[1+\frac{\hat{k}p}{2(1+\hat{k})}\right]\chi_1^2 \quad \text{and} \quad \chi_{f-1}^2,$$

as if $\hat{k}$ were a constant, is obviously not a valid procedure.

A new test statistic W is proposed, which is also invariant to orthogonal transformations and has an asymptotic null distribution χ_f^2 for all underlying elliptical distributions with finite fourth moments. The asymptotic non-null distribution under the sequence of local alternatives (13.16) is noncentral chi-square. It is asymptotically fully efficient at the normal distribution as compared to the modified likelihood ratio test.

Let $\mathbf{S}_n = \mathbf{I} + n^{-1/2}\mathbf{U}_n$, $\mathbf{S}_n = (s_{ij})$, $\mathbf{U}_n = (u_{ij})$. Then, when H_0 is true, the asymptotic distribution of

$$\mathbf{u}_n = (u_{11}/\sqrt{2}, \ldots, u_{pp}/\sqrt{2}, u_{12}, \ldots, u_{1p}, u_{23}, \ldots, u_{2p}, \ldots, u_{p-1,p})' \in \mathbb{R}^f$$

when the observations $\mathbf{x}_i$ are drawn from an elliptical distribution with kurtosis $3k$ is $N_f(\mathbf{0}, \mathbf{\Gamma})$, where

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Omega} & \mathbf{0} \\ \mathbf{0} & (1+k)\mathbf{I}_{f-p} \end{pmatrix}$$

with $\mathbf{\Omega} = (1+k)\mathbf{I}_p + \frac{1}{2}k\mathbf{1}\mathbf{1}'$, $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^p$. Then, from Corollary 5.1, we have under H_0,

$$\mathbf{u}_n'\mathbf{\Gamma}^{-1}\mathbf{u}_n = \tfrac{1}{2}(u_{11}, \ldots, u_{pp})\mathbf{\Omega}^{-1}\begin{pmatrix} u_{11} \\ \vdots \\ u_{pp} \end{pmatrix} + \sum_{i<j} u_{ij}^2/(1+k) \xrightarrow{d} \chi_f^2.$$

The test statistic proposed [Bentler (1983)] is

$$W = \frac{n}{2}(s_{11}-1, \ldots, s_{pp}-1)\hat{\mathbf{\Omega}}^{-1}\begin{pmatrix} s_{11}-1 \\ \vdots \\ s_{pp}-1 \end{pmatrix} + n\sum_{i<j} s_{ij}^2/(1+\hat{k}),$$

where $\hat{\mathbf{\Omega}} = (1+\hat{k})\mathbf{I}_p + \frac{1}{2}\hat{k}\mathbf{1}\mathbf{1}'$ and $\hat{k}$ is a consistent estimate of k.

Note that when $\hat{k} \equiv 0$, then W reduces to Nagao's (1973) test statistic

$$(n/2)\ \mathrm{tr}(\mathbf{S}_n - \mathbf{I})^2.$$

Asymptotic expansions of Nagao's test for elliptical distributions were derived by Purkayastha and Srivastava (1995). The test statistic W can be seen as a weighted form of Nagao's statistic with the diagonal and off-diagonal elements of the sample variance matrix, $\mathbf{S}_n$, being assigned different weights.

The identity (13.15) leads by the method of moments to the consistent and orthogonally invariant estimate

$$\hat{k} = pn \frac{\sum_{i=1}^n |\mathbf{x}_i - \bar{\mathbf{x}}|^4}{\left(\sum_{i=1}^n |\mathbf{x}_i - \bar{\mathbf{x}}|^2\right)^2} - 1. \tag{13.17}$$

Since $\hat{\mathbf{\Omega}}$ is $(1+\hat{k})\mathbf{I}$ perturbed by a rank 1 matrix, namely $\frac{1}{2}\hat{k}\mathbf{1}\mathbf{1}'$, it has a known inverse which leads to the equivalent expression

$$W = \frac{n}{2}\ \mathrm{tr}(\mathbf{S}_n - \mathbf{I})^2/(1+\hat{k}) - \frac{n}{2}\frac{\hat{k}}{(1+\hat{k})[2(1+\hat{k})+\hat{k}p]}(\mathrm{tr}\ \mathbf{S}_n - p)^2,$$

showing that W is invariant to orthogonal transformations, $\mathbf{x}_i \mapsto \mathbf{H}\mathbf{x}_i$ for any orthogonal matrix $\mathbf{H}$. Thus, without loss of generality, we can take for W the sequence of local alternatives (13.16) with a diagonal matrix $\mathbf{D}$. The following result was given in Bilodeau (1997b).

Proposition 13.17 *Under the sequence of local alternatives* $\mathbf{\Sigma}_n = \mathbf{I} + n^{-1/2}\mathbf{D}$, $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$, *the asymptotic distribution of* W *is noncentral chi-square,*

$$W \xrightarrow{d} \chi^2_f(\mathbf{d}'\mathbf{\Omega}^{-1}\mathbf{d}/4),$$

where

$$\begin{aligned} f &= \tfrac{1}{2}p(p+1), \\ \mathbf{d} &= (d_1, \ldots, d_p)', \\ \mathbf{\Omega} &= (1+k)\mathbf{I} + \tfrac{1}{2}k\mathbf{1}\mathbf{1}'. \end{aligned}$$

Proof. Let $\mathbf{x}_i = \mathbf{\Sigma}_n^{1/2}\mathbf{z}_i$, where $\mathbf{z}_i \sim E_p(\mathbf{0}, \mathbf{I})$. Also, let

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_n' \end{pmatrix}$$

be the sample matrices, and $\mathbf{S}_n(\mathbf{X})$ and $\mathbf{S}_n(\mathbf{Z})$ be the sample variance matrices obtained from $\mathbf{X}$ and $\mathbf{Z}$, respectively. Then, we have

$$\mathbf{U}_n(\mathbf{X}) \equiv n^{1/2}[\mathbf{S}_n(\mathbf{X}) - \mathbf{I}] = \mathbf{\Sigma}_n^{1/2}\mathbf{U}_n(\mathbf{Z})\mathbf{\Sigma}_n^{1/2} + n^{1/2}(\mathbf{\Sigma}_n - \mathbf{I}),$$

where $\mathbf{U}_n(\mathbf{Z}) \xrightarrow{d} N_p^p(\mathbf{0}, (1+k)(\mathbf{I}+\mathbf{K}) + k\ \mathrm{vec}(\mathbf{I})\mathrm{vec}(\mathbf{I})')$, $\mathbf{\Sigma}_n \to \mathbf{I}$, and $n^{1/2}(\mathbf{\Sigma}_n - \mathbf{I}) = \mathbf{D}$. Hence, the asymptotic result

$$\mathbf{U}_n(\mathbf{X}) \xrightarrow{d} N_p^p(\mathbf{D}, (1+k)(\mathbf{I}+\mathbf{K}) + k\ \mathrm{vec}(\mathbf{I})\mathrm{vec}(\mathbf{I})')$$

is obtained. Since W is a continuous function of $\mathbf{U}_n(\mathbf{X})$ and $\hat{k}$,

$$W = g(\mathbf{U}_n(\mathbf{X}), \hat{k}),$$

the conclusion follows from Lemma 6.3 and classical results on quadratic forms if $\hat{k}$ in (13.17) is consistent under the same sequence of local alternatives. This is now shown. From $\bar{\mathbf{x}} = \boldsymbol{\Sigma}_n^{1/2}\bar{\mathbf{z}}$ and since $\bar{\mathbf{z}} \xrightarrow{p} \mathbf{0}$, $\boldsymbol{\Sigma}_n \to \mathbf{I}$, we have $\bar{\mathbf{x}} \xrightarrow{p} \mathbf{0}$. Thus, the asymptotic equivalences

$$\frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i - \bar{\mathbf{x}}|^2 \sim \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i|^2,\ \ \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i - \bar{\mathbf{x}}|^4 \sim \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i|^4,$$

where $u \sim v$ means $u - v \xrightarrow{p} 0$, hold. But now, since

$$(1 + n^{-1/2}d_{(1)})^{j/2}\frac{1}{n}\sum_{i=1}^{n}|\mathbf{z}_i|^j \le \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i|^j \le (1 + n^{-1/2}d_{(p)})^{j/2}\frac{1}{n}\sum_{i=1}^{n}|\mathbf{z}_i|^j,$$

where $d_{(1)} = \min\{d_i\}$ and $d_{(p)} = \max\{d_i\}$, we also have the equivalences

$$\frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i|^2 \sim \frac{1}{n}\sum_{i=1}^{n}|\mathbf{z}_i|^2,\ \ \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i|^4 \sim \frac{1}{n}\sum_{i=1}^{n}|\mathbf{z}_i|^4.$$

Thus, $1 + \hat{k} \xrightarrow{p} pE|\mathbf{z}_i|^4/E^2|\mathbf{z}_i|^2 = 1 + k$, which completes the proof. □

When $k = 0$, the test statistic W is asymptotically distributed, under the sequence of local alternatives (13.16), as $\chi_f^2(\mathbf{d}'\mathbf{d}/4)$. Therefore, W is asymptotically fully efficient at the normal distribution as compared to the modified likelihood ratio test, $-2\ln\Lambda^*$. Sutradhar (1993) discusses the score test of the multivariate t.

For testing the hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$ against $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$ or $\boldsymbol{\Sigma} \neq \mathbf{I}$, consider the test statistic $W + n\bar{\mathbf{x}}'\bar{\mathbf{x}}$ under the sequence of local alternatives

$$\boldsymbol{\mu}_n = n^{-1/2}\boldsymbol{\tau},\ \ \boldsymbol{\Sigma}_n = \mathbf{I} + n^{-1/2}\mathbf{D},$$

where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$. Then, it can be established along the same lines

$$W + n\bar{\mathbf{x}}'\bar{\mathbf{x}} \xrightarrow{d} \chi_f^2(\delta),$$

where $f = \frac{1}{2}p(p+3)$, $\delta = \mathbf{d}'\boldsymbol{\Omega}^{-1}\mathbf{d}/4 + \boldsymbol{\tau}'\boldsymbol{\tau}/2$, and $\mathbf{d}$ and $\boldsymbol{\Omega}$ are as in Proposition 13.17. The test $W + n\bar{\mathbf{x}}'\bar{\mathbf{x}}$ is thus robust in the class of elliptical distributions with finite fourth moments. Its full efficiency at the normal distribution as compared to the likelihood ratio test follows immediately by comparing the asymptotic non-null distributions of the two tests [Khatri and Srivastava (1974)].

13.5.3 Relative efficiency of adjusted LRT

Under condition H, the adjusted LRT based on $\mathbf{S}_n$ has noncentrality parameter

$$\delta_{\mathbf{h}}(\boldsymbol{\Lambda}, \mathbf{B})/2(1 + k)$$

and for k moderately large, it is expected to have low power. A measure of efficiency can be derived by comparing the adjusted LRT with the exact LRT derived under a particular $E_p(\mathbf{0}, \mathbf{\Lambda})$ with known density defined by the function $g(\cdot)$. The likelihood for $\mathbf{\Lambda}$ built from the sample matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

is

$$L_g(\mathbf{\Lambda}) = |\mathbf{\Lambda}|^{-n/2} \prod_{i=1}^{n} g(\mathbf{x}_i' \mathbf{\Lambda}^{-1} \mathbf{x}_i).$$

Then,

$$\begin{aligned} \hat{\mathbf{\Lambda}}_n &= \arg \min_{\mathbf{\Lambda} > \mathbf{0}} L_g(\mathbf{\Lambda}), \\ \hat{\hat{\mathbf{\Lambda}}}_n &= \arg \min_{\mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}} L_g(\mathbf{\Lambda}) \end{aligned}$$

are respectively the restricted and unrestricted elliptical MLE of $\mathbf{\Lambda}$. Then, the optimal procedure is the LRT derived for a given $g(\cdot)$:

$$L_{\mathbf{h},g}(\mathbf{X}) = \frac{L_g(\hat{\hat{\mathbf{\Lambda}}}_n)}{L_g(\hat{\mathbf{\Lambda}}_n)}.$$

Then, Wald's classical formulation for this "elliptical" LRT is

$$-2 \ln L_{\mathbf{h},g}(\mathbf{X}) \sim n[\mathbf{h}(\hat{\mathbf{\Lambda}}_n)]'[C_{\mathbf{h}}(\hat{\mathbf{\Lambda}}_n)]^{-1} \mathbf{h}(\hat{\mathbf{\Lambda}}_n)/\sigma_{1,g}$$

under H_0 or under the sequence of alternatives $\mathbf{\Lambda}_n = \mathbf{\Lambda} + n^{-1/2}\mathbf{B}$. The parameter $\sigma_{1,g}$ is the value of σ_1 in the asymptotic variance of the MLE given in Proposition 13.7, i.e., $\sigma_{1,g} = p(p+2)/E[\psi^2(s)]$.

Corollary 13.4 *Assume condition H holds. Then:*

(i) *under* H_0, $-2 \ln L_{\mathbf{h},g}(\mathbf{X}) \xrightarrow{d} \chi^2_q$,

(ii) *under the sequence of contiguous alternatives* $\mathbf{\Lambda}_n = \mathbf{\Lambda} + n^{-1/2}\mathbf{B}$, *where* $\mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}$ *and* $\mathbf{B}$ *is a fixed symmetric matrix,*

$$-2 \ln L_{\mathbf{h},g}(\mathbf{X}) \xrightarrow{d} \chi^2_q \left(\delta_{\mathbf{h}}(\mathbf{\Lambda}, \mathbf{B})/2\sigma_{1,g} \right),$$

where, as before,

$$\delta_{\mathbf{h}}(\mathbf{\Lambda}, \mathbf{B}) = [\mathrm{vec}(\mathbf{B})]'[\mathbf{h}'(\mathbf{\Lambda})]'[C_{\mathbf{h}}(\mathbf{\Lambda})]^{-1} \mathbf{h}'(\mathbf{\Lambda}) \ \mathrm{vec}(\mathbf{B}).$$

When $g(\cdot)$ is known, another test which is first-order efficient and asymptotically distributed as chi-square is the minimum geodesic distance test [Berkane et al. (1997)].

The proof of Corollary 13.4 is identical to that of Corollary 13.2. The asymptotic efficiency of the adjusted LRT $-2 \ln L_{\mathbf{h}}(\mathbf{S}_n)/(1 + \hat{k})$ to the

	$\nu=5$	$\nu=6$	$\nu=7$	$\nu=8$	$\nu=30$
$q=1$	.26	.17	.13	.11	.06
$q=2$	.37	.22	.17	.14	.06
$q=3$	.46	.27	.20	.16	.06

Table 13.2. Asymptotic significance level of unadjusted LRT for $\alpha = 5\%$.

elliptical LRT can thus be measured by the ratio of the noncentrality parameters, i.e., $\sigma_{1,g}/(1+k)$. For the $t_{p,\nu}$ density, it was evaluated that $\sigma_{1,g} = 1 + 2/(p+\nu)$, whereas $1+k = (\nu-2)/(\nu-4)$ for a relative efficiency of $(\nu-4)(\nu+2)/[(\nu-2)(\nu+p)]$. For $p=2$ and $\nu=5$, this gives an efficiency of 33%. This is due to the poor robustness property of $\mathbf{S}_n$. This adjusted LRT cannot really be thought of as a robust test because of its low efficiency. To obtain a truly robust adjusted LRT, one has to replace $\mathbf{S}_n$ by an efficient robust estimate, i.e., one with a σ_1 close to $\sigma_{1,g}$.

We conclude this analysis by guarding the practitioner against assuming indiscriminantly the normality of the data and using the "optimal" test for normality. If the data came from an elliptical distribution with a kurtosis parameter k and the hypothesis (satisfying condition H) was $H_0 : \mathbf{h}(\mathbf{\Lambda}) = \mathbf{0}$, where $\mathbf{h}(\mathbf{\Lambda}) \in \mathbb{R}^q$, then what was supposed to be an $\alpha = 5\%$ significance level test would be, in fact, for large samples, a test of significance level:

$$\begin{aligned} P(-2\ln L_{\mathbf{h}}(\mathbf{S}_n) \geq \chi^2_{.95,q}) &= P(-2\ln L_{\mathbf{h}}(\mathbf{S}_n)/(1+k) \geq \chi^2_{.95,q}/(1+k)) \\ &\to P\left(\chi^2_q \geq \chi^2_{.95,q}/(1+k)\right). \end{aligned}$$

For a $t_{p,\nu}$ distribution with $1+k = (\nu-2)/(\nu-4)$, the significance level may be far from 5%, as evidenced by Table 13.2. The situation worsens as ν decreases, which means the tails become heavier or q increases, which is related to the complexity of the hypothesis. For $q=3$ and $\nu=5$, tossing a coin is nearly as reliable!

13.6 Problems

1. Demonstrate the following on normal mixture representation:

   (i) If $\mathbf{x} \stackrel{d}{=} w^{1/2}\mathbf{z}$, where $w \sim F$, $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$, and $w \perp\!\!\!\perp \mathbf{z}$, then

   $$f_{\mathbf{x}}(\mathbf{x}) = \int_0^\infty (2\pi w)^{-p/2} \exp(-\tfrac{1}{2} w^{-1}\mathbf{x}'\mathbf{x}) dF(w),$$

   where $F(\cdot)$ is a distribution function on $[0,\infty)$.

   (ii) If $\nu w^{-1} \sim \chi^2_\nu$, then $\mathbf{x} = w^{1/2}\mathbf{z} \sim t_{p,\nu}$ has density

   $$f_{\mathbf{x}}(\mathbf{x}) = c_{p,\nu}(1 + \mathbf{x}'\mathbf{x}/\nu)^{-(\nu+p)/2}, \ \mathbf{x} \in \mathbb{R}^p,$$

   where $c_{p,\nu} = (\nu\pi)^{-p/2}\Gamma\left(\frac{1}{2}(\nu+p)\right)/\Gamma\left(\frac{1}{2}\nu\right)$.

2. Assume $\mathbf{x} \sim t_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ where $\mathbf{x}$ is partitioned as $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$, $\mathbf{x}_i \in \mathbb{R}^{p_i}$, $i = 1, 2$, $p = p_1 + p_2$. Demonstrate the following:

   (i) $E\,\mathbf{x} = \boldsymbol{\mu}$, $\text{var}\,\mathbf{x} = [\nu/(\nu-2)]\boldsymbol{\Lambda}$, $\nu > 2$.
   (ii) $p^{-1}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim F(p,\nu)$.
   (iii) The marginal distribution is $\mathbf{x}_2 \sim t_{p_2,\nu}(\boldsymbol{\mu}_2, \boldsymbol{\Lambda}_{22})$, where

$$\begin{aligned} \boldsymbol{\mu} &= (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')', \\ \boldsymbol{\Lambda} &= \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \end{aligned}$$

   are partitioned in conformity.
   (iv) The conditional distribution is

$$\mathbf{x}_1|\mathbf{x}_2 \sim t_{p_1,\nu+p_2}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), h(\mathbf{x}_2)\boldsymbol{\Lambda}_{11.2}\right),$$

   where $h(\mathbf{x}_2) = [\nu/(\nu+p_2)] \cdot [1 + (\mathbf{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)/\nu]$. Determine $E(\mathbf{x}_1|\mathbf{x}_2)$, $\text{var}(\mathbf{x}_1|\mathbf{x}_2)$ and the condition for their existence.

3. Verify by differentiation of $\ln\phi(t_1^2 + t_2^2)$ the cumulants

$$\begin{aligned} k_2 &= -2\phi'(0), \\ k_4 &= 12(\phi''(0) - \phi'(0)^2), \\ k_{22} &= 4(\phi''(0) - \phi'(0)^2), \end{aligned}$$

   where $\phi(t_1^2 + t_2^2)$ is the characteristic function of a bivariate rotationally invariant vector.

4. Obtain $E\,\mathbf{u}\mathbf{u}' = p^{-1}\mathbf{I}$ and

$$\text{var}(\mathbf{u}\mathbf{u}') = \frac{1}{p(p+2)}(\mathbf{I} + \mathbf{K}_p) - \frac{2}{p^2(p+2)}\text{vec}(\mathbf{I})[\text{vec}(\mathbf{I})]',$$

   where $\mathbf{u} \sim \text{unif}(S^{p-1})$.

5. For the multivariate power exponential family (13.2), prove the following:

   (i) The normalizing constant is

$$c_{p,\alpha} = \frac{\alpha\Gamma(p/2)}{\pi^{p/2}2^{p/2\alpha}\Gamma(p/2\alpha)}.$$

   (ii) The variance of $\mathbf{x}$ is

$$\text{var}\,\mathbf{x} = \frac{2^{1/\alpha}\Gamma[(p/2+1)/\alpha]}{p\Gamma(p/2\alpha)}\boldsymbol{\Lambda}.$$

   (iii) For $\alpha = 1/2$, verify the assertion in Example 13.4.

   Hint: Use the representation in polar coordinates in Proposition 4.10 together with Problems 4.6.13 and 13.6.4.

6. Check that when $\hat{\mathbf{\Lambda}} = (\hat{\Lambda}_{ij})$ is affine equivariant with variance

$$\text{var } \hat{\mathbf{\Lambda}} = \sigma_1(\mathbf{I} + \mathbf{K}_p)(\mathbf{\Lambda} \otimes \mathbf{\Lambda}) + \sigma_2 \text{ vec}(\mathbf{\Lambda})[\text{vec}(\mathbf{\Lambda})]',$$

then

$$\text{cov}(\hat{\Lambda}_{ki}, \hat{\Lambda}_{lj}) = \sigma_1(\Lambda_{ij}\Lambda_{kl} + \Lambda_{kj}\Lambda_{il}) + \sigma_2\Lambda_{ki}\Lambda_{lj}.$$

7. Prove if $a_0 > p + 1$, $n \geq p + 1$, then condition D1 is satisfied w.p.1 when sampling from an absolutely continuous distribution.

8. Verify conditions M1 through M4 for u_1 and u_2 corresponding to the MLE under the $t_{p,\nu}$ distribution.

9. Verify conditions M1 through M4 for Huber's ψ function in Example 13.10.

10. Assume $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$ with density $g(|\mathbf{z}|^2)$. Define $u(s) = -2g'(s)/g(s)$ and $\psi(s) = su(s)$. Prove $E\ \psi(|\mathbf{z}|^2) = p$.
Hint: Integrate by parts.

11. For the $t_{p,\nu}$ distribution, verify that the asymptotic variance parameter σ_1 of the elliptical MLE in Proposition 13.7 is $\sigma_1 = 1 + 2/(p + \nu)$.

12. Prove $1 + k = pE(s^2)/[(p+2)E^2(s)]$, where $s = |\mathbf{z}|^2$ and $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$ has fourth-order moments.

13. Define $u(s) = -2g'(s)/g(s)$ and $\psi(s) = su(s)$. Let $s = |\mathbf{z}|^2$, where $\mathbf{z} \sim E_p(\mathbf{0}, \mathbf{I})$. Derive an upper bound for $\sigma_{1,g}/(1 + k)$, the index of relative efficiency, by going through the following steps:
(i) $E\ \psi^2(s) \geq [E\ \psi(s)]^2$,
(ii) $E(s^2)\ E[\psi^2(s)] \geq [Es\psi(s)]^2 = (p + 2)^2E^2(s)$,
(iii) $\sigma_{1,g}/(1 + k) \leq \min\{1, (1 + 2p^{-1})(1 + k)^{-1}\}$.
(iv) Interpret the bound in (iii).

14. Demonstrate that if $\rho(t) = t^2$ and $b_0 = p$ in the definition of the S estimate, then the solution is the normal MLE.

15. Demonstrate that the S estimate $(\boldsymbol{\mu}_n, \mathbf{V}_n)$ is necessarily a solution of the equations

$$\begin{aligned} \text{ave}\,[u(t_i)(\mathbf{x}_i - \boldsymbol{\mu}_n)] &= \mathbf{0} \\ \text{ave}\,[pu(t_i)(\mathbf{x}_i - \boldsymbol{\mu}_n)(\mathbf{x}_i - \boldsymbol{\mu}_n)' - v(t_i)\mathbf{V}_n] &= \mathbf{0}, \end{aligned}$$

where $t_i = \left[(\mathbf{x}_i - \boldsymbol{\mu}_n)'\mathbf{V}_n^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_n)\right]^{1/2}$ and $v(t) = t\psi(t) - \rho(t) + b_0$. M and S estimates are close relatives!

16. **Test for a given variance.**
This is a continuation of Problem 8.9.8. The LRT under the $N_p(\mathbf{0}, \mathbf{\Lambda})$ for $H_0 : \mathbf{\Lambda} = \mathbf{I}$ versus $H_1 : \mathbf{\Lambda} \neq \mathbf{I}$ is given by

$$L_{\mathbf{h}}(\mathbf{S}_n) = e^{pn/2}|\mathbf{S}_n|^{n/2} \text{ etr}(-\tfrac{1}{2}n\mathbf{S}_n),$$

where $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$. Suppose that, in fact, $\mathbf{x}_i \sim E_p(\mathbf{0}, \mathbf{\Lambda})$ and one decides to use a robust estimate $\hat{\mathbf{\Lambda}}_n$ satisfying condition E instead of $\mathbf{S}_n$. Then, demonstrate that under H_0,

$$-2\ln L_{\mathbf{h}}(\hat{\mathbf{\Lambda}}_n) \xrightarrow{d} (\sigma_1 + \tfrac{1}{2}\sigma_2 p)\chi_1^2 + \sigma_1\chi^2_{p(p+1)/2-1},$$

where $\chi_1^2 \perp\!\!\!\perp \chi^2_{p(p+1)/2-1}$ by following these steps:

(i) Write $\hat{\mathbf{\Lambda}}_n = \mathbf{I} + n^{-1/2}\mathbf{Z}_n$ and use the Taylor series for $\ln|\mathbf{I} + t\mathbf{A}|$ around $t = 0$ to show that under H_0,

$$-2\ln L_{\mathbf{h}}(\hat{\mathbf{\Lambda}}_n) \xrightarrow{d} \tfrac{1}{2}[\text{vec}(\mathbf{Z})]'\ \text{vec}(\mathbf{Z}),$$

where $\mathbf{Z} \sim N_p^p(\mathbf{0}, \sigma_1(\mathbf{I} + \mathbf{K}_p) + \sigma_2\ \text{vec}(\mathbf{I})[\text{vec}(\mathbf{I})]')$.

(ii) Demonstrate

$$\text{var}(z_{11}/\sqrt{2}, \ldots, z_{pp}/\sqrt{2}, z_{12}, \ldots, z_{1p}, z_{23}, \ldots, z_{2p}, \ldots, z_{p-1,p})'$$

is given by

$$\begin{pmatrix} \sigma_1\mathbf{I}_p + \frac{1}{2}\sigma_2\mathbf{1}\mathbf{1}' & \mathbf{0} \\ \mathbf{0} & \sigma_1\mathbf{I}_{p(p-1)/2} \end{pmatrix} \equiv \mathbf{\Omega}.$$

(iii) Verify the eigenvalues of $\mathbf{\Omega}$ are σ_1 of multiplicity $\frac{1}{2}p(p+1) - 1$ and $\sigma_1 + \frac{1}{2}\sigma_2 p$ of multiplicity 1.

17. **Test of multiple correlation.**
The LRT under $(x_1, \mathbf{x}_2')' \sim N_p(\mathbf{0}, \mathbf{\Lambda})$ for $H_0 : R^2 = 0$ versus $H_1 : R^2 \neq 0$ is given by

$$L_{\mathbf{h}}(\mathbf{S}_n) = (1 - \hat{R}^2(\mathbf{S}_n))^{n/2},$$

where $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$ and $\hat{R}^2(\mathbf{S}_n) = \mathbf{s}_{21}'\mathbf{S}_{22}^{-1}\mathbf{s}_{21}/s_{11}$ in terms of the partition

$$\mathbf{S}_n = \begin{pmatrix} s_{11} & \mathbf{s}_{21}' \\ \mathbf{s}_{21} & \mathbf{S}_{22} \end{pmatrix}.$$

Suppose, in fact, that $(x_1, \mathbf{x}_2')' \sim E_p(\mathbf{0}, \mathbf{\Lambda})$ and one decides to use a robust estimate $\hat{\mathbf{\Lambda}}_n$ satisfying condition E instead of $\mathbf{S}_n$. Then, demonstrate that under H_0,

$$-2\ln L_{\mathbf{h}}(\hat{\mathbf{\Lambda}}_n) \xrightarrow{d} \sigma_1\chi^2_{p-1},$$

by following these steps:

(i) Argue that one can assume $\mathbf{\Lambda} = \mathbf{I}$.

(ii) Using a Taylor series, prove that

$$-2\ln L_{\mathbf{h}}(\hat{\mathbf{\Lambda}}_n) \sim n\hat{R}^2(\hat{\mathbf{\Lambda}}_n).$$

(iii) Finally, prove that $n\hat{R}^2(\hat{\mathbf{\Lambda}}_n) \xrightarrow{d} \mathbf{z}'\mathbf{z}$, where $\mathbf{z} \sim N_{p-1}(\mathbf{0}, \sigma_1\mathbf{I})$ to conclude.

18. Let $\mathbf{l} = (l_1, \ldots, l_p)'$ be the eigenvalues of $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$, calculated from a sample of an $E_p(\mathbf{0}, \mathbf{\Lambda})$ distribution, $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$. If the population eigenvalues λ_α are all distinct, then prove that the joint limiting distribution is given by

$$n^{1/2}\mathbf{\Lambda}^{-1}(\mathbf{l} - \boldsymbol{\lambda}) \xrightarrow{d} N_p(\mathbf{0}, 2(1+k)\mathbf{I} + k\mathbf{1}\mathbf{1}').$$

19. Suppose the sample is taken from an elliptical distribution with kurtosis $3k$. Let $\mathbf{f} = (f_1, \ldots, f_p)'$ be the eigenvalues of the sample correlation matrix $\mathbf{R} = (r_{ij})$. If the eigenvalues γ_α of the population correlation matrix

$$\boldsymbol{\rho} = (\rho_{ij}) = \mathbf{G}\ \text{diag}(\gamma_1, \ldots, \gamma_p)\ \mathbf{G}',$$

where $\mathbf{G} = (g_{ij}) \in \mathbf{O}_p$, are all distinct, then prove that the joint limiting distribution is

$$n^{1/2}(\mathbf{f} - \boldsymbol{\gamma}) \xrightarrow{d} N_p(\mathbf{0}, (1+k)\mathbf{\Omega}),$$

where $\mathbf{\Omega} = (\omega_{\alpha\beta})$ is given by

$$\omega_{\alpha\beta} = 2\gamma_\alpha\gamma_\beta \left[\delta_{\alpha\beta} - (\gamma_\alpha + \gamma_\beta)\sum_{j=1}^{p} g_{j\alpha}^2 g_{j\beta}^2 + \sum_{j=1}^{p}\sum_{i=1}^{p} g_{j\alpha}^2 g_{i\beta}^2 \rho_{ji}^2\right].$$

14
Bootstrap confidence regions and tests

An important part of multivariate analysis deals with confidence regions and tests of hypotheses on the mean vector and variance matrix. The classical theoretical developments for such procedures rest mainly upon the multivariate normality assumption. Without multivariate normality, the asymptotic distribution of many tests becomes more complex and often leads to untabulated limit distributions. The bootstrap confidence regions and tests on the mean vector and variance matrix have the desired asymptotic levels under very mild conditions. We will present the bootstrap technique main ideas without formal proofs. The interested reader should consult the cited references. General references for the bootstrap are Efron (1982), who made the technique widely applicable by using modern computational power, Efron and Tibshirani (1993) and Hall (1992). The book by Davison and Hinkley (1997) has S-plus code which may prove useful.

14.1 Confidence regions and tests for the mean

Let $\mathbf{x} = (x_1, \ldots, x_p)' \sim F$ with mean $\boldsymbol{\mu}_F = (\mu_{F,i})$ and variance $\boldsymbol{\Sigma}_F = (\sigma_{F,ij})$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. F and

$$\begin{aligned} \bar{\mathbf{x}}_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \\ \mathbf{S}_n &= \frac{1}{(n-1)}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_n)(\mathbf{x}_i - \bar{\mathbf{x}}_n)', \end{aligned}$$

be the sample mean and sample variance, respectively. Define the "pivot"

$$w_n = n^{1/2}|\mathbf{S}_n^{-1/2}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_F)|.$$

Then, by the central limit theorem,

$$w_n \xrightarrow{d} |\mathbf{z}|, \ \mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I}).$$

The empirical distribution function of the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is denoted by

$$\hat{F}_n(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{x}_i \leq \mathbf{t}),$$

where $I(\mathbf{x}_i \leq \mathbf{t})$ is the indicator function. In other words, $\hat{F}_n$ is the discrete distribution function with equal probability $1/n$ at the points $\mathbf{x}_i$, $i = 1, \ldots, n$. Then, for $\mathbf{x}^* \sim \hat{F}_n$, we have

$$\begin{aligned} E\ \mathbf{x}^* &= \boldsymbol{\mu}_{\hat{F}_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \\ \text{var}\ \mathbf{x}^* &= \boldsymbol{\Sigma}_{\hat{F}_n} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_n)(\mathbf{x}_i - \bar{\mathbf{x}}_n)'. \end{aligned}$$

The nonparametric bootstrap estimate of the probability distribution of w_n under F, $J_n(F)$, is the bootstrap estimate $J_n(\hat{F}_n)$, which can be interpreted as follows. Let $\mathbf{x}_i^*$, $i = 1, \ldots, n$, be i.i.d. $\hat{F}_n$, and let $\bar{\mathbf{x}}_n^*$ and $\mathbf{S}_n^*$ be the sample mean and sample variance, respectively, of the $\mathbf{x}_i^*$'s. Then, $J_n(\hat{F}_n)$ is the probability law under $\hat{F}_n$ of

$$w_n^* = n^{1/2}|\mathbf{S}_n^{*\,-1/2}(\bar{\mathbf{x}}_n^* - \boldsymbol{\mu}_{\hat{F}_n})|.$$

In practice, $J_n(\hat{F}_n)$ may be approximated to any degree of accuracy with resampling by Monte Carlo methods. The consistency of the bootstrap was established by Beran (1984, example 3) using a triangular array version of the C.L.T.,

$$w_n^* \xrightarrow{d} |\mathbf{z}| \text{ w.p.1},$$

which means that w_n and w_n^* converge in distribution to the same limit. However, it was Singh (1981), and Bickel and Freedman (1981) who first established the consistency of the bootstrap in the univariate situation. Let $c_n(\alpha, \hat{F}_n)$ be a $(1-\alpha)$-quantile of the bootstrap distribution $J_n(\hat{F}_n)$. By the consistency of the bootstrap, if $D_n(\alpha)$ is the bootstrap confidence region for $\boldsymbol{\mu}_F$,

$$D_n(\alpha) = \{\boldsymbol{\mu}_F : \ n^{1/2}|\mathbf{S}_n^{-1/2}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_F)| \leq c_n(\alpha, \hat{F}_n)\},$$

then

$$\lim_{n\to\infty} P\left(\boldsymbol{\mu}_F \in D_n(\alpha)\right) = 1 - \alpha.$$

The bootstrap confidence region $D_n(\alpha)$ handles all norms, $|\cdot|$, on $\mathbb{R}^p$ with equal ease. Most often though, the euclidian norm is intended, and in that case, the ellipsoidal bootstrap confidence region can be written as

$$D_n(\alpha) = \{\boldsymbol{\mu}_F : \ n(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_F)'\mathbf{S}_n^{-1}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}_F) \le c_n^2(\alpha, \hat{F}_n)\}.$$

A $(1-\alpha)$-acceptance region, $A(\boldsymbol{\mu}_0)$, for testing the hypothesis $H_0 : \boldsymbol{\mu}_F = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu}_F \neq \boldsymbol{\mu}_0$ may be obtained by inverting the confidence region, $D_n(\alpha)$, in the usual way [Fraser (1976), p. 580]. Here, the test which rejects $H_0 : \boldsymbol{\mu}_F = \boldsymbol{\mu}_0$ iff $\boldsymbol{\mu}_0 \notin D_n(\alpha)$ is a test with asymptotic type I error probability α.

More generally, suppose a confidence region on $\mathbf{g}(\boldsymbol{\mu}_F) \in \mathbb{R}^k$, $k \le p$, is wanted where $\mathbf{g} : \ \mathbb{R}^p \to \mathbb{R}^k$ is a continuously differentiable function and has first derivative $\dot{\mathbf{g}} \in \mathbb{R}_p^k$. Let $u : \ \mathbb{R}^k \to \mathbb{R}$ be continuous on $\mathbb{R}^k$ such that

$$\{\mathbf{z} \in \mathbb{R}^k : \ u(\mathbf{z}) = c\}$$

has Lebesgue measure 0 for every $c \in \mathbb{R}$. Consider the statistic

$$w_{n,\mathbf{g}} = u\left[n^{1/2}\left(\mathbf{g}(\bar{\mathbf{x}}_n) - \mathbf{g}(\boldsymbol{\mu}_F)\right)\right].$$

The central limit theorem coupled with the delta method yields

$$w_{n,\mathbf{g}} \xrightarrow{d} u[\dot{\mathbf{g}}(\boldsymbol{\mu}_F)\ \mathbf{z}_F], \ \mathbf{z}_F \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_F).$$

The condition imposed on u ensures that the limit distribution is continuous. Using arguments as in Beran (1984), it can be established that the bootstrap estimate is consistent, i.e.,

$$w^*_{n,\mathbf{g}} = u\left[n^{1/2}\left(\mathbf{g}(\bar{\mathbf{x}}^*_n) - \mathbf{g}(\boldsymbol{\mu}_{\hat{F}_n})\right)\right] \xrightarrow{d} u[\dot{\mathbf{g}}(\boldsymbol{\mu}_F)\ \mathbf{z}_F] \text{ w.p.1.}$$

To construct the bootstrap confidence region, let $c_{n,\mathbf{g}}(\alpha, \hat{F}_n)$ be a $(1-\alpha)$-quantile of the bootstrap distribution. A bootstrap confidence region for $\mathbf{g}(\boldsymbol{\mu}_F)$ having asymptotic coverage probability $1-\alpha$ is

$$D_{n,\mathbf{g}}(\alpha) = \{\mathbf{g}(\boldsymbol{\mu}_F) : \ u\left[n^{1/2}\left(\mathbf{g}(\bar{\mathbf{x}}_n) - \mathbf{g}(\boldsymbol{\mu}_F)\right)\right] \le c_{n,\mathbf{g}}(\alpha, \hat{F}_n)\}.$$

In the examples to be considred, the function u has the additional property, $u(b\mathbf{z}) = bu(\mathbf{z})$, $\forall \mathbf{z} \in \mathbb{R}^k$, $\forall b > 0$. Then, the factor $n^{1/2}$ may be omitted and we may write equivalently

$$D_{n,\mathbf{g}}(\alpha) = \{\mathbf{g}(\boldsymbol{\mu}_F) : \ u\left[\mathbf{g}(\bar{\mathbf{x}}_n) - \mathbf{g}(\boldsymbol{\mu}_F)\right] \le c^*_{n,\mathbf{g}}(\alpha)\},$$

where $c^*_{n,\mathbf{g}}(\alpha)$ is a $(1-\alpha)$-quantile of the distribution of $u[\mathbf{g}(\bar{\mathbf{x}}^*_n) - \mathbf{g}(\boldsymbol{\mu}_{\hat{F}_n})]$ when $\hat{F}_n$ is fixed at its realized value and $\bar{\mathbf{x}}^*_n$ is the bootstrap sample mean. Judicious choices of u and $\mathbf{g}$ give interesting confidence regions, as the following examples show.

Example 14.1 *Let* $\mathbf{g}(\boldsymbol{\mu}_F) = \boldsymbol{\mu}_F$ *and* $u(\mathbf{z}) = |\mathbf{z}|_1 = \sum_{i=1}^p |z_i|$ *be the* l_1*-norm. Then, the bootstrap confidence region*

$$D_{n,\mathbf{g}}(\alpha) = \left\{ \boldsymbol{\mu}_F : \sum_{i=1}^p |\bar{x}_{n,i} - \mu_{F,i}| \le c^*_{n,\mathbf{g}}(\alpha) \right\},$$

has asymptotic coverage probability $1-\alpha$*, where* $c^*_{n,\mathbf{g}}(\alpha)$ *is a* $(1-\alpha)$*-quantile of the distribution of* $\sum_{i=1}^p |\bar{x}^*_{n,i} - \mu_{\hat{F}_n,i}|$ *when* $\hat{F}_n$ *is fixed at its realized value and* $\bar{\mathbf{x}}^*_n$ *is the bootstrap sample mean.*

Example 14.2 *Let* $\mathbf{g}(\boldsymbol{\mu}_F) = \boldsymbol{\mu}_F$ *and* $u(\mathbf{z}) = |\mathbf{z}|_\infty = \max_{1\le i\le p} |z_i|$ *be the* l_∞*-norm. The bootstrap simultaneous confidence intervals*

$$D_{n,\mathbf{g}}(\alpha) = \{\boldsymbol{\mu}_F : \ |\bar{x}_{n,i} - \mu_{F,i}| \le c^*_{n,\mathbf{g}}(\alpha), \ i = 1, \ldots, p\},$$

have asymptotic simultaneous coverage probability $1-\alpha$*, where* $c^*_{n,\mathbf{g}}(\alpha)$ *is a* $(1-\alpha)$*-quantile of the distribution of* $\max_{1\le i\le p} |\bar{x}^*_{n,i} - \mu_{\hat{F}_n,i}|$ *when* $\hat{F}_n$ *is fixed at its realized value and* $\bar{\mathbf{x}}^*_n$ *is the bootstrap sample mean.*

Example 14.3 *This example provides the bootstrap algorithm, easy to implement on a computer, to construct simultaneous confidence intervals on the means* $\mu_{F,i}$*,* $i = 1, \ldots, p$*. We are given a sample* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *from* F*.*

Bootstrap algorithm

- *Calculate* $\bar{\mathbf{x}}_n = (\bar{x}_{n,i})$*.*
- $b \leftarrow 1$
- $B \leftarrow 2000$ *(say)*
- *Do while* $b \le B$*.*
  - *Draw a bootstrap sample* $\mathbf{x}^*_1, \ldots, \mathbf{x}^*_n$ *from* $\hat{F}_n$*.*
  - *Calculate* $\bar{\mathbf{x}}^*_n = (\bar{x}^*_{n,i})$*.*
  - $u_b \leftarrow \max_{1\le i\le p} |\bar{x}^*_{n,i} - \bar{x}_{n,i}|$
  - $b \leftarrow b + 1$
- *End.*
- *Order the* u_b*'s:* $u_{(1)} \le u_{(2)} \le \cdots \le u_{(B)}$*.*
- $q \leftarrow \lfloor (1-\alpha)B \rfloor$ *(*$\lfloor\cdot\rfloor$ *is the integer part function)*
- *Simultaneous confidence intervals for* $\mu_{F,i}$ *with approximate simultaneous coverage probability* $1-\alpha$ *are*

$$\bar{x}_{n,i} - u_{(q)} \le \mu_{F,i} \le \bar{x}_{n,i} + u_{(q)}, \ i = 1, \ldots, p.$$

14.2 Confidence regions for the variance

This time, define

$$\mathbf{W}_n = n^{1/2}(\mathbf{S}_n - \boldsymbol{\Sigma}_F).$$

The asymptotic distribution of $\mathbf{S}_n$ was given in Section 6.3 for all underlying distribution F with finite fourth moments. The asymptotic distribution is

$$\mathbf{W}_n \xrightarrow{d} \mathbf{X}_F,\ \mathbf{X}_F \sim N_p^p(\mathbf{0}, \mathbf{\Omega}_F),$$

where $\mathbf{X}_F = (x_{F,ij})$, the elements of $\mathbf{\Omega}_F$ given by

$$\text{cov}(x_{F,ik}, x_{F,jl}) = k_{1111}^{ijkl} + k_{11}^{kl} k_{11}^{ij} + k_{11}^{il} k_{11}^{jk}$$

with the k's representing the cumulants of F. The nonparametric bootstrap estimate of the probability distribution under F of $\mathbf{W}_n$ is the probability distribution under $\hat{F}_n$ of

$$\mathbf{W}_n^* = n^{1/2} \left(\mathbf{S}_n^* - \mathbf{\Sigma}_{\hat{F}_n} \right).$$

Beran and Srivastava (1985) established the consistency of the bootstrap

$$\mathbf{W}_n^* \xrightarrow{d} \mathbf{X}_F \text{ w.p.1.}$$

A difficulty in deriving a confidence region for a function of $\mathbf{\Sigma}_F$ is the redundancy of elements due to the symmetry of $\mathbf{\Sigma}_F$. So let

$$\text{uvec}(\mathbf{S}) = (s_{11}, s_{12}, s_{22}, \ldots, s_{1p}, s_{2p}, \ldots, s_{pp})'$$

be the vec operator applied only to the upper triangular part of $\mathbf{S} \in \mathcal{S}_p$. Suppose a confidence region for $\mathbf{g}(\mathbf{\Sigma}_F) \in \mathbb{R}^k$ is desired where $\mathbf{g}$ is a function of $\text{uvec}(\mathbf{\Sigma}_F)$, which is continuously differentiable and has first derivative $\dot{\mathbf{g}} \in \mathbb{R}^k_{p(p+1)/2}$. Let $u: \mathbb{R}^k \to \mathbb{R}$ be continuous on $\mathbb{R}^k$ such that

$$\{\mathbf{z} \in \mathbb{R}^k : \ u(\mathbf{z}) = c\}$$

has Lebesgue measure 0 for every $c \in \mathbb{R}$ and $u(b\mathbf{z}) = bu(\mathbf{z})$, $\forall \mathbf{z} \in \mathbb{R}^k$, $\forall b > 0$. Let

$$\mathbf{W}_{n,\mathbf{g}} = u \left[n^{1/2} \left(\mathbf{g}(\mathbf{S}_n) - \mathbf{g}(\mathbf{\Sigma}_F) \right) \right].$$

The delta method (v. Proposition 6.2) immediately yields

$$\mathbf{W}_{n,\mathbf{g}} \xrightarrow{d} u[\dot{\mathbf{g}}(\mathbf{\Sigma}_F)\ \text{uvec}(\mathbf{X}_F)] \tag{14.1}$$

and

$$\mathbf{W}_{n,\mathbf{g}}^* \xrightarrow{d} u[\dot{\mathbf{g}}(\mathbf{\Sigma}_F)\ \text{uvec}(\mathbf{X}_F)] \text{ w.p.1,}$$

so the bootstrap is consistent [Beran and Srivastava (1985)]. The condition on u implies the limiting distribution in (14.1) is continuous. A bootstrap confidence region for $\mathbf{g}(\mathbf{\Sigma}_F)$ having asymptotic coverage probability $1 - \alpha$ is

$$D_{n,\mathbf{g}}(\alpha) = \{\mathbf{g}(\mathbf{\Sigma}_F) : \ u\left[\mathbf{g}(\mathbf{S}_n) - \mathbf{g}(\mathbf{\Sigma}_F)\right] \le c_{n,\mathbf{g}}^*(\alpha)\},$$

where $c_{n,\mathbf{g}}^*(\alpha)$ is a $(1-\alpha)$-quantile of the distribution of $u[\mathbf{g}(\mathbf{S}_n^*) - \mathbf{g}(\mathbf{\Sigma}_{\hat{F}_n})]$ when $\hat{F}_n$ is fixed at its realized value and $\mathbf{S}_n^*$ is the bootstrap sample variance.

Example 14.4 *Let*

$$\mathbf{g}(\mathbf{\Sigma}_F) = \rho_{F,ij} = \frac{\sigma_{F,ij}}{\sigma_{F,ii}^{1/2}\sigma_{F,jj}^{1/2}}.$$

Then, $\mathbf{g}(\mathbf{S}_n)$ *is the sample correlation coefficient* $r_{n,ij}$ *and bootstrap confidence regions based on* $|r_{n,ij} - \rho_{F,ij}|$ *have the correct asymptotic coverage probability. Also covered is the Fisher z-transform*

$$g(\mathbf{\Sigma}_F) = \tfrac{1}{2}\ln\left(\frac{1+\rho_{F,ij}}{1-\rho_{F,ij}}\right) = \tanh^{-1}(\rho_{F,ij}).$$

Example 14.5 *This example provides the bootstrap algorithm, in an easily programmable form, to construct a confidence interval for the correlation coefficient* $\rho_{F,ij}$ *using the Fisher z-transformation to stabilize the variance. We are given a sample* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *from* F.

Bootstrap algorithm

- *Calculate* $\mathbf{S}_n = (s_{n,ij})$.
- *Calculate* $r_{n,ij} = s_{n,ij}/[s_{n,ii}^{1/2}s_{n,jj}^{1/2}]$.
- $b \leftarrow 1$
- $B \leftarrow 2000$ *(say)*
- *Do while* $b \le B$.
  - *Draw a bootstrap sample* $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$ *from* $\hat{F}_n$.
  - *Calculate* $\mathbf{S}_n^* = (s_{n,ij}^*)$.
  - *Calculate* $r_{n,ij}^* = s_{n,ij}^*/[s_{n,ii}^{*1/2}\, s_{n,jj}^{*1/2}]$.
  - $u_b \leftarrow \left|\tanh^{-1}(r_{n,ij}^*) - \tanh^{-1}(r_{n,ij})\right|$
  - $b \leftarrow b + 1$
- *End.*
- *Order the* u_b*'s:* $u_{(1)} \le u_{(2)} \le \cdots \le u_{(B)}$.
- $q \leftarrow \lfloor (1-\alpha)B \rfloor$
- *An approximate* $(1-\alpha)$ *confidence interval for* $\rho_{F,ij}$ *is*

$$\tanh[\tanh^{-1}(r_{n,ij}) - u_{(q)}] \le \rho_{F,ij} \le \tanh[\tanh^{-1}(r_{n,ij}) + u_{(q)}].$$

Example 14.6 *Let*

$$\phi_1(\mathbf{\Sigma}_F) > \phi_2(\mathbf{\Sigma}_F) > \cdots > \phi_p(\mathbf{\Sigma}_F) > 0$$

be the ordered eigenvalues of $\mathbf{\Sigma}_F$ *assumed distinct. The vector*

$$\boldsymbol{\phi}(\mathbf{\Sigma}_F) = (\phi_1(\mathbf{\Sigma}_F), \ldots, \phi_p(\mathbf{\Sigma}_F))'$$

is a continuously differentiable function of $\mathrm{uvec}(\mathbf{\Sigma}_F)$ *[Kato* (1982)*, Section* 6 *of Chapter* 2*]. The ordered sample eigenvalues are*

$$\boldsymbol{\phi}(\mathbf{S}_n) = (\phi_1(\mathbf{S}_n), \ldots, \phi_p(\mathbf{S}_n))'.$$

The bootstrap confidence region based on

$$\max_{1\le i\le p} |\ln\phi_i(\mathbf{S}_n) - \ln\phi_i(\mathbf{\Sigma}_F)| \tag{14.2}$$

has the correct asymptotic coverage probability. Here, $u(\mathbf{z}) = \max_{1 \le i \le p} |z_i|$, $\mathbf{z} \in \mathbb{R}^p$. *The logarithmic transformation stabilizes the variance in the normal model asymptotic for sample eigenvalues (v. Problem* 8.9.15*). The bootstrap confidence region for* $\boldsymbol{\phi}(\boldsymbol{\Sigma}_F)$ *corresponding to* (14.2) *is*

$$\{\phi_i(\boldsymbol{\Sigma}_F) : \ \phi_i(\mathbf{S}_n)/A_n \le \phi_i(\boldsymbol{\Sigma}_F) \le \phi_i(\mathbf{S}_n)A_n, \ i = 1, \ldots, p\},$$

where

$$A_n = e^{c^*_{n,\mathbf{g}}(\alpha)}$$

and $c^*_{n,\mathbf{g}}(\alpha)$ *is a* $(1-\alpha)$*-quantile of the distribution of*

$$\max_{1 \le i \le p} |\ln \phi_i(\mathbf{S}^*_n) - \ln \phi_i(\boldsymbol{\Sigma}_{\hat{F}_n})|$$

when $\hat{F}_n$ *is fixed at its realized value and* $\mathbf{S}^*_n$ *is the bootstrap sample variance.*

The problem of efficiently bootstrapping sample eigenvalues when $\boldsymbol{\Sigma}_F$ may have multiple eigenvalues is still an unresolved problem [Beran and Srivastava (1987), Eaton and Tyler (1991)].

14.3 Tests on the variance

Rather than inverting a confidence region, it is sometimes possible to construct bootstrap tests directly from test statistics. This approach [Beran and Srivastava (1985)] to testing structural hypotheses about $\boldsymbol{\Sigma}_F$ is the subject of this section.

Assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. F with finite fourth moments. Let $\pi : \mathcal{P}_p \to \mathcal{P}_p$ be a linear projection ($\pi^2 = \pi$), not the identity map. Suppose $T_n(\mathbf{S}_n) = n\, h(\mathbf{S}_n)$ is a test statistic for the null hypothesis,

$$H_0 : \ \boldsymbol{\Sigma}_F = \pi(\boldsymbol{\Sigma}_F).$$

Let F_0 be any distribution function satisfying H_0.

Example 14.7 *Define the constant linear projection* $\pi(\boldsymbol{\Sigma}_F) = \mathbf{I}$. *Then, the hypothesis* $H_0 : \ \boldsymbol{\Sigma}_F = \mathbf{I}$ *is equivalent to* $H_0 : \ \boldsymbol{\Sigma}_F = \pi(\boldsymbol{\Sigma}_F)$.

Example 14.8 *Partition* $\boldsymbol{\Sigma}_F$ *as*

$$\boldsymbol{\Sigma}_F = \begin{pmatrix} \sigma_{F,11} & \boldsymbol{\sigma}'_{F,21} \\ \boldsymbol{\sigma}_{F,21} & \boldsymbol{\Sigma}_{F,22} \end{pmatrix}$$

and define the linear projection

$$\pi(\boldsymbol{\Sigma}_F) = \begin{pmatrix} \sigma_{F,11} & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Sigma}_{F,22} \end{pmatrix}.$$

The hypothesis on the multiple correlation, $H_0 : \ R = 0$, *or* $H_0 : \ \boldsymbol{\sigma}_{12} = \mathbf{0}$ *is equivalent to* $H_0 : \ \boldsymbol{\Sigma}_F = \pi(\boldsymbol{\Sigma}_F)$.

Example 14.9 *Define the linear map* $\pi(\mathbf{\Sigma}_F) = (\sum_{i=1}^{p} \sigma_{F,ii}/p)\,\mathbf{I}$. *Then, the sphericity hypothesis* $H_0: \mathbf{\Sigma}_F = \gamma\mathbf{I},\ \gamma > 0$ *is equivalent to* $H_0: \mathbf{\Sigma}_F = \pi(\mathbf{\Sigma}_F)$.

The function h defining the test statistic $T_n(\mathbf{S}_n)$ is twice continuously differentiable at $\text{uvec}(\mathbf{\Sigma}_{F_0}) \in \mathbb{R}^{p(p+1)/2}$, with $h(\mathbf{\Sigma}_{F_0}) = 0$ and $\dot{h}(\mathbf{\Sigma}_{F_0}) = \mathbf{0}$. This formulation includes the normal model likelihood ratio test in particular. Let $\ddot{h} \in \mathbb{R}^{p(p+1)/2}_{p(p+1)/2}$ denote the second derivative of h and $\mathbf{x}_{F_0} \stackrel{d}{=} \text{uvec}(\mathbf{X}_{F_0})$. Then, using the Taylor series,

$$T_n(\mathbf{S}_n)|F_0 \stackrel{d}{\to} \mathbf{x}'_{F_0}\ddot{h}(\mathbf{\Sigma}_{F_0})\mathbf{x}_{F_0}.$$

We can construct a bootstrap estimate for the null distribution of $T_n(\mathbf{S}_n)$ as follows. Let

$$\mathbf{V}_{n,F} = [\pi(\mathbf{\Sigma}_F)]^{1/2}\mathbf{\Sigma}_F^{-1/2}\mathbf{S}_n\mathbf{\Sigma}_F^{-1/2}[\pi(\mathbf{\Sigma}_F)]^{1/2}.$$

The bootstrap estimate for the null distribution of $T_n(\mathbf{S}_n)$ is defined to be that of $T_n(\mathbf{V}_{n,\hat{F}_n})$. Let $d_{n,h}(\alpha, \hat{F}_n)$ be a $(1-\alpha)$-quantile of $T_n(\mathbf{V}_{n,\hat{F}_n})$. Beran and Srivastava (1985) established the consistency of the bootstrap,

$$T_n(\mathbf{V}_{n,\hat{F}_n}) \stackrel{d}{\to} \mathbf{x}'_{F_0}\ddot{h}(\mathbf{\Sigma}_{F_0})\mathbf{x}_{F_0} \text{ w.p.1.}$$

Hence, the test which rejects H_0 whenever $T_n(\mathbf{S}_n) > d_{n,h}(\alpha, \hat{F}_n)$ has asymptotic size α, provided $\ddot{h}(\mathbf{\Sigma}_{F_0}) \neq \mathbf{0}$.

In practice the bootstrap null distribution can be constructed as follows. Let

$$\mathbf{y}_i = [\pi(\mathbf{S}_n)]^{1/2}\mathbf{S}_n^{-1/2}\mathbf{x}_i,\ i = 1, \ldots, n.$$

Let $\hat{F}_{n,\mathbf{y}}$ be the empirical distribution function of the $\mathbf{y}_i$'s. Note that $\mathbf{\Sigma}_{\hat{F}_{n,\mathbf{y}}} = \pi(\mathbf{\Sigma}_{\hat{F}_n})$, which satisfies H_0 since $\pi = \pi^2$. If $\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*$ are i.i.d. $\hat{F}_{n,\mathbf{y}}$ and $\mathbf{S}^*_{n,\mathbf{y}}$ is the sample variance of the $\mathbf{y}_i^*$'s, then $T_n(\mathbf{V}_{n,\hat{F}_n}) \stackrel{d}{=} T_n(\mathbf{S}^*_{n,\mathbf{y}})$ whose distribution can be approximated by Monte Carlo methods.

Example 14.10 *We wish to test the hypothesis* $H_0: \mathbf{\Sigma}_{F,12} = \mathbf{0}$ *using the invariant test statistic (v. Section* 11.3*)*

$$\begin{aligned} T_n &= n\ \text{tr}[\mathbf{S}_{n,12}\mathbf{S}_{n,22}^{-1}\mathbf{S}_{n,21}\mathbf{S}_{n,11}^{-1}] \\ &= n\sum_{i=1}^{p_1} r_{n,i}^2, \end{aligned}$$

where $r_{n,i}^2$ *are the squared sample canonical correlations. The linear projection in this case is defined by*

$$\pi(\mathbf{S}_n) = \begin{pmatrix} \mathbf{S}_{n,11} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{n,22} \end{pmatrix}$$

with its square root

$$[\pi(\mathbf{S}_n)]^{1/2} = \begin{pmatrix} \mathbf{S}_{n,11}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{n,22}^{1/2} \end{pmatrix}.$$

We are given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *from* F.

Bootstrap algorithm

- *Calculate* $\mathbf{S}_n$ *and partition*

$$\mathbf{S}_n = \begin{pmatrix} \mathbf{S}_{n,11} & \mathbf{S}_{n,12} \\ \mathbf{S}_{n,21} & \mathbf{S}_{n,22} \end{pmatrix}.$$

- *Calculate the square roots* $\mathbf{S}_n^{1/2}$, $\mathbf{S}_{n,11}^{1/2}$, *and* $\mathbf{S}_{n,22}^{1/2}$ *and the inverse* $\mathbf{S}_n^{-1/2}$.
- *Transform* $\mathbf{y}_i = [\pi(\mathbf{S}_n)]^{1/2}\mathbf{S}_n^{-1/2}\mathbf{x}_i$, $i = 1, \ldots, n$.
- $b \leftarrow 1$
- $B \leftarrow 2000$ *(say)*
- *Do while* $b \leq B$.
  - *Draw a bootstrap sample* $\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*$ *from* $\hat{F}_{n,\mathbf{y}}$.
  - *Calculate* $\mathbf{S}_n^*$ *and partition*

$$\mathbf{S}_n^* = \begin{pmatrix} \mathbf{S}_{n,11}^* & \mathbf{S}_{n,12}^* \\ \mathbf{S}_{n,21}^* & \mathbf{S}_{n,22}^* \end{pmatrix}.$$

  - $u_b \leftarrow n \ \mathrm{tr}[\mathbf{S}_{n,12}^* \mathbf{S}_{n,22}^{*^{-1}} \mathbf{S}_{n,21}^* \mathbf{S}_{n,11}^{*^{-1}}]$
  - $b \leftarrow b + 1$
- *End.*
- *Order the* u_b*'s:* $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(B)}$.
- $q \leftarrow \lfloor (1-\alpha)B \rfloor$
- *An approximate size* α *test rejects* $H_0 : \mathbf{\Sigma}_{F,12} = \mathbf{0}$ *whenever* $T_n > u_{(q)}$.

It is an easy matter to modify this bootstrap algorithm to bootstrap the test statistic

$$\begin{aligned} T_n &= n \ \mathrm{tr}\left[\mathbf{S}_{n,12}\mathbf{S}_{n,22}^{-1}\mathbf{S}_{n,21}\mathbf{S}_{n,11}^{-1}\left(\mathbf{I} - \mathbf{S}_{n,12}\mathbf{S}_{n,22}^{-1}\mathbf{S}_{n,21}\mathbf{S}_{n,11}^{-1}\right)^{-1}\right] \\ &= n\sum_{i=1}^{p_1} r_{n,i}^2/(1-r_{n,i}^2). \end{aligned}$$

However, the test based on the largest sample canonical correlation, $T_n = n\, r_{n,1}^2$, *should not be bootstraped unless the user is sure the largest population canonical correlation is distinct. In case of multiplicity the population canonical correlations are not a differentiable function of* $\mathbf{\Sigma}_F$ *[Kato* (1982)*, Section* 6 *of Chapter* 2*].*

Bootstrap algorithms for estimating the power function of a test statistic can be found in Beran (1986). Nagao and Srivastava (1992) considered high-order asymptotic expansions to the distribution of some test criteria on the variance matrix under local alternatives. For the test of sphericity

in dimension $p = 3$, they compared these expansions to the bootstrap approximations for both the normal model likelihood ratio test and Nagao's test when the distribution is actually multivariate normal or multivariate t.

14.4 Problem

1. John (1971) showed that the test based on

$$J = \text{tr } \mathbf{V}^2/(\text{tr } \mathbf{V})^2,$$

where $\mathbf{V} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, is LBI for the hypothesis of sphericity, $H_0 : \mathbf{\Sigma}_F = \gamma\mathbf{I}$, $\gamma > 0$, when the underlying distribution F is multivariate normal. Write down a detailed bootstrap algorithm to evaluate the α critical point of the test J but when F is multivariate student, $t_{p,\nu}(\mathbf{0}, \mathbf{I})$.
Hint: A $t_{p,\nu}(\mathbf{0}, \mathbf{I})$ distribution can be simulated with Problem 13.6.1.

Appendix A
Inversion formulas

Assume $\mathbf{x} \sim F$, $\mathbf{y} \sim G$, $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ on $\mathbb{R}^n$. Then, $\mathbf{z} = \mathbf{x} + \mathbf{y}$ has a d.f., $\mathbf{z} \sim H$, given by

$$\begin{aligned} H(\mathbf{t}) &= P(\mathbf{x} + \mathbf{y} \leq \mathbf{t}) \\ &= E\ P(\mathbf{x} + \mathbf{y} \leq \mathbf{t} | \mathbf{y}) \\ &= E\ F(\mathbf{t} - \mathbf{y}) = \int_{\mathbb{R}^n} F(\mathbf{t} - \mathbf{y}) dG(\mathbf{y}). \end{aligned}$$

Similarly, inverting the roles of $\mathbf{x}$ and $\mathbf{y}$, we also have

$$H(\mathbf{t}) = E\ G(\mathbf{t} - \mathbf{x}) = \int_{\mathbb{R}^n} G(\mathbf{t} - \mathbf{x}) dF(\mathbf{x}).$$

This leads to the smoothing lemma on convolution.

Lemma A.1 (Smoothing lemma) *If $\mathbf{x}$ is absolutely continuous with p.d.f. $f(\mathbf{t})$, then $\mathbf{z} = \mathbf{x} + \mathbf{y}$, where $\mathbf{y} \sim G$ and $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$, is absolutely continuous with p.d.f.*

$$h(\mathbf{t}) = E\ f(\mathbf{t} - \mathbf{y}).$$

Proof. It follows readily that

$$\begin{aligned} H(\mathbf{t}) &= \int_{\mathbb{R}^n} F(\mathbf{t} - \mathbf{y}) dG(\mathbf{y}) \\ &= \int_{\mathbb{R}^n} \left[\int_{(-\boldsymbol{\infty}, \mathbf{t} - \mathbf{y}]} f(\mathbf{x}) d\mathbf{x} \right] dG(\mathbf{y}) \end{aligned}$$

$$= \int_{\mathbb{R}^n} \left[\int_{(-\infty,\mathbf{t}]} f(\mathbf{x}-\mathbf{y}) d\mathbf{x} \right] dG(\mathbf{y}).$$

By Tonelli's theorem, it is posible to interchange the order of integration whereby

$$H(\mathbf{t}) = \int_{(-\infty,\mathbf{t}]} \left[\int_{\mathbb{R}^n} f(\mathbf{x}-\mathbf{y}) dG(\mathbf{y}) \right] d\mathbf{x} = \int_{(-\infty,\mathbf{t}]} E\ f(\mathbf{x}-\mathbf{y}) d\mathbf{x}.$$

□

We can now establish the inversion formula on $\mathbb{R}^n$. The proof resembles that of Feller (1966, p. 480) for $n = 1$.

Proposition A.1 (Inversion formula) *The probability measure $P_{\mathbf{x}}$ is given in terms of the characteristic function $c(\mathbf{t}) = c_{\mathbf{x}}(\mathbf{t})$ by*

$$P_{\mathbf{x}}(\mathbf{a},\mathbf{b}] = \lim_{N\to\infty} \frac{1}{(2\pi)^n} \int_{(\mathbf{a},\mathbf{b}]} \int_{\mathbb{R}^n} e^{-i\mathbf{t}'\mathbf{x}} c(\mathbf{t}) e^{-\mathbf{t}'\mathbf{t}/2N^2} d\mathbf{t} d\mathbf{x},$$

$\forall \mathbf{a}, \mathbf{b}$ *such that* $P_{\mathbf{x}}(\partial(\mathbf{a},\mathbf{b}]) = 0$.

Proof. Take any random $\mathbf{t}$ such that $\mathbf{t} \perp\!\!\!\perp \mathbf{x}$. Then, conditioning yields

$$E\ e^{i\mathbf{x}'\mathbf{t}} = E\ E(e^{i\mathbf{x}'\mathbf{t}}|\mathbf{x}) = E\ c_{\mathbf{t}}(\mathbf{x}) = E\ c_{\mathbf{x}}(\mathbf{t}).$$

Replace $\mathbf{x}$ by $\mathbf{x} - \mathbf{s}$ for any fixed value of $\mathbf{s}$ to find Parseval's relation:

$$E\ c_{\mathbf{t}}(\mathbf{x}-\mathbf{s}) = E\ e^{-i\mathbf{s}'\mathbf{t}} c_{\mathbf{x}}(\mathbf{t}).$$

However, letting $\mathbf{t} \sim N_n(\mathbf{0}, \sigma^{-2}\mathbf{I})$ with $c_{\mathbf{t}}(\mathbf{s}) = \exp(-|\mathbf{s}|^2/2\sigma^2)$,

$$\begin{aligned} E\ \exp(-|\mathbf{s}-\mathbf{x}|^2/2\sigma^2) &= E\ e^{-i\mathbf{t}'\mathbf{s}} c(\mathbf{t}) \\ &= \left(\frac{\sigma^2}{2\pi}\right)^{n/2} \int_{\mathbb{R}^n} e^{-i\mathbf{t}'\mathbf{s}} c(\mathbf{t}) \exp\left(-\sigma^2|\mathbf{t}|^2/2\right) d\mathbf{t}. \end{aligned}$$

Divide by $(2\pi\sigma^2)^{n/2}$ to obtain

$$E\ \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-|\mathbf{s}-\mathbf{x}|^2/2\sigma^2) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\mathbf{t}'\mathbf{s}} c(\mathbf{t}) \exp\left(-\sigma^2|\mathbf{t}|^2/2\right) d\mathbf{t}.$$

This is of the form $E\ g(\mathbf{s}-\mathbf{x}) = h(\mathbf{s})$ in the smoothing lemma where $g(\mathbf{s})$ is the p.d.f. for a $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. Thus, $h(\mathbf{s})$ is the p.d.f. of $\mathbf{x} + \sigma\mathbf{z}$, where $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I})$, and if we let P_σ denote the probability measure for $\mathbf{x} + \sigma\mathbf{z}$,

$$P_\sigma(\mathbf{a},\mathbf{b}] = \frac{1}{(2\pi)^n} \int_{(\mathbf{a},\mathbf{b}]} \int_{\mathbb{R}^n} e^{-i\mathbf{t}'\mathbf{s}} c(\mathbf{t}) e^{-\sigma^2|\mathbf{t}|^2/2} d\mathbf{t} d\mathbf{s},$$

whereby Slutsky's theorem with $\sigma = 1/N$ gives the result. □

An immediate corollary is the inversion formula for absolutely continuous distribution.

Corollary A.1 *If $c(\mathbf{t})$ is integrable with respect to Lebesgue measure, then*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\mathbf{t}'\mathbf{x}} c(\mathbf{t}) d\mathbf{t}.$$

Proof. If $c(\mathbf{t})$ is integrable, then the integrand in Proposition A.1 is dominated by an integrable function. By the D.C.T., we can interchange the limit and the integral, which gives the result. □

Appendix B
Multivariate cumulants

B.1 Definition and properties

The *moments* of a univariate random variable x, $\mu_r = E\ x^r$, are the coefficients of $(it)^r/r!$ in the Taylor series of the characteristic function,

$$c_x(t) = \sum_{r=0}^{\infty} \mu_r (it)^r / r!$$

whereas the *cumulants* are the coefficients in the series for $K_x(t) \equiv \ln[c_x(t)]$,

$$K_x(t) = \sum_{r=0}^{\infty} k_r (it)^r / r!,$$

provided the expansions are valid. The function $K_x(t)$ is the cumulant generating function. Relations between moments and cumulants are thus obtained by equating the coefficients in the Taylor series of $\exp(\cdot)$ in the equation

$$\sum_{r=0}^{\infty} \mu_r (it)^r / r! = \exp\left(\sum_{r=0}^{\infty} k_r (it)^r / r! \right).$$

We require only the relations between the first four moments and cumulants (assuming they exist):

$$\begin{aligned} \mu_1 &= k_1, \\ \mu_2 &= k_2 + k_1^2, \end{aligned}$$

$$\begin{aligned}\mu_3 &= k_3 + 3k_2k_1 + k_1^3,\\ \mu_4 &= k_4 + 4k_3k_1 + 3k_2^2 + 6k_2k_1^2 + k_1^4,\end{aligned}$$

$$\begin{aligned}k_1 &= \mu_1,\\ k_2 &= \mu_2 - \mu_1^2,\\ k_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3,\\ k_4 &= \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4.\end{aligned}$$

When x is centered, i.e., $E\,x = \mu_1 = k_1 = 0$, these simplify to

$$\begin{aligned}\mu_2 &= k_2, & k_2 &= \mu_2,\\ \mu_3 &= k_3, & k_3 &= \mu_3,\\ \mu_4 &= k_4 + 3k_2^2, & k_4 &= \mu_4 - 3\mu_2^2.\end{aligned}$$

For a random vector $\mathbf{x} \in \mathbb{R}^p$, product-moments $\mu_{r_1,\ldots,r_p} = E(x_1^{r_1}\cdots x_p^{r_p})$ and multivariate cumulants $k_{r_1,\ldots,r_p}$ of order $r = \sum_{i=1}^p r_i$ are defined similarly,

$$\begin{aligned}c_{\mathbf{x}}(\mathbf{t}) &= \sum_{r_1,\ldots,r_p=0}^{\infty} \mu_{r_1,\ldots,r_p} \frac{(it_1)^{r_1}}{r_1!}\cdots\frac{(it_p)^{r_p}}{r_p!},\\ K_{\mathbf{x}}(\mathbf{t}) = \ln[c_{\mathbf{x}}(\mathbf{t})] &= \sum_{r_1,\ldots,r_p=0}^{\infty} k_{r_1,\ldots,r_p} \frac{(it_1)^{r_1}}{r_1!}\cdots\frac{(it_p)^{r_p}}{r_p!}.\end{aligned}$$

Example B.1 *For* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, we have* $K_{\mathbf{x}}(\mathbf{t}) = i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}$*, a quadratic function of* $\mathbf{t}$*, and, thus, all multivariate cumulants of order* $r > 2$ *are null. Multivariate cumulants of order* 1 *are the means,* μ_i*, and those of order* 2 *are the variances,* σ_{ii}*, and covariances,* σ_{ij}*.*

Obtaining product-moments in terms of cumulants, and vice versa, is a laborious task which can be greatly simplified with a "symbolic differential operator" [Kendall et al. (1987)]. For example, when $E\,\mathbf{x} = \mathbf{0}$, consider the relation $\mu_4 = k_4 + 3k_2^2$, which we write symbolically as

$$\mu(r_1^4) = k(r_1^4) + 3k^2(r_1^2).$$

To obtain a relation between fourth-order product-moments and cumulants of a bivariate distribution, consider the operator $r_2\partial(\cdot)/\partial r_1$. When applied to $\mu(r_1^4)$, it yields

$$4\mu(r_1^3 r_2) = 4k(r_1^3 r_2) + 12k(r_1^2)k(r_1 r_2),$$

which means, after dividing by 4,

$$\mu_{31} = k_{31} + 3k_{20}k_{11}.$$

Example B.2 *The same method can be used to obtain cumulants in terms of product-moments. Considering the relation*

$$k_4 = \mu_4 - 3\mu_2^2$$

in symbolic form

$$k(r_1^4) = \mu(r_1^4) - 3\mu^2(r_1^2),$$

and applying the operator $r_2\partial(\cdot)/\partial r_1$*, we get*

$$4k(r_1^3 r_2) = 4\mu(r_1^3 r_2) - 12\mu(r_1^2)\mu(r_1 r_2)$$

or

$$k_{31} = \mu_{31} - 3\mu_{20}\mu_{11}.$$

Continuing this process, it is possible to obtain relations for trivariate distributions with either operator, $r_3\partial(\cdot)/\partial r_1$ or $r_3\partial(\cdot)/\partial r_2$. The operator $r_3\partial(\cdot)/\partial r_1$ applied to the last symbolic equation yields

$$12\mu(r_1^2 r_2 r_3) = 12k(r_1^2 r_2 r_3) + 24k(r_1 r_3)k(r_1 r_2) + 12k(r_1^2)k(r_2 r_3),$$

which is equivalent to

$$\mu_{211} = k_{211} + 2k_{101}k_{110} + k_{200}k_{011}.$$

The operator $r_4\partial(\cdot)/\partial r_1$ finally gives the relation for fourth-order product-moments and cumulants of a four-dimensional distribution

$$24\mu(r_1 r_2 r_3 r_4) = 24k(r_1 r_2 r_3 r_4) + 24k(r_3 r_4)k(r_1 r_2) + 24k(r_1 r_3)k(r_2 r_4)$$

$$+24k(r_1 r_4)k(r_2 r_3),$$

or

$$\mu_{1111} = k_{1111} + k_{0011}k_{1100} + k_{1010}k_{0101} + k_{1001}k_{0110}.$$

For fourth-order product-moments of a p-dimensional, $p > 4$, distribution, we need only specify which four variables enter. For example, $\mu_{1111}^{ijkl} = E(x_i x_j x_k x_l)$ satisfies

$$\mu_{1111}^{ijkl} = k_{1111}^{ijkl} + k_{0011}^{ijkl}k_{1100}^{ijkl} + k_{1010}^{ijkl}k_{0101}^{ijkl} + k_{1001}^{ijkl}k_{0110}^{ijkl}.$$

A zero subscript means the superscript variable does not enter, so we can rewrite

$$\mu_{1111}^{ijkl} = k_{1111}^{ijkl} + k_{11}^{kl}k_{11}^{ij} + k_{11}^{ik}k_{11}^{jl} + k_{11}^{il}k_{11}^{jk}.$$

When a variable is repeated, the indices can be amalgamated. For example, the equation where $i = j$,

$$\mu_{1111}^{iikl} = k_{1111}^{iikl} + k_{11}^{kl}k_{11}^{ii} + k_{11}^{ik}k_{11}^{il} + k_{11}^{il}k_{11}^{ik},$$

becomes

$$\mu_{211}^{ikl} = k_{211}^{ikl} + k_{11}^{kl}k_{2}^{i} + k_{11}^{ik}k_{11}^{il} + k_{11}^{il}k_{11}^{ik},$$

and if $i = j = k = l$, then we recover the initial equation $\mu_4^i = k_4^i + 3(k_2^i)^2$.

Departures from normality is often assessed with the coefficients of *skewness*, γ_1, and *kurtosis*, γ_2. For a centered variable x, they are defined as

$$\begin{aligned}\gamma_1 &= \frac{\mu_3}{\mu_2^{3/2}} = \frac{k_3}{k_2^{3/2}},\\ \gamma_2 &= \frac{\mu_4}{\mu_2^2} - 3 = \frac{k_4}{k_2^2}.\end{aligned}$$

For a normal variable, $\gamma_1 = \gamma_2 = 0$.

Cumulants of random symmetric matrices can also be defined. For a description of miminal moments and cumulants of symmetric matrices with an application to the Wishart distribution, the reader is referred to Kollo and von Rosen (1995).

B.2 Application to asymptotic distributions

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\mathbf{x} \in \mathbb{R}^p$ which has finite fourth-order moments and $E\ \mathbf{x} = \mathbf{0}$ and var $\mathbf{x} = \mathbf{\Sigma}$. The asymptotic distribution of $\mathbf{S} = \frac{1}{(n-1)}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ was derived generally in Section 6.3:

$$n^{1/2}(\mathbf{S} - \mathbf{\Sigma}) \xrightarrow{d} N_p^p(\mathbf{0}, \text{var } \mathbf{W}),$$

where $\mathbf{W} = \mathbf{x}\mathbf{x}'$. The only problem is to calculate var $\mathbf{W}$. This can now be done in terms of multivariate cumulants. The block (i, j) of var $\mathbf{W}$ is

$$E(x_i x_j \mathbf{x}\mathbf{x}') - E(x_i \mathbf{x})E(x_j \mathbf{x}')$$

and the element (k, l) of the block (i, j) becomes

$$\begin{aligned}E(x_i x_j x_k x_l) - E(x_i x_k)E(x_j x_l) &= \mu_{1111}^{ijkl} - \mu_{11}^{ik}\mu_{11}^{jl}\\ &= k_{1111}^{ijkl} + k_{11}^{kl}k_{11}^{ij} + k_{11}^{il}k_{11}^{jk}.\end{aligned}$$

The general result is thus

$$\text{cov}(w_{ik}, w_{jl}) = k_{1111}^{ijkl} + k_{11}^{kl}k_{11}^{ij} + k_{11}^{il}k_{11}^{jk}.$$

B.3 Problems

1. Establish the following:

   (i) $\mu_{11} = k_{11}$ and $\mu_{21} = k_{21}$.
   (ii) $\mu_{22} = k_{22} + k_{20}k_{02} + 2k_{11}^2$.
   (iii) Given $\mu_5 = k_5 + 10k_3k_2$, calculate μ_{32} and μ_{41}.
   (iv) Obtain μ_{301} in terms of lower-order cumulants.

2. Demonstrate the kurtosis γ_2 of a symmetric contaminated normal density

$$(1-\epsilon)(2\pi)^{-1/2}\exp(-\tfrac{1}{2}x^2)+\epsilon(2\pi\sigma)^{-1/2}\exp(-\tfrac{1}{2}x^2/\sigma^2)$$

is

$$\gamma_2 = 3\frac{[1+\epsilon(\sigma^4-1)]}{[1+\epsilon(\sigma^2-1)]^2}-3.$$

3. Evaluate the kurtosis of a Student's t_ν distribution as $\gamma_2 = 6/(\nu-4)$, $\nu > 4$.

Appendix C
S-plus functions

This appendix describes three S-plus programs which the reader can download from the World Wide Web site www.dms.umontreal.ca/∼bilodeau. Simply download the file named multivariate and, at the S-plus prompt, type: source("multivariate") to compile the functions.

1. $U(p; m, n)$ **distribution function**.
   Usage: $pu(\zeta, p, m, n)$
   Value: The function returns $P(U(p; m, n) \leq \zeta)$.

2. $U(p; m, n)$ **quantiles**.
   Usage: $qu(\alpha, p, m, n)$
   Value: The function returns the α-quantile , $U_\alpha(p; m, n)$ say, satisfying

   $$P(U(p; m, n) \leq U_\alpha(p; m, n)) = \alpha.$$

   It returns as well a $Cfactor$, frequently used by people relying on the asymptotic result

   $$-[n - \tfrac{1}{2}(p - m + 1)] \ln U(p; m, n) \xrightarrow{d} \chi^2_{pm},$$

   to make the approximate χ^2_{pm} quantile an exact quantile of $-[n - \frac{1}{2}(p - m + 1)] \ln U(p; m, n)$. More precisely,

   $$Cfactor \cdot \chi^2_{1-\alpha,pm} = -[n - \tfrac{1}{2}(p - m + 1)] \ln U_\alpha(p; m, n).$$

   Note that lower quantiles of $U(p; m, n)$ correspond to upper quantiles of χ^2_{pm}.

3. **Beta Q-Q plot for multivariate normality**.
Usage: *qqbeta*(x)
The input x is the $n \times p$ sample matrix.
Value: The function returns the Q-Q plot of the points

$$\left(d^2_{(i)},\ [(n-1)^2/n]\ \text{beta}_{\gamma_i}\left(\tfrac{1}{2}p; \tfrac{1}{2}(n-p-1)\right)\right),\ i = 1, \ldots, n,$$

as described in Section 11.4.1. The graphic device must be activated before using this function.

4. **Robust S estimate**.
Usage: *s.estimate*$(x, r, nr, Nsamp)$
The input x denotes the $n \times p$ sample matrix. The input r in the interval $(0, .5)$ is the asymptotic breakdown point, nr and $Nsamp$ are positive integer parameters of the numerical algorithm [Ruppert (1992)]. Values of $nr = 3$ and $Nsamp = 80p$ are appropriate for most purposes. The user is urged to experiment with other values of nr and $Nsamp$ to certify that the *s.estimate* function returned the global minimum.
Value: The function returns the S estimate of location and scatter, $\boldsymbol{\mu}_n$ and $\mathbf{V}_n$, the Mahalanobis distances, *distance.mahalanobis*, for outlier detection and the objective function, *determinant*, which the S estimate seeks to minimize. Points with a Mahalanobis distance greater than $(\chi^2_{.95,p})^{1/2}$ should be checked for outliers [Rousseeuw and van Zomeren (1990)].
The implementation uses the biweight $\rho(\cdot)$ function of Section 13.4.2 and determines c_0 such that $E\ \rho(|\mathbf{z}|)/(c_0^2/6) = r$, where $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$, to achieve the desired breakdown point.

5. **Asymptotic variance of S estimate**.
Usage: *asymp*(p, r)
The input p is the number of variables, whereas r is the breakdown point.
Value: The function returns the asymptotic variance constants, at the normal distribution, in Proposition 13.11: $\lambda = \alpha/\beta^2$, σ_1, and σ_2. The constants λ^{-1} and σ_1^{-1}, in particular, serve as measures of relative efficiencies of the location and scatter estimates, respectively, at the normal distribution.

References

[1] Ali, M.M., and R. Ponnapalli (1990). An optimal property of the Gauss-Markoff estimator. *Journal of Multivariate Analysis* **32**, 171-176.

[2] Anderson, G.A. (1965). An asymptotic expansion for the distribution of the latent roots of the estimated covariance matrix. *Annals of Mathematical Statistics* **36**, 1153-1173.

[3] Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* **34**, 122-148.

[4] Anderson, T.W. (1965). An asymtotic expansion for the distribution of the latent roots of the estimated covariance matrix. *Annals of Mathematical Statistics* **36**, 1153-1173.

[5] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. 2nd ed. John Wiley & Sons, New York.

[6] Andrews, D.F., R. Gnanadesikan, and J.L. Warner (1971). Transformations of multivariate data. *Biometrics* **27**, 825-840.

[7] Andrews, D.F., R. Gnanadesikan, and J.L. Warner (1973). Methods for assessing multivariate normality. *Multivariate Analysis*, ed. P.K. Krishnaiah. Academic Press, New York, 95-116.

[8] Ash, R. (1972). *Real Analysis and Probability.* Academic Press, New York.

[9] Baringhaus, L., and N. Henze(1991). Limit distributions for measures of multivariate skewness and kurtosis based on projections. *Journal of Multivariate Analysis* **38**, 51-69.

[10] Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society. London. Series A.* **160**, 268-282.

[11] Bartlett, M.S. (1938). Further aspects of the theory of multiple regression. *Proceedings of the Cambridge Philosophical Society* **34**, 33-40.

[12] Bellman, R. (1960). *Introduction to Matrix Analysis*. McGraw-Hill, New York.

[13] Bentler, P.M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika* **48**, 493-517.

[14] Beran, R. (1984). Bootstrap methods in statistics. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **86**, 14-30.

[15] Beran, R. (1986). Simulated power functions. *Annals of Statistics* **14**, 151-173.

[16] Beran, R. (1987). Prepivoting to reduce level error in confidence sets. *Biometrika* **74**, 457-468.

[17] Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* **83**, 687-697.

[18] Beran, R., and M.S. Srivastava (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Annals of Statistics* **13**, 95-115.

[19] Beran, R., and M.S. Srivastava (1987). Correction: Bootstrap tests and confidence regions for functions of a covariance matrix. *Annals of Statistics* **15**, 470-471.

[20] Berk, R., and J.T. Hwang (1989). Optimality of the least squares estimator. *Journal of Multivariate Analysis* **30**, 245-254.

[21] Berkane, M., K. Oden, and P.M. Bentler (1997). Geodesic estimation in elliptical distributions. *Journal of Multivariate Analysis* **63**, 35-46.

[22] Bhat, B.R. (1981). *Modern Probability Theory*. John Wiley & Sons, New York.

[23] Bickel, P.J., and D.A. Freedman (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**, 1196-1217.

[24] Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons, New York.

[25] Billingsley, P. (1995). *Probability and Measure*. 3rd ed. John Wiley & Sons, New York.

[26] Bilodeau, M. (1988). On the simultaneous estimation of scale parameters. *The Canadian Journal of Statistics* **16**, 169-174.

[27] Bilodeau, M. (1990). On the choice of the loss function in covariance estimation. *Statistics & Decisions* **8**, 131-139.

[28] Bilodeau, M. (1995). Minimax estimators of the mean vector in normal mixed linear models. *Journal of Multivariate Analysis* **52**, 73-82.

[29] Bilodeau, M.(1996). Some remarks on $U(p; m, n)$ distributions. *Statistics and Probability Letters* **31**, 41-43.

[30] Bilodeau M. (1997a). Estimating a multivariate treatment effect under a biased allocation rule. *Communications in Statistics, Theory and Methods* **26**, 1119-1124.

[31] Bilodeau, M. (1997b). Robust test for a given variance. Technical Report, Université de Montréal.

[32] Bilodeau, M. (1998). Multivariate flattening for better predictions. Technical Report, Université de Montréal.

[33] Bilodeau, M., and M.S. Srivastava (1989a). Estimation of the MSE matrix of the Stein estimator. *The Canadian Journal of Statistics* **16**, 153-159.

[34] Bilodeau, M., and M.S. Srivastava (1989b). Stein estimation under elliptical distributions. *Journal of Multivariate Analysis* **28**, 247-259.

[35] Bilodeau, M., and M.S. Srivastava (1992). Estimation of the eigenvalues of $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}$. *Journal of Multivariate Analysis* **41**, 1-13.

[36] Bilodeau, M., and T. Kariya (1989). Minimax estimators in the normal MANOVA model. *Journal of Multivariate Analysis* **28**, 260-270.

[37] Bilodeau, M., and T. Kariya (1994). LBI tests of independence in bivariate exponential distributions. *Annals of the Institute of Statistical Mathematics* **46**, 127-136.

[38] Blom, G. (1958). *Statistical Estimates and Transformed Beta-variables*. John Wiley & Sons, New York.

[39] Boente, G. (1987). Asymptotic theory for robust principal components. *Journal of Multivariate Analysis* **21**, 67-78.

[40] Boulerice, B., and G.R. Ducharme (1997). Smooth tests of goodness-of-fit for directional and axial data. *Journal of Multivariate Analysis* **60**, 154-175.

[41] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika* **36**, 317-346.

[42] Box, G.E.P., and D.R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* **26**, 211-252.

[43] Breiman, L., and J.H. Friedman (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society B* **59**, 3-54.

[44] Brown, P.J. (1980). Aspects of multivariate regression (with discussion). *Bayesian Statistics*. eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith. Valencia University Press, Valencia.

[45] Browne, M.W., and A. Shapiro (1987). Adjustments for kurtosis in factor analysis with elliptically distributed errors. *Journal of the Royal Statistical Society B* **49**, 346-352.

[46] Carrière, J.F. (1994). Dependent decrement theory. *Transactions* **XLVI**, Society of Actuaries, 45-65.

[47] Casella, G., and R.L. Berger (1990). *Statistical Inference*, Duxbury Press, Belmont, California.

[48] Chattopadhyay, A.K., and K.C.S. Pillai (1973). Asymptotic expansions for the distributions of characteristic roots when the parameter matrix has several multiple roots. *Multivariate analysis III*. Academic Press, New York, 117-127.

[49] Chikuse, Y. (1976). Asymptotic distributions of the latent roots of the covariance matrix with multiple population roots. *Journal of Multivariate Analysis* **6**, 237-249.

[50] Cléroux, R., and G.R. Ducharme (1989). Vector correlation for elliptical distributions. *Communications in Statistics, Theory and Methods* **18**, 1441-1454.

[51] Coelho, C.A. (1998). The generalized integer gamma distribution–A basis for distributions in multivariate statistics. *Journal of Multivariate Analysis* **64**, 86-102.

[52] Cook, R.D., M.E. Johnson (1981). A family of distributions for modelling nonelliptically symmetric multivariate data. *Journal of the Royal Statistical Society B* **43**, 210-218.

[53] Copas, J.B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika* **62**, 701-704.

[54] Courant, R. (1936). *Differential and Integral Calculus II.* John Wiley & Sons, New York.

[55] Cox, D.R., and N.J.H. Small (1978). Testing multivariate normality. *Biometrika* **65**, 263-272.

[56] Cuadras, C.M. (1992). Probability distributions with given multivariate marginals and given dependence structure. *Journal of Multivariate Analysis* **42**, 51-66.

[57] Datta, S., N. Mukhopadhyay (1997). On sequential fixed-size confidence regions for the mean vector. *Journal of Multivariate Analysis* **60**, 233-251.

[58] Davies, P.L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics* **15**, 1269-1292.

[59] Davis, A.W. (1971). Percentile approximations for a class of likelihood ratio criteria. *Biometrika* **58**, 349-356.

[60] Davison, A.C., and D.V. Hinkley (1997). *Bootstrap Methods and their Application.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

[61] Donoho, D.L. (1982). Breakdown Properties of Multivariate Location Estimators. Qualifying paper, Harvard University.

[62] Ducharme, G.R., P. Milasevic (1987). Spatial median and directional data. *Biometrika* **74**, 212-215.

[63] Dümbgen, L. (1998). Perturbation inequalities and confidence sets for functions of a scatter matrix. *Journal of Multivariate Analysis* **65**, 19-35.

[64] Dykstra, R.L. (1970). Establishing the positive definiteness of the sample covariance matrix. *Annals of Mathematical Statistics* **41**, 2153-2154.

[65] Eaton, M.L. (1983). *Multivariate Statistics, a Vector Space Approach.* John Wiley & Sons, New York.

[66] Eaton, M.L. (1988). Concentration inequalities for Gauss-Markov estimators. *Journal of Multivariate Analysis* **25**, 119-138.

[67] Eaton, M.L., and M.D. Perlman (1973). The non-singularity of generalized sample covariance matrices. *Annals of Statistics* **1**, 710-717.

[68] Eaton, M.L., and D.E. Tyler (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Annals of Statistics* **19**, 260-271.

[69] Eaton, M.L., and D.E. Tyler (1994). The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis. *Journal of Multivariate Analysis* **50**, 238-264.

[70] Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association* **64**, 1278-1302.

[71] Efron B. (1982). *The Jacknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia. bibitemefrm Efron B., and C. Morris (1976). Multivariate empirical Bayes and estimation of covariance matrix. *Annals of Statistics* **4**, 22-32. indexaiMorris, C.

[72] Efron, B., and R.J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York. indexaiTibshirani, R.J.

[73] Erdélyi, A., W. Magnus, F. Oberhettinger, and F.G. Tricomi (1953). *Higher Transcendental Functions*. McGraw-Hill, New York.

[74] Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* **29**, 751-760.

[75] Fan, Y. (1997). Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis* **62**, 36-63.

[76] Fang, K.T., S. Kotz, and K.W. Ng (1991). *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London.

[77] Fang, K.T., L.-X. Zhu, and P.M. Bentler (1993). A necessary test of goodness of fit for sphericity. *Journal of Multivariate Analysis* **45**, 34-55.

[78] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications (Vol. II)*. John Wiley & Sons, New York.

[79] Fisher, R.A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society. London. Series A.* **217**, 295-305.

[80] Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer-Verlag, New York.

[81] Frank, M.J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Math.* **19**, 194-226.

[82] Fraser, D.A.S. (1976). *Probability and Statistics: Theory and Applications*. DAI Press, Toronto.

[83] Fraser, D.A.S., I. Guttman, and M.S. Srivastava (1991). Conditional inference for treatment and error in multivariate analysis. *Biometrika* **78**, 565-572.

[84] Fujikoshi, Y. (1970). Asymptotic expansions of the distributions of test statistics in multivariate analysis. *Journal of Science of the Hiroshima University. Series A, Mathematics* **34**, 73-144.

[85] Fujikoshi, Y. (1977). An asymptotic expansion for the distributions of the latent roots of the Wishart matrix with multiple population roots. *Annals of the Institute of Statistical Mathematics* **29**, 379-387.

[86] Fujikoshi, Y. (1978). Asymptotic expansions for the distributions of some functions of the latent roots of matrices in three situations. *Journal of Multivariate Analysis* **8**, 63-72.

[87] Fujikoshi, Y. (1988). Comparison of powers of a class of tests for multivariate linear hypothesis and independence. *Journal of Multivariate Analysis* **26**, 48-58.

[88] Fujikoshi, Y. (1997). An asymptotic expansion for the distribution of Hotelling's T^2-statistic under nonnormality. *Journal of Multivariate Analysis* **61**, 187-193.

[89] Fujikoshi, Y., and Y. Watamori (1992). Tests for the mean direction of the Langevin distribution with large concentration parameter. *Journal of Multivariate Analysis* **42**, 210-225.

[90] Fujisawa, H. (1997). Improvement on chi-squared approximation by monotone transformation. *Journal of Multivariate Analysis* **60**, 84-89.

[91] Fujisawa, H. (1997). Likelihood ratio criterion for mean structure in the growth curve model with random effects. *Journal of Multivariate Analysis* **60**, 90-98.

[92] Genest, C., and R.J. MacKay (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics* **14**, 145-159.

[93] Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika* **74**, 549-555.

[94] Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses.* Addison-Wesley, Reading, Massachusetts.

[95] Giri, N.C. (1996). *Multivariate Statistical Analysis.* Marcel Dekker, New York.

[96] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* John Wiley & Sons, New York.

[97] Gnanadesikan, R., and J.R. Kettenring (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81-124.

[98] Grübel, R., and D.M. Rocke (1990). On the cumulants of affine equivariant estimators in elliptical families. *Journal of Multivariate Analysis* **35**, 203-222.

[99] Gunderson, B.K., and R.J. Muirhead (1997). On estimating the dimensionality in canonical correlation analysis. *Journal of Multivariate Analysis* **62**, 121-136

[100] Gupta, A.K., and D. Song (1997). Characterization of p-generalized normality. *Journal of Multivariate Analysis* **60**, 61-71.

[101] Gupta, A.K., and D. St. P. Richards (1990). The Dirichlet distributions and polynomial regression. *Journal of Multivariate Analysis* **32**, 95-102.

[102] Gupta, A.K., and T. Varga (1992). Characterization of matrix variate normal distributions. *Journal of Multivariate Analysis* **41**, 80-88.

[103] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[104] Hendriks, H., Z. Landsman, and F. Ruymgaart (1996). Asymptotic behavior of sample mean direction for spheres. *Journal of Multivariate Analysis* **59**, 141-152.

[105] Henze, N., and T. Wagner (1997). A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis* **62**, 1-23.

[106] Henze, N., and B. Zirkler (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics, Theory and Methods* **19**, 3595-3617.

[107] Hsu, P.L. (1941). On the limiting distribution of the canonical correlations. *Biometrika* **32**, 38-45.

[108] Huber, P.J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

[109] Iwashita, T. (1997). Asymptotic null and nonnull distribution of Hotelling's T^2-statistic under the elliptical distribution. *Journal of Statistical Planning and Inference* **61**, 85-104..

[110] Iwashita, T., and M. Siotani (1994). Asymptotic distributions of functions of a sample covariance matrix under the elliptical distribution. *The Canadian Journal of Statistics* **22**, 273-283.

[111] James, A.T. (1954). Normal multivariate analysis and the orthogonal group. *Annals of Mathematical Statistics* **25**, 40-75.

[112] James, A.T. (1969). Tests of equality of latent roots of the covariance matrix. *Multivariate Analysis II*. Academic Press, New York, 205-218.

[113] John, S. (1971). Some optimal multivariate tests. *Biometrika* **58**, 123-127.

[114] John, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika* **59**, 169-174.

[115] Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* **36**, 149-176.

[116] Johnson, N.L., S. Kotz, and A.W. Kemp (1992). *Univariate Discrete Distributions*. 2nd ed. John Wiley & Sons, New York.

[117] Johnson, R.A., and D.W. Wichern (1992). *Applied Multivariate Statistical Analysis*. 3rd ed. Prentice-Hall, Englewood Cliffs, New Jersey.

[118] Jolliffe, I.T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

[119] Jordan, S.M., and K. Krishnamoorthy (1995). Confidence regions for the common mean vector of several multivariate normal populations. *The Canadian Journal of Statistics* **23**, 283-297.

[120] Kano, Y. (1994). Consistency property of elliptical probability density function. *Journal of Multivariate Analysis* **51**, 139-147.

[121] Kano, Y. (1995). An asymptotic expansion of the distribution of Hotelling's T^2-statistic under general condition. *American Journal of Mathematical and Management Sciences* **15**, 317-341.

[122] Kariya, T. (1985). *Testing in the Multivariate General Linear Model*. Kinokunia, Tokyo.

[123] Kariya, T., and B.K. Sinha (1989). *Robustness of Statistical Tests*. Academic Press, San Diego.

[124] Kariya, T., and M.L. Eaton (1977). Robust tests for spherical symmetry. *Annals of Statistics* **5**, 206-215.

[125] Kariya, T., R.S. Tsay, N. Terui, and Hong Li (1999). Tests for multinormality with applications to time series. *Communications in Statistics, Theory and Methods* **28**, 519-536.

[126] Kariya, T., Y. Fujikoshi, and P.R. Krishnaiah (1987). On tests for selection of variables and independence under multivariate regression models. *Journal of Multivariate Analysis* **21**, 207-237.

[127] Kato, T. (1982). *A Short Introduction to Perturbation Theory for Linear Operators*. Springer-Verlag, New York.

[128] Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā A* **32**, 419-430.

[129] Kendall, M., A. Stuart, and J.K. Ord (1987). *Kendall's Advanced Theory of Statistics*. 5th ed. Vol. 1. Oxford University Press, New York.

[130] Kent, J.T., and D.E. Tyler (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics* **19**, 2102-2019.

[131] Khatri, C.G., and M.S. Srivastava (1974). Asymptotic expansions of the non-null distributions of the likelihood ratio criteria for covariance matrices II. Proc. Carleton University, Ottawa. *Metron* **36**, 55-71.

[132] Khatri, C. G., and M.S. Srivastava (1978). Asymptotic expansions for distributions of characteristic roots of covariance matrices. *South African Statistical Journal* **12**, 161-186.

[133] Ko, D., and T. Chang (1993). Robust M-estimators on spheres. *Journal of Multivariate Analysis* **45**, 104-136.

[134] Koehler, K.J., and J.T. Symanowski (1995). Constructing multivariate distributions with specific marginal distributions. *Journal of Multivariate Analysis* **55**, 261-282.

[135] Kollo, T., and H. Neudecker (1993). Asymptotics of eigenvalues and unit-length eigenvectors of sample variance and correlation matrices. *Journal of Multivariate Analysis* **47**, 283-300.

[136] Kollo, T., and D. von Rosen (1995). Minimal moments and cumulants of symmetric matrices: an application to the Wishart distribution. *Journal of Multivariate Analysis* **55**, 149-164.

[137] Koltchinskii, V.I., and L. Li (1998). Testing for spherical symmetry of a multivariate distribution. *Journal of Multivariate Analysis* **65**, 228-244.

[138] Konishi, S. (1979). Asymptotic expansions for the distributions of statistics based on the sample correlation matrix in principal component analysis. *Hiroshima Mathematical Journal* **9**, 647-700.

[139] Konishi, S., and C.G. Khatri (1990). Inferences on interclass and intraclass correlations in multivariate familial data. *Annals of the Institute of Statistical Mathematics* **42**, 561-580.

[140] Konishi, S., and C.R. Rao (1991). Inferences on multivariate measures of interclass and intraclass correlations in familial data. *Journal of the Royal Statistical Society B* **53**, 649-659.

[141] Konishi, S., and C.R. Rao (1992). Principal component analysis for multivariate familial data. *Biometrika* **79**, 631-641.

[142] Kotz, S., and I. Ostrovskii (1994). Characteristic functions of a class of elliptical distributions. *Journal of Multivariate Analysis* **49**, 164-178.

[143] Kres, H. (1983). *Statistical Tables for Multivariate Analysis, a Handbook with References to Applications*. Springer-Verlag, New York.

[144] Kshirsagar, A.M. (1972). *Multivariate Analysis*. Marcel Dekker, New York.

[145] Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* **50**, 403-418.

[146] Kuwana, Y., and T. Kariya (1991). LBI tests for multivariate normality in exponential power distributions. *Journal of Multivariate Analysis* **39**, 117-134.

[147] Lee, Y.-S. (1972). Some results on the distribution of Wilk's likelihood-ratio criterion. *Biometrika* **59**, 649-664.

[148] Lehmann, E.L. (1983). *Theory of Point Estimation*. John Wiley & Sons, New York.

[149] Li, Haijun, M. Scarsini, and M. Shaked (1996). Linkages: A tool for construction of multivariate distributions with given nonoverlapping multivariate marginals. *Journal of Multivariate Analysis* **56**, 20-41.

[150] Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis* **53**, 139-158.

[151] Liu, C. (1997). ML estimation of the multivariate t distribution and the EM algorithm. *Journal of Multivariate Analysis* **63**, 296-312.

[152] Looney, S.W. (1995). How to use test for univariate normality to assess multivariate normality. *The American Statistician* **49**, 64-70.

[153] Lopuhaä, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics* **17**, 1662-1683.

[154] Lopuhaä, H.P. (1991). Multivariate τ-estimators for location and scatter. *The Canadian Journal of Statistics* **19**, 307-321.

[155] Lopuhaä, H.P., and P.J. Rousseeuw (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics* **19**, 229-248.

[156] MacDuffy, C.C. (1943). *Vectors and Matrices*. The Mathematical Association of America, Providence, Rhode Island.

[157] Magnus, J.R., and H. Neudecker (1979). The commutation matrix: Some properties and applications. *Annals of Statistics* **7**, 381-394.

[158] Malkovich, J.F., and A.A. Afifi (1973). On tests for multivariate normality. *Journal of the American Statistical Association* **68**, 176-179.

[159] Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519-530.

[160] Mardia, K.V. (1972). *Statistics of Directional Data*. Academic Press, London.

[161] Mardia, K.V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test. *Applied Statistics* **24**, 163-171.

[162] Mardia, K.V., J.T. Kent, and J.M. Bibby (1979). *Multivariate Analysis*. Academic Press, New York.

[163] Märkeläinen, T., K. Schmidt, and G.P.H. Styan (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed sample sizes. *Annals of Statistics* **9**, 758-767.

[164] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics* **4**, 51-67.

[165] Marshall, A.W., and I. Olkin(1988). Families of multivariate distributions. *Journal of the American Statistical Association* **83**, 834-841.

[166] Martin, M.A. (1990). On bootstrap iteration for coverage correction in confidence intervals. *Journal of the American Statistical Association* **85**, 1105-1118. bibitemmat Mathew, T., and K. Nordström (1997). Wishart and chi-square distributions associated with matrix quadratic forms. *Journal of Multivariate Analysis* **61**, 129-143.

[167] Mauchly, J.W. (1940). Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics* **11**, 204-209.

[168] Muirhead, R.J. (1970). Asymptotic distributions of some multivariate tests. *Annals of Mathematical Statistics* **41**, 1002-1010.

[169] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, New York.

[170] Muirhead, R.J., and Y. Chikuse (1975). Asymptotic expansions for the joint and marginal distributions of the latent roots of the covariance matrix. *Annals of Statistics* **3**, 1011-1017.

[171] 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.

[172] Nagao, H. (1973). On some test criteria for covariance matrix. *Annals of Statistics* **1**, 700-709.

[173] Nagao, H., and M.S. Srivastava (1992). On the distributions of some test criteria for a covariance matrix under local alternatives and bootstrap approximations. *Journal of Multivariate Analysis* **43**, 331-350.

[174] Naito, K. (1998). Approximation of the power of kurtosis test for multinormality. *Journal of Multivariate Analysis* **65**, 166-180.

[175] Nelsen, R.B. (1986). Properties of a one-parameter family of bivariate distributions with specified marginals. *Communications in Statistics* **15**, 3277-3285.

[176] Nguyen, T.T. (1997). A note on matrix variate normal distribution. *Journal of Multivariate Analysis* **60**, 148-153.

[177] Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society B* **44**, 414-442.

[178] Olkin, I., and J.W. Pratt (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics* **29**, 201-211.

[179] Olkin, I., and S.N. Roy (1954). On multivariate distribution theory. *Annals of Mathematical Statistics* **25**, 329-339.

[180] Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics* **40**, 549-567; Correction, *Annals of Mathematical Statistics* **42** (1971), 1777.

[181] 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.

[182] Press, W.H. (1992). *Numerical Recipies in C: The Art of Scientific Computing*. 2nd ed. Cambridge University Press, New York.

[183] Purkayastha, S., and M.S. Srivastava (1995). Asymptotic distributions of some test criteria for the covariance matrix in elliptical distributions under local alternatives. *Journal of Multivariate Analysis* **55**, 165-186.

[184] Rao, B.V., and B.K. Sinha (1988). A characterization of Dirichlet distributions. *Journal of Multivariate Analysis* **25**, 25-30.

[185] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. 2nd ed. John Wiley & Sons, New York.

[186] Redfern, D. (1996). *Maple V Release 4*. 3rd ed. Springer-Verlag, New York.

[187] Reeds, J.A. (1985). Asymptotic number of roots of Cauchy likelihood equations. *Annals of Statistics* **13**, 778-784.

[188] Rocke, D.M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics* **24**, 1327-1345.

[189] Romeu, J.L., and A. Ozturk (1993). A comparative study of goodness-of-fit tests for multivariate normality. *Journal of Multivariate Analysis* **46**, 309-334.

[190] Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*. eds. W. Grossmann, G. Pflug, I. Vincze and W. Wertz. Vol. 8. Reidel, Dordrecht, 283-297.

[191] Rousseeuw, P.J., and B.C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* **85**, 633-639.

[192] Rousseeuw, P.J., and V.J. Yohai (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis*. Lecture Notes in Statistics Vol. 26. Springer, New York, 256-272.

[193] Royston, J.F. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* **31**, 115-124.

[194] Royston, J.F. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Applied Statistics* **32**, 121-133.

[195] Ruppert, D. (1992). Computing S estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics* **1**, 253-270.

[196] Saw, J.G. (1978). A family of distributions on the m-sphere and some hypothesis tests. *Biometrika* **65**, 69-73.

[197] Schoenberg, I.J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics* **39**, 811-841.

[198] Sepanski, S.J. (1994). Asymptotics for multivariate t-statistic and Hotelling's T^2-statistic under infinite second moments via bootstrapping. *Journal of Multivariate Analysis* **49**, 41-54.

[199] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

[200] Shapiro, A., and M.W. Browne (1987). Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association* **82**, 1092-1097.

[201] Shapiro, S.S., and M.B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591-611.

[202] Shapiro, S.S., and R.S. Francia (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association* **67**, 215-216.

[203] Silvapulle, M.J. (1995). A Hotelling's T^2-type statistic for testing against one-sided hypotheses. *Journal of Multivariate Analysis* **55**, 312-319.

[204] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics* **9**, 1187-1195.

[205] Siotani, M., T. Hayakawa, and Y. Fujikoshi (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*. American Sciences Press, Columbus, Ohio.

[206] Small, N.J.H. (1978). Plotting squared radii. *Biometrika* **65**, 657-658.

[207] Spivak, M. (1965). *Calculus on Manifolds*. Addison-Wesley, New York.

[208] Srivastava, M.S. (1967). On fixed-width confidence bounds for regression parameters and the mean vector. *Journal of the Royal Statistical Society B* **29**, 132-140.

[209] Srivastava, M.S. (1984). Estimation of interclass correlations in familial data. *Biometrika* **71**, 177-185.

[210] Srivastava, M.S., and E.M. Carter (1980). Asymptotic expansions for hypergeometric functions. *Multivariate analysis V*. North-Holland, Amsterdam-New York, 337-347.

[211] Srivastava, M.S., and E.M. Carter (1983). *An Introduction to Applied Multivariate Statistics*. North-Holland, New York.

[212] Srivastava, M.S., and T.K. Hui (1987). On assessing multivariate normality based on Shapiro-Wilk W statistic. *Statistics & Probability Letters* **5**, 15-18.

[213] Srivastava, M.S., K.J. Keen, and R.S. Katapa (1988). Estimation of interclass and intraclass correlations in multivariate familial data. *Biometrics* **44**, 141-150.

[214] Srivastava, M.S., and C.G. Khatri (1979). *An Introduction to Multivariate Statistics*. North-Holland, New York.

[215] Srivastava, M.S., C.G. Khatri, and E.M. Carter (1978). On monotonicity of the modified likelihood ratio test for the equality of two covariances. *Journal of Multivariate Analysis* **8**, 262-267.

[216] Srivastava, M.S., and D. von Rosen (1998). Outliers in multivariate regression models. *Journal of Multivariate Analysis* **65**, 195-208.

[217] Srivastava, M.S., and W.K. Yau (1989). Saddlepoint method for obtaining tail probability of Wilks' likelihood ratio test. *Journal of Multivariate Analysis* **31**, 117-126.

[218] Stadje, W. (1993). ML characterization of the multivariate normal distribution. *Journal of Multivariate Analysis* **46**, 131-138.

[219] Stahel, W.A. (1981). *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. Ph. D. thesis, ETH Zürich.

[220] Statistical Sciences (1995). *S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.3*. StatSci, a division of MathSoft, Inc., Seattle, Washington.

[221] Stein, C. (1969). Multivariate Analysis I. Technical Report No. 42, Stanford University.

[222] Steyn, H.S. (1993). On the problem of more than one kurtosis parameter in multivariate analysis. *Journal of Multivariate Analysis* **44**, 1-22.

[223] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, **36**, 111-147.

[224] Strang, G. (1980). *Linear Algebra and its Applications*. 2nd ed. Academic Press, New York.

[225] Sugiura, N. (1973). Derivatives of the characteristic root of a symmetric or a hermitian matrix with two applications in mutivariate analysis. *Communications in Statistics* **1**, 393-417.

[226] Sugiura, N. (1976). Asymptotic expansions of the distributions of the latent roots and the latent vector of the Wishart and multivariate F matrices. *Journal of Multivariate Analysis* **6**, 500-525.

[227] Sugiura, N., and H. Nagao (1968). Unbiasedness of some test criteria for the equality of one or two covariance matrices. *Annals of Mathematical Statistics* **39**, 1686-1692.

[228] Sugiura, N., and H. Nagao (1971). Asymptotic expansion of the distribution of the generalized variance for noncentral Wishart matrix, when $\Omega = O(n)$. *Annals of the Institute of Statistical Mathematics* **23**, 469-475.

[229] Sutradhar, B.C. (1993). Score test for the covariance matrix of the elliptical t-distribution. *Journal of Multivariate Analysis* **46**, 1-12.

[230] Szablowski, P.J. (1998). Uniform distributions on spheres in finite dimensional L_α and their generalizations. *Journal of Multivariate Analysis* **64**, 103-117.

[231] Tang, D. (1994). Uniformly more powerful tests in a one-sided multivariate problem. *Journal of the American Statistical Association* **89**, 1006-1011.

[232] Tang, D. (1996). Erratum:"Uniformly more powerful tests in a one-sided multivariate problem" [*Journal of the American Statistical Association* **89** (1994), 1006-1011]. *Journal of the American Statistical Association* **91**, 1757.

[233] Tyler, D.E. (1982). Radial estimates and the test for sphericity. *Biometrika* **69**, 429-436.

[234] Tyler, D.E. (1983a). Robustness and efficiency properties of scatter matrices. *Biometrika* **70**, 411-420.

[235] Tyler, D.E. (1983b). The asymptotic distribution of principal components roots under local alternatives to multiple roots. *Annals of Statistics* **11**, 1232-1242.

[236] Tyler, D.E. (1986). Breakdown properties of the M-estimators of multivariate scatter. Technical report, Department of Statistics, Rutgers University.

[237] Uhlig, H. (1994). On singular Wishart and singular multivariate beta distributions. *Annals of Statistics* **22**, 395-405.

[238] van der Merwe, A., and J.V. Zidek (1980). Multivariate regression analysis and canonical variates. *Canadian Journal of Statistics*, **8**, 27-39.

[239] von Mises, R. (1918). Uber die "Ganzahligkeit" der Atomegewicht und verwante Fragen. *Physikalische Zeitschrift* **19**, 490-500.

[240] Wakaki, H., S. Eguchi, and Y. Fujikoshi (1990). A class of tests for a general covariance structure. *Journal of Multivariate Analysis* **32**, 313-325.

[241] Wang, Y., and M.P. McDermott (1998a). A conditional test for a nonnegative mean vector based on a Hotelling's T^2-type statistic. *Journal of Multivariate Analysis* **66**, 64-70.

[242] Wang, Y., and M.P. McDermott (1998b). Conditional likelihood ratio test for a nonnegative normal mean vector. *Journal of the American Statistical Association* **93**, 380-386.

[243] Waternaux, C.M. (1976). Asymptotic distributions of the sample roots for a non-normal population. *Biometrika* **63**, 639-664.

[244] Watson, G.S. (1983). *Statistics on Spheres*. The University of Arkansas Lecture Notes in Mathematical Sciences. John Wiley & Sons, New York.

[245] Wielandt, H. (1967). *Topics in the Analytic Theory of Matrices* (Lecture notes prepared by R.R. Meyer.) University of Wisconsin Press, Madison.

[246] Wilks, S.S. (1963). Multivariate statistical outliers. *Sankhyā: Series A* **25**, 407-426.

[247] Wolfram, S. (1996). *The Mathematica Book*. 3rd ed. Wolfram Media, Inc. and Cambridge University Press, New York.

[248] Wong, C.S., and D. Liu (1994). Moments for left elliptically contoured random matrices. *Journal of Multivariate Analysis* **49**, 1-23.

[249] Yamato, H. (1990). Uniformly minimum variance unbiased estimation for symmetric normal distributions. *Journal of Multivariate Analysis* **34**, 227-237.

Author Index

Subject Index

Springer Texts in Statistics *(continued from page ii)*

Noether: Introduction to Statistics: The Nonparametric Way
Peters: Counting for Something: Statistical Principles and Personalities
Pfeiffer: Probability for Applications
Pitman: Probability
Rawlings, Pantula and Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: A Decision-Theoretic Motivation
Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics
Terrell: Mathematical Statistics: A Unified Introduction
Whittle: Probability via Expectation, Third Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods