MULTIVARIATE DATA ANALYSIS

S. B. Provost
Statistical and Actuarial Sciences, The University of Western Ontario, London, Ontario, Canada

E. M. Rudiuk
Physical and Applied Sciences, University College of Cape Breton, Sydney, Nova Scotia, Canada

Keywords: Multivariate statistical inference, elliptically contoured distributions, quadratic forms, multivariate normal distribution, parameter estimation, tests of hypotheses, linear models, multiple regression, partial correlation, multivariate analysis of variance, discriminant analysis, covariance models, principal components, factor analysis.

Contents

UNESCO – EOLSS
SAMPLE CHAPTERS

Summary

Decisions in various spheres of activity including those of environmental or geo-political nature are often made on the basis of statistical analyses involving several variables. Certain basic results on multivariate distributions as well as some of the main techniques utilized in multivariate statistical inference are presented in this chapter. Some numerical examples illustrate the theory.

1. Introduction

Multivariate methods analyze several variables simultaneously, unlike the more familiar univariate or bivariate methods which deal with only one or two variables. A variable can be *independent* (or explanatory) in which case it is a quantity used to explain or predict the values of other variables which are called *dependent* (or response) variables. Multivariate statistical inference is often based on a data table (also called a data matrix) consisting of rows and columns. In this chapter, the rows contain the variables and the columns represent the objects or individuals being studied. Many of the results derived in multivariate analysis rely on matrix algebra, and it is assumed that reader has some knowledge of linear algebra in addition to being familiar with the basic concepts of mathematical statistics.

Section 2 introduces certain distributional properties of multivariate normal and elliptically contoured random vectors; some basic results are also presented in connection with quadratic forms and the Wishart distribution is defined. Estimators of the mean and covariance matrix of a ultivariate normal distribution are given in Section 3, while useful tests of hypotheses are enumerated in Section 4. Section 5 covers multiple regression and correlation as well as one- and two-way multivariate analysis of variance. Discriminant analysis is introduced in Section 6. Certain multivariate covariance structures are discussed in Section 7 and Section 8 which, respectively, deal with principal components and factor analysis.

2. Multivariate Distributions

The multivariate normal distribution is the most widely used distribution in multivariate statistical inference. Some of its main properties are summarized in Section 2.1 where certain distributional results on quadratic forms in normal random vectors are also given. The normal distribution belongs to the more general class of elliptically contoured distributions which is described in Section 2.2. A brief definition of the Wishart distribution is given in Section 2.3.

2.1 The Multivariate Normal Distribution

A p-variate random vector **X** is said to have a real nonsingular normal distribution with mean $\boldsymbol{\mu}$ = E(**X**) and real positive definite covariance matrix **Σ** = E[(**X** – E(**X**)) (**X** – E(**X**))'], if its density function is given by

$$f(\mathbf{x}) = \frac{\exp\{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\}}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}} \qquad (2.1.1)$$

The notation **X** ~ N_p ($\boldsymbol{\mu}$, **Σ**) will indicate that the p-variate random **X** is normally distributed with the parameters $\boldsymbol{\mu}$ and **Σ**. When $\boldsymbol{\mu} = \mathbf{0}$ (the null vector) and **Σ** = **I** (where **I** is a diagonal matrix whose diagonal elements are all equal to one), **X** is said to have a *standard normal* distribution. If the covariance matrix **Σ** of a random vector **X** is diagonal, then the components of **X** are independently distributed.

Let $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ where $\mathbf{X}_1$ has q components and $\mathbf{X}_2$ has p–q components, and **X** ~ N_p $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

UNESCO – EOLSS SAMPLE CHAPTERS

$$\boldsymbol{\mu} = \begin{pmatrix} \underset{q\times 1}{\boldsymbol{\mu}_1} \\ \underset{(p-q)\times 1}{\boldsymbol{\mu}_2} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \underset{q\times q}{\boldsymbol{\Sigma}_{11}} & \underset{q\times(p-q)}{\boldsymbol{\Sigma}_{12}} \\ \underset{(p-q)\times q}{\boldsymbol{\Sigma}_{21}} & \underset{(p-q)\times(p-q)}{\boldsymbol{\Sigma}_{22}} \end{pmatrix} ;$$

then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independently and normally distributed with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices Σ_{11} and Σ_{22} , respectively, whenever $\Sigma_{12} = \mathbf{O}$ and $\Sigma_{21} = \mathbf{O}$, $\mathbf{O}$ denoting the null matrix whose elements are all equal to zero. Moreover, letting $\mathbf{Y}_1 = \mathbf{X}_1$ and $\mathbf{Y}_2 = \mathbf{X}_2 - \Sigma_{21}\,\Sigma_{11}^{-1}\,\mathbf{X}_1$, it can be shown that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independently distributed with $\mathbf{Y}_1 \sim N_q\,(\boldsymbol{\mu}_1, \Sigma_{11})$ and $\mathbf{Y}_2 \sim N_{p\text{-}q}\,(\boldsymbol{\mu}_2 - \Sigma_{21}\,\Sigma_{11}^{-1}\boldsymbol{\mu}_1\,,\, \Sigma_{22.1})$ where $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\,\Sigma_{11}^{-1}\,\Sigma_{12}$. It follows that the conditional distribution of $\mathbf{X}_2$ given $\mathbf{X}_1 = \mathbf{x}_1$ is a multivariate normal with mean $\boldsymbol{\mu}_2 + \Sigma_{21}\,\Sigma_{11}^{-1}\,(\mathbf{x}_1 - \boldsymbol{\mu}_1)$ and covariance matrix $\Sigma_{22.1}$.

Let $\mathbf{A}$ be a $q \times p$ matrix of rank q with $q \times p$, $\mathbf{b}$ be a q-dimensional vector of constants and $\mathbf{X} \sim N_p\,(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{AX} + \mathbf{b} \sim N_q\,(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}')$. As an application, letting $p = q$, $\mathbf{A} = \boldsymbol{\Sigma}^{-1/2}$ and $\mathbf{b} = -\boldsymbol{\Sigma}^{-1/2}\,\boldsymbol{\mu}$ where $\boldsymbol{\Sigma}^{-1/2}$ denotes the symmetric square root of Σ^{-1}, one can standardize the vector $\mathbf{X} \sim N_p\,(\boldsymbol{\mu}, \Sigma)$ by means of the transformation $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, and then $\mathbf{Z} \sim N_p\,(\mathbf{0}, \mathbf{I})$. When q=1 (the case of a linear combination of the components of $\mathbf{X}$), one has $\mathbf{a}'\mathbf{X} \sim N\,(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$ for any real vector $\mathbf{a}$. Moreover, if $\mathbf{b}'\mathbf{X}$ is univariate normal for every real vector $\mathbf{b}$, then $\mathbf{X}$ is distributed as a multivariate normal vector; this is a characterization of the multivariate normal distribution.

If Σ, the covariance matrix of $\mathbf{X}$, has rank $r \le p$, then one can write $\Sigma = \mathbf{BB}'$, where $\mathbf{B}$ is $p \times r$. Whether Σ is singular or nonsingular, there exists a standard normal vector $\mathbf{Y} \sim N_r(\mathbf{0}, \mathbf{I})$ such that

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{BY} \qquad (2.1.2)$$

In the *singular normal* case, the density of $\mathbf{X}$ does not exist but all the properties of $\mathbf{X}$ can be studied through the vector $\mathbf{Y}$ of (2.1.2). In the nonsingular case, the matrix $\mathbf{B}$ in (2.1.2) has dimension $p \times p$ and is of rank p.

The *characteristic function* of $\mathbf{X} \sim N_p\,(\boldsymbol{\mu}, \Sigma)$ is

$$M_{\mathbf{X}}(\mathbf{t}) = E(\exp\{i\mathbf{t}'\mathbf{X}\}) = \exp\{i\mathbf{t}'\boldsymbol{\mu} - (\mathbf{t}'\Sigma\mathbf{t})/2\} \qquad (2.1.3)$$

where $\mathbf{t}$ is a p-dimensional real vector and $i = \sqrt{-1}$.

Quadratic Forms in Normal Vectors

Consider the *quadratic form* $\mathbf{X}'\,\mathbf{BX}$ where $\mathbf{X} \sim N_p\,(\mathbf{0}, \Sigma)$ with $\Sigma > 0$ and $\mathbf{B}$ is a real symmetric matrix. Writing $\mathbf{X}$ as $\Sigma^{1/2}\,\mathbf{Z}$ where $\Sigma^{1/2}$ is such that $\Sigma^{1/2}\,\Sigma^{1/2} = \Sigma$ and $\mathbf{Z} \sim$

UNESCO – EOLSS
SAMPLE CHAPTERS

N_p (**0**, **I**), one has **X**' **BX = Z**' ($\Sigma^{1/2}$ **B** $\Sigma^{1/2}$) **Z .** Let **Q** be an *orthogonal* matrix (with the property that **Q'Q = QQ' = I**) which diagonalizes $\Sigma^{1/2}$ **B** $\Sigma^{1/2}$. Then **Q**'($\Sigma^{1/2}$**B**$\Sigma^{1/2}$)**Q** = Λ, where Λ is a diagonal matrix whose diagonal elements λ_i are the characteristic roots of $\Sigma^{1/2}$ **B** $\Sigma^{1/2}$, and **X'BX** = **Z**' **QΛQ**' **Z** . Note that on letting **W** = **Q**' **Z,** one has **W**~ N_p(**0**, **Q'IQ**) ~ N_p (**0**, **I**). Thus,

$$\mathbf{X}'\,\mathbf{BX} = \mathbf{W}'\,\Lambda\mathbf{W} = (\,W_1\;\;W_2\;\;\ldots\;\;W_p\,)\begin{pmatrix}\lambda_1 & 0 & . & 0\\ 0 & \lambda_2 & & 0\\ . & & . & .\\ 0 & 0 & . & \lambda_p\end{pmatrix}\begin{pmatrix}W_1\\ W_2\\ .\\ W_p\end{pmatrix} = \sum_{i=1}^{p}\lambda_i W_i^2\ ,$$

where the W_i's are independent standard normal variables, that is, the quadratic form **X'BX** is distributed as a linear combination of independent chi-square variables each having one degree of freedom.

We now state two basic results in connection with quadratic forms.

1. Let **X** ~ N_p (**0**, **I**); then **X'BX** ~ χ^2_r if and only if **B** = $\mathbf{B}^2$ (that is, **B** is idempotent) and the rank of **B** is r .
2. Let $\mathbf{Q}_1$ = **X'AX** and $\mathbf{Q}_2$ = **X' BX** where **X** ~ $\mathrm{N_p}$ (**0, I**) , **A** = **A**' and **B** = **B**' ; then $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are independently distributed if and only if **AB** = **BA** = **O** (Craig’s theorem).

2.2. Elliptically Contoured Distributions

We shall first present some of the main properties characterizing the class of elliptically contoured distributions via the multivariate normal distribution.

Consider the equation f(**x**) = c_1 where c_1 is a constant and f(**x**) is the density function of a multivariate normal distribution given in Equation 2.1.1. Then

$$-\ln f(\mathbf{x}) = -\ln c_1 \Rightarrow (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c \qquad (2.2.1)$$

where c is a constant. Thus the contours of constant density are ellipsoids whenever **X** ~ N_p (**μ** , **Σ**) When **μ** = **0** and **Σ** = **I**, the identity matrix, these ellipsoids are spheres centered at the origin and **X** is said to have a spherical distribution. Moreover, the characteristic function of **Y** ~ N_p (**0, I**) as well that of **Z** = **PY**, where **P** is an orthogonal matrix, are one and the same, which means that there is invariance under orthogonal transformations. If the factor $\exp\{-(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})/2\}$ of (2.1.3) is replaced by a general nonnegative function of **t**' **Σ t** , the resulting distribution is said to belong to the class of elliptically contoured distributions. Elliptically contoured and spherical distributions are formally defined below.

UNESCO – EOLSS SAMPLE CHAPTERS

Let $\boldsymbol{\mu}$ be a *p*-dimensional real vector, **Σ** be a $p \times p$ nonnegative definite matrix and $\xi(.)$ be a nonnegative function; then the *p*-dimensional vector **X** is said to have an *elliptical* or *elliptically contoured* distribution if its characteristic function $\phi(\mathbf{t})$ can be expressed as $\exp(i\mathbf{t}'\boldsymbol{\mu})\xi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$, and we write $\mathbf{X} \sim C_p\ (\xi\ ; \boldsymbol{\mu},\ \boldsymbol{\Sigma})$. If some components of **t** in the characteristic function of **X** are set to zero, the resulting characteristic function still has the same form, which means that if **X** has an elliptically contoured distribution then all the marginal distributions are also elliptical. When $\boldsymbol{\mu}$ is the null vector and **Σ** is the identity matrix of order *p*, the notation $\mathbf{X} \sim C_p\ (\xi\ ; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is shortened to $\mathbf{X} \sim S_p\ (\xi)$, and **X** is said to have a *spherical* or *spherically symmetric* distribution.

Let **Y** be spherically distributed with p.d.f. $g(\mathbf{y}'\mathbf{y})$ and **X** be distributed as $\boldsymbol{\mu} + \mathbf{BY}$ where $\boldsymbol{\mu}$ is a *p*-dimensional vector and **B** is a $p \times p$ matrix such that $\boldsymbol{\Sigma} = \mathbf{BB}'$; then **X** has the p.d.f. $|\,\boldsymbol{\Sigma}\,|^{-1/2}\ g[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$. Clearly, the contours of constant density of **X** are ellipsoids as in the multivariate normal case.

Elliptically contoured distributions can also be defined as follows: Let **Σ** be $p \times p$ matrix of rank $q \le p$; then there exists a $q \times p$ matrix **L** such that $\boldsymbol{\Sigma} = \mathbf{L}'\mathbf{L}$. Now, letting *W* be a nonnegative random variable which is identically distributed in every direction along radii from the point $\boldsymbol{\mu}$ and $\mathbf{U}^{(\alpha)}$ denote a random vector which is uniformly distributed on the unit sphere in $\Re^{\alpha}$ and whose distribution is independent of that of *W*, one has $\mathbf{X} \sim C_p\ (\xi\ ;\ \boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if $\mathbf{X} \sim \boldsymbol{\mu} + \mathbf{L}'\ W\ \mathbf{U}^{(q)}$ and $F(w)$, the distribution function of *W*, is such that

$$\xi(\mathbf{t}'\mathbf{t}) = E_W\,[E_{\mathbf{U}^{(q)}|W=w}(\exp\{i\mathbf{t}'w\mathbf{U}^{(q)}\})] = \int_0^{\infty} E(\exp\{iw\mathbf{t}'\mathbf{U}^{(q)}\})dF(w)$$

where $\xi(\mathbf{t}'\,\mathbf{t})$ is the characteristic function of $W\mathbf{U}^{(q)} \sim S_q\ (\xi)$.

2.3 The Wishart Distribution

Letting $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_\mathrm{n}$ be a random sample of size n from a $N_p\ (\mathbf{0}, \boldsymbol{\Sigma})$ population,

$$\underset{p\times n}{\mathbf{X}} = \begin{pmatrix} \underset{p\times 1}{\mathbf{X}_1} & \mathbf{X}_2 & \ldots & \mathbf{X}_n \end{pmatrix} \qquad \text{and} \qquad \underset{p\times p}{\mathbf{A}} = \underset{p\times n}{\mathbf{X}}\ \underset{n\times p}{\mathbf{X}'} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ . & . & . & . \\ a_{p1} & a_{p2} & \ldots & a_{pp} \end{pmatrix},$$

the random matrix **A** is said to follow a Wishart distribution with *n* degrees of freedom and covariance parameter **Σ** , and we write $\mathbf{A} \sim W_p\ (n\ , \boldsymbol{\Sigma})$.

3. Parameter Estimation for a Multivariate Normal Population

UNESCO – EOLSS
SAMPLE CHAPTERS

A multivariate normal vector $\mathbf{X} \sim N_p\,(\boldsymbol{\mu}, \Sigma)$ is specified completely by its mean $\boldsymbol{\mu}$ and covariance matrix Σ. Given a random sample $\mathbf{X}_1$, $\mathbf{X}_2$,…, $\mathbf{X}_N$ with $N > p$ from this distribution, the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained by maximizing the likelihood function of the sample, which is

$$\frac{\exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}[\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})'])\}}{(2\pi)^{pN/2}\,|\,\boldsymbol{\Sigma}\,|^{N/2}}\ .$$

They are respectively

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})'\ .$$

The sample mean $\bar{\mathbf{X}}$ is normally distributed as $N_p\,(\boldsymbol{\mu}, \Sigma/N)$, and so $N(\bar{\mathbf{X}} - \boldsymbol{\mu})'\,\Sigma^{-1}\,(\bar{\mathbf{X}} - \boldsymbol{\mu})$ has a chi-square distribution with p degrees of freedom. It can also be shown that $N\hat{\boldsymbol{\Sigma}}$ is distributed as $\sum_{i=1}^{N-1}$ $\mathbf{Z}_i\,\mathbf{Z}_i'$ where the $\mathbf{Z}_i$'s are independently distributed as $N_p(\mathbf{0}, \Sigma)$, i = 1, 2 ,…, N–1. Consequently, letting the sample covariance S = $\sum_{i=1}^{N}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})'/(N-1)$, $(N - 1)S = N\hat{\boldsymbol{\Sigma}}$ is seen to be distributed as a Wishart random matrix with (N – 1) degrees of freedom. The estimators $\bar{\mathbf{X}}$ and S are independently distributed and respectively unbiased for $\boldsymbol{\mu}$ and Σ .

4. Tests of Hypotheses for Mean Vectors and Covariance Matrices

Some useful test statistics for the mean vectors and covariance matrices of multivariate normal populations are enumerated below.

1. *Testing that a mean is equal to a given vector when the covariance matrix is known*.
Let $\mathbf{X}_1$, … , $\mathbf{X}_N$ be a random sample from $N_p\,(\boldsymbol{\mu}, \Sigma)$ and Σ be a known covariance matrix; then a rejection region of size α to test the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ **,** where $\boldsymbol{\mu}_0$ is a specified vector, is given by

$$N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\,\Sigma^{-1}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) > \chi_p^{\,2}(\alpha)$$

where $\chi_p^{\,2}(\alpha)$ is a critical value such that $P[\chi_p^{\,2} > \chi_p^{\,2}(\alpha)] = \alpha$, and a $100\ (1 - \alpha)\%$ confidence region for the mean vector $\boldsymbol{\mu}$ is specified by the set of points $\mathbf{m}$ satisfying the following inequality:

$$N(\bar{\mathbf{X}} - \mathbf{m})'\,\Sigma^{-1}\,(\bar{\mathbf{X}} - \mathbf{m}) \le \chi_p^{\,2}(\alpha)\ .$$

2. *Testing equality of the mean vectors in two populations with known common covariance matrix*.

UNESCO – EOLSS
SAMPLE CHAPTERS

Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ be random samples from $N_p\,(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma})$ and $N_p\,(\boldsymbol{\mu}_{(2)}, \boldsymbol{\Sigma})$, respectively, with sample means $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$, and assume that $\boldsymbol{\Sigma}$ is known. Then, $(\overline{\mathbf{X}} - \overline{\mathbf{Y}})$ is distributed as
$N_p\,(\boldsymbol{\mu}_{(1)} - \boldsymbol{\mu}_{(2)}, [(1/N) + (1/M)]\boldsymbol{\Sigma})$

and a rejection region for testing the hypothesis $\boldsymbol{\mu}_{(1)} = \boldsymbol{\mu}_{(2)}$, is given by

$$\frac{NM}{N+M}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})'\,\boldsymbol{\Sigma}^{-1}\,(\overline{\mathbf{X}} - \overline{\mathbf{Y}}) > \chi_p{}^2(\alpha)\,.$$

3. *Testing that a mean is equal to a given vector when the covariance matrix is unknown*

The likelihood ratio test of the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ for a $N_p\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population is based on $T^2 = N(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\,S^{-1}\,(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)$, where $\overline{X}$ and S are respectively the sample mean and the sample covariance. A 100 (1 - α)% confidence region for the mean vector $\boldsymbol{\mu}$ is given by

$$N(\overline{\mathbf{X}} - \mathbf{m})'\,S^{-1}\,(\overline{\mathbf{X}} - \mathbf{m}) \le \frac{(N-1)p}{(N-p)}\,F_{p,\,N-p}\,(\alpha)$$

as an inequality with respect to $\mathbf{m}$ where $F_{p,\,N-p}\,(\alpha)$ is such that $P(F_{p,\,N-p} > F_{p,\,N-p}\,(\alpha)) = \alpha$, $F_{p,\,N-p}$ denoting the F distribution with p and $N-p$ degrees of freedom.

4. *Testing equality of means in two populations with unknown common covariance matrix*

Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ be random samples from $N_p\,(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma})$ and $N_p\,(\boldsymbol{\mu}_{(2)}, \boldsymbol{\Sigma})$, respectively, $\boldsymbol{\Sigma}$ being unknown, $T^2 = NM\,(\overline{\mathbf{X}} - \overline{\mathbf{Y}})'\,S^{-1}\,(\overline{\mathbf{X}} - \overline{\mathbf{Y}})/(N+M)$, and
$S = \dfrac{1}{N+M-2}\left\{\sum_{i=1}^{N}(\mathbf{X}_\mathbf{i} - \overline{\mathbf{X}})(\mathbf{X}_\mathbf{i} - \overline{\mathbf{X}})' + \sum_{i=1}^{M}(\mathbf{Y}_\mathbf{i} - \overline{\mathbf{Y}})(\mathbf{Y}_\mathbf{i} - \overline{\mathbf{Y}})'\right\}$. A rejection region for testing the hypothesis $\boldsymbol{\mu}_{(1)} = \boldsymbol{\mu}_{(2)}$, is given by

$$T^2 > \frac{(N+m-2)p}{N+M}\,F_{p,\,N+M-p-1}(\alpha)\,.$$

5. *Testing the hypothesis that a mean vector and a covariance matrix are equal to a given vector and matrix.*

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ be a random sample from $N_p\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite covariance matrix. The likelihood ratio criterion for testing the hypothesis H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are given, is

UNESCO – EOLSS
SAMPLE CHAPTERS

$$\lambda = \left(\frac{e}{N}\right)^{\frac{1}{2}pN} \mid \mathbf{A}\boldsymbol{\Sigma}_0^{-1} \mid^{\frac{1}{2}N} \ \exp\{-\frac{1}{2}[\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_0^{-1}) + N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)]\}$$

where $\mathbf{A} = \sum_{i=1}^{N}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'$. Under the null hypothesis, – 2log λ is asymptotically distributed as a chi-square random variable with $p(p + 1)/2 + p$ degrees of freedom.

6. *Testing the hypothesis that a covariance matrix is equal to a given matrix*.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a positive definite covariance matrix. The likelihood ratio criterion for testing the hypothesis H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, where $\boldsymbol{\Sigma}_0$ is a given covariance matrix, is

$$\lambda = \left(\frac{e}{N}\right)^{\frac{1}{2}pN} \mid \mathbf{A}\boldsymbol{\Sigma}_0^{-1} \mid^{\frac{1}{2}N} \ \exp\{-\frac{1}{2}[\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_0^{-1})]\}$$

where $\mathbf{A} = \sum_{i=1}^{\mathbf{N}}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'$. Under the null hypothesis, – 2log λ is asymptotically distributed as a chi-square random variable with $p(p + 1)/2$ degrees of freedom.

7. *Testing the hypothesis that a covariance matrix is proportional to an identity matrix*.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a positive definite covariance matrix. The likelihood ratio criterion for testing the hypothesis H_0: $\boldsymbol{\Sigma} = \sigma^2\,\mathbf{I}$, where σ^2 is not specified , is

$$\lambda = \frac{\mid \mathbf{A} \mid^{\frac{1}{2}N}}{(\mathrm{tr}(\mathbf{A})/p)^{\frac{1}{2}pN}}$$

where $\mathbf{A} = \sum_{i=1}^{N}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'$. Under the null hypothesis, – 2log λ is asymptotically distributed as a chi-square random variable with $p(p + 1)/2 - p$ degrees of freedom. This test is called the *sphericity test*. It determines whether the components of a random vector are uncorrelated and they all have equal variances.

-
-
-

UNESCO – EOLSS
SAMPLE CHAPTERS

TO ACCESS ALL THE **23 PAGES** OF THIS CHAPTER,
Visit: http://www.eolss.net/Eolss-sampleAllChapter.aspx

Bibliography

Anderson T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 675 pp. New York, Wiley.[A well-written textbook which introduces the essential aspects of the theory of multivariate statistical inference]

Flury B. and Riedwyl H. (1988). *Multivariate Statistics*, *A Practical Approach*. 296 pp. London, Chapman and Hall.[Revised translation of 'Angewandte multivariate Statistik'. Stuttgart : Gustav Fischer Verlag, 1983]

Giri, N.C. (1977). *Mutivariate Statistical Inference*. 320 pp. New York, Academic Press.[An introductory textbook on the topic, which also discusses group invariance]

Green P.E. (1978). *Analyzing Multivariate Data*, 520 pp. Hinsdale, The Dryden Press.[An applied textbook aimed at users of multivariate statistical techniques who are not professional statisticians]

Johnson R.A. and Wichern D.W. (1998). *Applied Multivariate Statistical Analysis*, *Fourth Edition*, 816 pp. New Jersey, Prentice Hall.[An excellent textbook which offers a clear presentation of the main theoretical concepts in multivariate statistical inference and contains a wealth of illustrative examples]

Kres H. (1983). *Statistical Tables for Multivariate Analysis*, *A Handbook with References to Applications*, 504 pp. New York, Springer Verlag.[A reference containing numerous tables to be used in conjunction with tests of hypotheses performed on the basis of multivariate data. Translation of 'Statistische Tafeln zur multivariaten Analysis']

Krzanowski W.J. and Marriott F.H.C. (1994). *Multivariate Analysis Part* 1, *Distributions*, *Ordination and Inference*, 280 pp. London, Halsted Press.[This text emphasizes the distributional aspects involved in multivariate analysis]

Mathai A.M. and Provost S.B. (1992). *Quadratic Forms in Random* Variables, *Theory and Applications*, 367 pp. New York, Marcel Dekker, Inc. [This monograph is a compendium of useful results on the distribution of quadratic forms in normal random variables]

Mardia K.V., Kent J.T. and Bibby J.M. (1979). *Multivariate Analysis*, 522 pp. New York, Academic Press. [A mathematical treatment of the theory of multivariate statistical inference]

Rao C.R. (1965). *Linear Statistical Inference and Its Applications*, 522 pp. New York, Wiley.[This text offers a classical treatment of the theory of linear models and also includes numerous key results in multivariate analysis]

Seber G.A.F. (1984). *Multivariate Observations*, 686 pp. New York, Wiley.[A useful textbook which includes data oriented techniques as well as a sound coverage of classical methods]

Siotani M., Hayakawa T. and Fujikoshi Y. (1985). *Modern Multivariate Statistical Analysis A* Graduate *Course and Handbook*, 760 pp. Columbus, Ohio, American Sciences Press, Inc.[A solid theoretical treatment of the theory of multivariate analysis]

Srivastava M.S. and Carter E.M. (1983). *An Introduction to Applied Multivariate Statistics*, 394 pp. New York, North Holland.[A well-written textbook which features the main methodologies in use in multivariate statistics and includes several worked examples on each topic]

Biographical Sketches

UNESCO – EOLSS SAMPLE CHAPTERS

Serge B. Provost is Professor in the Department of Statistical and Actuarial Sciences at the University of Western Ontario. He received his B.Sc and M.Sc degrees from l'université de Montréal and his Ph.D. degree in Mathematics and Statistics from McGill University in 1984. He is a Fellow of the *Royal Statistical Society*, a Chartered Statistician, an elected member of the *International Statistical Institute*, and an Associate of the *Society of Actuaries*. He is a member of the *American Statistical Association*, the *Institute of Mathematical Statistics* and the *Statistical Society of Canada*. He also belongs to the Education sections of the *American Statistical Association*, the *International Statistical Institute* and the *Society of Actuaries*. He has supervised seventeen Master's students and three Ph.D. students. He edited issues of the *International Journal of Mathematical and Statistical Sciences* and the *Pakistan Journal of Statistics*, co-edited three books and is Editor-at-Large for the Statistics series published by Marcel Dekker, Inc. He was an invited research scholar at the Centre for Mathematical Studies, Trivandrum, India, in December 2000 and the Korean Advanced Institute for Science and Technology, Taejon, Korea, in August 2001. He also served on the Bilingualism Committee of the *Statistical Society of Canada* and is currently an academic consultant for *STATLAB*, the department statistical consulting unit. He is the author of two research monographs, four encyclopedia articles and some thirty-five papers published in peer-reviewed publications. The joint paper with E. M. Rudiuk entitled ``The sampling distribution of the serial correlation coefficient" and published in 1995 in The *American Journal of Mathematical and Management Sciences* was awarded the Jacob Wolfowitz Prize in 1997. His research interests include distribution theory, time series, density estimation, order statistics and multivariate analysis with applications to the environmental sciences, econometrics and engineering, among other disciplines.

Edmund M. Rudiuk received an M.Sc. degree in Mathematics from the University of Maria Curie-Sklodowska, Lublin, Poland as well as M.Sc. and Ph.D. degrees in Statistics from the University of Western Ontario in 1992 and 1993, respectively. He is currently an Associate Professor of Statistics at the University College of Cape Breton, Sydney, Canada. His research interests include the distribution of ratios of quadratic forms and in particular that of the sample serial correlation coefficient as well as applications in distribution theory of the inverse Mellin transform technique and the generalized hypergeometric functions. On the practical side, he was involved in the analysis of cancer data and of pollution data at the tar ponds and coke oven site in Sydney, Cape Breton. In 1997, he was awarded the Jacob Wolfowitz Prize by the Editorial Board of the *American Journal of Mathematical and Management Sciences*.

UNESCO – EOLSS
SAMPLE CHAPTERS