MT4609 Multivariate Analysis

This module introduces some of the ideas and techniques of multivariate statistical analysis.

Lecturer & Module co-ordinator: Dr Giorgos Minas
e-mail address: gm256@st-andrews.ac.uk
Office: Room 310, Maths Institute

Prerequisite: MT3507 or MT3606

Lectures: At 10 a.m. on Wednesdays, Fridays and odd Mondays in Room 3B of the Maths Institute.
Style of lectures: Notes with gaps.
Tutorials/Feedback sessions: In Room 1C at 2 p.m. on odd Tuesdays, starting in week 3.

Assessment: 2 Hour Examination = 100%.

Attendance: As with all 3000-5000 Level taught modules in the School of Mathematics and Statistics attendance will be monitored at various points over the semester. This is to comply with UKVI regulations and to monitor student wellbeing. For MT4609 attendance will be taken at the lectures of Wednesday/13.02.2019 and Wednesday/13.03.2019.

MMS: See MMS for lecture notes, tutorial sheets and solutions, and past papers (later in the semester) and additional information.

Lecture notes: Lecture notes will be added to MMS before lectures (typically the Friday before each week) and a hard copy for each student will also be available in class. As you will notice, some derivations will be left blank and you will be asked to complete them during lectures. Lecture notes with filled gaps will be added to MMS at a later time.

Fire Exit Route: Room 3B: Go down the adjacent south stairwell, and use the exit under the archway.
Room 1C: Turn right as you leave the room and use the exit under the archway.
Fire Assembly Point: Grass on the south side of the Computer Science Building.

Syllabus 2018-19

[The numbers in square brackets give the approximate number of lectures.]

1. **Introduction** [3]
   What is multivariate analysis?; Multivariate distributions; population moments; summary statistics.
2. **The Multivariate Normal (MVN) distribution** [4]
   Definition; principal axes; properties of the MVN distribution.
3. **Correlation** [2.5]
   The matrix $\mathbf{\Sigma}_{11\cdot 2}$; partial correlation; multiple correlation.

Multivariate Normal Inference

4. **Estimation** [2.5]
   Method of moments; some mathematical preliminaries; maximum likelihood estimators (MLEs); distributions of the MLEs.
5. **Tests on means** [2.5]
   Known variance matrix $\mathbf{\Sigma}$; Unknown variance matrix $\mathbf{\Sigma}$; testing the difference between two means.
6. **Tests on variance matrices** [3]
   Likelihood ratio tests; Testing whether $\mathbf{\Sigma} = \mathbf{\Sigma}_0$; Union–intersection test of independence.
7. **Discriminant Analysis** [2]
   Fisher's linear discriminant function; using R for discriminant analysis; maximum likelihood (ML) discriminant rule for known MVN populations; ML discriminant rule when parameters are unknown.

Data-analytic methods

8. **Preliminary data analysis using R.** [1]
   Data inspection; checking multivariate normality.
9. **Principal components analysis** [1.5]
   Principal components and eigenvectors of **S**; using R for principal components analysis.
10. **Canonical correlation** [2]
    Finding canonical variates; using R for canonical correlation analysis.

Tutorials

We will work on five tutorial sheets that will cover:

1. Matrix analysis revision, population moments, principal axes
2. Moments of linear combinations of variables, principal axes, correlation.
3. Multiple correlation coefficient, Estimation, Testing on means
4. Hypothesis testing on variance matrices,
5. Discriminant analysis, PCA, Canonical correlation

Principal references

Multivariate Statistical Methods (4th ed.) : D. F. Morrison; Thomson/Brooks/Cole; 2005. (M)
Multivariate Analysis : K. V. Mardia, J. T. Kent & J. M. Bibby; Academic Press; 1979; (MKB)
Applied Multivariate Statistical Analysis (6th ed.) : R. A. Johnson & D. W. Wichern; Pearson; 2007. (JW)

Notation

The notes will use the following notation:

Vectors We denote the $n \times 1$ dimensional (column) real vector or n-vector by $\mathbf{a} = (a_1, a_2, \ldots, a_n)'$. The indicator i, as in a_i, indicate that a_i is the i^{th} entry of $\mathbf{a}$. The vector $\mathbf{a}'$ is the transpose of the vector $\mathbf{a}$. That is, for

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \qquad \mathbf{a}' = (a_1, a_2, \ldots, a_n).$$

We call *unit* vectors or *normalised* vectors, the vectors $\mathbf{a}$ that for some norm function $\|\cdot\|$ have $\|\mathbf{a}\| = 1$.

Dot product The dot product of two n-vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)'$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)'$ is denoted as

$$\mathbf{a}'\mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

Matrices We denote the $n \times m$ matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & & a_{nm} \end{pmatrix}$$

by bold capital letters and, when necessary, write $\mathbf{A} = (a_{ij})$ or $\mathbf{A} = (a_{ij})_{i,j=1}^{n,m}$ to display its entries. The transpose of matrix $\mathbf{A}$, $\mathbf{A}'$, is the $m \times n$ matrix with $(ij)^{th}$ entry a_{ji}, that is $\mathbf{A}' = (a_{ji})_{j,i=1}^{m,n}$. We will write the columns of matrix $\mathbf{A}$ as the vectors $\mathbf{a}_{\cdot j} = (a_{1j}, a_{2j}, \ldots, a_{nj})'$, $j = 1, \ldots, m$, and its rows as the vectors $\mathbf{a}_{i\cdot} = (a_{i1}, a_{i2}, \ldots, a_{im})'$, $i = 1, \ldots, n$. We will sometimes use this notation to write matrices in terms of their columns, that is

$$\mathbf{A} = (\mathbf{a}_{\cdot\mathbf{1}}\ \mathbf{a}_{\cdot\mathbf{2}} \ldots \mathbf{a}_{\cdot\mathbf{m}}),$$

or their rows

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}'_{1\cdot} \\ \mathbf{a}'_{2\cdot} \\ \vdots \\ \mathbf{a}'_{n\cdot} \end{pmatrix}.$$

This notation is particularly useful for matrix multiplication. If $\mathbf{B}$ is an $m \times k$ matrix, then the $n \times k$ matrix $\mathbf{AB}$ can be written as

$$\mathbf{AB} = \begin{pmatrix} \mathbf{a}'_{1\cdot}\mathbf{b}_{\cdot 1} & \mathbf{a}'_{1\cdot}\mathbf{b}_{\cdot 2} & \cdots & \mathbf{a}'_{1\cdot}\mathbf{b}_{\cdot k} \\ \mathbf{a}'_{2\cdot}\mathbf{b}_{\cdot 1} & \mathbf{a}'_{2\cdot}\mathbf{b}_{\cdot 2} & \cdots & \mathbf{a}'_{2\cdot}\mathbf{b}_{\cdot k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}'_{n\cdot}\mathbf{b}_{\cdot 1} & \mathbf{a}'_{n\cdot}\mathbf{b}_{\cdot 2} & & \mathbf{a}'_{n\cdot}\mathbf{b}_{\cdot k} \end{pmatrix} = (\mathbf{Ab}_{\cdot\mathbf{1}}\ \mathbf{Ab}_{\cdot\mathbf{2}} \cdots \mathbf{Ab}_{\cdot\mathbf{k}}) = \begin{pmatrix} \mathbf{a}'_{1\cdot}\mathbf{B} \\ \mathbf{a}'_{2\cdot}\mathbf{B} \\ \vdots \\ \mathbf{a}'_{n\cdot}\mathbf{B} \end{pmatrix},$$

where $\mathbf{Ab}_{\cdot\mathbf{j}}$, $j = 1, \ldots, k$, are $n \times 1$ vectors and $\mathbf{a}'_{i\cdot}\mathbf{B}$ are $1 \times k$ vectors. We will sometimes write the column vector $\mathbf{a}_{\cdot j} = \mathbf{a}_j$ (i.e. without the $\cdot$ symbol in subscript).

It is useful to note that we can always think of vectors as matrices with one dimension equal to one and scalars as matrices with both dimensions equal to one. Scalars correspond to a linear transformation (just like all matrices correspond to linear transformations) that scales the number that is applied to.

Symmetric matrices A symmetric matrix $\mathbf{A}$ is a square $p \times p$ matrix that has entries $a_{ij} = a_{ji}$, $i, j = 1, \dots, p$. Then $\mathbf{A} = \mathbf{A}'$.

Diagonal matrices A diagonal matrix $\mathbf{A}$ is a square $p \times p$ matrix with all non-diagonal elements equal to 0. That is, a diagonal matrix $\mathbf{A}$ is of the form

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{pp} \end{pmatrix}.$$

We write a diagonal matrix $A = \text{diag}(a_{11}, \dots, a_{pp})$. Notice that diagonal matrices are symmetric.

Identity matrix A diagonal matrix with all elements of the main diagonal equal to 1. The $p \times p$ identity matrix is denoted by $\mathbf{I}_p$.

Random variables with equal probability distribution We write $X \sim Y$ to state that the two random variables, X and Y, have the same probability distribution, and $X \overset{H}{\sim} Y$ to state that the two random variables have the same distribution under the condition H.

Normal distribution We write $X \sim N(\mu, \sigma^2)$ to state that the random variable X is Normally distributed with mean μ and variance σ^2. The probability density function (p.d.f.) of X is then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

If we set the mean $\mu = 0$ and the variance $\sigma^2 = 1$ then we derive the standard Normal distribution $N(0, 1)$. The p.d.f. of the standard Normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty.$$

1 Introduction

1.1 What is multivariate analysis (MVA)?

In univariate analysis, a common sampling scheme is to measure a *single* random variable on each of n subjects. Most statistical problems that you have met so far have been of this type, except for regression problems where the focus is still on a *single* response variable. In contrast, it is very often the case when encountering real life problems that a number of random variables are measured for each subject and the relationship between these variables is of interest.
We live in the era of "big-data" and the number of measured variables in nowadays datasets can vary from tens to hundreds (e.g. answers to questions of a survey), to thousands (e.g. expression of every gene on a tissue sample of an organism, colour intensity in every voxel of an MRI image), or even millions of measured variables (e.g. preferences expressed on social networks). The computational aspects need to be carefully considered for such large datasets (even storing or performing simple operations can be a challenge), but statistical principles still apply and also need to be considered carefully. This module provides the mathematical underpinnings of statistical methods for inference and analysis of datasets that involve multiple correlated random variables.

In multivariate analysis, we typically assume that p **measurements/variables** are taken for each of our n **subjects**. For example, a physician may have seen n patients and has measured their weight, height, temperature, blood pressure, and oxygen level ($p = 5$). The different measurements on any one individual may well be correlated, e.g. height and weight, temperature and blood pressure. These dependencies can be hard to describe.

The multivariate data set is usually recorded as an $n \times p$ data matrix $\mathbf{D}$ in which the $(i, j)^{\text{th}}$ entry is x_{ij}, the reading on the i^{th} subject of the j^{th} variable.

$$\mathbf{D} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{pmatrix} \qquad = (\mathbf{x}_{\cdot 1} \cdots \mathbf{x}_{\cdot p}) = \begin{pmatrix} \mathbf{x}'_{1\cdot} \\ \vdots \\ \mathbf{x}'_{n\cdot} \end{pmatrix}.$$

Therefore rows correspond to subjects and columns correspond to variables. The convention in this module will be to write vectors as column vectors (see **Notation**). The measurements on the i^{th} subject form the i^{th} row of $\mathbf{D}$, and we will therefore write this row as the vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, where prime denotes the transpose.

1.2 Multivariate distributions

Defn. A **random vector X** is a vector $\mathbf{X} = (X_1, \dots, X_p)'$ in which the elements X_i are random variables.

It is often reasonable to assume that the data in matrix $\mathbf{D}$ form a random sample from the p-variate distribution of $\mathbf{X}$. Thus we regard $\mathbf{x}_1, \dots, \mathbf{x}_n$, with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, $i = 1, \dots, n$, as n independent observations on $\mathbf{X} = (X_1, \dots, X_p)'$.

N.B. Although this says that the n rows of the matrix $\mathbf{D}$ are independent observations, the random variables $X_1, \dots, X_p$ in the p columns are usually dependent.

We can specify the distribution of the random vector $\mathbf{X} = (X_1, \dots, X_p)'$, by its cumulative distribution function (c.d.f.) F, given by

$$F(\mathbf{x}) = P(X_1 \le x_1, \ldots, X_p \le x_p), \qquad \text{where } \mathbf{x} = (x_1, \ldots, x_p)'.$$

If the random variables X_i are continuous, the random vector $\mathbf{X}$ has a probability density function (p.d.f.) $f(\mathbf{x})$, given by

$$f(\mathbf{x}) \;=\; \frac{\partial^p F(\mathbf{x})}{\partial x_1 \ldots \partial x_p},$$

where $\partial F/\partial x_i$ denotes the partial derivative of F with respect to x_i.
Then, for any region A in $\mathbb{R}^p$,

$$P(\mathbf{X} \in A) \;=\; \int_A f(\mathbf{x})\, d\mathbf{x} \;=\; \int \cdots \int_A f(\mathbf{x})\, dx_1 \ldots dx_p.$$

Notes
(i) A function $f(\mathbf{x})$ is a p.d.f. if and only if

(a) $f(\mathbf{x}) \ge 0$ for all $\mathbf{x} \in \mathbb{R}^p$ and (b) $\int_{\mathbb{R}^p} f(\mathbf{x})\, d\mathbf{x} = 1.$ (Honesty Condition)

(ii) The properties for bivariate distributions given in MT2504 are a special case of those stated above.

Example
If $X_1, \ldots, X_p$ are independent $N(0, 1)$ random variables with p.d.f. $\phi(x)$, then the p.d.f. of the random vector $\mathbf{X} = (X_1, \ldots, X_p)'$ at $\mathbf{x} = (x_1, \ldots, x_p)'$, where $-\infty < x_i < \infty$ $(i = 1, \ldots, p)$, is

$$\begin{aligned} f(\mathbf{x}) &= \phi(x_1) \;\ldots\; \phi(x_p) \qquad \text{(by independence of } X_1, \ldots, X_p\text{)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \cdots \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_p^2}{2}\right) \\ &= \frac{1}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}\left(x_1^2 + \ldots + x_p^2\right)\right] \\ &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{x}\right), \end{aligned} \tag{1}$$

This is a special case of the multivariate normal distribution. (See §2.1.)

1.3 Moments of multivariate random variables

Defn. A **random matrix X** is a matrix $\mathbf{X} = (X_{ij})$, in which the elements X_{ij} are random variables.

Defn. The expected value $E[\mathbf{X}]$ of a random matrix $\mathbf{X}$ is the matrix $(E[X_{ij}])$.

Thus to take the expectation of the matrix, you simply take an expectation of each element. In particular:

Defn. The **mean** of the random vector $\mathbf{X} = (X_1, \ldots, X_p)'$ is the p-vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ given by

$$\boldsymbol{\mu} = E[\mathbf{X}] = (E(X_1), \ldots, E(X_p))'. \tag{2}$$

Prop. Suppose that $\mathbf{X}$ is a random p-vector, $\mathbf{a}$ is a q-vector and $\mathbf{B} = (b_{ij})$ is a $q \times p$ matrix. If $\mathbf{X}$ has mean $\boldsymbol{\mu}$, then

$$E[\mathbf{a} + \mathbf{B}\mathbf{X}] = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}. \tag{3}$$

Proof. The expectation of $\mathbf{a} + \mathbf{B}\boldsymbol{\mu}$ is

$$E[\mathbf{a}+\mathbf{BX}] = E\left[\begin{pmatrix} a_1 \\ \vdots \\ a_q \end{pmatrix} + \begin{pmatrix} b_{11} & \dots & b_{1p} \\ \vdots & & \vdots \\ b_{q1} & \dots & b_{qp} \end{pmatrix}\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\right] = E\left[\begin{pmatrix} a_1 + b_{11}X_1 + \ldots + b_{1p}X_p \\ \vdots \\ a_q + b_{q1}X_1 + \ldots + b_{qp}X_p \end{pmatrix}\right]$$

$$\overset{(*)}{=} \begin{pmatrix} a_1 + b_{11}\mu_1 + \ldots + b_{1p}\mu_p \\ \vdots \\ a_q + b_{q1}\mu_1 + \ldots + b_{qp}\mu_p \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_q \end{pmatrix} + \begin{pmatrix} b_{11} & \dots & b_{1p} \\ \vdots & & \vdots \\ b_{q1} & \dots & b_{qp} \end{pmatrix}\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \mathbf{a} + \mathbf{B}\boldsymbol{\mu},$$

where $(*)$ follows by (2).□ More generally:

Prop. Let $\mathbf{X}$ be a random matrix, and $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ be matrices of constants. If the dimensions of these matrices are such that $\mathbf{A} + \mathbf{BXC}$ is defined, then

$$E[\mathbf{A} + \mathbf{BXC}] = \mathbf{A} + \mathbf{B}\,E[\mathbf{X}]\,\mathbf{C} \qquad \text{(See T1Q1 - i.e. Tutorial 1, Question 1).} \qquad (4)$$

Defn. If $\mathbf{X}$ is a random p-vector and $\mathbf{Y}$ is a random q-vector, the **covariance** (or cross-covariance) of $\mathbf{X}$ and $\mathbf{Y}$ is the p x q matrix given by

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = E[\{\mathbf{X} - E(\mathbf{X})\}\{\mathbf{Y} - E(\mathbf{Y})\}'].$$

Note that, in this definition, the dimensions p and q of the vectors may differ. The $(i, j)^{\text{th}}$ element of $\text{cov}(\mathbf{X}, \mathbf{Y})$ is

$$E[\{X_i - E(X_i)\}\{Y_j - E(Y_j)\}] = \text{cov}(X_i, Y_j).$$

Setting $\mathbf{X} = \mathbf{Y}$ gives the important special case of the variance matrix (a.k.a. the covariance matrix, the variance-covariance matrix or the dispersion matrix).

Defn. If the random vector $\mathbf{X}$ has mean $\boldsymbol{\mu}$, the **variance matrix** of $\mathbf{X}$ is the p x p matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma} = \text{var}(\mathbf{X}) = \text{cov}(\mathbf{X}, \mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'].$$

Multiplying out the product gives an alternative expression for $\boldsymbol{\Sigma}$:-

$$\begin{aligned} \boldsymbol{\Sigma} &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X}' - \boldsymbol{\mu}')] \\ &= E[\mathbf{XX}' - \mathbf{X}\boldsymbol{\mu}' - \boldsymbol{\mu}\mathbf{X}' + \boldsymbol{\mu}\boldsymbol{\mu}'] \\ &= E(\mathbf{XX}') - E(\mathbf{X})\boldsymbol{\mu}' - \boldsymbol{\mu}E(\mathbf{X}') + \boldsymbol{\mu}\boldsymbol{\mu}' && \text{by T1Q1} \\ &= E(\mathbf{XX}') - \boldsymbol{\mu}\boldsymbol{\mu}' - \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}' \quad = \quad E(\mathbf{XX}') - \boldsymbol{\mu}\boldsymbol{\mu}'. && (5) \end{aligned}$$

If $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$, this could alternatively have been shown by noting that the $(i, j)^{\text{th}}$ element of $\boldsymbol{\Sigma} = \text{var}(\mathbf{X})$ is

$$E[(X_i - \mu_i)(X_j - \mu_j)] = \text{cov}(X_i, X_j) = \sigma_{ij}, \text{ say,}$$

and, by a standard result,

$$\text{cov}(X_i, X_j) = E(X_i X_j) - \mu_i \mu_j.$$

These terms are the $(i, j)^{\text{th}}$ elements of the matrices in (5).

Prop. If $\mathbf{X}$ is a random p-vector with variance matrix $\boldsymbol{\Sigma}$, the vector $\mathbf{a}$ is a q-vector of constants and $\mathbf{B} = (b_{ij})$ is a $q \times p$ matrix of constants, then

$$\text{var}[\mathbf{a} + \mathbf{BX}] = \mathbf{B}\,\boldsymbol{\Sigma}\,\mathbf{B}' \qquad (6)$$

Proof. By the above definition of the variance matrix, we have

$$\begin{aligned}
\mathrm{var}(\mathbf{a}+\mathbf{BX}) &= E[\{\mathbf{a}+\mathbf{BX}-(\mathbf{a}+\mathbf{B}\boldsymbol{\mu})\}\{\mathbf{a}+\mathbf{BX}-(\mathbf{a}+\mathbf{B}\boldsymbol{\mu})\}'] \\
&= E[\{\mathbf{B}(\mathbf{X}-\boldsymbol{\mu})\}\{\mathbf{B}(\mathbf{X}-\boldsymbol{\mu})\}'] \\
&= E[\mathbf{B}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\mathbf{B}'] \\
&= \mathbf{B}E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})']\mathbf{B}' \qquad \text{by (4)} \\
&= \mathbf{B}\,\boldsymbol{\Sigma}\,\mathbf{B}'.
\end{aligned}$$

We noted above that the $(i, j)^{\text{th}}$ element of $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X})$ is $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$. In the case where $j = i$, we have $\sigma_{ii} = \mathrm{var}(X_i) = \sigma_i^2$, say.

Defn. The **correlation matrix P** is given by

$$\mathbf{P} = (\rho_{ij}), \qquad \text{where } \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\,\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i\,\sigma_j}.$$

The correlation matrix can be obtained from the variance matrix using the diagonal matrix

$$\boldsymbol{\Delta} = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_p}\right)$$

that has the reciprocals of the standard deviations on the diagonal and zeroes elsewhere.

Example

In the case where $p = 2$, if we write $\rho = \mathrm{corr}(X_1, X_2) = \mathrm{corr}(X_2, X_1)$, we have

$$\boldsymbol{\Delta} = \begin{pmatrix} {\sigma_1}^{-1} & 0 \\ 0 & {\sigma_2}^{-1} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

$$\begin{aligned}
\text{Hence } \boldsymbol{\Delta\Sigma\Delta} &= \begin{pmatrix} {\sigma_1}^{-1} & 0 \\ 0 & {\sigma_2}^{-1} \end{pmatrix}\begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\begin{pmatrix} {\sigma_1}^{-1} & 0 \\ 0 & {\sigma_2}^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1 & \rho\,\sigma_2 \\ \rho\,\sigma_1 & \sigma_2 \end{pmatrix}\begin{pmatrix} {\sigma_1}^{-1} & 0 \\ 0 & {\sigma_2}^{-1} \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \mathbf{P}.
\end{aligned}$$

It is clear that it will also be true in the p dimensional case that

$$\boldsymbol{\Delta\Sigma\Delta} = \mathbf{P}. \tag{7}$$

Observe that both $\boldsymbol{\Sigma}$ and $\mathbf{P}$ are *symmetric* matrices.

Defns. The symmetric $p \times p$ matrix $\mathbf{A}$ is $\begin{cases} \textbf{positive definite} \text{ (written } \mathbf{A} > 0\text{) if the quadratic form } \mathbf{c}'\mathbf{Ac} > 0 \text{ for any} \\ \text{non-zero } p\text{-vector } \mathbf{c}; \\ \textbf{positive semi-definite} \text{ (written } \mathbf{A} \geq 0\text{) if } \mathbf{c}'\mathbf{Ac} \geq 0 \text{ for any } p\text{-vector } \mathbf{c}. \end{cases}$

For $p = 1$, the positive definite $\mathbf{A} = a_{11} > 0$ is a positive scalar number.

Prop. Variance matrices and correlation matrices are positive semi-definite.

Proof. If $\mathrm{var}(\mathbf{X}) = \boldsymbol{\Sigma}$, setting $\mathbf{a} = \mathbf{0}$ and $\mathbf{B} = \mathbf{c}'$ in (6), we have that, for any p-vector $\mathbf{c}$,

$$\mathbf{c}'\,\boldsymbol{\Sigma}\,\mathbf{c} = \mathrm{var}(\mathbf{c}'\,\mathbf{X}) \geq 0,$$

as $Y = \mathbf{c}'\,\mathbf{X}$ is a (scalar) random variable and therefore it has non-negative variance. Thus

$$\boldsymbol{\Sigma} \geq 0. \tag{8}$$

Moreover, by (7),

$$\begin{aligned}
\mathbf{c}'\mathbf{P}\,\mathbf{c} &= \mathbf{c}'\,\boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta}\,\mathbf{c} && \text{where } \boldsymbol{\Delta} \text{ is a diagonal matrix;}\\
&= \mathbf{c}'\boldsymbol{\Delta}'\,\boldsymbol{\Sigma}\,\boldsymbol{\Delta}\,\mathbf{c} && \text{since } \boldsymbol{\Delta} \text{ is a diagonal matrix;}\\
&= (\boldsymbol{\Delta}\mathbf{c})'\,\boldsymbol{\Sigma}\,(\boldsymbol{\Delta}\mathbf{c})\\
&\geq 0 && \text{as } \boldsymbol{\Sigma} \geq 0.\\
\Rightarrow \quad \mathbf{P} &\geq 0.
\end{aligned}$$

Note. We shall usually assume that $\boldsymbol{\Sigma} > 0$ (i.e. that there is no non-trivial **c** for which $\mathbf{c}'\,\mathbf{X}$ is degenerate). This assumption implies that $\boldsymbol{\Sigma}$ is invertible (T1Q5(iii)).

1.4 Summary Statistics

Recall that the j^{th} column of an $n \times p$ data matrix **D** gives the observations on the j^{th} variable for each subject. We denote the mean of the j^{th} column by

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

For the observed p-vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the **sample mean** vector is the *column* vector defined by

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = (\bar{x}_1, \ldots, \bar{x}_p)'.$$

We denote the sample variance of the j^{th} variable by

$$s_j^2 = s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

and the sample covariance of the j^{th} and k^{th} variables by

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

Note that $s_{jk} = s_{kj}$.

The variance/covariance structure of the data is given by the **sample variance matrix** (a.k.a. sample covariance matrix)

$$\mathbf{S} = (s_{jk}) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

We denote the sample correlation between j^{th} and k^{th} variables by

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}\,s_{kk}}} = \frac{s_{jk}}{s_j\,s_k}.$$

The correlation structure of the data as a whole is given by the **sample correlation matrix** R defined by

$$\mathbf{R} = (r_{jk}).$$

Observe that the relationship between **S** and **R** mirrors that between $\boldsymbol{\Sigma}$ and **P**, since, if

$$\mathbf{L} = \operatorname{diag}\left(\frac{1}{s_1}, \ldots, \frac{1}{s_p}\right),$$

then

$$\mathbf{R} = \mathbf{L}\,\mathbf{S}\,\mathbf{L}. \qquad \text{(Check)} \qquad (9)$$

Example Fisher reported the body weights $x_{1,1}, \ldots, x_{47,1}$ in kilograms and the heart weights $x_{1,2}, \ldots, x_{47,2}$ in grams of each of 47 female cats. So here we have $p = 2$ and the data matrix $\mathbf{D}$ has 47 rows and 2 columns.
The summary statistics were $\sum_{i=1}^{47} x_{i1} = 110.9$, $\sum_{i=1}^{47} x_{i2} = 432.5$,
$\sum_{i=1}^{47} x_{i1}^2 = 265.13$, $\sum_{i=1}^{47} x_{i2}^2 = 4064.71$, $\sum_{i=1}^{47} x_{i1}x_{i2} = 1029.62$.
So the sample mean is

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} \frac{110.9}{47} \\ \frac{432.5}{47} \end{pmatrix} = \begin{pmatrix} 2.36 \\ 9.20 \end{pmatrix}$$

Also we have

$$s_1^2 = s_{11} = \frac{1}{46}\left[265.13 - \frac{110.9^2}{47}\right] = 0.075.$$

Similarly $s_2^2 = s_{22} = 1.843$. The sample covariance of the two variables is

$$s_{12} = \frac{1}{46}\left[1029.62 - \frac{110.9 \text{ x } 432.5}{47}\right] = 0.198.$$

Thus the sample variance matrix and the sample correlation matrix are

$$\mathbf{S} = \begin{pmatrix} 0.075 & 0.198 \\ 0.198 & 1.843 \end{pmatrix} \qquad \text{and} \qquad \mathbf{R} = \begin{pmatrix} 1 & 0.53 \\ 0.53 & 1 \end{pmatrix}.$$

If $\mathbf{c}$ and $\mathbf{d}$ are p-vectors of constants, the sample mean of the (univariate) random sample $\mathbf{c}'\mathbf{x}_1, \ldots \mathbf{c}'\mathbf{x}_n$ is $\mathbf{c}'\bar{\mathbf{x}}$ and the sample covariance of $\mathbf{c}'\mathbf{x}_1, \ldots \mathbf{c}'\mathbf{x}_n$ and $\mathbf{d}'\mathbf{x}_1, \ldots \mathbf{d}'\mathbf{x}_n$ is

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{c}'\mathbf{x}_i - \mathbf{c}'\bar{\mathbf{x}})(\mathbf{d}'\mathbf{x}_i - \mathbf{d}'\bar{\mathbf{x}}) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{c}'(\mathbf{x}_i - \bar{\mathbf{x}})\left[\mathbf{d}'(\mathbf{x}_i - \bar{\mathbf{x}})\right] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{c}'(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{d} = \mathbf{c}'\mathbf{S}\mathbf{d} \qquad (10) \end{aligned}$$

In particular, the sample variance of $\mathbf{c}'\mathbf{x}_1, \ldots \mathbf{c}'\mathbf{x}_n$ is $\mathbf{c}'\mathbf{S}\mathbf{c}$. (11)

From its definition, the sample variance of $\mathbf{c}'\mathbf{x}_1, \ldots \mathbf{c}'\mathbf{x}_n$ is non-negative, so it follows that $\mathbf{c}'\mathbf{S}\mathbf{c} \geq 0$. Thus the sample variance matrix $\mathbf{S}$ is positive semi-definite.

References for Chapter 1
M: §1.3 MV distribs., Popn. moments.
MKB: §1.4 Summary stats.; §2.2.1, §2.2.2 Popn. moments.
JW: §1.3 Summary stats.; §2.5, §2.6 Popn. moments.

---

2. The Multivariate Normal Distribution

2.1 Definition

In this module we consider only *continuous* multivariate probability distributions (for the discrete case, see log-linear models for categorical data in the GLMs and Data Analysis module.) In fact, for most of these lectures we only consider the Multivariate Normal (MVN) probability distribution. Other important multivariate probability distributions include the Wishart distribution for covariance matrices, the Multivariate Gamma and Student's t distributions, the von Mises distribution for spherical observations (see Mardia and Jupp book "Directional statistics") and many others.

The MVN probability distribution plays a central role in multivariate statistics for the following reasons:

(i) The Central Limit Theorem holds for multivariate distributions. As the distribution of the sample average of independent and identically distributed (i.i.d) random vectors tends to the MVN distribution, as the sample size increases, situations frequently arise in which MVN is an adequate approximation.

(ii) It is analytically tractable. Being a straightforward generalisation of the univariate Normal distribution, it allows for functions such as the likelihood function, the Fisher information, or distributions of test statistics, to have analytical expressions and be fairly easily computed.

(iii) Marginal and conditional distributions, and linear combinations of MVN random vectors are all MVN.

(iv) If $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ is MVN, then $\mathbf{X}_1, \mathbf{X}_2$ are uncorrelated if and only if $\mathbf{X}_1, \mathbf{X}_2$ are independent. [See (20) below.]

(v) It makes parsimonious use of parameters, being completely specified by the mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$. More precisely, the number of parameters required to specify the components of $\boldsymbol{\mu}$ is p. As $\boldsymbol{\Sigma}$ is symmetric, the number of parameters required to specify its elements is

$$1 + 2 + \ldots + p = \tfrac{1}{2}p(p+1).$$

$$\Rightarrow \text{ The total number of parameters is } p + \frac{p(p+1)}{2} = \frac{p(p+3)}{2}.$$

Now suppose that $X_1, \ldots, X_p$ are independent random variables such that $X_i \sim N(\mu_i, \sigma_i^2)$. The p.d.f. of $\mathbf{X} = (X_1, \ldots, X_p)'$ for $\mathbf{x} = (x_1, \ldots, x_n)'$ where $-\infty < x_i < \infty$ $(i = 1, \ldots, p)$ is

$$f(\mathbf{x}) = \prod_{i=1}^{p} \left[\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ \frac{-(x_i - \mu_i)^2}{2\sigma_i^2} \right\} \right] = \frac{1}{\sigma_1 \cdots \sigma_p \, (2\pi)^{p/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{p} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right]. \tag{12}$$

Now, if $\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1{}^2, \ldots, \sigma_p{}^2\right)$, we have $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}\left(\sigma_1{}^{-2}, \ldots, \sigma_p{}^{-2}\right)$ and

$2\pi\boldsymbol{\Sigma} = \mathrm{diag}\left(2\pi\sigma_1{}^2, \ldots, 2\pi\sigma_p{}^2\right)$, $\Rightarrow$ $|2\pi\boldsymbol{\Sigma}| = (2\pi)^p \, \sigma_1^2 \ldots \sigma_p^2$.

Now recall the standard result that, if $\mathbf{A}$ is a p x p matrix and $\mathbf{x} = (x_1, \ldots, x_p)'$, then the matrix product $\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} x_i x_j$.

Applying this result, we get that, if $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and $\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1{}^2, \ldots, \sigma_p{}^2\right)$, then

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \ldots + \frac{(x_p - \mu_p)^2}{\sigma_p^2}.$$

It follows that the p.d.f. of $\mathbf{X}$ may be written

$$f(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right] \qquad \mathbf{x} = (x_1, \ldots, x_p)', -\infty < x_i < \infty \ (i = 1, \ldots, p). \tag{13}$$

Due to our particular choice of $\boldsymbol{\Sigma}$, this is a special case of the general definition:-

Defn. The p-vector $\mathbf{X}$ has the multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (or just $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) if it has p.d.f.

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right], \quad \mathbf{x} = (x_1, \ldots, x_p)', \quad -\infty < x_i < \infty \ (i = 1, \ldots, p), \tag{14}$$

where $\boldsymbol{\Sigma}$ is a p x p symmetric, *positive definite* matrix.

Example

The case $p = 2$ gives the bivariate Normal distribution, for which, as on p. 5, we may write

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \Rightarrow |\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2(1-\rho^2)$$

$$\Rightarrow |2\pi\mathbf{\Sigma}| = (2\pi)^2\sigma_1^2\sigma_2^2(1-\rho^2) \qquad \text{and} \qquad \mathbf{\Sigma}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \dfrac{1}{\sigma_1^2} & -\dfrac{\rho}{\sigma_1\sigma_2} \\ -\dfrac{\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} \end{pmatrix}.$$

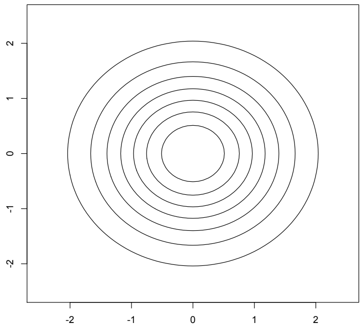
So, for $-\infty < x_i < \infty$ $(i = 1, 2)$, the bivariate Normal p.d.f. $f(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ is given by

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

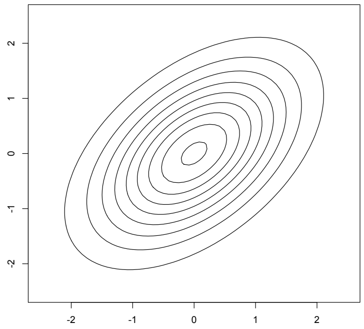
In particular, when $\boldsymbol{\mu} = \mathbf{0}, \sigma_1 = \sigma_2 = 1$, we get

$$f(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(x_1^2 - 2\rho\, x_1 x_2 + x_2^2\right)\right] \qquad -\infty < x_i < \infty \ (i = 1, 2).$$

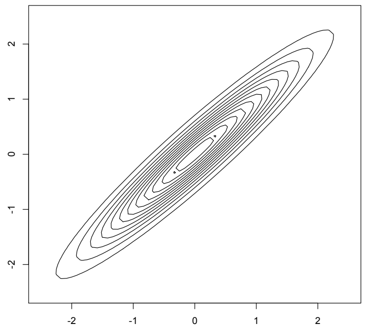
For three values of ρ, contour plots and perspective plots of this p.d.f. are shown on the next figure.



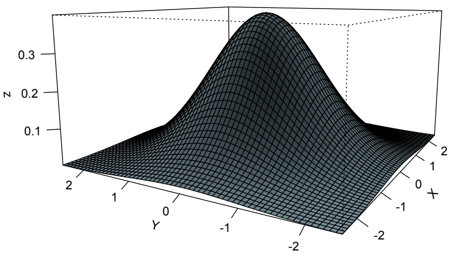
Bivariate Normal with $\rho = 0$

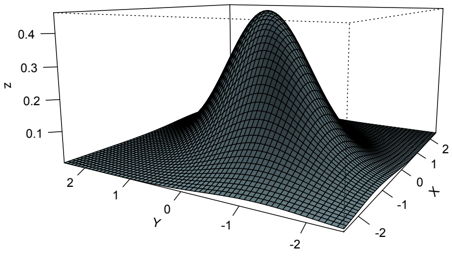


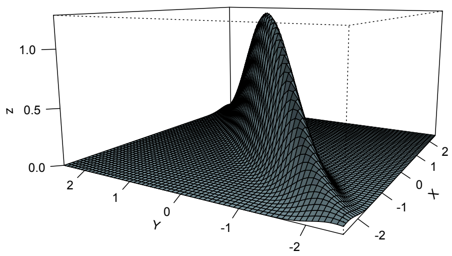
Bivariate Normal with $\rho = 0.5$



Bivariate Normal with $\rho = 0.95$







2.2 Principal axes

The above figure shows that, for $\rho \neq 0$, there are directions in $\mathbb{R}^p$ (here $p = 2$) where the variance of the MVN distribution is larger. So, it is interesting to derive those directions.

For $\rho \neq 0$, the contours of equal density of the MVN distribution are ellipses, and for $\rho = 0$ they are circles. For the general MVN p.d.f. (14), the contours of equal density have the form

$$(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c, \tag{15}$$

where c is constant. [Note that for values of $\mathbf{x}$ satisfying equation (15), the p.d.f. $f(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ is constant.] This is the equation of an ellipsoid in p dimensions. [Note that c must be positive as we are assuming that

$\boldsymbol{\Sigma}$ is positive definite, and hence, by T1Q5(iii), $\boldsymbol{\Sigma}^{-1}$ is also.] Varying c gives concentric ellipsoids centred at $\boldsymbol{\mu}$.

Defn. The **first principal axis** of the MVN distribution (14) is the line through $\boldsymbol{\mu}$ that is the major axis of any ellipsoid of constant density.

Before seeking the first principal axis, we first consider the differentiation of quadratic forms of a $p \times p$ matrix $\mathbf{A}$. The quadratic form

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j = \sum_{i=1}^{p} a_{ii}x_i^2 + \sum_{i=1}^{p}\sum_{j\neq i} a_{ij}x_i x_j.$$

Then, assuming that $\mathbf{A}$ is symmetric,

$$\frac{\partial f}{\partial x_k} = 2a_{kk}x_k + \sum_{j\neq k} a_{kj}x_j + \sum_{i\neq k} a_{ik}x_i = 2\sum_{j=1}^{p} a_{kj}x_j = 2\mathbf{a}'_{k.}\mathbf{x} \quad \text{where } a'_{k.} \text{ as in } \textbf{Notation}.$$

Thus, when $\mathbf{A}$ is symmetric, the gradient of f is given by

$$\left(\frac{\partial f}{\partial \mathbf{x}}\right) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_p}\right)' = 2(\mathbf{a}'_{1.}\mathbf{x}, \cdots, \mathbf{a}'_{p.}\mathbf{x}) = 2\mathbf{A}\mathbf{x}. \tag{16}$$

To find the first principal axis of the MVN distribution (14), we maximise the squared distance from $\boldsymbol{\mu}$,

$(\mathbf{x}-\boldsymbol{\mu})'(\mathbf{x}-\boldsymbol{\mu})$, subject to $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c$, where c is positive [as in (15)].

Equivalently, if we put $\mathbf{y} = \mathbf{x}-\boldsymbol{\mu}$, we wish to maximise

$$\mathbf{y}'\mathbf{y} \qquad \text{subject to} \qquad \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = c. \tag{17}$$

This can be done by maximising the Lagrange form $L = \mathbf{y}'\mathbf{y} - \lambda(\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - c)$, or $L = \mathbf{y}'\mathbf{y} - \lambda(\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y})$, where λ is a Lagrange multiplier. Two applications of (16) yield

$$\left(\frac{\partial L}{\partial \mathbf{y}}\right) = 2\mathbf{y} - \lambda(2\boldsymbol{\Sigma}^{-1}\mathbf{y}).$$

Equating to $\mathbf{0}$ gives

$$\mathbf{0} = \mathbf{y} - \lambda\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{18}$$

Pre-multiplying by $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma}\mathbf{y} = \lambda\mathbf{y}$$

Thus λ is an eigenvalue of $\boldsymbol{\Sigma}$, with associated eigenvector $\mathbf{y}$. Moreover, by (18), this eigenvector satisfies

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'(\lambda\boldsymbol{\Sigma}^{-1}\mathbf{y}) = \lambda(\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}) = \lambda c, \qquad \text{by (17)}$$

which is a maximum when λ is the largest eigenvalue of $\boldsymbol{\Sigma}$. Thus the first principal axis is given by the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$.

We now subsume the above definition of the first principal axis within a more general definition:-

Defn. If c is a constant, the i^{th} **principal axis** of the MVN distribution (14) is given by a vector $\pm\mathbf{v}_i$, where $\mathbf{v}_i$ is a value of $\mathbf{y} = \mathbf{x}-\boldsymbol{\mu}$ that maximises $\mathbf{y}'\mathbf{y}$ subject to $\begin{cases} \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = c \text{ when } i = 1; \\ \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = c \text{ and } \mathbf{y}'\mathbf{v}_j = 0\,(j < i) \text{ when } i \geq 2. \end{cases}$

Note. The i^{th} principal axis will not be uniquely defined where circular symmetry is present.

Suppose that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the eigenvalues of $\boldsymbol{\Sigma}$. It can be shown that the vector $\mathbf{v}_i$, giving the i^{th} principal axis, is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to λ_i.

Example

Suppose that $\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. The eigenvalues of $\mathbf{\Sigma}$ satisfy

$$0 = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1).$$

So the (ordered) eigenvalues are $\lambda_1 = 3, \quad \lambda_2 = 1.$ The eigenvector $(u_1, u_2)'$ corresponding to λ_1 satisfies

$$\mathbf{0} = (\mathbf{\Sigma} - \lambda_1 \mathbf{I}) \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \Rightarrow \quad u_2 = u_1.$$

Normalising the eigenvector, the first principal axis is given by $\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The second principal axis could be found similarly, but, as it must be orthogonal to $\mathbf{v}_1$, it can clearly be taken as $\mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

2.3 Properties of the Multivariate Normal distribution

Example

If $\mathbf{X} \sim N_2(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, what is the distribution of $\mathbf{Y} = \mathbf{A}\,\mathbf{X}$, where $\mathbf{A} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$?

From (14), the p.d.f. of $\mathbf{X}$, for $-\infty < x_i < \infty$ $(i = 1, 2)$ is

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{3}} \exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}\right), \qquad \text{where } \mathbf{\Sigma}^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Inverting the transformation, we have

$$\mathbf{X} = \mathbf{A}^{-1}\,\mathbf{Y} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \mathbf{Y} = \begin{pmatrix} (Y_1 + Y_2)/\sqrt{2} \\ (Y_1 - Y_2)/\sqrt{2} \end{pmatrix} \quad \Rightarrow \quad \left|\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right| = \begin{vmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{vmatrix} = -1.$$

$$\begin{aligned} \Rightarrow f(\mathbf{y}) &= \frac{1}{2\pi\sqrt{3}} \exp\left[-\frac{1}{12}\mathbf{y}'\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\mathbf{y}\right] \qquad -\infty < y_i < \infty \; (i = 1, 2) \\ &= \frac{1}{2\pi\sqrt{3}} \exp\left[-\frac{1}{12}\mathbf{y}'\begin{pmatrix} 1 & 1 \\ 3 & -3 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\mathbf{y}\right] \\ &= \frac{1}{2\pi\sqrt{3}} \exp\left[-\frac{1}{2}\mathbf{y}'\begin{pmatrix} 1/3 & 0 \\ 0 & 1 \end{pmatrix}\mathbf{y}\right] \qquad = \frac{1}{\sqrt{6\pi}} \exp\left[-\frac{y_1^2}{6}\right] . \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y_2^2}{2}\right]. \end{aligned}$$

Thus Y_1 and Y_2 are independent random variables such that $Y_1 \sim N(0, 3)$ and $Y_2 \sim N(0, 1)$.

Note that the previous example and the information in T1Q4 imply that, for this orthogonal matrix $\mathbf{A}$, the product $\mathbf{A}'\mathbf{\Sigma}\mathbf{A} = \text{diag}(3, 1)$.

So we could have deduced that $\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A} = \mathbf{A}^{-1}\mathbf{\Sigma}^{-1}\mathbf{A} = (\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A})^{-1} = (\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-1} = \text{diag}(1/3, 1)$.

More generally:-

Prop. Suppose that $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. If $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, where $\mathbf{a}$ is an q−vector $(q \leq p)$ and $\mathbf{B}$ is a $q \times p$ matrix of rank q, then

$$\mathbf{Y} \sim N_q(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{\Sigma}\mathbf{B}') \tag{19}$$

Proof. We consider here *only* the case where $q = p$, implying that $\mathbf{B}$ is a $p \times p$ matrix of full rank.
If $\mathbf{B}$ is of full rank, then $\mathbf{C} = (c_{ij}) = \mathbf{B}^{-1}$ exists, and we may write $\mathbf{X} = \mathbf{C}(\mathbf{Y} - \mathbf{a})$. Thus

$$x_i = \sum_{j=1}^{p} c_{ij}(y_j - a_j) \qquad \Rightarrow \qquad \frac{\partial x_i}{\partial y_k} = c_{ik} \qquad (k = 1, \ldots, p) \qquad \Rightarrow \qquad |J| = \left| \frac{\partial(x_1, \ldots, x_p)}{\partial(y_1, \ldots, y_p)} \right| = |\mathbf{C}|.$$

We have from (14), that

$$f(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \qquad \mathbf{x} \in \mathbb{R}^p.$$

$$\begin{aligned}
\Rightarrow\ f(\mathbf{y}) &= \{|\mathbf{C}|^2\}^{1/2}.|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{Cy} - \mathbf{Ca} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Cy} - \mathbf{Ca} - \boldsymbol{\mu})\right] \qquad \begin{array}{c} -\infty < y_i < \infty \\ (i = 1, \ldots, p) \end{array} \\
&= \{|\mathbf{B}|^2\}^{-1/2}.|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{C}(\mathbf{y} - \mathbf{a} - \mathbf{B}\boldsymbol{\mu})\}'\,\boldsymbol{\Sigma}^{-1}\{\mathbf{C}(\mathbf{y} - \mathbf{a} - \mathbf{B}\boldsymbol{\mu})\}\right] \\
&= |2\pi\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{a} - \mathbf{B}\boldsymbol{\mu})'(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})(\mathbf{y} - \mathbf{a} - \mathbf{B}\boldsymbol{\mu})\right]
\end{aligned}$$

Now $(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1} = \mathbf{C}^{-1}\boldsymbol{\Sigma}(\mathbf{C}')^{-1} = \mathbf{C}^{-1}\boldsymbol{\Sigma}(\mathbf{C}^{-1})' = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'$.

$$\Rightarrow\ f(\mathbf{y}) = |2\pi\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{a} - \mathbf{B}\boldsymbol{\mu})'(\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')^{-1}(\mathbf{y} - \mathbf{a} - \mathbf{B}\boldsymbol{\mu})\right].$$

for $-\infty < y_i < \infty$ $(i = 1, \ldots, p)$. So, by comparison with (14), we have $\mathbf{Y} \sim N_p(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$.

Note. The key element in this proposition is the multivariate Normality of $\mathbf{Y}$. The moments of $\mathbf{Y}$ were already known from results (3) and (6).

Cor. 1 If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for all p-vectors $\mathbf{b} \neq \mathbf{0}$, we have $\mathbf{b}'\mathbf{X} \sim N(\mathbf{b}'\boldsymbol{\mu}, \mathbf{b}'\boldsymbol{\Sigma}\mathbf{b})$. **Note.** The reverse implication is also true. An alternative definition of the MVN distribution is to say that $\mathbf{X}$ has a p-variate Normal distribution if and only if $\mathbf{b}'\mathbf{X}$ is univariate Normal for all non-zero p-vectors $\mathbf{b}$. (cf. MKB, p. 60.)

Cor. 2 The joint distribution of any set of components of $\mathbf{X}$ (such as X_1 and X_3) is MVN.

By analogy with the univariate case, it has been easy to think of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as the mean and variance matrix respectively, and we saw on p. 8 that this is so when $X_1, \ldots, X_p$ are independent. We now prove that it is true in general:-

Prop. If $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E[\mathbf{X}] = \boldsymbol{\mu}$ and $\text{var}(\mathbf{X}) = \boldsymbol{\Sigma}$.

Proof. By (19), we have that $\mathbf{X} - \boldsymbol{\mu} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Then, since $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix, the matrix $\boldsymbol{\Sigma}^{-1/2}$ is uniquely defined and symmetric and

$$\boldsymbol{\Sigma}^{-1/2}\,\boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^{-1/2} = \mathbf{I}_p.$$

Again by (19), we have that $\mathbf{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I}_p)$. So, by (14), the p.d.f. of $\mathbf{Y}$ is

$$(2\pi)^{-p/2} \exp\left(-\tfrac{1}{2}\mathbf{y}'\mathbf{y}\right) = \phi(y_1)\ \ldots\ \phi(y_p)$$

and therefore $Y_1, \ldots, Y_p$ are i.i.d. $N(0, 1)$ random variables.

Thus $E[\mathbf{Y}] = \mathbf{0}$ and $\text{var}[\mathbf{Y}] = \mathbf{I}_p$.

Hence, by (3) and (6), $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Y}$ has mean and variance matrix given by

$$E[\mathbf{X}] = \boldsymbol{\mu} \qquad \text{and} \qquad \text{var}[\mathbf{X}] = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{I}_p\,(\boldsymbol{\Sigma}^{\frac{1}{2}})' = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{I}_p\,\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{\Sigma}.$$

Prop. If $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$, then

$$\mathbf{X}_1 \text{ and } \mathbf{X}_2 \text{ are independent} \quad \Leftrightarrow \quad \boldsymbol{\Sigma}_{12} = \mathbf{0}. \tag{20}$$

Proof. See T1Q2(iv), for proof that if $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent then $\boldsymbol{\Sigma}_{12} = \text{cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$.

Conversely, if $\boldsymbol{\Sigma}_{12} = 0$, then, by T1Q2(ii), $\quad \boldsymbol{\Sigma}_{21} = \text{cov}(\mathbf{X}_2, \mathbf{X}_1) = \text{cov}(\mathbf{X}_1, \mathbf{X}_2)' = \mathbf{0}.$

Hence $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \quad \Rightarrow \quad |2\pi\boldsymbol{\Sigma}| = |2\pi\boldsymbol{\Sigma}_{11}|\,|2\pi\boldsymbol{\Sigma}_{22}|$ and $\quad \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}.$

Moreover, if $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 - \boldsymbol{\mu}_1 \\ \mathbf{X}_2 - \boldsymbol{\mu}_2 \end{pmatrix}$, then $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, by (19),

$$\text{and} \quad \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = (\mathbf{y}_1' \;\; \mathbf{y}_2')\begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \mathbf{y}_1'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}_1 + \mathbf{y}_2'\boldsymbol{\Sigma}_{22}^{-1}\mathbf{y}_2. \tag{21}$$

Now $\quad f(\mathbf{y}; \mathbf{0}, \boldsymbol{\Sigma}) \;=\; |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\right].$ So, using the expression (21) for the quadratic form, we get

$$f(\mathbf{y}; \mathbf{0}, \boldsymbol{\Sigma}) \;=\; |2\pi\boldsymbol{\Sigma}_{11}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{y}_1'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{y}_1\right] \cdot |2\pi\boldsymbol{\Sigma}_{22}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{y}_2'\boldsymbol{\Sigma}_{22}^{-1}\mathbf{y}_2\right] \quad = \quad \prod_{i=1}^{2} f(\mathbf{y}_i; \mathbf{0}, \boldsymbol{\Sigma}_{ii}).$$

Thus $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent and hence so too are $\mathbf{X}_1$ and $\mathbf{X}_2$.

The **principal components** of $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the components $\mathbf{v}_i'\mathbf{X}$ along the principal axes. For $i \neq j$, result (19) implies that the joint distribution of $\mathbf{v}_i'\mathbf{X}$ and $\mathbf{v}_j'\mathbf{X}$ is MVN. Moreover, by T1Q2(i)

$$\text{cov}(\mathbf{v}_i'\mathbf{X},\, \mathbf{v}_j'\mathbf{X}) = \mathbf{v}_i'\text{cov}(\mathbf{X},\, \mathbf{X})\mathbf{v}_j = \mathbf{v}_i'(\boldsymbol{\Sigma}\mathbf{v}_j) = \mathbf{v}_i'(\lambda_j\mathbf{v}_j) = \lambda_j(\mathbf{v}_i'\mathbf{v}_j) = 0,$$

where we have used the fact that $\mathbf{v}_j$ is an eigenvector of $\boldsymbol{\Sigma}$ with associated eigenvalue λ_j. Thus $\mathbf{v}_i'\mathbf{X}$ and $\mathbf{v}_j'\mathbf{X}$ are uncorrelated. By the last proposition they are also independent.

A better way to reach the same conclusion is to make $\mathbf{v}_i$ a *normalised* vector along the i^{th} principal axis and let $\mathbf{V}$ be a matrix with $\mathbf{v}_i$ as its i^{th} column. As the principal axes are orthogonal, so too is the matrix $\mathbf{V}$. It then follows from (19) that $\mathbf{V}'\mathbf{X} = (\mathbf{v}_1'\mathbf{X}, \ldots, \mathbf{v}_p'\mathbf{X})'$ has an MVN distribution, with mean $\mathbf{V}'\boldsymbol{\mu}$ and variance matrix

$$\mathbf{V}'\,\boldsymbol{\Sigma}\,\mathbf{V} = \boldsymbol{\Lambda}, \text{ where } \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p) \text{ and } \lambda_1, \ldots, \lambda_p \text{ are the eigenvalues of } \boldsymbol{\Sigma}.$$

Since $(\mathbf{V}'\,\boldsymbol{\Sigma}\,\mathbf{V})^{-1} = \mathbf{V}'\,\boldsymbol{\Sigma}^{-1}\,\mathbf{V} = \boldsymbol{\Lambda}^{-1}$ is also diagonal, it can then be seen that the joint density of $\mathbf{v}_1'\mathbf{X}, \ldots, \mathbf{v}_p'\mathbf{X}$ will factorise into a product of densities, showing that these random variables are independent. T1Q6 provides an example of the matrix operations in this approach.

References for Chapter 2
M: §1.4 MVN distrib., princ. axes; §1.5 MVN properties.
MKB: §2.5.1 MVN distrib.; §2.5.2 princ. axes.
JW: §4.2 MVN distrib. and properties.

3. Correlation

3.1 The matrix $\boldsymbol{\Sigma}_{11\cdot 2}$

Let $\mathbf{X}_1$ be a random p-vector and $\mathbf{X}_2$ be a random q-vector. Let $\mathbf{X}_i$ $(i = 1, 2)$ have mean $\boldsymbol{\mu}_i$ and set

$$\boldsymbol{\Sigma}_{11} = \text{var}(\mathbf{X}_1), \qquad \boldsymbol{\Sigma}_{22} = \text{var}(\mathbf{X}_2),$$

$$\boldsymbol{\Sigma}_{12} = \text{cov}(\mathbf{X}_1, \mathbf{X}_2), \qquad \boldsymbol{\Sigma}_{21} = \text{cov}(\mathbf{X}_2, \mathbf{X}_1) = \text{cov}(\mathbf{X}_1, \mathbf{X}_2)' = \boldsymbol{\Sigma}_{12}' \qquad \text{(using T1Q2(ii))}.$$

Now let $\mathbf{X}$ be the $(p+q)$-vector $(\mathbf{X}_1' \ \mathbf{X}_2')'$. Then the mean and variance matrix of $\mathbf{X}$ are

$$E(\mathbf{X}) = E\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \boldsymbol{\mu}, \text{ say,} \qquad \text{and} \qquad \text{var}(\mathbf{X}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \boldsymbol{\Sigma}, \text{ say.}$$

By T1Q5(iii), a real, symmetric, positive definite matrix is invertible, so the following definition is valid:-

Defn. If $\boldsymbol{\Sigma}_{22} > 0$, then $\quad \boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}$.COM

Examples (i) If $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ then $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11}$.

(ii) If $p = q = 1$ then, as on pp. 7 and 11CHECK, $\boldsymbol{\Sigma}$ may be written

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad \text{and so} \quad \boldsymbol{\Sigma}_{11\cdot 2} = \sigma_1^2 - \rho\sigma_1\sigma_2\left(\frac{1}{\sigma_2^2}\right)\rho\sigma_1\sigma_2 = \left(1-\rho^2\right)\sigma_1^2.$$

The matrix $\boldsymbol{\Sigma}_{11\cdot 2}$ turns out to be surprisingly important. It occurs in at least four contexts in multivariate analysis:-

(i) **Uncorrelated variables**

Prop. Suppose that $\mathbf{B}$ is a p x q matrix of constants and $\boldsymbol{\Sigma}_{22} > 0$. Then

$$(a) \quad \text{cov}(\mathbf{X}_1 - \mathbf{B}\mathbf{X}_2,\ \mathbf{X}_2) = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{B} = \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}, \quad COM \tag{22}$$

$$(b) \quad \text{var}(\mathbf{X}_1 - \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}\,\mathbf{X}_2) = \boldsymbol{\Sigma}_{11\cdot 2}. \tag{23}$$

Proof

(a) If $\boldsymbol{\Sigma}_{22}$ is positive definite then it is invertible (T1Q5(iii)). Using T1Q2 (v) and (i), we have

$$\begin{aligned} \text{cov}(\mathbf{X}_1 - \mathbf{B}\mathbf{X}_2, \mathbf{X}_2) &= \text{cov}(\mathbf{X}_1, \mathbf{X}_2) + \text{cov}\{(-\mathbf{B})\mathbf{X}_2,\ \mathbf{X}_2\} \\ &= \text{cov}(\mathbf{X}_1, \mathbf{X}_2) + (-\mathbf{B})\,\text{cov}(\mathbf{X}_2, \mathbf{X}_2) \ = \ \boldsymbol{\Sigma}_{12} - \mathbf{B}\boldsymbol{\Sigma}_{22}, \end{aligned}$$

and $\qquad \mathbf{0} = \boldsymbol{\Sigma}_{12} - \mathbf{B}\boldsymbol{\Sigma}_{22} \qquad \Leftrightarrow \ \mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$.

(b) We may write $\mathbf{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2 = \begin{pmatrix} \mathbf{I}_p & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$. Noting that $\left(\boldsymbol{\Sigma}_{22}^{-1}\right)' = \left(\boldsymbol{\Sigma}_{22}'\right)^{-1} = \boldsymbol{\Sigma}_{22}^{-1}$, we have, by (6),

$$\begin{aligned} \text{var}(\mathbf{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2) &= \begin{pmatrix} \mathbf{I}_p & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{I}_p \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{11\cdot 2} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{I}_p \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{pmatrix} = \boldsymbol{\Sigma}_{11\cdot 2}.COM \end{aligned}$$

Cor. The matrix $\boldsymbol{\Sigma}_{11\cdot 2} \geq 0$.

Proof

The proposition shows that $\boldsymbol{\Sigma}_{11\cdot 2}$ may be regarded as a variance matrix, and we know (see (8)) that such matrices are positive semi-definite.

(ii) **Inverses and determinants of partitioned matrices**

Prop. Assuming that the relevant inverses exist,

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11\cdot 2}^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\,\boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22\cdot 1}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}\,\boldsymbol{\Sigma}_{11\cdot 2}^{-1} & \boldsymbol{\Sigma}_{22\cdot 1}^{-1} \end{pmatrix}. \qquad \text{(See T2Q4(i).)}COM \tag{24}$$

We have seen that we may view $\mathbf{\Sigma}_{11\cdot 2}$, and therefore $\mathbf{\Sigma}_{22\cdot 1}$, as variance matrices. If we assume that these variance matrices, and also the variance matrices $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$, are positive definite then they are also invertible, ensuring that the condition in the proposition will be satisfied.

In fact the proposition need not be stated in terms of variance and covariance matrices, as it applies to a more general partitioned matrix, though the submatrices on the diagonal must, of course, be square if they are to be invertible.

Suppose now that the matrix $\mathbf{A}$ is partitioned in the form

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

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are matrices of dimensions p x p and q x q respectively. We note for later use that, if $\mathbf{A}_{22}$ is invertible and the definition of $\mathbf{A}_{11.2}$ mirrors that of $\mathbf{\Sigma}_{11.2}$ above, then

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{22}|\,|\mathbf{A}_{11.2}|\,. \qquad \text{(See T2Q4(ii) for the symmetric case.)} \qquad (25)$$

(iii) **Conditional variance**

Prop. If $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$,

$$\text{then (i) } \mathbf{X}_2 \sim N(\boldsymbol{\mu}_2,\ \mathbf{\Sigma}_{22}),$$
$$(ii)\ \mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2 \sim N\left(\boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\,\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2),\ \ \mathbf{\Sigma}_{11\cdot 2}\right). \qquad (26)$$

Proof

The trick here is to set up a transformation from $\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ to $\begin{pmatrix} \mathbf{Y} \\ \mathbf{X}_2 \end{pmatrix}$ where $\mathbf{Y}$ is chosen so to be independent of $\mathbf{X}_2$. It turns out that such a $\mathbf{Y}$ is easily found using the results above. That is,

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 - \mathbf{\Sigma}_{12}\,\mathbf{\Sigma}_{22}^{-1}\,\mathbf{X}_2 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & -\mathbf{\Sigma}_{12}\,\mathbf{\Sigma}_{22}^{-1}) \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}.$$

Then we have from (19) that both $\mathbf{Y}$ and $\mathbf{X}_2$ are MVN.
For (i) it only remains to compute the mean and covariance matrix of $\mathbf{X}_2$ and for this we can use the results at the start of §3.1 on p. 13. CHECK.
For (ii), first note that $\mathbf{X}_1 = \mathbf{Y} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{X}_2$ and therefore

$$(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) \sim (\mathbf{Y} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{X}_2|\mathbf{X}_2 = \mathbf{x}_2) \sim (\mathbf{Y}|\mathbf{X}_2 = \mathbf{x}_2) + (\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{X}_2|\mathbf{X}_2 = \mathbf{x}_2) \sim \mathbf{Y} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_2,$$

where the last step follows from that, by (22), $\mathbf{Y}$ and $\mathbf{X}_2$ are uncorrelated and therefore, by (20), independent. But $\mathbf{Y}$, from above, is MVN and therefore, by (19), $(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2)$ is also MVN. We can also easily compute the mean and covariance matrix of $\mathbf{Y}$

$$E(\mathbf{Y}) = E\left(\mathbf{X}_1 - \mathbf{\Sigma}_{12}\,\mathbf{\Sigma}_{22}^{-1}\,\mathbf{X}_2\right) = \boldsymbol{\mu}_1 - \mathbf{\Sigma}_{12}\,\mathbf{\Sigma}_{22}^{-1}\,\boldsymbol{\mu}_2 = \boldsymbol{\mu}_{1\cdot 2} \qquad \text{by (4)}$$

$$\text{and} \qquad \mathrm{var}(\mathbf{Y}) = \mathbf{\Sigma}_{11\cdot 2} \qquad \text{by (23).}$$

and from this deduce the mean and covariance matrix of the conditional distribution $(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2)$ in (ii).

This result shows that the conditional mean of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is a linear function of $\mathbf{x}_2$, but the conditional variance matrix is the same for all $\mathbf{x}_2$. It gives the adjustment of the distribution of $\mathbf{X}_1$ once the value of $\mathbf{X}_2$ is observed.

(iv) Linear regression

The result in (26) indicates that

$$E[\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2). \tag{27}$$

Considered as a function of $\mathbf{x}_2$, this conditional mean is called the multivariate regression of $\mathbf{X}_1$ on $\mathbf{X}_2$. It can be derived by minimisation of the mean squared error (MSE).

More precisely, if we predict $\mathbf{X}_1$ using a linear function $\mathbf{a}+\mathbf{B}\mathbf{X}_2$ of $\mathbf{X}_2$, and we choose $\mathbf{a}$ and $\mathbf{B}$ to minimise the MSE

$$S(\mathbf{a},\mathbf{B}) = E\left[\{\mathbf{X}_1 - (\mathbf{a}+\mathbf{B}\mathbf{X}_2)\}'\{\mathbf{X}_1 - (\mathbf{a}+\mathbf{B}\mathbf{X}_2)\}\right],$$

we obtain the following proposition:-

Prop.
(a) The $p \times q$ matrix $\mathbf{B}$ and the p-vector $\mathbf{a}$ minimising $S(\mathbf{a},\mathbf{B})$ are $\hat{\mathbf{B}} = \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}$ and $\hat{\mathbf{a}} = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}\,\boldsymbol{\mu}_2$.
(b) The minimum value of $S(\mathbf{a},\mathbf{B})$ is $S(\hat{\mathbf{a}},\hat{\mathbf{B}}) = \mathrm{tr}\,(\boldsymbol{\Sigma}_{11\cdot 2})$. (Proof omitted.)

Then, the **regression function** of $\mathbf{X}_1$ on $\mathbf{X}_2$

$$\hat{\mathbf{a}} + \hat{\mathbf{B}}\mathbf{X}_2 = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}\,\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \qquad \text{(cf.(27))}$$

The matrix $\hat{\mathbf{B}} = \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}$ is called the **matrix of regression coefficients** of $\mathbf{X}_1$ on $\mathbf{X}_2$.

3.2 Partial Correlation

Another (related) context in which the matrix $\boldsymbol{\Sigma}_{11\cdot 2}$ arises is in the notion of partial correlation, which looks at the correlation between some variables after the effects of others have been eliminated.

As in the last section, we let $\mathbf{X} = (\mathbf{X}_1' \;\; \mathbf{X}_2')'$, and we now assume that $\mathbf{X}_1 = (X_1, \ldots, X_p)'$ and $\mathbf{X}_2 = (X_{p+1}, \ldots, X_{p+q})'$.

Defn. For $1 \le i \le p$ and $1 \le j \le p$, the **partial correlation** of the (scalar) random variables X_i and X_j, given $\mathbf{X}_2$, is

$$\rho_{ij\cdot p+1,\ldots,p+q} = \frac{(\boldsymbol{\Sigma}_{11\cdot 2})_{ij}}{\sqrt{(\boldsymbol{\Sigma}_{11\cdot 2})_{ii}(\boldsymbol{\Sigma}_{11\cdot 2})_{jj}}}.$$

Thus X_i and X_j are two components of $\mathbf{X}_1$, and $\rho_{ij\cdot p+1,\ldots,p+q}$ is the correlation coefficient calculated as if the variance matrix of $\mathbf{X}_1$ were $\boldsymbol{\Sigma}_{11\cdot 2}$ rather than $\boldsymbol{\Sigma}_{11}$ (cf. p. 5CHECK).

In the case where $\mathbf{X}$ has a multivariate normal distribution, we have seen that the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is MVN with variance matrix $\boldsymbol{\Sigma}_{11\cdot 2}$. So $\rho_{ij\cdot p+1,\ldots,p+q}$ is the correlation between X_i and X_j conditional on $\mathbf{X}_2$. Note that, in the notation $\rho_{ij\cdot p+1,\ldots,p+q}$, the subscripts after the dot indicate the component random variables of the vector $(X_{p+1}, \ldots, X_{p+q})' = \mathbf{X}_2$ on which we are conditioning. A similar comment applies to the conditioning on $\mathbf{X}_2$ in the notation for the variance matrix $\boldsymbol{\Sigma}_{11\cdot 2}$.

Sample versions of these quantities are also useful. Similarly to the way in which we partitioned the (population) variance matrix, we may write the sample variance matrix $\mathbf{S}$ as

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

Assuming the invertibility of $\mathbf{S}_{22}$, for a sample-based equivalent of $\boldsymbol{\Sigma}_{11\cdot 2}$ we set

$$\mathbf{S}_{11\cdot 2} = \mathbf{S}_{11} - \mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}. \tag{28}$$

This is a partial sample variance matrix which allows for the effect of observing $\mathbf{X}_2$ on the sample variance of $\mathbf{X}_1$. It can also be viewed as a conditional sample variance matrix for $\mathbf{X}_1$ given $\mathbf{X}_2$. The elements of $\mathbf{S}_{11\cdot 2}$ are also used in the following sample-based equivalent of $\rho_{i\,j\,\cdot\, p+1,\ldots,p+q}$.

Defn. The **sample partial correlation** between X_i and X_j is

$$\Gamma_{ij\,\cdot\, p+1,\ldots,p+q} \quad = \quad \frac{(\mathbf{S}_{11\cdot 2})_{ij}}{\sqrt{(\mathbf{S}_{11\cdot 2})_{ii}(\mathbf{S}_{11\cdot 2})_{jj}}}.$$

Example
Measurements made of the intelligence X_1, weight X_2, and age X_3 of schoolchildren yielded the following sample correlation matrix,

$$\mathbf{R} = \begin{pmatrix} 1 & 0.6162 & 0.8267 \\ 0.6162 & 1 & 0.7321 \\ 0.8267 & 0.7321 & 1 \end{pmatrix},$$

with the $(1,2)^{\text{th}}$ element of 0.6162 suggesting some correlation between intelligence and weight. It may be verified, however, (T2Q3) that the partial correlation between intelligence and weight given age is 0.0286. So, after the effect of age has been taken into account, the correlation between intelligence and weight is close to zero.

3.3 Multiple Correlation

Now suppose that the dimension p of the first part of the partitioned vector $\mathbf{X}$ is one. We may then write this $(1+q)$-vector as $\mathbf{X} = (X_1,\ \mathbf{X}_2')'$, where X_1 is scalar, and the variance matrix becomes

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

Defn. The (**population**) **multiple correlation** (coefficient) between the random variable X_1 and the random vector $\mathbf{X}_2$ is

$$\rho_{1.2,\ldots,q+1)} = \frac{\sqrt{\mathbf{\Sigma}_{12}\,\mathbf{\Sigma}_{22}^{-1}\,\mathbf{\Sigma}_{21}}}{\sigma_1}. \tag{29}$$

It can be shown that this expression arises from regressing the random variable X_1 on the random variables $X_2,\ldots,X_{q+1}$. More precisely, the multiple correlation coefficient can be characterised as maximising the correlation between X_1 and linear functions $b_2X_2+\ldots+b_{q+1}X_{q+1}$. We have:-

Prop. If $\mathbf{b} = (b_2,\cdots,b_{q+1})'$, then $\quad \max_{\mathbf{b}}\{\text{corr}(X_1,\mathbf{b}'\mathbf{X}_2)\} = \rho_{1\cdot 2,\ldots,q+1}$. (See T2Q6)

The definition (29) for the multiple correlation coefficient can alternatively be expressed in terms of correlation matrices. To see this, partition the correlation matrix $\mathbf{P}$ in a similar manner to get

$$\mathbf{P} = \begin{pmatrix} \rho_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix}$$

and (cf. p.5CHECK) let $\Delta_2 = \text{diag}\left(\sigma_2^{-1},\ldots,\sigma_{q+1}^{-1}\right)$. Then, (cf. (7)), we have $\Delta_2\,\mathbf{\Sigma}_{22}\,\Delta_2 = \mathbf{P}_{22}$

$$\Rightarrow \quad \mathbf{\Sigma}_{22} = \Delta_2^{-1}\,\mathbf{P}_{22}\,\Delta_2^{-1} \qquad \Rightarrow \quad \mathbf{\Sigma}_{22}^{-1} = \Delta_2\,\mathbf{P}_{22}^{-1}\,\Delta_2.$$

Also $\quad \sigma_1\,\mathbf{P}_{12}\,\Delta_2^{-1} \quad = \quad \sigma_1\,(\rho_{12},\ \ldots\ ,\ \rho_{1,q+1})\begin{pmatrix} \sigma_2 & 0 & \cdots & 0 \\ 0 & \sigma_3 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \cdots & \sigma_{q+1} \end{pmatrix}$

$$= \quad \sigma_1 \, (\sigma_2 \rho_{12}, \quad \ldots \; , \quad \sigma_{q+1} \rho_{1,\,q+1})$$

$$= \quad (\sigma_{12} \, , \quad \ldots \; , \quad \sigma_{1,\,q+1} \,) \;=\; \mathbf{\Sigma}_{12}$$

$$\Rightarrow \qquad \rho_{1\cdot 2 \ldots q+1} = \sigma_1^{-1} \sqrt{\mathbf{\Sigma}_{12}\, \mathbf{\Sigma}_{22}^{-1}\, \mathbf{\Sigma}_{21}} \;=\; \sigma_1^{-1} \sqrt{\left(\sigma_1 \, \mathbf{P}_{12}\, \Delta_2^{-1}\right) \left(\Delta_2 \, \mathbf{P}_{22}^{-1}\, \Delta_2\right) \left(\Delta_2^{-1} \mathbf{P}_{21}\, \sigma_1\right)}$$

$$= \sqrt{\mathbf{P}_{12} \left(\Delta_2^{-1}\, \Delta_2\right) \mathbf{P}_{22}^{-1} \left(\Delta_2 \Delta_2^{-1}\right) \mathbf{P}_{21}}$$

$$= \sqrt{\mathbf{P}_{12}\, \mathbf{P}_{22}^{-1}\, \mathbf{P}_{21}} \tag{30}$$

As was the case with partial correlation, if we assume the invertibility of $\mathbf{S}_{22}$, there is a sample version as well as a population version:-

Defn. The (**sample**) **multiple correlation** coefficient between X_1 and $\mathbf{X}_2$ is

$$r_{1.2,\ldots,q+1} = \frac{\sqrt{\mathbf{S}_{12}\, \mathbf{S}_{22}^{-1}\, \mathbf{S}_{21}}}{s_1}.$$

For an alternative expression, we can partition the sample correlation matrix $\mathbf{R}$ in a similar way:-

$$\mathbf{R} = \begin{pmatrix} r_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}$$

Then, analogously to equation (30), the sample multiple correlation coefficient can be expressed in terms of sample correlation matrices:-

$$r_{1\cdot 2,\ldots,q+1} = \sqrt{\mathbf{R}_{12}\, \mathbf{R}_{22}^{-1}\, \mathbf{R}_{21}} \tag{31}$$

Example (MKB, p. 170.) Reconsider the example on the relationship between intelligence X_1, weight X_2, and age X_3. The sample correlation matrix $\mathbf{R}$ was given on p. 17. The blocks of the corresponding partitioned matrix are:-

$$\mathbf{R}_{12} = \mathbf{R}_{21}' = (0.6162, \;\; 0.8267). \qquad \mathbf{R}_{22} = \begin{pmatrix} 1 & 0.7321 \\ 0.7321 & 1 \end{pmatrix} \qquad \mathbf{R}_{22}^{-1} = \frac{1}{0.4640} \begin{pmatrix} 1 & -0.7321 \\ -0.7321 & 1 \end{pmatrix},$$

By equation (31), the square of the multiple correlation of intelligence on weight and age is

$$r_{1\cdot 23}^2 = \left(\frac{1}{0.4640}\right) (0.6162, \;\; 0.8267) \begin{pmatrix} 1 & -0.7321 \\ -0.7321 & 1 \end{pmatrix} \begin{pmatrix} 0.6162 \\ 0.8267 \end{pmatrix} \;=\; 0.6837$$

It follows that $r_{1\cdot 23} = 0.8269$. Observe that this multiple correlation coefficient is only very marginally greater than the simple correlation coefficient, $r_{13} = 0.8267$, between intelligence and age. Thus weight provides almost no additional assistance in explaining intelligence.

References for Chapter 3

M: Cond. MVN pp. 14-16, Partial & multiple correlation pp.17-20 (popn.), pp.29-30 (sample).
MKB: §6.5.2 Multiple correlation; §6.5.3 Partial correlation.
JW: Multiple correlation pp. 402-3; Partial correlation pp. 409-10.

---

4. Estimation

So far we have considered $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributions in which $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were both known. In practice, these parameters are, of course, usually unknown, and need to be estimated.

4.1 Unbiased estimators

A desired property is that the estimator function, $\hat{\theta}$, of parameter θ is an unbiased estimator, i.e. $E(\hat{\theta}) = \theta.$ If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. random vectors, each with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and variance matrix $\boldsymbol{\Sigma}$, we have

$$E(\bar{\mathbf{X}}) = \boldsymbol{\mu} \qquad \text{and} \qquad \operatorname{var}(\bar{\mathbf{X}}) = (1/n)\,\boldsymbol{\Sigma}. \qquad \text{(See T3Q4(a).)}$$

It can then be shown (T3Q4(a)) that the sample variance matrix has mean $E(\mathbf{S}) = \boldsymbol{\Sigma}$. Thus $\bar{\mathbf{X}}$ and $\mathbf{S}$ are unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, Observe that this conclusion is not based on any assumption of Normality, and is therefore true for other distributions.

The proof that $E(\mathbf{S}) = \boldsymbol{\Sigma}$ uses an alternative expression (cf. the univariate case) for the corrected sum of squares and products (SSP) matrix:-

$$\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' = \sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' - n(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'.$$

The data-based version of this equation is also of interest. After rearranging, it reads

$$\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' = (n-1)\mathbf{S} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'. \tag{32}$$

This can be used to re-express the exponential factor in the MVN p.d.f.. If we let $\boldsymbol{\Psi} = \boldsymbol{\Sigma}^{-1}$, we have

$$\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Psi}(\mathbf{x}_i - \boldsymbol{\mu}) = (n-1)\operatorname{tr}(\boldsymbol{\Psi}\mathbf{S}) + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Psi}(\bar{\mathbf{x}} - \boldsymbol{\mu}), \qquad \text{(See T3Q4(b)),} \tag{33}$$

that we will use next.

4.2 Some mathematical preliminaries

An often preferable approach for estimating the parameters of the MVN distribution is to use the maximum likelihood estimators (MLEs). This implies the use of calculus, and, in this context, some additional tools are required.

In discussing principal axes on p. 9CHECK, we made use of vector calculus, considering grad$f$ given by

$$\left(\frac{\partial f}{\partial \mathbf{x}}\right) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_p}\right)',$$

where $\mathbf{x}$ was a p-vector. Here we will need to use matrix calculus:-

Defn. The derivative of f with respect to an n x p matrix $\mathbf{X}$ is defined to be

$$\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}\right) = \left(\frac{\partial f(\mathbf{X})}{\partial x_{ij}}\right) \qquad \text{(a matrix in which the } (i, j)^{\text{th}} \text{ element is the partial derivative w.r.t. } x_{ij}\text{).}$$

Prop. Suppose that $\mathbf{A} = (a_{ij})$ is an n x p matrix and $\mathbf{B} = (b_{ij})$ is a p x n matrix. Let diag($\mathbf{B}$) be a diagonal matrix in which the diagonal elements are the same as those of $\mathbf{B}$. Then

$$\left(\frac{\partial \operatorname{tr}(\mathbf{AB})}{\partial \mathrm{A}}\right) = \begin{cases} \mathbf{B}' & \text{if all the elements of the } n \text{ x } p \text{ matrix } \mathbf{A} \text{ are distinct;} \\ \mathbf{B} + \mathbf{B}' - \operatorname{diag}(\mathbf{B}) & \text{if } \mathbf{A} \text{ is a } p \text{ x } p \text{ symmetric matrix.} \end{cases}$$

Proof

We have

$$\frac{\partial\,\mathrm{tr}(\mathbf{AB})}{\partial a_{hk}} = \frac{\partial}{\partial a_{hk}}\left[\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}\, b_{ji}\right] \qquad \text{[cf. the solution of T1Q3(ii)]}$$

$$= \begin{cases} b_{kh} & \text{if all the elements of } \mathbf{A} \text{ are distinct;} \\ b_{kh} + b_{hk} & \text{if } \mathbf{A} \text{ is symmetric and } h \neq k. \\ b_{hh} & \text{if } \mathbf{A} \text{ is symmetric and } h = k. \end{cases}$$

The required result follows immediately in the case where all the elements of **A** are distinct. When **A** is symmetric, we get

$$\left(\frac{\partial\,\mathrm{tr}(\mathbf{AB})}{\partial \mathbf{A}}\right) = \begin{pmatrix} b_{11} & b_{12} + b_{21} & \cdots & b_{1p} + b_{p1} \\ b_{21} + b_{12} & b_{22} & & \vdots \\ \vdots & & \ddots & b_{p-1,p} + b_{p,p-1} \\ b_{p1} + b_{1p} & \cdots & \cdots & b_{pp} \end{pmatrix} = \mathbf{B} + \mathbf{B}' - \mathrm{diag}(\mathbf{B}).$$

If the matrices **A** and **B** are both symmetric, we may conclude that

$$\left(\frac{\partial\,\mathrm{tr}(\mathbf{AB})}{\partial \mathbf{A}}\right) = 2\mathbf{B} - \mathrm{diag}(\mathbf{B}). \tag{34}$$

Suppose now that $\mathbf{A} = (a_{ij})$ is a p x p matrix in which all elements are *distinct*. If A_{ij} is the cofactor of a_{ij}, we have that, for any i $(i = 1, \ldots, p)$,

$$|\mathbf{A}| = \sum_{j=1}^{p} a_{ij} A_{ij} \qquad \text{(applying a standard result for determinants).}$$

$$\Rightarrow \quad \frac{\partial\,|\mathbf{A}|}{\partial a_{ik}} = \frac{\partial}{\partial a_{ik}}\left(\sum_{j=1}^{p} a_{ij}\, A_{ij}\right) = A_{ik}, \qquad \text{(none of the cofactors in the sum is a function of } a_{ik}\text{).} \tag{35}$$

Consider next a p x p matrix **B** in which $b_{ij} = \beta_{ij}(c_1, c_2, \ldots, c_n)$, where β_{ij} is some function and its arguments $c_1, c_2, \ldots, c_n$ are distinct. If B_{ij} is the cofactor of b_{ij}, then, by the chain rule,

$$\frac{\partial\,|\mathbf{B}|}{\partial c_k} = \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial\,|\mathbf{B}|}{\partial b_{ij}} . \frac{\partial \beta_{ij}(c_1, c_2, \ldots, c_n)}{\partial c_k} = \sum_{i=1}^{p}\sum_{j=1}^{p} B_{ij} \frac{\partial \beta_{ij}(c_1, c_2, \ldots, c_n)}{\partial c_k}, \tag{36}$$

since, if we apply the approach that led to (35), none of the relevant cofactors is a function of b_{ij}.

To find the derivative corresponding to (35) in the case where $\mathbf{A} = (a_{ij})$ is a p x p *symmetric* matrix, we apply (36), taking the arguments of β_{ij} to be the distinct elements $\{a_{ij} : i \leq j\}$ that lie on or above the diagonal of **A**. Specifically, let $b_{ij} = a_{ij}$ for $i \leq j$ and $b_{ij} = a_{ji}$ for $i > j$. Then $|\mathbf{B}| = |\mathbf{A}|$ and $\partial\beta_{ij}/\partial a_{hk}$ is only non-zero when $(i, j) = (h, k)$ or (k, h). Hence

$$\frac{\partial\,|\mathbf{A}|}{\partial a_{hk}} = \frac{\partial\,|\mathbf{B}|}{\partial a_{hk}} = \begin{cases} B_{hh} = A_{hh} & \text{when } h = k; \\ B_{hk} + B_{kh} = 2B_{kh} = 2A_{kh}. & \text{when } h \neq k \text{ (using the symmetry of } \mathbf{B}\text{).} \end{cases}$$

It follows that

$$\frac{\partial\, \log|\mathbf{A}|}{\partial a_{hk}} = \frac{1}{|\mathbf{A}|}\frac{\partial\,|\mathbf{A}|}{\partial a_{hk}} = \begin{cases} A_{hh}/|\mathbf{A}| & \text{when } h = k; \\ 2A_{kh}/|\mathbf{A}| & \text{when } h \neq k. \end{cases}$$

Thus, for a symmetric matrix $\mathbf{A}$, we have

$$\left(\frac{\partial \log|\mathbf{A}|}{\partial \mathbf{A}}\right) = 2\mathbf{A}^{-1} - \operatorname{diag}(\mathbf{A}^{-1}). \tag{37}$$

4.3 Maximum likelihood estimation

Suppose now that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. random vectors such that $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$. The joint p.d.f. of $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$ is

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n}\left[|2\pi\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}\right] \qquad \mathbf{x}_i \in \mathbb{R}^p \ (i = 1, \ldots, n).$$

Thus, if $c = (2\pi)^{-n/2}$, the likelihood of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ based on $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = c\,|\boldsymbol{\Sigma}|^{-n/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right].$$

Setting $\boldsymbol{\Psi} = \boldsymbol{\Sigma}^{-1}$ and $c' = \log(c)$, the corresponding log-likelihood is

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Psi}(\mathbf{x}_i - \boldsymbol{\mu}) + c'$$

$$= \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{(n-1)}{2}\operatorname{tr}(\boldsymbol{\Psi}\mathbf{S}) - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Psi}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + c' \qquad \text{[by T1Q5(i) and (33)].} \tag{38}$$

Thus maximising $l(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ over $\boldsymbol{\mu}$ implies minimising the quadratic form $(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Psi}(\bar{\mathbf{x}} - \boldsymbol{\mu})$, which is non-negative since $\boldsymbol{\Sigma}$, and hence $\boldsymbol{\Psi}$, is positive definite [cf. (14) and T1Q5(iii)]. So we conclude that

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}. \tag{39}$$

Differentiating (38), with $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$, with respect to $\boldsymbol{\Psi}$, and applying results (37) and (34) gives

$$\frac{\partial l}{\partial \boldsymbol{\Psi}} = \frac{n}{2}[2\boldsymbol{\Sigma} - \operatorname{diag}(\boldsymbol{\Sigma})] - \frac{(n-1)}{2}[2\mathbf{S} - \operatorname{diag}(\mathbf{S})] = 2\mathbf{M} - \operatorname{diag}(\mathbf{M}),$$

where

$$\mathbf{M} = \frac{n}{2}\boldsymbol{\Sigma} - \frac{(n-1)}{2}\mathbf{S}.$$

Now

$$\frac{\partial l}{\partial \boldsymbol{\Psi}} = \mathbf{0} \quad \Rightarrow \quad 2\hat{\mathbf{M}} - \operatorname{diag}(\hat{\mathbf{M}}) = \mathbf{0} \quad \Rightarrow \quad \hat{\mathbf{M}} = \mathbf{0}. \quad \Rightarrow \quad \hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S}. \tag{40}$$

To show that this value of $\boldsymbol{\Sigma}$ *maximises* the likelihood, first note that, as $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, we have from (38)

$$l(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{(n-1)}{2}\operatorname{tr}(\boldsymbol{\Psi}\mathbf{S}) + c'$$

$$\Rightarrow \ l(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = -\frac{n}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{(n-1)}{2}\operatorname{tr}\left[\hat{\boldsymbol{\Sigma}}^{-1}.\frac{n}{n-1}\hat{\boldsymbol{\Sigma}}\right] + c' \qquad \text{(assuming } \mathbf{S} \text{ is invertible)}$$

$$= -\frac{n}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{n}{2}\operatorname{tr}(\mathbf{I}_p) + c' \quad = \quad -\frac{n}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{np}{2} + c'. \tag{41}$$

$$\begin{aligned}
\Rightarrow l(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) - l(\hat{\boldsymbol{\mu}},\boldsymbol{\Sigma}) &= -\frac{n}{2}\left[-\log|\boldsymbol{\Sigma}| + \log|\hat{\boldsymbol{\Sigma}}|\right] - \frac{np}{2} + \frac{(n-1)}{2}\mathrm{tr}(\boldsymbol{\Psi}\mathbf{S}) \\
&= -\frac{n}{2}\left[\log|\boldsymbol{\Sigma}^{-1}| + \log|\hat{\boldsymbol{\Sigma}}|\right] - \frac{np}{2} + \frac{n}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\,\hat{\boldsymbol{\Sigma}}) && \text{as } n\hat{\boldsymbol{\Sigma}} = (n-1)\mathbf{S} \\
&= -\frac{n}{2}\log|\mathbf{B}| - \frac{np}{2} + \frac{n}{2}\mathrm{tr}(\mathbf{B}) && \text{where } \mathbf{B} = \boldsymbol{\Sigma}^{-1}\,\hat{\boldsymbol{\Sigma}}.
\end{aligned}$$

Thus if a and g are the arithmetic and geometric means respectively of the eigenvalues of $\mathbf{B}$, it follows, using T1Q3(iii), that

$$l(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) - l(\hat{\boldsymbol{\mu}},\boldsymbol{\Sigma}) = -\frac{n}{2}\log(g^p) - \frac{np}{2} + \frac{n}{2}(pa) = (a - 1 - \log g)\,\frac{np}{2}. \qquad (42)$$

It can be shown (T3Q5) that the eigenvalues of $\mathbf{B}$ are positive and that $a - 1 - \log g \geq 0$.
Hence $l(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) - l(\hat{\boldsymbol{\mu}},\boldsymbol{\Sigma}) \geq 0$ for all $\boldsymbol{\Sigma}$, and thus the stationary point of the likelihood is indeed a maximum.

Note It is clear from the results of §4.1 that $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ is an unbiased estimator of $\boldsymbol{\mu}$. It also follows that

$$\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S} = \left(1 - \frac{1}{n}\right)\mathbf{S}$$

is biased, but asymptotically unbiased.

4.4 The distributions of the maximum likelihood estimators

Here, as in the last section, we assume that $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are i.i.d. random vectors such that $\mathbf{X}_i \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ for $i = 1,\ldots,n$. If $\mathbf{X} = (\mathbf{X}_1',\ldots,\mathbf{X}_n')'$, then the np-vector $\mathbf{X}$ is MVN with its mean and variance matrix being given respectively by (cf. T2Q1)

$$\begin{pmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \boldsymbol{\Sigma} \end{pmatrix}.$$

Since we may write

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \begin{pmatrix} \mathbf{I}_p/n, & \mathbf{I}_p/n, & \cdots & \mathbf{I}_p/n \end{pmatrix}\mathbf{X},$$

it follows from result (19) that the p-vector $\bar{\mathbf{X}}$ is MVN. More precisely, either by using that proposition or from the general results in §4.1, we have

$$\bar{\mathbf{X}} \sim N_p\left(\boldsymbol{\mu},\ \frac{\boldsymbol{\Sigma}}{n}\right). \qquad (43)$$

In order to discuss the distribution of $\hat{\boldsymbol{\Sigma}}$, we need a generalisation of the chi-squared distribution:-

Defn. A random p x p positive definite symmetric matrix $\mathbf{M}$ has a **Wishart distribution** $W(\boldsymbol{\Sigma},k)$ if

$$\mathbf{M} = \mathbf{X}_1\mathbf{X}_1' + \cdots + \mathbf{X}_k\mathbf{X}_k',$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_k$ are i.i.d. $N_p(\mathbf{0},\boldsymbol{\Sigma})$ random vectors.

Example

Suppose, when $p = 1$, we have i.i.d. random variables $X_1,\ldots,X_n$ such that $X_i \sim N(0,\sigma^2)$ for $i = 1,\ldots,n$.
If we set $\sigma^2 = 1$, then the random variable $M = X_1^2 + \cdots + X_n^2 \sim \chi_n^2$.
Hence, when $p = 1$, the above definition implies that $W(1,n) \equiv \chi_n^2$. Recall also that, for $p = 1$,

$$S = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{and} \qquad \frac{(n-1)\,s^2}{\sigma^2} \sim \chi_{n-1}^2.$$

So, if we again set $\sigma^2 = 1$, we have $(n-1)\,S \sim \chi^2_{n-1} \equiv W(1, n-1)$.

Prop. If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vectors then

$$(n-1)\,\mathbf{S} \sim W(\boldsymbol{\Sigma}, n-1). \qquad \text{(Proof omitted.)}$$

Reference for Chapter 4

MKB: Appendix A.9 Matrix Differentiation; §4.4.2 MLEs; §3.4 Wishart distribution.

5. Tests on Means

5.1 Known variance matrix Σ

If we have a random sample of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$, we may wish to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ v $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. This could be approached as p tests, one on each component of $\boldsymbol{\mu}$. However, in this case, we will need to deal with the increased chance of false positive results (i.e. the *multiple-testing* problem). In multivariate analysis, we can use one multivariate test, thus obtaining reliable significance levels.

We consider first the case where the variance matrix $\boldsymbol{\Sigma}$ is known. Under the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, we have from (43)

$$\bar{\mathbf{X}} \sim N_p\left(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma}}{n}\right). \qquad \Rightarrow \qquad \sqrt{n}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \sim N_p\,(\mathbf{0},\ \boldsymbol{\Sigma}),$$

and hence, recalling (from p. 12CHECK) that $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ is well-defined and symmetric, we have from (19),

$$\begin{aligned} \mathbf{Z} = \sqrt{n}\,\boldsymbol{\Sigma}^{-\frac{1}{2}}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \ &\sim\ N_p\left(\mathbf{0},\ \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-\frac{1}{2}})'\right) \\ &\sim\ N_p\left(\mathbf{0},\ \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\right) \ \equiv\ N_p(\mathbf{0},\ \mathbf{I}_p). \end{aligned}$$

Thus, if $\mathbf{Z} = (Z_1, \cdots, Z_p)'$, the random variables $Z_1, \cdots, Z_p$ are i.i.d., each distributed as $N(0,1)$.

We will take as test statistic

$$\begin{aligned} U = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \ &=\ \{(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\,(\boldsymbol{\Sigma}^{-\frac{1}{2}})'\,\sqrt{n}\}\{\sqrt{n}\,\boldsymbol{\Sigma}^{-\frac{1}{2}}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)\} \\ &=\ \mathbf{Z}'\mathbf{Z}\ =\ Z_1^2 + \cdots + Z_p^2. \end{aligned}$$

Thus, under H_0, we have $U \sim \chi^2_p$. Now when H_0 is true, we expect U to be small. i.e. we reject H_0 for large values of U.

To find confidence regions for $\boldsymbol{\mu}$, we can exploit the standard relationship with tests, namely that the region comprises those values of the parameter that would not lead to rejection of the null hypothesis in a test at the appropriate level. For instance, when $p = 2$, we have that, under H_0,

$$0.95 = P(U < \chi^2_{2;0.05} = 5.99) = P[n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) < 5.99.]$$

So a 95% confidence region for $\boldsymbol{\mu}$ is given by $\{\boldsymbol{\mu} : n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) < 5.99\}$.

Example (See MKB.)

For a sample of 25 families, measurements X_{i1} and X_{i2} were made of the head length (in millimetres) of both the first and second sons respectively of the i^{th} family ($i = 1, \ldots, 25$). The sample mean vector was

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 185.72 \\ 183.84 \end{pmatrix}.$$

Suppose it is assumed that each son's measurement is Normally distributed with a variance of 100, and that his reading is independent of his brother's. So we have $\boldsymbol{\Sigma} = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}$.

A test is desired of the hypotheses $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{v} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0 = (182,\ 182)'$.
Testing at the 5% level, we would reject the null hypothesis if $U > \chi^2_{2;0.05} = 5.99$.
In fact, the value of the test statistic U is

$$\begin{aligned} U = 25(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) &= 25\begin{pmatrix} 3.72 & 1.84 \end{pmatrix}\begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}\begin{pmatrix} 3.72 \\ 1.84 \end{pmatrix} \\ &= \tfrac{1}{4}[3.72^2 + 1.84^2] = 4.31 \end{aligned}$$

So we accept H_0 in a 5% test.
Based on the above data, a 95% confidence region for $\boldsymbol{\mu}$ is given by the values of $\boldsymbol{\mu}$ that satisfy

$$\begin{aligned} 5.99 &> 25\begin{pmatrix} \bar{x}_1 - \mu_1, & \bar{x}_2 - \mu_2 \end{pmatrix}\boldsymbol{\Sigma}^{-1}\begin{pmatrix} \bar{x}_1 - \mu_1 \\ \bar{x}_2 - \mu_2 \end{pmatrix} \\ &= 25\begin{pmatrix} 185.72 - \mu_1, & 183.84 - \mu_2 \end{pmatrix}\begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}\begin{pmatrix} 185.72 - \mu_1 \\ 183.84 - \mu_2 \end{pmatrix} \\ &= \tfrac{1}{4}[(185.72 - \mu_1)^2 + (183.84 - \mu_2)^2]. \end{aligned}$$

Thus our assumption that the brothers' measurements are independent and have a common variance results in the confidence region being circular.

5.2 Unknown variance matrix $\boldsymbol{\Sigma}$
Two general methods of deriving hypothesis tests are the likelihood ratio method and the union-intersection method. The latter, which we use here, derives multivariate tests from the corresponding univariate ones:-

The union-intersection principle
The idea here is to see what the multivariate hypothesis H_0 says about scalar random variables $Y = \mathbf{a}'\mathbf{X}$, where $\mathbf{a} = (a_1, \ldots, a_p)' \neq \mathbf{0}$, formed from linear combinations of the components of the p-vector $\mathbf{X}$ that we observe. The multivariate hypothesis H_0 implies a corresponding univariate hypothesis $H_0(\mathbf{a})$ about the linear combinations, and H_0 is true if and only if $H_0(\mathbf{a})$ holds for *all* $\mathbf{a} \neq \mathbf{0}$. Thus we write

$$H_0 \equiv \bigcap_{\mathbf{a}\neq\mathbf{0}} H_0(\mathbf{a}).$$

Thus H_0 should be rejected if any one of the hypotheses $H_0(\mathbf{a})$ is rejected. Hence the critical region R for testing H_0 is given by

$$R \equiv \bigcup_{\mathbf{a}\neq\mathbf{0}} R(\mathbf{a}),$$

where $R(\mathbf{a})$ is the critical region for testing $H_0(\mathbf{a})$.

Returning to the MVN sample that we were considering in the last section, we now assume that the variance matrix $\boldsymbol{\Sigma}$ is unknown. Again we wish to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \ \text{v} \ H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.
Applying the above approach, for each $\mathbf{a} \neq \mathbf{0}$, we consider $\mathbf{a}'\mathbf{X} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$.

Now $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \Leftrightarrow \quad H_0(\mathbf{a}) : \mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0$ for all $\mathbf{a} \neq \mathbf{0}$. Thus $H_0 \equiv \bigcap_{\mathbf{a}\neq\mathbf{0}} H_0(\mathbf{a})$.
So, considering the random sample $\mathbf{a}'\mathbf{x}_1, \ldots, \mathbf{a}'\mathbf{x}_n$, we conduct the univariate test

$$H_0(\mathbf{a}) : \mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0 \ \text{ v } \ H_1(\mathbf{a}) : \mathbf{a}'\boldsymbol{\mu} \neq \mathbf{a}'\boldsymbol{\mu}_0.$$

We do this using the t-statistic

$$t(\mathbf{a}) = \frac{\mathbf{a}'\bar{\mathbf{X}} - \mathbf{a}'\boldsymbol{\mu}_0}{\sqrt{\dfrac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}} = \frac{\mathbf{a}'(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)\sqrt{n}}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}} \overset{H_0(\mathbf{a})}{\sim} t_{n-1}, \tag{44}$$

and reject $H_0(\mathbf{a})$ for large values of $|t(\mathbf{a})|$, or equivalently the critical region $R(\mathbf{a})$ comprises large values of $\{t(\mathbf{a})\}^2$.

By the union-intersection principle, for testing H_0 we take the critical region $R = \bigcup_{\mathbf{a}\neq\mathbf{0}} R(\mathbf{a})$. Thus we reject H_0 for large values of

$$\max_{\mathbf{a}\neq\mathbf{0}}\{t(\mathbf{a})\}^2.$$

We now derive a more explicit expression for this test statistic. Firstly, note from (44) that, if c is a positive constant, $t(c\mathbf{a}) = t(\mathbf{a})$. Thus we may assume without loss of generality that $\mathbf{a}'\mathbf{S}\mathbf{a} = 1$.

With this assumption,

$$\{t(\mathbf{a})\}^2 = n\,[\mathbf{a}'(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{a}]. \qquad (45)$$

In order to maximise $\{t(\mathbf{a})\}^2$ subject to $\mathbf{a}'\mathbf{S}\mathbf{a} = 1$, consider the Lagrange form

$$L(\mathbf{a}) = \mathbf{a}'\mathbf{B}\mathbf{a} - \lambda\,\mathbf{a}'\mathbf{S}\mathbf{a}, \qquad \text{where } \mathbf{B} = n\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'.$$

Then

$$\left(\frac{\partial L}{\partial \mathbf{a}}\right) = 2\,\mathbf{B}\,\mathbf{a} - 2\lambda\,\mathbf{S}\mathbf{a}, \qquad \text{by (16).}$$

For a value of $\mathbf{a}$ at which L is a maximum, we have

$$\mathbf{0} = \mathbf{B}\,\mathbf{a} - \lambda\,\mathbf{S}\mathbf{a} \qquad (46)$$

$$\Leftrightarrow \quad \mathbf{0} = (\mathbf{B}\,\mathbf{S}^{-1})\,\mathbf{S}\mathbf{a} - \lambda\,\mathbf{S}\mathbf{a} \qquad \text{(assuming } \mathbf{S} \text{ is invertible)}$$

so that λ is an eigenvalue of $\mathbf{B}\,\mathbf{S}^{-1}$ with associated eigenvector $\mathbf{S}\,\mathbf{a}$. Moreover, from (45) and (46),

$$\{t(\mathbf{a})\}^2 = \mathbf{a}'(\mathbf{B}\mathbf{a}) = \mathbf{a}'(\lambda\mathbf{S}\mathbf{a}) = \lambda(\mathbf{a}'\,\mathbf{S}\,\mathbf{a}) = \lambda, \qquad (47)$$

indicating that λ is positive. As we wish to maximise $\{t(\mathbf{a})\}^2$, note that we want the largest such eigenvalue. By its definition, however, the matrix $\mathbf{B} = n\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'$ has rank 1, and hence $\mathbf{B}\,\mathbf{S}^{-1}$ also has rank 1, and therefore the required λ is its <u>only</u> non-zero eigenvalue. Hence, since the trace equals the sum of the eigenvalues (T1Q3(iii)), we have

$$\begin{aligned}
\max_{\mathbf{a}\neq\mathbf{0}}\{t(\mathbf{a})\}^2 = \lambda &= \operatorname{tr}(\mathbf{B}\,\mathbf{S}^{-1}) \\
&= \operatorname{tr}\left[n\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\right] \\
&= n\operatorname{tr}\left[(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)\right] && \text{(by T1Q3(ii))} \\
&= n\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0), && \text{as the quadratic form is scalar.}
\end{aligned}$$

Defn. Hotelling's one-sample T^2 statistic is $T^2 = n\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$.

Given, as above, a random sample of size n of p dimensional MVN observations, it can be shown that, under $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, we have

$$\frac{n-p}{p(n-1)}T^2 \sim F_{p,n-p}. \qquad \text{(Proof omitted.)} \qquad (48)$$

Notes

(i) If $p = 1$, we have

$$T^2 = \frac{(\bar{X} - \mu_0)^2}{S/n} = \left[\frac{\bar{X} - \mu_0}{\sqrt{s^2/n}}\right]^2,$$

the square of the usual t-statistic, which, under H_0, is distributed as t_{n-1}. The standard characterisations confirm that the squared variable is distributed as $F_{1,n-1}$.

(ii) Hotelling's one-sample T^2 statistic can also be derived as a likelihood ratio test. Thus, in this case, the two procedures yield the same test, but that is not true in general.

A 95% confidence region for $\boldsymbol{\mu}$ can again be found from the test at the 5% level. It is given by

$$\left\{\boldsymbol{\mu} : \frac{n-p}{p(n-1)}T^2 = \frac{n(n-p)}{p(n-1)}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}}-\boldsymbol{\mu}) < F_{p,n-p,0.05}\right\}.$$

Example

For the head length data from 25 families considered in §5.1, the sample variance matrix was

$$\mathbf{S} = \begin{pmatrix} 95.293 & 69.661 \\ 69.661 & 100.81 \end{pmatrix} = \begin{pmatrix} 0.0212 & -0.0147 \\ -0.0147 & 0.0200 \end{pmatrix}^{-1}. \qquad (49)$$

Taking $\boldsymbol{\Sigma}$ as unknown, to test at the 5% level the hypotheses $H_0 : \boldsymbol{\mu} = (182,\ 182)'$ v $H_1 : \boldsymbol{\mu} \neq (182,\ 182)'$,

we would use Hotelling's one-sample T^2 statistic and reject H_0 if $\frac{23}{48}T^2 > F_{2,23,0.05} = 3.42$.

$$\begin{aligned}\text{In fact,}\quad \frac{23}{48}T^2 &= \left(\frac{23}{48}\right)[25\,(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)] \\ &= \frac{575}{48}\begin{pmatrix} 3.72 & 1.84 \end{pmatrix}\begin{pmatrix} 0.0212 & -0.0147 \\ -0.0147 & 0.0200 \end{pmatrix}\begin{pmatrix} 3.72 \\ 1.84 \end{pmatrix} \\ &= 11.98\left[0.0212(3.72)^2 + 0.0200(1.84)^2 - 2(0.0147)(3.72)(1.84)\right] \\ &= 1.91\end{aligned}$$

Hence we again accept H_0 in a 5% test.

Assuming $\boldsymbol{\Sigma}$ is unknown, a 95% elliptical confidence region for $\boldsymbol{\mu}$ is given by the values of $\boldsymbol{\mu}$ that satisfy

$$\begin{aligned} 3.42 &> \frac{575}{48}\begin{pmatrix} \bar{x}_1-\mu_1, & \bar{x}_2-\mu_2 \end{pmatrix}\mathbf{S}^{-1}\begin{pmatrix} \bar{x}_1-\mu_1 \\ \bar{x}_2-\mu_2 \end{pmatrix} \\ &= 11.98\begin{pmatrix} 185.72-\mu_1, & 183.84-\mu_2 \end{pmatrix}\begin{pmatrix} 0.0212 & -0.0147 \\ -0.0147 & 0.0200 \end{pmatrix}\begin{pmatrix} 185.72-\mu_1 \\ 183.84-\mu_2 \end{pmatrix}.\end{aligned}$$

5.3 Testing the difference between two means

Suppose that we have n_1 readings on a random vector $\mathbf{X}_1$ and n_2 readings on a random vector $\mathbf{X}_2$, with $\mathbf{X}_1$ being independent of $\mathbf{X}_2$. We assume that $\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where the *common* variance matrix $\boldsymbol{\Sigma}$ is unknown. Suppose we wish to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 + \boldsymbol{\delta}$ against $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is some specified vector of constants. Thus the case $\boldsymbol{\delta} = \mathbf{0}$ provides a test of the equality of the two means.

Applying the union-intersection principle, for each $\mathbf{a} \neq \mathbf{0}$, we consider the random variable $\mathbf{a}'\mathbf{X}_i$.

For $i = 1$ and 2, we have $\mathbf{a}'\mathbf{X}_i \sim N(\mathbf{a}'\boldsymbol{\mu}_i, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$.

Now $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 + \boldsymbol{\delta} \quad \Leftrightarrow \quad H_0(\mathbf{a}) : \mathbf{a}'\boldsymbol{\mu}_1 = \mathbf{a}'\boldsymbol{\mu}_2 + \mathbf{a}'\boldsymbol{\delta}$ for all $\mathbf{a} \neq \mathbf{0}$. $\quad \Rightarrow \ H_0 = \bigcap_{\mathbf{a}\neq\mathbf{0}} H_0(\mathbf{a})$.

Now let $\mathbf{S}$ be the pooled sample variance matrix, given in terms of those from the two samples by

$$\mathbf{S} = \frac{(n_1-1)\,\mathbf{S}_1 + (n_2-1)\,\mathbf{S}_2}{n-2},$$

where $n = n_1 + n_2$ is the size of the pooled sample. Note that $\mathbf{S}$ is an unbiased estimator of $\boldsymbol{\Sigma}$. To test

$H_0(\mathbf{a}) : \mathbf{a}'\boldsymbol{\mu}_1 = \mathbf{a}'\boldsymbol{\mu}_2 + \mathbf{a}'\boldsymbol{\delta}$ against $H_1(\mathbf{a}) : \mathbf{a}'\boldsymbol{\mu}_1 \neq \mathbf{a}'\boldsymbol{\mu}_2 + \mathbf{a}'\boldsymbol{\delta}$, we would use the two-sample t-statistic

$$t(\mathbf{a}) = \frac{\mathbf{a}'\bar{\mathbf{X}}_1 - \mathbf{a}'\bar{\mathbf{X}}_2 - \mathbf{a}'\boldsymbol{\delta}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{a}'\mathbf{S}\mathbf{a}}} = \frac{\mathbf{a}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta})}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}\sqrt{\frac{n_1n_2}{n}} \overset{H_0(\mathbf{a})}{\sim} t_{n-2}. \qquad (50)$$

As in the one-sample case, the critical region $R(\mathbf{a})$ for testing $H_0(\mathbf{a})$ comprises large values of $|t(\mathbf{a})|$, or equivalently of $\{t(\mathbf{a})\}^2$. Again the union-intersection principle implies that the critical region for testing H_0 should be $R = \bigcup_{\mathbf{a}\neq\mathbf{0}} R(\mathbf{a})$, leading to rejection of H_0 for large values of $\max_{\mathbf{a}\neq\mathbf{0}}\{t(\mathbf{a})\}^2$. From (50), we can confirm that it is again true that, for positive c, we have $t(c\mathbf{a}) = t(\mathbf{a})$, enabling us to assume without loss of generality that $\mathbf{a}'\mathbf{S}\mathbf{a} = 1$. With this assumption,

$$\{t(\mathbf{a})\}^2 = \left(\frac{n_1n_2}{n}\right)\mathbf{a}'(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2-\boldsymbol{\delta})(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2-\boldsymbol{\delta})'\mathbf{a} = \mathbf{a}'\,\mathbf{B}\,\mathbf{a}, \quad \text{where } \mathbf{B} = \left(\frac{n_1n_2}{n}\right)(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2-\boldsymbol{\delta})(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2-\boldsymbol{\delta})'.$$

To carry out the constrained maximisation of $\{t(\mathbf{a})\}^2$, we consider the Lagrange form

$$L(\mathbf{a}) = \mathbf{a}'\,\mathbf{B}\,\mathbf{a} - \lambda\,\mathbf{a}'\,\mathbf{S}\,\mathbf{a}.$$

If we again assume the invertibility of $\mathbf{S}$, this problem is essentially the same as the one-sample case considered on p. 25CHECK. The proof given there shows not only that λ must be a positive eigenvalue of $\mathbf{B}\,\mathbf{S}^{-1}$, but also that, if $\{t(\mathbf{a})\}^2$ is to be maximised, it must be the largest such eigenvalue. The matrix $\mathbf{B}$ is not the same here as it was in the one-sample case, but it is again obvious from its definition that it has rank 1. So it again follows that $\mathbf{B}\,\mathbf{S}^{-1}$ has rank 1, and that the required λ is its unique positive eigenvalue. So, similarly to the one-sample case, since the trace equals the sum of the eigenvalues (T1Q3(iii)), we can conclude that $T^2 = \max_{\mathbf{a}\neq\mathbf{0}}\{t(\mathbf{a})\}^2$ is given by

$$\begin{aligned} T^2 = \lambda = \operatorname{tr}(\mathbf{B}\,\mathbf{S}^{-1}) &= \operatorname{tr}\left[\left(\frac{n_1n_2}{n}\right)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta})(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta})'\,\mathbf{S}^{-1}\right] \\ &= \left(\frac{n_1n_2}{n}\right)\operatorname{tr}\left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta})'\,\mathbf{S}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta})\right] \qquad \text{(by T1Q3(ii))} \\ &= \left(\frac{n_1n_2}{n}\right)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta})'\,\mathbf{S}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}), \qquad \text{as the quadratic form is scalar.} \end{aligned}$$

It can be shown that, under $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 + \boldsymbol{\delta}$, we have

$$\frac{n-p-1}{p(n-2)}T^2 \sim F_{p,\,n-p-1}. \qquad \text{(Proof omitted.)}$$

Defn. In the case $\boldsymbol{\delta} = \mathbf{0}$, where we are testing the equality of the means, the above test statistic is called **Hotelling's two-sample T^2 statistic**. So this is given by

$$\left(\frac{n_1n_2}{n}\right)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\,\mathbf{S}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2).$$

Similarly to the one sample case, a 95% confidence region for $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ can be found from the test at the 5% level. It is given by

$$\left\{\boldsymbol{\delta} : \frac{n-p-1}{p(n-2)}T^2 = \frac{n-p-1}{p(n-2)}\left(\frac{n_1n_2}{n}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta})'\,\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}) < F_{p,n-p-1,0.05}\right\}.$$

Example (JW)
In Wisconsin, electricity usage was measured for a first sample of $n_1 = 45$ homeowners with air conditioning and for a second sample of $n_2 = 55$ homeowners who did not have air conditioning. During

July, two measurements were taken at each home, the first being the total on-peak consumption (in kilowatt hours) and the second being the total off-peak consumption. The summary statistics were

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 204.4 \\ 556.6 \end{pmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 130.0 \\ 355.0 \end{pmatrix}, \quad \mathbf{S}_1 = \begin{pmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{pmatrix}, \quad \mathbf{S}_2 = \begin{pmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{pmatrix}$$

So the pooled sample variance matrix is

$$\mathbf{S} = \frac{44}{98}\mathbf{S}_1 + \frac{54}{98}\mathbf{S}_2 = \begin{pmatrix} 10963.6 & 21505.4 \\ 21505.4 & 63661.3 \end{pmatrix} \quad \Rightarrow \quad \frac{(45)(55)}{100}\mathbf{S}^{-1} = \begin{pmatrix} 0.00669 & -0.00226 \\ -0.00226 & 0.00115 \end{pmatrix}$$

To test at the 5% level the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (that the population mean vectors for the two types of homeowner are equal) against $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, we would use Hotelling's two-sample T^2 statistic and reject H_0 if $\frac{97}{196}T^2 > F_{2,97,0.05} = 3.09$.

In fact,

$$\begin{aligned} \frac{97}{196}T^2 &= \left(\frac{97}{196}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \,(24.75\,\mathbf{S}^{-1})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\ &= \frac{97}{196}\begin{pmatrix} 74.4, & 201.6 \end{pmatrix}\begin{pmatrix} 0.00669 & -0.00226 \\ -0.00226 & 0.00115 \end{pmatrix}\begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} \\ &= 0.4949\left[0.00669(74.4)^2 + 0.00115(201.6)^2 - 2(0.00226)(74.4)(201.6)\right] \\ &= 7.91 \qquad \Rightarrow \quad \text{we reject } H_0 \text{ in a 5\% test.} \end{aligned}$$

A 95% elliptical confidence region for $\boldsymbol{\delta} = (\delta_1, \delta_2)' = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is given by

$$\begin{aligned} 3.09 &> \left(\frac{97}{196}\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta})' \,(24.75\,\mathbf{S}^{-1})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\delta}) \\ &= \frac{97}{196}\begin{pmatrix} 74.4 - \delta_1, & 201.6 - \delta_2 \end{pmatrix}\begin{pmatrix} 0.00669 & -0.00226 \\ -0.00226 & 0.00115 \end{pmatrix}\begin{pmatrix} 74.4 - \delta_1 \\ 201.6 - \delta_2 \end{pmatrix}. \end{aligned}$$

References for Chapter 5

M: §2.2 Hotelling's one-sample T^2; §2.4 Hotelling's two-sample T^2.
MKB: §5.2.2 Union–intersection principle; §5.2.2b Hotelling's one-sample T^2.
JW: §6.3 Two-sample T^2.

6. Tests on variance matrices

6.1 Likelihood ratio tests

In the last chapter we used the union-intersection principle to derive multivariate tests. Assuming that $\mathbf{X}$ has a distribution depending on $\boldsymbol{\theta}$, procedures based on likelihood ratios provide another approach to testing $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ v $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, where one or both the hypotheses is composite. Assuming that the relevant maxima exist, the (generalised) **likelihood ratio statistic** is

$$W(\mathbf{x}) = \frac{L(\hat{\boldsymbol{\theta}}; \mathbf{x})}{L(\tilde{\boldsymbol{\theta}}; \mathbf{x})},$$

where $\hat{\boldsymbol{\theta}}$ is the unrestricted m.l.e. of $\boldsymbol{\theta}$ over $\boldsymbol{\Theta} \equiv \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$; and $\tilde{\boldsymbol{\theta}}$ is the restricted m.l.e. of $\boldsymbol{\theta}$ over $\boldsymbol{\Theta}_0$ (i.e. under H_0). Thus $W(\mathbf{x})$ is a statistic (i.e. it depends only on the data and not on any parameter value) and it compares the overall maximum value of the likelihood with the maximum value when $\boldsymbol{\theta}$ is restricted to the set given by the null hypothesis. Moreover it is clear that $W(\mathbf{x}) \geq 1$, since the maximum in the denominator is taken over a subset of the set used in the numerator.

A test of H_0 v H_1 is a (generalised) **likelihood ratio test** if it has critical region $\{\mathbf{x} : W(\mathbf{x}) > k\}$, for some constant k. The distribution of $W(\mathbf{x})$ under H_0 can be intractable, but a large-sample approximation exists:-

Prop. Under regularity conditions, when the sample size is large, we have approximately that, under H_0,

$$w = 2\log[W(\mathbf{x})] = 2\,l(\hat{\boldsymbol{\theta}};\mathbf{x}) - 2\,l(\tilde{\boldsymbol{\theta}};\mathbf{x}) \sim \chi^2_d \qquad \text{where } d = \dim(\boldsymbol{\Theta}) - \dim(\boldsymbol{\Theta}_0). \qquad \text{(Proof omitted)}$$

The degrees d of freedom of the asymptotic distribution may equivalently be regarded as the number of constraints on $\boldsymbol{\Theta}$ required to define $\boldsymbol{\Theta}_0$.

6.2 Testing whether $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$

Consider a random sample of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$. We wish to test $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ v $H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$, where under the null hypothesis the elements of $\boldsymbol{\Sigma}_0$ are all specified.

In §4.3 we looked in detail at the likelihood arising from a random sample of MVN observations. Setting $\boldsymbol{\Psi} = \boldsymbol{\Sigma}^{-1}$, we saw in equation (38) that the log likelihood for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ could be written as

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{(n-1)}{2}\mathrm{tr}(\boldsymbol{\Psi}\mathbf{S}) - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Psi}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + c'.$$

We also saw that, for *any* $\boldsymbol{\Sigma}$, this log likelihood is maximised over $\boldsymbol{\mu}$ by choosing $\boldsymbol{\mu}$ to make the value of the quadratic form in this expression equal to zero. As this is true in particular for the value $\boldsymbol{\Sigma}_0$ of $\boldsymbol{\Sigma}$ specified by H_0, we can maximise $l(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ by taking $\tilde{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. In §4.3, we went on to show in equation (42) that

$$l(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) - l(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = (a - 1 - \log g)\frac{np}{2},$$

where a and g were the arithmetic and geometric means respectively of the eigenvalues of $\mathbf{B} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$. Since $\tilde{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}$ are identical, it follows that the test statistic for the likelihood ratio test is given by

$$w = 2\,l(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) - 2\,l(\tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_0) = (a - 1 - \log g)\,np,$$

where a and g are now interpreted as the arithmetic and geometric means respectively of the eigenvalues of $\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}$.

In this test we have $\boldsymbol{\Theta} = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma})\}$, which (cf. p. 8CHECK) has dimension $p + \frac{1}{2}p(p+1)$, whilst $\boldsymbol{\Theta}_0 = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)\}$, has dimension p, implying that $d = \frac{1}{2}p(p+1)$. Thus, under the null hypothesis, the approximate distribution of w is given by $w \sim \chi^2_d$ where $d = \frac{1}{2}p(p+1)$.

Example (Head Lengths)

Reconsider the data on head lengths discussed in Chapter 5. In §5.1, we tested whether the mean head length was 182 for each brother, assuming a known variance matrix, which we will now denote by $\boldsymbol{\Sigma}_0 = \mathrm{diag}(100, 100)$. To examine whether this was a valid assumption, we test $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. Noting the sample variance matrix $\mathbf{S}$ given by (49) and recalling that $n = 25$, we consider the matrix

$$\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_0^{-1}\left(\frac{24}{25}\mathbf{S}\right) = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}\begin{pmatrix} 91.481 & 66.875 \\ 66.875 & 96.778 \end{pmatrix} = \begin{pmatrix} 0.9148 & 0.6688 \\ 0.6688 & 0.9678 \end{pmatrix}.$$

The characteristic equation of this matrix is

$$0 = \begin{vmatrix} 0.9148 - \lambda & 0.6688 \\ 0.6688 & 0.9678 - \lambda \end{vmatrix} = \lambda^2 - 1.8826\lambda + 0.4381.$$

Hence the arithmetic and geometric means of the eigenvalues are, respectively,

$$a = 1.8826/2 = 0.941 \qquad \text{and} \qquad g = \sqrt{0.4381} = 0.662.$$

Thus the statistic for the likelihood ratio test takes the value

$$w = (a - 1 - \log g)\, np = \{0.941 - 1 - \log(0.662)\}(50) = 17.7.$$

Under H_0 this statistic has approximately a chi-squared distribution on 3 d.f., so we would reject H_0 at the 5% level when $w > \chi^2_{3,0.05} = 7.8$, and at the 1% level when $w > \chi^2_{3,0.01} = 11.3.$

Hence we reject H_0 at the 1% level. i.e. there is strong evidence that the assumption was not valid.

6.3 Union-intersection test of independence

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be random vectors of dimensions p and q, respectively, and suppose that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad \text{where} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Assume that we obtain n independent observations of $\mathbf{X}$, and that the sample variance matrix $\mathbf{S}$ is correspondingly partitioned to give

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

We wish to examine whether $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. Equivalently (cf. (20)) we need to test

$$H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0} \qquad \text{v} \qquad H_1 : \boldsymbol{\Sigma}_{12} \neq \mathbf{0}.$$

We will use the union-intersection principle to derive a test. For $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$, consider the random variables $U = \mathbf{a}'\mathbf{X}_1$ and $V = \mathbf{b}'\mathbf{X}_2$. The variance matrix of $(U, V)'$ is

$$\begin{pmatrix} \mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a} & \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} \\ \mathbf{b}'\boldsymbol{\Sigma}_{21}\mathbf{a} & \mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b} \end{pmatrix} \qquad \text{by (6) and T1Q2(i),}$$

and the correlation of U and V is given by

$$\rho(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{(\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a})\,(\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b})}}.$$

Now $H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0} \quad \Leftrightarrow \quad H_0(\mathbf{a}, \mathbf{b}) : \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} = 0$ for all $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$.

To test $\quad H_0(\mathbf{a}, \mathbf{b}) : \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} = 0 \qquad$ against $\quad H_1(\mathbf{a}, \mathbf{b}) : \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} \neq 0,$

we would estimate $\rho(\mathbf{a}, \mathbf{b})$ by the sample correlation coefficient

$$r(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}'\mathbf{S}_{12}\mathbf{b}}{\sqrt{(\mathbf{a}'\,\mathbf{S}_{11}\,\mathbf{a})\,(\mathbf{b}'\,\mathbf{S}_{22}\,\mathbf{b})}} \tag{51}$$

and use this as the test statistic, letting the critical region $R(\mathbf{a}, \mathbf{b})$ comprise large values of $\{r(\mathbf{a}, \mathbf{b})\}^2$. The union-intersection principle implies that the critical region for testing H_0 should be $R = \bigcup_{\mathbf{a}\neq\mathbf{0},\,\mathbf{b}\neq\mathbf{0}} R(\mathbf{a}, \mathbf{b})$, leading to rejection of H_0 for large values of $\max_{\mathbf{a}\neq\mathbf{0},\,\mathbf{b}\neq\mathbf{0}}\{r(\mathbf{a}, \mathbf{b})\}^2$.

Again a more convenient expression for the test statistic can be found. First note from (51) that $r(c\mathbf{a}, d\mathbf{b}) = r(\mathbf{a}, \mathbf{b})$ for all $c > 0$ and $d > 0$. Thus we may assume without loss of generality that

$$\mathbf{a}'\mathbf{S}_{11}\mathbf{a} = 1 = \mathbf{b}'\mathbf{S}_{22}\mathbf{b}. \tag{52}$$

With this assumption,

$$\{r(\mathbf{a}, \mathbf{b})\}^2 = (\mathbf{a}'\mathbf{S}_{12}\mathbf{b})^2\,. \tag{53}$$

To carry out the required maximisation, we first note that if $\mathbf{x}$ is a p-vector, $\mathbf{y}$ a q-vector, $\mathbf{A}$ a p x q matrix, and $f(\mathbf{x}) = \mathbf{x}'\,\mathbf{A}\,\mathbf{y}$ then

$$f(\mathbf{x}) = \sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}\, x_i\, y_j = \sum_{i=1}^{p} x_i \sum_{j=1}^{q} a_{ij}\, y_j \quad \Rightarrow \quad \frac{\partial f}{\partial x_i} = \sum_{j=1}^{q} a_{ij}\, y_j \qquad \Rightarrow \left(\frac{\partial f}{\partial \mathbf{x}}\right) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_p}\right)'$$

$$= \mathbf{Ay}. \qquad (54)$$

In order to maximise $\{r(\mathbf{a},\mathbf{b})\}^2$ subject to $\mathbf{a}'\mathbf{S}_{11}\mathbf{a} = 1 = \mathbf{b}'\mathbf{S}_{22}\mathbf{b}$, consider the Lagrange form

$$L = (\mathbf{a}'\mathbf{S}_{12}\mathbf{b})^2 - \lambda\, \mathbf{a}'\mathbf{S}_{11}\mathbf{a} - \theta\, \mathbf{b}'\mathbf{S}_{22}\mathbf{b} = (\mathbf{b}'\mathbf{S}_{21}\mathbf{a})^2 - \lambda\, \mathbf{a}'\mathbf{S}_{11}\mathbf{a} - \theta\, \mathbf{b}'\mathbf{S}_{22}\mathbf{b}.$$

$$\Rightarrow \left(\frac{\partial L}{\partial \mathbf{a}}\right) = 2\,(\mathbf{a}'\mathbf{S}_{12}\mathbf{b})\,\mathbf{S}_{12}\mathbf{b} - 2\lambda\,\mathbf{S}_{11}\mathbf{a} \qquad \text{and} \quad \left(\frac{\partial L}{\partial \mathbf{b}}\right) = 2\,(\mathbf{b}'\mathbf{S}_{21}\mathbf{a})\,\mathbf{S}_{21}\mathbf{a} - 2\theta\,\mathbf{S}_{22}\mathbf{b}.$$

Thus for values of **a** and **b** at which L is a maximum, we have

$$[\mathbf{a}'\mathbf{S}_{12}\mathbf{b}]\,\mathbf{S}_{12}\mathbf{b} = \lambda\,\mathbf{S}_{11}\,\mathbf{a} \qquad (55)$$

$$\text{and} \quad [\mathbf{b}'\mathbf{S}_{21}\mathbf{a}]\,\mathbf{S}_{21}\mathbf{a} = \theta\,\mathbf{S}_{22}\,\mathbf{b}. \qquad (56)$$

Using (52), and pre-multiplying (55) by $\mathbf{a}'$ and (56) by $\mathbf{b}'$ gives

$$\lambda = \lambda\,(\mathbf{a}'\mathbf{S}_{11}\mathbf{a}) = (\mathbf{a}'\mathbf{S}_{12}\mathbf{b})^2 = (\mathbf{b}'\mathbf{S}_{21}\mathbf{a})^2 = \theta\,(\mathbf{b}'\mathbf{S}_{22}\mathbf{b}) = \theta.$$

$$\Rightarrow \quad \mathbf{a}'\mathbf{S}_{12}\mathbf{b} = \sqrt{\lambda} = \sqrt{\theta}. \qquad (57)$$

We make the usual assumption that $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are positive-definite matrices, and therefore invertible (T1Q5(iii)). Substituting (57) in (55) and (56) gives

$$\begin{aligned} -\lambda\,\mathbf{S}_{11}\,\mathbf{a} + \sqrt{\lambda}\,\mathbf{S}_{12}\,\mathbf{b} &= \mathbf{0} \\ \sqrt{\lambda}\,\mathbf{S}_{21}\,\mathbf{a} - \lambda\,\mathbf{S}_{22}\,\mathbf{b} &= \mathbf{0}, \end{aligned} \qquad \Rightarrow \qquad \begin{pmatrix} -\lambda\,\mathbf{S}_{11} & \sqrt{\lambda}\,\mathbf{S}_{12} \\ \sqrt{\lambda}\,\mathbf{S}_{21} & -\lambda\,\mathbf{S}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

Since $\mathbf{a} \neq \mathbf{0} \neq \mathbf{b}$, we have

$$0 = \begin{vmatrix} -\lambda\,\mathbf{S}_{11} & \sqrt{\lambda}\,\mathbf{S}_{12} \\ \sqrt{\lambda}\,\mathbf{S}_{21} & -\lambda\,\mathbf{S}_{22} \end{vmatrix} = |-\lambda\,\mathbf{S}_{22}|\left|-\lambda\,\mathbf{S}_{11} + \sqrt{\lambda}\,\mathbf{S}_{12}\,(\lambda\,\mathbf{S}_{22})^{-1}\,\sqrt{\lambda}\,\mathbf{S}_{21}\right|. \qquad \text{by (25).}$$

Hence, as $|\mathbf{S}_{22}| > 0$,

$$0 = \left|-\lambda\,\mathbf{S}_{11} + \mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\right| \qquad \Rightarrow \qquad 0 = \left|\mathbf{S}_{11}^{-1}\right|\left|-\lambda\,\mathbf{S}_{11} + \mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\right|$$

$$= \left|-\lambda\,\mathbf{S}_{11}^{-1}\mathbf{S}_{11} + \mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\right|$$

$$= \left|\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21} - \lambda\,\mathbf{I}_p\right|.$$

Hence λ is an eigenvalue of the matrix $\mathbf{M}_1 = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$. Now we know, from (53) and (57), that $r(\mathbf{a},\mathbf{b})^2 = [\mathbf{a}'\mathbf{S}_{12}\mathbf{b}]^2 = \lambda$, so, since we are maximising $\{r(\mathbf{a},\mathbf{b})\}^2$, λ must be the largest eigenvalue of $\mathbf{M}_1$. Thus our test statistic is λ_1, the largest eigenvalue of $\mathbf{M}_1$, and H_0 is rejected if λ_1 is large.

It can be shown that the largest eigenvalue (or characteristic root) of $\mathbf{M}_1$ has what is called a greatest root distribution:-

Defn. The **greatest root distribution** $\Theta(p, n_1, n_2)$ is the distribution of the largest eigenvalue of the matrix $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$, where the $p \times p$ matrices $\mathbf{A} \sim W(\mathbf{I}, n_1)$ and $\mathbf{B} \sim W(\mathbf{I}, n_2)$ are independent and $n_1 \geq p$.

Specifically (MKB, p. 136) the largest eigenvalue of $\mathbf{M}_1$ has the distribution $\Theta(q, n-1-p, p)$.

Critical values of the greatest root distribution are not specified in closed form. They are, however, tabulated in M. pp. 426-442. (See also the charts on pp. 418-425.) The critical values depend on three other constants specified in terms of n, p and q - see p. 252, noting that M. uses N for the sample size.

Notes

(i) If $p = 1$, we have $\mathbf{S}_{11} = s_1^2$, and the matrix $\mathbf{M}_1$ is also scalar. Its only eigenvalue is itself, namely

$$\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21} = \frac{1}{s_1^2}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21} = r^2_{1\cdot 2,\ldots,q+1},$$

the square of the sample multiple correlation coefficient (cf. p. 18).
(ii) Suppose that λ_1, the largest eigenvalue of $\mathbf{M}_1$, has $\mathbf{u}$ as its associated eigenvector.
Let $\mathbf{M}_2 = \mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\,\mathbf{S}_{11}^{-1}$, and $\mathbf{v} = \mathbf{S}_{11}\mathbf{u}$. Then

$$\mathbf{M}_2\mathbf{v} = \mathbf{S}_{11}\,\mathbf{M}_1\mathbf{S}_{11}^{-1}\mathbf{v} = \mathbf{S}_{11}\mathbf{M}_1\mathbf{u} = \mathbf{S}_{11}\lambda_1\mathbf{u} = \lambda_1\mathbf{v}.$$

So λ_1 is also an eigenvalue of $\mathbf{M}_2$. Suppose that $\mathbf{M}_2$ has a larger eigenvalue ϕ with $\tilde{\mathbf{u}}$ as its associated eigenvector, and let $\tilde{\mathbf{v}} = \mathbf{S}_{11}^{-1}\tilde{\mathbf{u}}$. Then

$$\mathbf{M}_1\tilde{\mathbf{v}} = \mathbf{S}_{11}^{-1}\mathbf{M}_2\mathbf{S}_{11}\tilde{\mathbf{v}} = \mathbf{S}_{11}^{-1}\mathbf{M}_2\tilde{\mathbf{u}} = \mathbf{S}_{11}^{-1}\phi\tilde{\mathbf{u}} = \phi\tilde{\mathbf{v}}.$$

So ϕ, which is larger than λ_1, must be an eigenvalue of $\mathbf{M}_1$, contradicting the fact that λ_1 is the largest one. Hence λ_1 is also the largest eigenvalue of $\mathbf{M}_2$. It can similarly be shown that λ_1 is also the largest eigenvalue of the other cyclic permutations of $\mathbf{M}_1$, namely $\mathbf{M}_3 = \mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}$ and $\mathbf{M}_4 = \mathbf{S}_{21}\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}$.

(iii) Consider an invertible transformation

$$\begin{aligned}\mathbf{Y}_1 &= \mathbf{A}\mathbf{X}_1 + \mathbf{a}\\ \mathbf{Y}_2 &= \mathbf{B}\mathbf{X}_2 + \mathbf{b}.\end{aligned}$$

It may be verified that the sample variance matrix for the new variables is

$$\tilde{\mathbf{S}} = \begin{pmatrix} \mathbf{A}\mathbf{S}_{11}\mathbf{A}' & \mathbf{A}\mathbf{S}_{12}\mathbf{B}' \\ \mathbf{B}\mathbf{S}_{21}\mathbf{A}' & \mathbf{B}\mathbf{S}_{22}\mathbf{B}' \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{S}}_{11} & \tilde{\mathbf{S}}_{12} \\ \tilde{\mathbf{S}}_{21} & \tilde{\mathbf{S}}_{22} \end{pmatrix}, \text{say.}$$

The eigenvalues of $\tilde{\mathbf{M}}_1 = \tilde{\mathbf{S}}_{11}^{-1}\tilde{\mathbf{S}}_{12}\tilde{\mathbf{S}}_{22}^{-1}\tilde{\mathbf{S}}_{21}$ satisfy

$$\begin{aligned} 0 = \left|\tilde{\mathbf{M}}_1 - \lambda\,\mathbf{I}_p\right| &= \left|\left\{(\mathbf{A}')^{-1}\mathbf{S}_{11}^{-1}\mathbf{A}^{-1}\right\}\mathbf{A}\mathbf{S}_{12}\mathbf{B}'\left\{(\mathbf{B}')^{-1}\mathbf{S}_{22}^{-1}\mathbf{B}^{-1}\right\}\mathbf{B}\mathbf{S}_{21}\mathbf{A}' - \lambda\,\mathbf{I}_p\right| \\ &= \left|(\mathbf{A}')^{-1}\mathbf{M}_1\mathbf{A}' - \lambda\,\mathbf{I}_p\right| \\ &= \left|(\mathbf{A}')^{-1}\left\{\mathbf{M}_1 - \lambda\,\mathbf{I}_p\right\}\mathbf{A}'\right| \\ &= \left|(\mathbf{A}')^{-1}\right|\left|\mathbf{M}_1 - \lambda\,\mathbf{I}_p\right|\left|\mathbf{A}'\right| = \left|\mathbf{M}_1 - \lambda\,\mathbf{I}_p\right|. \end{aligned}$$

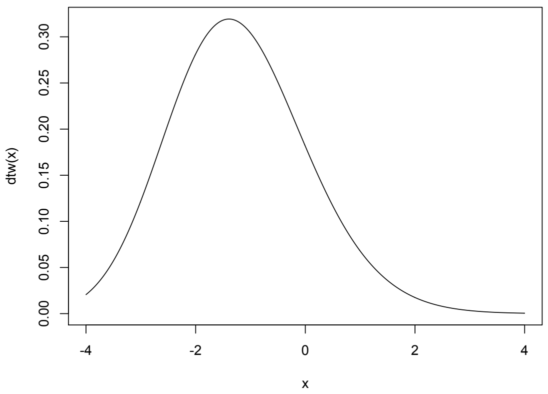
So the eigenvalues of $\mathbf{M}_1 = \mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}$ are unaltered by invertible transformations.

(iv) In particular, if we choose $\mathbf{A} = [\text{diag}(\mathbf{S}_{11})]^{-1/2}$, $\mathbf{B} = [\text{diag}(\mathbf{S}_{22})]^{-1/2}$ and $\mathbf{a} = \mathbf{b} = \mathbf{0}$, then

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \mathbf{K}\mathbf{X}, \text{ where } \mathbf{K} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} = [\text{diag}(\mathbf{S})]^{-1/2},$$

and so, in this case, [cf. (9)] $\tilde{\mathbf{S}} = \mathbf{R}$, the sample correlation matrix. Partitioning $\mathbf{R}$ similarly, we deduce that the largest eigenvalue of $\mathbf{M}_1$ is also the largest eigenvalue of $\tilde{\mathbf{M}}_1 = \mathbf{R}_{11}^{-1}\,\mathbf{R}_{12}\,\mathbf{R}_{22}^{-1}\,\mathbf{R}_{21}$.

(v) I.M.Johnstone (2009) has shown that, *when suitably transformed*, the greatest root distribution $\Theta(q,\ n-1-p,\ p)$ (whatever the values of its parameters) can be approximated by a *single* distribution, namely the **Tracy-Widom** distribution that is standardly (potentially confusingly!) denoted F_1. (The plot shows its p.d.f..) He describes this approximation as "reasonably accurate" in the sense that there is, for example, less than 10% relative error in the 95th percentile, which is calculated using $f_{0.95} = 0.9793$.



Explicitly, the approximate 95th percentile of $\Theta(q, n-1-p, p)$ is $\Theta_{0.95} = \dfrac{\exp(\mu + \sigma f_{0.95})}{1 + \exp(\mu + \sigma f_{0.95})}$, where

$$\mu = 2\log\tan\left(\frac{\phi+\gamma}{2}\right), \qquad \sigma = \left[\frac{16(n-2)^{-2}}{\sin^2(\phi+\gamma)\sin\phi\sin\gamma}\right]^{1/3},$$

$$\gamma = 2\sin^{-1}\left[\sqrt{\frac{\min(p,q)-0.5}{n-2}}\right], \qquad \phi = 2\sin^{-1}\left[\sqrt{\frac{\max(p,q)-0.5}{n-2}}\right].$$

Observe that $\gamma = \phi$ when $p = q$.

Example

Blood glucose levels were measured on 52 pregnant women. The vector $\mathbf{X}_1$ gave readings after fasting on three occasions, whilst $\mathbf{X}_2$ gave three readings one hour after sugar intake. The sample variance matrix was

$$\mathbf{S} = \left(\begin{array}{ccc|ccc} 93.751 & 18.048 & 12.653 & 60.618 & 27.618 & 52.979 \\ 18.048 & 71.465 & 14.087 & 9.241 & 61.222 & -5.875 \\ 12.653 & 14.087 & 74.575 & 50.677 & 37.560 & 65.665 \\ \hline 60.618 & 9.241 & 50.677 & 790.935 & 311.033 & 209.071 \\ 27.618 & 61.222 & 37.560 & 311.033 & 503.072 & 139.894 \\ 52.979 & -5.875 & 65.665 & 209.071 & 139.894 & 516.760 \end{array}\right).$$

$$\Rightarrow \quad \mathbf{M}_1 = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} = \begin{pmatrix} 0.0751 & -0.0275 & 0.0741 \\ -0.0405 & 0.1399 & -0.0227 \\ 0.1028 & -0.0031 & 0.1191 \end{pmatrix},$$

which has 0.2016 as its largest eigenvalue. From Chart 11 in M., the hypothesis $H_0 : \mathbf{\Sigma}_{12} = \mathbf{0}$ is rejected in a 5% test if the largest eigenvalue exceeds about 0.242. (Since, in *his* notation, $s = \min(p,q) = 3,$ $m = \frac{1}{2}[|p-q|-1] = -0.5,$ and $n = \frac{1}{2}[N-p-q-2] = 22.$) Hence we do not reject H_0.

Using the Tracy-Widom approximation, we have $\gamma = \phi = 2\sin^{-1}\left[\sqrt{2.5/50}\right] = 0.45103,$

$\mu = 2\log\tan(0.45103) = -1.45001, \quad \sigma = \left[16(50)^{-2}/\{\sin^2(0.90206)\sin^2(0.45103)\}\right]^{1/3} = 0.37964.$

This implies an approximate 5% critical value of $\{\exp(-1.07823)\}/\{1+\exp(-1.07823)\} = 0.254$.

References for Chapter 6

MKB: §5.2.1c $\mathbf{\Sigma} = \mathbf{\Sigma}_0$; §5.3.2b (p. 136) $\mathbf{\Sigma}_{12} = \mathbf{0}$.
M: §5.5 $\mathbf{\Sigma}_{12} = \mathbf{0}$.

---

7. Discriminant Analysis

7.1 Fisher's linear discriminant function

A common statistical problem is that of assigning (classifying) an object to one of k defined populations or categories. Examples include - assignment of patients to disease categories;
- assignment of stem cells to different stages of development
- assignment of a social network users to groups of different interests

If we have seen p-dimensional observations, we do the assignment using a **discriminant rule** which is a partition of $\mathbb{R}^p$ into k regions $R_1, \ldots, R_k$. We assign a new observation $\mathbf{x} \in \mathbb{R}^p$ to the ith population ($i = 1, \ldots, k$) if and only if $\mathbf{x} \in R_i$.

Suppose that, for $i = 1, \ldots, k$, we have a sample of size n_i from the ith population, given by the random p-vectors $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$. The ith sample can be summarised by its sample mean $\bar{\mathbf{x}}_i$ and its sample variance matrix $\mathbf{S}_i$. The combined sample of size $n = n_1 + \cdots + n_k$ has sample mean $\bar{\mathbf{x}}$ and sample variance $\mathbf{S}$. Then (cf. ANOVA) we make the following definitions:-

Defns.
The **between groups sum of squares and products** (SSP) is $\quad \mathbf{B} = \sum_{i=1}^{k} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$.

The **within groups SSP** is $\quad \mathbf{W} = \sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$.

The **total SSP** is $\quad \mathbf{T} = (n-1)\mathbf{S} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})'$.

It may be verified that these SSPs satisfy the ANOVA decomposition:-

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

Observe that if you take the linear combination $\mathbf{a}'\mathbf{x}_{ij}$ of each observed vector, then the within groups SS for the $\mathbf{a}'\mathbf{x}_{ij}$ is

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{a}'\mathbf{x}_{ij} - \mathbf{a}'\bar{\mathbf{x}}_i)(\mathbf{a}'\mathbf{x}_{ij} - \mathbf{a}'\bar{\mathbf{x}}_i)' \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{a}'(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'\mathbf{a}$$

$$= \quad \mathbf{a}'\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'\right]\mathbf{a} = \mathbf{a}'\,\mathbf{W}\,\mathbf{a}.$$

Similarly the between groups SS for the $\mathbf{a}'\mathbf{x}_{ij}$ is $\mathbf{a}'\,\mathbf{B}\,\mathbf{a}$.

Fisher's approach to discriminant analysis is to
(i) choose $\mathbf{a}$ to maximise the ratio

$$\frac{\text{between groups SS of } \mathbf{a}'\mathbf{X}}{\text{within groups SS of } \mathbf{a}'\mathbf{X}} = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}.$$

(ii) assign a new observation $\mathbf{x}$ on the basis of the discriminant score $\mathbf{a}'\mathbf{x}$. In particular, $\mathbf{x}$ is assigned to population i if

$$|\,\mathbf{a}'\,(\mathbf{x} - \bar{\mathbf{x}}_i)\,| \leq |\,\mathbf{a}'\,(\mathbf{x} - \bar{\mathbf{x}}_j)\,| \qquad \text{for all } j. \tag{58}$$

i.e. if $\mathbf{a}'\mathbf{x}$ is at least as close to $\mathbf{a}'\bar{\mathbf{x}}_i$ as to any other $\mathbf{a}'\bar{\mathbf{x}}_j$.

To find $\mathbf{a}$ first observe that replacing $\mathbf{a}$ by $c\mathbf{a}$, where c is positive, leaves this ratio unchanged. So we may assume that $\mathbf{a}'\mathbf{W}\mathbf{a} = 1$.
We seek therefore to maximise $\mathbf{a}'\,\mathbf{B}\,\mathbf{a}$ subject to $\mathbf{a}'\,\mathbf{W}\,\mathbf{a} = 1$. So consider the Lagrange form

$$L = \mathbf{a}'\,\mathbf{B}\,\mathbf{a} - \lambda\,\mathbf{a}'\,\mathbf{W}\,\mathbf{a}. \tag{59}$$

$$\Rightarrow \qquad \left(\frac{\partial L}{\partial \mathbf{a}}\right) = 2\,\mathbf{B}\,\mathbf{a} - 2\,\lambda\,\mathbf{W}\,\mathbf{a}.$$

Equating to $\mathbf{0}$, $\qquad \mathbf{0} = \mathbf{B}\,\mathbf{a} - \lambda\,\mathbf{W}\,\mathbf{a}. \qquad (60)$

Assuming that $\mathbf{W}$ is invertible, we obtain

$$\mathbf{0} = (\mathbf{W}^{-1}\,\mathbf{B})\,\mathbf{a} - \lambda\,\mathbf{a},$$

showing that λ is an eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ with corresponding eigenvector $\mathbf{a}$. (More precisely, as $\mathbf{W}^{-1}\mathbf{B}$ is not in general symmetric, we should say it is a right eigenvector.) Pre-multiplying (60) by $\mathbf{a}'$ gives

$$\mathbf{a}'\,\mathbf{B}\,\mathbf{a} = \mathbf{a}'\,(\lambda\mathbf{W}\,\mathbf{a}) = \lambda\,(\mathbf{a}'\mathbf{W}\,\mathbf{a}) = \lambda,$$

implying that, as $\mathbf{a}'\,\mathbf{B}\,\mathbf{a}$ is to be maximised, we seek the largest such eigenvalue λ.

For the case where $k = 2$, put $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, which we assume is non-zero. Then

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}} = \bar{\mathbf{x}}_1 - \frac{n_1\bar{\mathbf{x}}_1 + n_2\bar{\mathbf{x}}_2}{n_1 + n_2} = \frac{n_2}{n}\mathbf{d} \quad \text{and similarly} \quad \bar{\mathbf{x}}_2 - \bar{\mathbf{x}} = -\frac{n_1}{n}\mathbf{d}.$$

Then,

$$\begin{aligned} \mathbf{B} &= n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})' + n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})'. \\ &= n_1\left(\frac{n_2}{n}\right)^2 \mathbf{d}\mathbf{d}' + n_2\left(\frac{n_1}{n}\right)^2 \mathbf{d}\mathbf{d}' = \frac{n_1 n_2}{n}\mathbf{d}\mathbf{d}'. \qquad (61) \end{aligned}$$

Thus $\mathbf{B}$ has rank one, and so has one non-zero eigenvalue. The same is therefore true of $\mathbf{W}^{-1}\mathbf{B}$. Also,

$$\mathbf{W} = (n_1 - 1)\,\mathbf{S}_1 + (n_2 - 1)\,\mathbf{S}_2 = (n - 2)\mathbf{S}, \qquad (62)$$

where $\mathbf{S}$, the pooled sample variance matrix, is an unbiased estimator of $\mathbf{\Sigma}$. Assuming that $\mathbf{S}$ is positive definite, we have that $\mathbf{S}$ and $\mathbf{W}$ are invertible. Hence, from (61) and (62),

$$\mathbf{W}^{-1}\mathbf{B} = \frac{n_1 n_2}{n(n-2)}\mathbf{S}^{-1}\mathbf{d}\mathbf{d}'. \qquad (63)$$

Now the matrix product here satisfies

$$\mathbf{S}^{-1}\mathbf{d}\mathbf{d}'(\mathbf{S}^{-1}\mathbf{d}) = \mathbf{S}^{-1}\mathbf{d}(\mathbf{d}'\mathbf{S}^{-1}\mathbf{d}) = (\mathbf{d}'\mathbf{S}^{-1}\mathbf{d})\mathbf{S}^{-1}\mathbf{d},$$

so $\mathbf{S}^{-1}\mathbf{d}$ is a right eigenvector of $\mathbf{S}^{-1}\mathbf{d}\mathbf{d}'$ with eigenvalue $\mathbf{d}'\mathbf{S}^{-1}\mathbf{d} > 0$ as $\mathbf{S}^{-1} > 0$.
Explicitly

$$\mathbf{d}'\mathbf{S}^{-1}\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D^2,$$

where D^2 is called the **squared sample Mahalanobis distance** between $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$. By (63), we know that $\mathbf{S}^{-1}\mathbf{d}\mathbf{d}'$ is a multiple of $\mathbf{W}^{-1}\mathbf{B}$, so it follows that $\mathbf{S}^{-1}\mathbf{d}$ is also a right eigenvector of $\mathbf{W}^{-1}\mathbf{B}$. As it is associated with the unique non-zero eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$, (which is the largest eigenvalue as it is positive) it is the required right eigenvector $\mathbf{a}$.

Therefore assignment of the observation $\mathbf{x}$ is to be made on the basis of the discriminant score

$$\mathbf{a}'\,\mathbf{x} = (\mathbf{S}^{-1}\,\mathbf{d})'\,\mathbf{x} = \mathbf{d}'\,\mathbf{S}^{-1}\,\mathbf{x}.$$

As $\mathbf{S}^{-1} > 0$, we have $0 < \mathbf{d}'\,\mathbf{S}^{-1}\,\mathbf{d} = \mathbf{d}'\,\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \quad \Rightarrow \quad \mathbf{d}'\,\mathbf{S}^{-1}\bar{\mathbf{x}}_1 > \mathbf{d}'\,\mathbf{S}^{-1}\bar{\mathbf{x}}_2$.
Thus we assign $\mathbf{x}$ to population 1 if

$$(\mathbf{d}'\mathbf{S}^{-1})\,\mathbf{x} > (\mathbf{d}'\mathbf{S}^{-1})\left[\frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2}\right].$$

Equivalently, if $K(\mathbf{x})$ is **Fisher's linear discriminant function** given by

$$K(\mathbf{x}) = (\mathbf{d}'\mathbf{S}^{-1})\left[\mathbf{x} - \frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2}\right],$$

then **Fisher's linear discriminant rule** assigns $\mathbf{x}$ to population 1 if $K(\mathbf{x}) > 0$, and to population 2 otherwise.

Thus $\mathbf{R}^p$ is cut into 2 regions R_1 and R_2 by a plane through $\frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$ normal to $\mathbf{S}^{-1}\mathbf{d}$.

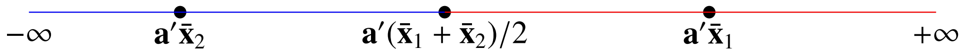

$-\infty$
$\mathbf{a}'\bar{\mathbf{x}}_2$
$\mathbf{a}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$
$\mathbf{a}'\bar{\mathbf{x}}_1$
$+\infty$

Figure 1: Diagram of the Fisher's linear discriminant function scores. Points that fall into the red region are assigned to population 1 and points that fall into the blue region are assigned to population 2.

Notes

(i) Observe that the derivation of Fisher's linear discriminant rule requires no distributional assumptions.
(ii) For the case where $k = 2$, the squared sample Mahalanobis distance D^2 is a multiple of Hotelling's two-sample T^2 statistic.

Example (MKB)

Fisher considered data on different species of iris. In particular, he had $n_1 = 50$ measurements (in centimetres) of sepal length and sepal width for *Iris setosa* and $n_2 = 50$ similar measurements for *Iris versicolor*. The summary statistics were

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 5.936 \\ 2.770 \end{pmatrix}, \quad \mathbf{S}_1 = \begin{pmatrix} 0.12425 & 0.09922 \\ 0.09922 & 0.14369 \end{pmatrix}, \quad \mathbf{S}_2 = \begin{pmatrix} 0.26643 & 0.08518 \\ 0.08518 & 0.09847 \end{pmatrix}$$

So the pooled sample variance matrix is

$$\mathbf{S} = \frac{49}{98}\mathbf{S}_1 + \frac{49}{98}\mathbf{S}_2 = \begin{pmatrix} 0.19534 & 0.09220 \\ 0.09220 & 0.12108 \end{pmatrix} \quad \Rightarrow \quad \mathbf{S}^{-1} = \begin{pmatrix} 7.9916 & -6.0854 \\ -6.0854 & 12.8929 \end{pmatrix}$$

Thus $\mathbf{d}'\,\mathbf{S}^{-1} = (-0.930,\ 0.658)\begin{pmatrix} 7.9916 & -6.0854 \\ -6.0854 & 12.8929 \end{pmatrix} = (-11.4364,\ 14.1430)\,.$

So here, if $\mathbf{x} = (x_1,\ x_2)'$, Fisher's linear discriminant function is

$$K(\mathbf{x}) = (-11.4364,\ 14.1430)\begin{pmatrix} x_1 - 5.471 \\ x_2 - 3.099 \end{pmatrix} = -11.4364x_1 + 14.1430x_2 + 18.739.$$

Fisher's linear discriminant rule classifies a new observation $\mathbf{x}$ as *Iris setosa* if and only if $K(\mathbf{x}) > 0$. For example, if $\mathbf{x} = (5.5,\ 3)'$,

$$K(\mathbf{x}) = (\mathbf{d}'S^{-1})\left[\mathbf{x} - \frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2}\right] = -11.4364(5.5) + 14.1430(3) + 18.739 = -1.732.$$

$\Rightarrow$ $\mathbf{x}$ is assigned to *Iris versicolor*.

The scaling of the eigenvector $\mathbf{a} = \mathbf{S}^{-1}\,\mathbf{d}$, is arbitrary, but one approach is to scale so that the pooled variance of the two groups of discriminant scores is one. Here, with $\mathbf{a}' = (-11.4364,\ 14.1430)$, the discriminant scores for *Iris setosa* have sample variance 12.8967 and those for *Iris versicolor* have sample variance 26.9873, giving a pooled variance of 19.942. This can be scaled to one by taking $\tilde{\mathbf{a}}' = \mathbf{a}'/\sqrt{19.942} = (-2.561,\ 3.167)$.

7.2 Using R for discriminant analysis

Linear discriminant analysis is available within the MASS package using the function `lda`. Fisher's iris data, giving 50 measurements on 4 variables for each of 3 species, is available within the base package. As above, we will select the first two species given in the first 100 rows:-

```
> iris2 <- iris[1:100,]
> iris2
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1           5.1         3.5          1.4         0.2     setosa
2           4.9         3.0          1.4         0.2     setosa
. . .

50          5.0         3.3          1.4         0.2     setosa
51          7.0         3.2          4.7         1.4 versicolor

. . .
```

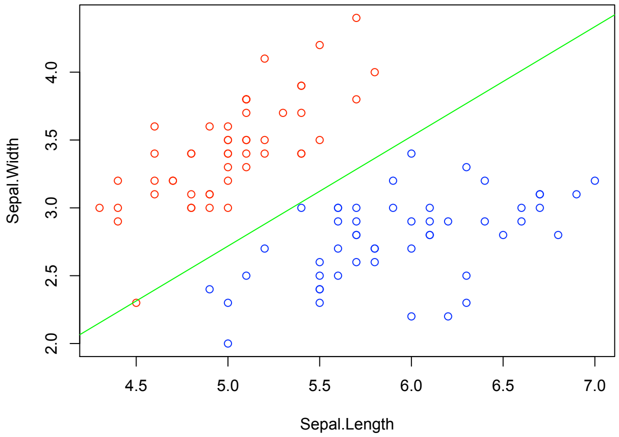
```
100          5.7          2.8          4.1          1.3 versicolor
```

The output from lda, which refers to prior probabilities, assumes the adoption of a Bayesian approach. In fact, however, the Bayesian method, allocating to the population with the highest posterior probability, can be shown to yield the same procedure if it is assumed that the populations are MVN with common variance matrix and equal prior probabilities.

```
> library(MASS)
> dis <- lda(Species ~ Sepal.Length + Sepal.Width, data = iris2)
> dis
Call:
lda(Species ~ Sepal.Length + Sepal.Width, data = iris2)

 Prior probabilities of groups:
     setosa versicolor
        0.5        0.5

 Group means:
            Sepal.Length Sepal.Width
 setosa            5.006       3.428
 versicolor        5.936       2.770

 Coefficients of linear discriminants:
                    LD1
 Sepal.Length  2.560968
 Sepal.Width  -3.167079
```

We may plot sepal width against sepal length, showing *Iris setosa* observations in red and *Iris versicolor* in blue, and the line (in green) cutting $\mathbf{R}^2$ into the 2 regions as follows:-

```
> a1<-dis$scaling[1];    a2<-dis$scaling[2]
> xbar<-mean(iris2[,1]);   ybar<-mean(iris2[,2])
> Z<-c(rep("red",50),rep("blue",50))
> plot(iris2[, 1:2], col=Z)
> slope<-  -a1/a2
> intercept<-  ybar - slope*xbar
> abline(a=intercept, b =slope, col = "green")
```



The new observation can be classified as indicated:-

```
> newobs<- data.frame(Sepal.Length=5.5,Sepal.Width=3)
> predict(dis,newobs)$class
[1] versicolor
Levels: setosa versicolor virginica
```

7.3 The maximum likelihood discriminant rule for known MVN populations

Suppose that we have k known populations, where the i^{th} population ($i = 1, \ldots, k$) has an $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ distribution. Thus the populations have a common variance matrix. The principle of maximum likelihood discriminant rules is to assign an observation $\mathbf{x}$ to the population with the largest (or larger) likelihood.

In the case of $k = 2$ known populations, discrimination becomes equivalent to a test of simple hypotheses. For this situation, the Neyman-Pearson lemma (covered in MT4606) says that the most powerful test is a likelihood ratio test.

After observing $\mathbf{x}$, the likelihood ratio in favour of population 1 equals

$$\frac{L(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}; \mathbf{x})}{L(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}; \mathbf{x})} = \frac{|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_1)\right]}{|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_2)\right]},$$

where $\Psi = \boldsymbol{\Sigma}^{-1}$. So the log-likelihood ratio

$$l(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}; \mathbf{x}) - l(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}; \mathbf{x}) = -\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_2)\right]$$

is a multiple of the difference of the **Mahalanobis distance** of $\mathbf{x}$ to $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Then, the point $\mathbf{x}$ is assigned to the closest, according to the Mahalanobis distance, population. That is, $\mathbf{x}$ is assigned to population 1 if

$$(\mathbf{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_1) < (\mathbf{x}-\boldsymbol{\mu}_2)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_2).$$

We can also easily see that

$$\begin{aligned} l(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}; \mathbf{x}) - l(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}; \mathbf{x}) &= \frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_2)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_2) - (\mathbf{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Psi} (\mathbf{x}-\boldsymbol{\mu}_1)\right] \\ &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Psi}\mathbf{x} - (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Psi}\,\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}. \end{aligned} \tag{64}$$

In general, when there are k populations, set

$$U_{ij} = l(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}; \mathbf{x}) - l(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}; \mathbf{x}) = \Delta(\mathbf{x}) - \Delta\left(\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_j}{2}\right), \qquad \text{where } \Delta(\mathbf{x}) = (\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)'\,\boldsymbol{\Sigma}^{-1}\,\mathbf{x}. \tag{65}$$

This function $\Delta(\mathbf{x})$ is called the **maximum likelihood discriminant function.** The maximum likelihood discriminant rule then assigns $\mathbf{x}$ to the i^{th} population if

$$U_{ij} \geq 0 \qquad \text{for all } j \neq i.$$

Note For $i, j = 1, \ldots, k$, we have $U_{ij} = -U_{ji}$ $(j \neq i)$ and, if i, j and k are distinct, $U_{ij} + U_{jk} = U_{ik}$.

7.4 The maximum likelihood discriminant rule when parameters are unknown

Suppose now that we have k samples, where the i^{th} sample $(i = 1, \ldots, k)$ is of size n_i and has sample mean $\bar{\mathbf{x}}_i$ and sample variance matrix $\mathbf{S}_i$. These samples are assumed to come from MVN populations with a common variance matrix $\boldsymbol{\Sigma}$, of which we can find the pooled estimate

$$\mathbf{S} = \frac{\sum_{i=1}^k (n_i-1)\mathbf{S}_i}{\sum_{i=1}^k (n_i-1)}.$$

If we take expression (65) for U_{ij} that we used when the parameters were known, and replace $\boldsymbol{\mu}_i$ by $\bar{\mathbf{x}}_i$ and $\boldsymbol{\Sigma}$ by $\mathbf{S}$, we get the **Wald-Anderson statistic**

$$W_{ij} = (\bar{\mathbf{x}}_i-\bar{\mathbf{x}}_j)'\mathbf{S}^{-1}\mathbf{x} - (\bar{\mathbf{x}}_i-\bar{\mathbf{x}}_j)'\mathbf{S}^{-1}\left(\frac{\bar{\mathbf{x}}_i+\bar{\mathbf{x}}_j}{2}\right).$$

Reversing the derivation of (64), replacing 1 and 2 by i and j respectively, and using the above substitutions, gives

$$W_{ij} = \tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_j)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_j) - \tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) = \tfrac{1}{2}\left(D_j^2 - D_i^2\right),$$

where $D_i^2 = (\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)$ is the squared sample Mahalanobis distance from $\mathbf{x}$ to $\bar{\mathbf{x}}_i$ (cf. p. 35).

The (sample) maximum likelihood rule assigns $\mathbf{x}$ to population i if $W_{ij} \geq 0$ for all $j \neq i$, or equivalently if $D_i^2 \leq D_j^2$ for all $j \neq i$. Thus it is a minimum sample Mahalanobis distance rule.

Note Comparing W_{12} with $K(\mathbf{x})$ on p. 35, it is clear that for $k = 2$, the maximum likelihood rule coincides with Fisher's linear discriminant rule. This is NOT true, however, for larger k.

References for Chapter 7

MKB: Fisher's LDF §11.5; ML rule for known population §11.2.1; for unknown parameters §11.3.1.
M: LDF for two groups §4.2; known parameters case §4.3; unknown parameters §4.6.
JW: Fisher's method pp. 590-2.

8. Preliminary data analysis using R

8.1 Data Inspection

It is wise to inspect a data set before attempting to analyse it. A range of useful tools, particularly graphical ones, is available in R.

(i) **Summary statistics** for each variable can be found using the function `summary()` in R.

Example

Summary statistics for the data on *Iris setosa* can be obtained by:-

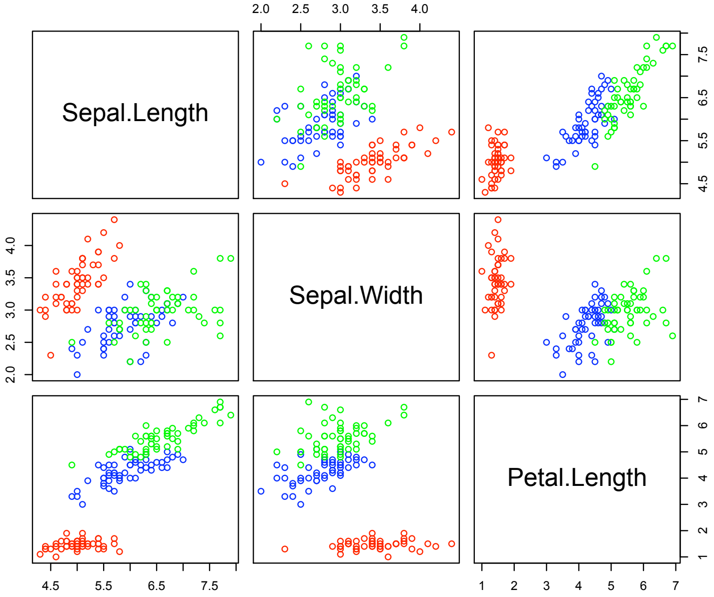
```
> summary(iris[1:50,])
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200   versicolor: 0
 Median :5.000   Median :3.400   Median :1.500   Median :0.200   virginica : 0
 Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
 3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
 Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
```

(ii) **Scatter plots** for each pair of variables can be found using the function `pairs()` in R.

Example

Consider the first three variables (sepal length, sepal width and petal length) of the Iris data. Plotting *Iris setosa* in red, *Iris versicolor* in blue and *Iris virginica* in green, we obtain:-

```
> Y <- c(rep("red",50),rep("blue",50),
rep("green",50))
> pairs(iris[,1:3],col=Y)
```

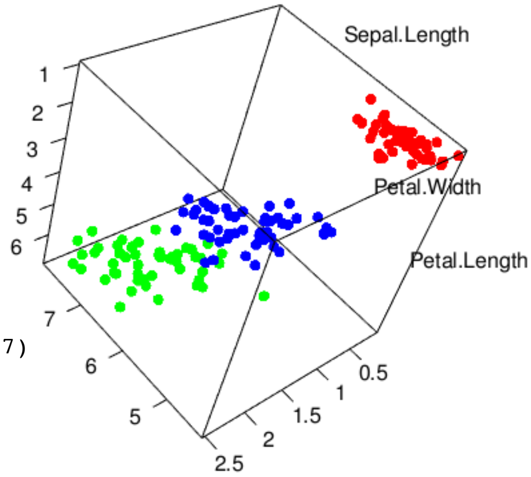


(iii) **Three-dimensional plots** for subsets of three variables can be generated within various packages in R. One option is the function `plot3d()` in the package `rgl` . This produces 3-d plots that can be rotated using the mouse or trackpad.

Example

If we omit the data on sepal width, we may plot the other three variables in the Iris data, using the same colour scheme as above, as follows:-

```
 > library(rgl)
 > attach(iris)
 > Y<- c(rep("red",50),rep("blue",50),rep("green",50) )
 > plot3d(Sepal.Length, Petal.Length, Petal.Width, col=Y, size=7)
```



8.2 Checking Multivariate Normality

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be p-vectors. There are various ways of checking whether they may plausibly be regarded as a random sample from a MVN distribution.

(i) **Univariate marginal Normality**

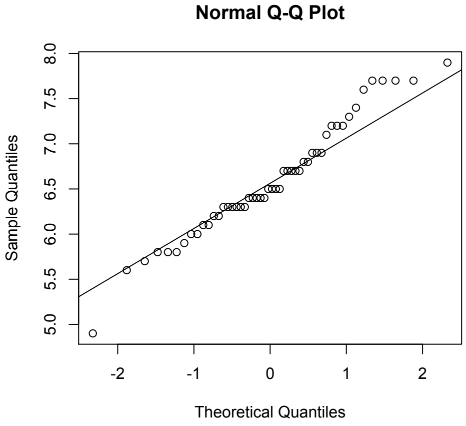
For each variable, obtain a Normal probability plot and assess its linearity.

Example

A Normal probability plot for the sepal length observations for *Iris virginica* can be obtained by:-

```
        > qqnorm(iris[101:150,1])
        > qqline(iris[101:150,1])
```

The plot suggests that the upper tail of the underlying distribution is longer than that of the Normal distribution.



(ii) **Bivariate marginal Normality**

For each pair of variables, reconsider the plot found using the function `pairs()` previously. If the assumption of bivariate Normality is sound, the plot would be expected to reflect the elliptic contours of the bivariate Normal. For instance, the plot of sepal length against sepal width for *Iris setosa* might be deemed to have roughly elliptical contours.

(iii) **Joint Normality**

Suppose that $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If, in §5.1, we replace $\boldsymbol{\mu}_0$ by $\boldsymbol{\mu}$ and apply the derivation in the case where $n = 1$, we get that

$$U = (\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p. \tag{66}$$

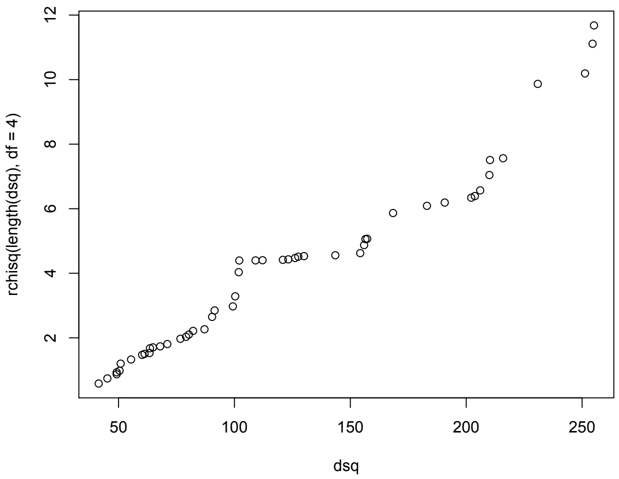
Turning to the p-vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ that we have observed, suppose that we now define

$$U_i = (\mathbf{X}_i - \bar{\mathbf{X}})'\mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \qquad (i = 1, \ldots, n).$$

Note that the "observed" value $u_i = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$ is the squared sample Mahalanobis distance from $\mathbf{x}_i$ to $\bar{\mathbf{x}}$. Since we have replaced $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (66) by their estimates, if $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variables, then $U_1, \ldots, U_n$ should be approximately i.i.d. χ^2_p random variables. This can be assessed by a quantile plot of the order statistics of $u_1, \ldots, u_n$ against simulated quantiles of χ^2_p.

Example

Suppose we consider the logs of the measurements for *Iris virginica*:-



```
> X<-data.matrix(log(iris[101:150,1:4]))
> S<-var(X)
> IS<-solve(S)
> xbar<-apply(X,2,mean)
> xcen<-X-xbar
> dsq<-diag(xcen %*% IS %*% t(xcen))
> qqplot(dsq,rchisq(length(dsq),df=4))
```

Here `xbar` is the sample mean vector, and, in `xcen`, the relevant component of the sample mean has been subtracted from each component of every observation. The vector `dsq`, the diagonal of the matrix product, gives the values of $u_1, \ldots, u_n$. The resulting quantile plot is as shown.

References for Chapter 8

JW: Graphical techniques pp. 11-19, Assessing the assumption of Normality §4.6.
M: Testing the Normality assumption §2.11.

9. Principal components analysis

9.1 Principal components and eigenvectors of S

If we are dealing with a large number p of variables, we may wish to reduce these to a smaller number of 'factors' (or composite variables or latent variables). For simplicity, we consider only *linear* combinations of the original variables. We aim to choose these factors so that we retain an adequate summary of the data, including its covariance structure. Principal components analysis (PCA) provides one method for doing this data dimensionality reduction.

Suppose that we have observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ on a random p-vector $\mathbf{X}$, and that the sample variance matrix is $\mathbf{S}$. We seek to rotate our first axis so that it aligns with the direction in which the observations have the greatest variance. Then, holding this first axis fixed, we rotate a second axis about it to get the alignment that gives the next largest component of the variance. We continue in this manner until all p axes have been aligned. More precisely, we pursue a formulation similar to that for the principal axes of a MVN distribution on p. 10CHECK.

Defn. For $j = 1, \ldots, p$, the j^{th} **principal component** of a sample is the linear combination $Y_j = \mathbf{a}'_j\mathbf{X}$, such that $\mathbf{a}'_j\mathbf{a}_j = 1$ and $Y_j = \mathbf{a}'_j\mathbf{X}$ maximises the sample variance

$$\begin{cases} \mathbf{a}'\mathbf{S}\mathbf{a} \text{ of } \mathbf{a}'\mathbf{X} \text{ when } j = 1; \\ \mathbf{a}'\mathbf{S}\mathbf{a} \text{ of } \mathbf{a}'\mathbf{X} \text{ subject to the sample correlation } \operatorname{corr}(Y_i, Y_j) = 0 \text{ for } i < j \text{ when } j \geq 2. \end{cases}$$

Prop. The vector $\mathbf{a}_j$ giving the j^{th} principal component of the sample is a normalised eigenvector of $\mathbf{S}$ corresponding to the j^{th} largest eigenvalue.

Proof

First note that, when $i < j$, we have the sample correlation $\operatorname{corr}(Y_i, Y_j) = 0$, or equivalently the sample covariance satisfies

$$0 = \operatorname{cov}(Y_i, Y_j) = \operatorname{cov}(\mathbf{a}'_i\mathbf{X}, \mathbf{a}'_j\mathbf{X}) = \mathbf{a}'_i\mathbf{S}\mathbf{a}_j \qquad \text{[by (10)].} \qquad (67)$$

Once we have chosen $\mathbf{a}_i$, for $i < j$, we will wish to maximise the sample variance $\operatorname{var}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\mathbf{S}\mathbf{a},$

subject to $\qquad 0 = \mathbf{a}'_i\mathbf{S}\mathbf{a} = \mathbf{a}'\mathbf{S}\mathbf{a}_i \qquad$ for $i < j$, and also $\mathbf{a}'\mathbf{a} = 1$.

Therefore we maximise the Lagrange form

$$L(\mathbf{a}) = \mathbf{a}'\mathbf{S}\mathbf{a} - \lambda\,\mathbf{a}'\mathbf{a} - \sum_{i=1}^{j-1} \mu_i\,\mathbf{a}'\mathbf{S}\mathbf{a}_i,$$

where, by a standard convention, the sum is omitted when $j = 1$. Then, by (16) and (54),

$$\left(\frac{\partial L}{\partial \mathbf{a}}\right) = 2\mathbf{S}\mathbf{a} - 2\lambda\,\mathbf{a} - \sum_{i=1}^{j-1} \mu_i\,\mathbf{S}\mathbf{a}_i.$$

At the maximum, $\mathbf{a} = \mathbf{a}_j$ and $\mathbf{0} = 2\mathbf{S}\mathbf{a}_j - 2\lambda\,\mathbf{a}_j - \sum_{i=1}^{j-1} \mu_i\,\mathbf{S}\mathbf{a}_i.$ (68)

Pre-multiplying by $\mathbf{a}'_j$,

$$\begin{aligned} 0 &= 2\mathbf{a}'_j\mathbf{S}\mathbf{a}_j - 2\lambda\,\mathbf{a}'_j\mathbf{a}_j - \sum_{i=1}^{j-1} \mu_i\,\mathbf{a}'_j\mathbf{S}\mathbf{a}_i \\ &= 2\mathbf{a}'_j\mathbf{S}\mathbf{a}_j - 2\lambda \qquad \text{(using the transpose of (67)).} \end{aligned}$$

$$\Rightarrow \qquad \mathbf{a}'_j\,\mathbf{S}\,\mathbf{a}_j = \lambda. \qquad (69)$$

Now denote the eigenvalues of $\mathbf{S}$ by $\lambda_1, \ldots, \lambda_p$, where $\lambda_1 \geq \cdots \geq \lambda_p$,. and consider the inductive hypothesis

P_l : for $i < l$, $\mathbf{a}_i$ is a normalised eigenvector of $\mathbf{S}$ corresponding to λ_i.

Now, when $j = 1$, as the sum in (68) is omitted, that equation gives $\mathbf{S}\mathbf{a}_1 = \lambda\,\mathbf{a}_1$, showing that $\mathbf{a}_1$ is an eigenvector of $\mathbf{S}$. Moreover, as $\mathbf{a}'_1\,\mathbf{S}\,\mathbf{a}_1$ gives a maximum, equation (69) shows that we want the largest eigenvalue λ_1. Hence hypothesis P_2 holds.

Suppose now that P_j ($j \geq 2$) holds. $\Rightarrow$ $\mathbf{S}\mathbf{a}_k = \lambda_k\mathbf{a}_k$ for $k < j$. Pre-multiplying (68) by $\mathbf{a}'_k$, where $k < j$,

$$\begin{aligned} 0 = 2\mathbf{a}'_k\mathbf{S}\mathbf{a}_j - 2\lambda\mathbf{a}'_k\mathbf{a}_j - \sum_{i=1}^{j-1} \mu_i\mathbf{a}'_k\mathbf{S}\mathbf{a}_i &= -2\lambda\mathbf{a}'_j\mathbf{a}_k - \mu_k\mathbf{a}'_k\mathbf{S}\mathbf{a}_k \\ &\qquad \text{(by (67) or its transpose)} \\ &= -2\frac{\lambda}{\lambda_k}\mathbf{a}'_j\mathbf{S}\mathbf{a}_k - \mu_k\mathbf{a}'_k\mathbf{S}\mathbf{a}_k \\ &= -\mu_k\mathbf{a}'_k\mathbf{S}\mathbf{a}_k. \end{aligned}$$

$\Rightarrow$ $\mu_k = 0$ for $k < j$ (assuming $\mathbf{S} > 0$) and (68) then implies that $\mathbf{S}\mathbf{a}_j = \lambda\mathbf{a}_j$, so that $\mathbf{a}_j$ is an eigenvector of $\mathbf{S}$ with eigenvalue λ. Again equation (69) shows that we want the largest permissible eigenvalue. If the constraints (67) are to be satisfied, the largest permissible eigenvalue is $\lambda = \lambda_j$ and thus P_{j+1} holds.

In the last proposition, when seeking the j^{th} principal component $Y_j = \mathbf{a}'_j\,\mathbf{X}$ of the sample, we proved in equation (69) that $\mathbf{a}'_j\,\mathbf{S}\,\mathbf{a}_j$ was equal to the Lagrange multiplier λ, before going on to prove that λ had to equal the j^{th} largest eigenvalue λ_j of $\mathbf{S}$. It follows that the sample variance

$$\text{var}(Y_j) = \text{var}(\mathbf{a}'_j\,\mathbf{X}) = \mathbf{a}'_j\mathbf{S}\mathbf{a}_j = \lambda_j.$$

The sum of the sample variances of all p principal components is therefore

$$\lambda_1 + \ldots + \lambda_p = \text{tr}(\mathbf{S}) \qquad \text{[by T1Q3(iii)].}$$

Thus the proportion of the total accounted for by the j^{th} principal component (PC) is

$$\frac{\lambda_j}{\lambda_1 + \ldots + \lambda_p}.$$

Clearly the cumulative proportion of the total accounted for by the first j PCs increases as j increases, and equals one when j reaches p.

As the sequence $\{\lambda_j\}$ is non-increasing, the proportion of the sum of the sample variances accounted for by each successive PC is also non-increasing. Thus the first few PCs are the most important ones, and, as our aim is to find a briefer summary of the data, we may choose to discard later PCs that account for very little of the sum of the sample variances.

There is no cut-and-dried answer to the question of how many PCs to retain. One consideration is the proportion of the sum of the sample variances that is accounted for by those that are retained. A target of around 80% to 90% will often be reasonable, but the values of the eigenvalues in any particular problem will determine what options are available. One tool that can be used to assist the decision is the **scree plot**, in which the sample variance of the j^{th} PC, namely λ_j, is plotted against the ordinal number j of the eigenvalue. The informal interpretation of this diagram is to look for an "elbow" point in the plot and retain only those PCs that lie to the left of it. The line becomes flatter at the elbow, and PCs located at or to the right of the elbow will only explain relatively small proportions of the sum of the sample variances. The interpretations of the PCs (see below) in terms of the subject matter of the problem may also influence the decision on which ones are kept.

Although the above derivation of PCs was based on the sample variance matrix **S**, the same approach can also be applied to the sample correlation matrix **R**. Despite the relationship (9) between **S** and **R**, the PCs provided by the two different approaches are NOT the same.

Which then of the two matrices should be used? Use of **S** has been recommended when all the variables are of the same type (e.g. all are lengths) *and* the variances are reasonably similar. If there are variables measured on different scales (e.g. some lengths, some weights) or if the variances differ widely, then use of the sample correlation matrix **R** is to be preferred.

9.2 Using `R` for principal components analysis

Principal components analysis can be carried out using the function `prcomp`. This function has a logical argument `scale`, which, when set to TRUE, will first scale all the variables to have unit variances.

Example (Morrison, p.127)
A random sample of 15 girls was drawn from a third grade class in a Connecticut school, and Metropolitan Achievement Test scores were obtained for each girl. The test has six sections, namely vocabulary (V), word recognition skills (WREC), reading comprehension (RC), mathematical concepts (MCON), mathematical computation (MCOMP) and mathematical problem solving (MPS).

The data were entered in `R` as a data frame X with the scores for the i^{th} girl ($i = 1, \ldots, 15$) along row i.

```
> X
     V WREC  RC MCON MCOMP MPS
1  638  617 631  615   524 541
2  638  607 631  591   508 572
3  530  576 569  573   501 496
4  480  501 541  521   501 496
5  571  576 610  539   470 561
6  638  617 717  581   516 586
7  614  558 604  591   548 576
8  638  630 631  602   508 572
9  638  558 616  581   516 561
10 587  564 585  573   470 510
11 515  546 548  533   501 541
12 614  673 616  591   575 586
13 638  617 652  581   548 561
14 638  617 694  685   595 633
15 638  630 652  615   608 586
```

If the variances of the scores for the six topics are evaluated, it may be verified that the largest is about twice that of the smallest. The principal components analysis was therefore carried out as follows:-

```
> PCA <-  prcomp(X, scale = TRUE)
> PCA
Standard deviations:
[1] 2.0888371 0.7984329 0.6247014 0.5532648 0.3941500 0.3841309
Rotation:
             PC1        PC2         PC3        PC4         PC5        PC6
V     -0.4221973  0.4247254 -0.03853355  0.2331711 -0.68270376 -0.3455716
WREC  -0.3953361  0.1830423  0.84658080 -0.1386573  0.14707216  0.2294553
RC    -0.4240913  0.3597394 -0.30921314 -0.1664569  0.64424756 -0.3903314
MCON  -0.4168901 -0.2168655 -0.16213293  0.7536356  0.19559596  0.3829629
MCOMP -0.3612956 -0.7769191  0.11872447 -0.1366787 -0.07161379 -0.4774486
MPS   -0.4257622 -0.0778298 -0.38186484 -0.5586142 -0.23204337  0.5485930

> summary(PCA)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     2.0888 0.7984 0.62470 0.55326 0.39415 0.38413
Proportion of Variance 0.7272 0.1062 0.06504 0.05102 0.02589 0.02459
Cumulative Proportion  0.7272 0.8335 0.89850 0.94951 0.97541 1.00000
```

Although principal components analysis is driven by mathematics rather than contextual considerations, it is worth seeing whether each PC can be interpreted in terms of the real world. Here
PC1 is ≈ -0.4(total score) for each individual.
i.e. a measure of overall ability.

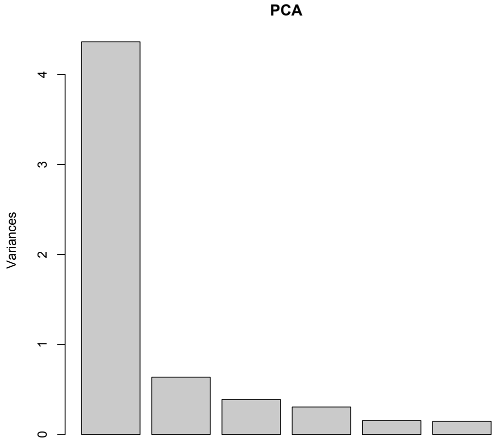
PC2 contrasts ability at English and ability at mathematical computation.

PC3 contrasts ability at WREC with ability at RC and MPS.

PC4 is primarily a contrast between mathematical concepts and problem solving.

A scree plot can be obtained using:-

```
> screeplot(PCA)
```



How many of these of these PCs should be retained? The first PC explains 72.7% of the variance, with the second, third and fourth PCs taking us to 83.3%, 89.8% and 95% respectively.

This scree plot has a pronounced elbow at $j = 2$, which would suggest retaining only the first PC. That would, however, retain a rather low proportion of the total variance. The argument for keeping the second PC with 10.6% of the variability and a useful practical interpretation looks quite strong, and, depending on further uses envisaged, there could be a case for keeping more.

References for Chapter 9
JW: Summarizing sample variation by PCs. §8.3.
M: PCs of multivariate observations §6.2; Interpretation of PCs §6.4
MKB: Sample PCs §8.2.2; Properties §8.2.3; Ignoring components §8.2.5.

---

10. Canonical correlation

10.1 Finding canonical variates

Suppose we have n observations on a random vector $(\mathbf{X}_1', \mathbf{X}_2')'$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ have dimensions p and q respectively. Let the corresponding partition of the sample variance matrix $\mathbf{S}$ be

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

where we assume that $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are both invertible. In §6.3 we considered testing the independence of $\mathbf{X}_1$ and $\mathbf{X}_2$ (assuming Normality), and, using the union-intersection principle, took as our test statistic $\max_{\mathbf{a}\neq\mathbf{0},\, \mathbf{b}\neq\mathbf{0}}\{r(\mathbf{a},\mathbf{b})\}^2$, where $r(\mathbf{a},\mathbf{b})$ was the sample correlation between $\mathbf{a}'\mathbf{X}_1$ and $\mathbf{b}'\mathbf{X}_2$.

Here, in contrast, we accept that there is dependence between $\mathbf{X}_1$ and $\mathbf{X}_2$, and aim to describe the relationship parsimoniously. We consider pairs of linear combinations of the form $U = \mathbf{a}'\mathbf{X}_1$ and $V = \mathbf{b}'\mathbf{X}_2$, and, if $s = \min(p,q)$, obtain s successive pairs such that the components within each pair are as highly correlated as possible, but uncorrelated with previous choices for that component. More precisely, for the j^{th} $(j = 1,\ldots,s)$ **pair of canonical variates** $(U_j, V_j) = (\mathbf{a}_j'\mathbf{X}_1, \mathbf{b}_j'\mathbf{X}_2)$, the vectors $\mathbf{a}_j$ and $\mathbf{b}_j$ give the respective values of $\mathbf{a}$ and $\mathbf{b}$ that yield the maximum value r_j^2 of $\{r(\mathbf{a},\mathbf{b})\}^2 = [\text{corr}(U,V)]^2$ subject to (a) $\text{var}(U_j) = 1 = \text{var}(V_j)$,

(b) U_j is uncorrelated with U_i for $i < j$, and V_j is uncorrelated with V_i for $i < j$.

The sample correlation $r_j = \text{corr}(U_j, V_j)$ is called the j^{th} **canonical correlation**.

Prop.

For $j = 1,\ldots,s$, the vectors $\mathbf{a}_j$ and $\mathbf{b}_j$ which maximise $\{r(\mathbf{a},\mathbf{b})\}^2$ subject to

$$\text{var}(U) = \mathbf{a}'\,\mathbf{S}_{11}\,\mathbf{a} = 1 = \mathbf{b}'\,\mathbf{S}_{22}\,\mathbf{b} = \text{var}(V), \tag{70}$$

$$\text{cov}(U,U_i) = \mathbf{a}'\,\mathbf{S}_{11}\,\mathbf{a}_i = 0 = \mathbf{b}'\,\mathbf{S}_{22}\,\mathbf{b}_i = \text{cov}(V,V_i) \qquad (i < j) \tag{71}$$

are eigenvectors of $\mathbf{M}_1 = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ and of $\mathbf{M}_3 = \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ respectively, with associated eigenvalue $\lambda_j = r_j^2$, where λ_j is the j^{th} largest eigenvalue of both $\mathbf{M}_1$ and $\mathbf{M}_3$.

Proof

When the constraint (70) holds, the sample correlation

$$r(\mathbf{a},\mathbf{b}) = \text{corr}(U,V) = \text{cov}(U,V) = \mathbf{a}'\,\mathbf{S}_{12}\,\mathbf{b} = \mathbf{b}'\,\mathbf{S}_{21}\,\mathbf{a} \qquad \text{by (10).} \tag{72}$$

Once we have chosen (U_i, V_i), for $i < j$, we will wish to maximise $\{r(\mathbf{a},\mathbf{b})\}^2$ subject to

$$\mathbf{a}'\,\mathbf{S}_{11}\,\mathbf{a} = 1 = \mathbf{b}'\,\mathbf{S}_{22}\,\mathbf{b} \qquad \text{and} \qquad \mathbf{a}'\,\mathbf{S}_{11}\,\mathbf{a}_i = 0 = \mathbf{b}'\,\mathbf{S}_{22}\,\mathbf{b}_i \ \text{ for } i < j.$$

Therefore we maximise the Lagrange form

$$L(\mathbf{a},\mathbf{b}) = (\mathbf{a}'\,\mathbf{S}_{12}\,\mathbf{b})^2 - \lambda\,\mathbf{a}'\,\mathbf{S}_{11}\,\mathbf{a} - \theta\,\mathbf{b}'\,\mathbf{S}_{22}\,\mathbf{b} - \sum_{i=1}^{j-1}\mu_i\,\mathbf{a}'\,\mathbf{S}_{11}\,\mathbf{a}_i - \sum_{i=1}^{j-1}\gamma_i\,\mathbf{b}'\,\mathbf{S}_{22}\,\mathbf{b}_i.$$

Then

$$\left(\frac{\partial L}{\partial \mathbf{a}}\right) = 2(\mathbf{a}'\,\mathbf{S}_{12}\,\mathbf{b})\,\mathbf{S}_{12}\,\mathbf{b} - 2\lambda\,\mathbf{S}_{11}\,\mathbf{a} - \sum_{i=1}^{j-1}\mu_i\,\mathbf{S}_{11}\,\mathbf{a}_i \qquad \text{[by (16) and (54)],}$$

$$\left(\frac{\partial L}{\partial \mathbf{b}}\right) = 2(\mathbf{a}'\,\mathbf{S}_{12}\,\mathbf{b})\,\mathbf{S}_{21}\,\mathbf{a} - 2\theta\,\mathbf{S}_{22}\,\mathbf{b} - \sum_{i=1}^{j-1}\gamma_i\,\mathbf{S}_{22}\,\mathbf{b}_i \qquad \text{as } \mathbf{a}'\,\mathbf{S}_{12}\,\mathbf{b} = \mathbf{b}'\,\mathbf{S}_{21}\,\mathbf{a}.$$

At the maximum, $\mathbf{a} = \mathbf{a}_j$, $\mathbf{b} = \mathbf{b}_j$ and $r(\mathbf{a},\mathbf{b}) = r_j$, where from (72)

$$r_j = \mathbf{a}_j'\,\mathbf{S}_{12}\,\mathbf{b}_j = \mathbf{b}_j'\,\mathbf{S}_{21}\,\mathbf{a}_j. \tag{73}$$

Therefore, at the maximum, we have

$$\mathbf{0} = 2(\mathbf{a}_j'\,\mathbf{S}_{12}\,\mathbf{b}_j)\,\mathbf{S}_{12}\,\mathbf{b}_j - 2\lambda\,\mathbf{S}_{11}\,\mathbf{a}_j - \sum_{i=1}^{j-1}\mu_i\,\mathbf{S}_{11}\,\mathbf{a}_i \tag{74}$$

$$\mathbf{0} = 2(\mathbf{a}_j'\,\mathbf{S}_{12}\,\mathbf{b}_j)\,\mathbf{S}_{21}\,\mathbf{a}_j - 2\theta\,\mathbf{S}_{22}\,\mathbf{b}_j - \sum_{i=1}^{j-1}\gamma_i\,\mathbf{S}_{22}\,\mathbf{b}_i. \tag{75}$$

Suppose we knew that the summations in (74) and (75) were zero. We could then premultiply (74) and (75) by $\mathbf{a}_j'$ and $\mathbf{b}_j'$ respectively to get

$$\begin{array}{rcllll} 0 & = & (\mathbf{a}_j'\,\mathbf{S}_{12}\,\mathbf{b}_j)^2 - \lambda\,\mathbf{a}_j'\,\mathbf{S}_{11}\,\mathbf{a}_j & \Rightarrow & \lambda = r_j^2 & \text{by (70) and (73)} \\ 0 & = & (\mathbf{a}_j'\,\mathbf{S}_{12}\,\mathbf{b}_j)\,(\mathbf{b}_j'\,\mathbf{S}_{21}\,\mathbf{a}_j) - \theta\,\mathbf{b}_j'\,\mathbf{S}_{22}\,\mathbf{b}_j & \Rightarrow & \theta = r_j^2 = \lambda & \text{by (70) and (73).} \end{array}$$

Equations (74) and (75) would then imply that

$$\begin{array}{rclcll} \mathbf{0} & = & \mathbf{S}_{12}\,\mathbf{b}_j - r_j\,\mathbf{S}_{11}\,\mathbf{a}_j & \Rightarrow & \mathbf{a}_j = \mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{b}_j/r_j & (76) \\ \mathbf{0} & = & \mathbf{S}_{21}\,\mathbf{a}_j - r_j\,\mathbf{S}_{22}\,\mathbf{b}_j & \Rightarrow & \mathbf{b}_j = \mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\,\mathbf{a}_j/r_j. & (77) \end{array}$$

Substituting for $\mathbf{b}_j$ in the first equation of (76) and for $\mathbf{a}_j$ in the first equation of (77),

$$\begin{array}{rclcll} r_j^2\,\mathbf{S}_{11}\,\mathbf{a}_j & = & \mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\,\mathbf{a}_j & \Rightarrow & \mathbf{M}_1\,\mathbf{a}_j = \lambda\,\mathbf{a}_j \quad \text{where } \mathbf{M}_1 = \mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}, & (78) \\ r_j^2\,\mathbf{S}_{22}\,\mathbf{b}_j & = & \mathbf{S}_{21}\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{b}_j & \Rightarrow & \mathbf{M}_3\,\mathbf{b}_j = \lambda\,\mathbf{b}_j \quad \text{where } \mathbf{M}_3 = \mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}. & (79) \end{array}$$

Thus λ would be an eigenvalue of $\mathbf{M}_1$ and of $\mathbf{M}_3$.

To *confirm* the suppositions made above, we consider the inductive hypothesis

$$H_j : \text{for } k < j, \begin{cases} \mathbf{S}_{12}\,\mathbf{b}_k = r_k\,\mathbf{S}_{11}\,\mathbf{a}_k \text{ and } \mathbf{S}_{21}\mathbf{a}_k = r_k\,\mathbf{S}_{22}\,\mathbf{b}_k, \\ \mathbf{a}_k \text{ and } \mathbf{b}_k \text{ are eigenvectors of } \mathbf{M}_1 \text{ and } \mathbf{M}_3 \text{ respectively with eigenvalue } \lambda_k. \end{cases}$$

Now, when $j = 1$, the sums are omitted automatically from (74) and (75), so the results (76) to (79) all hold. Moreover, as $\lambda = r_1^2 = (\mathbf{a}_1'\mathbf{S}_{12}\mathbf{b}_1)^2$, which is the maximum value, we must have $\lambda = \lambda_1$. So H_2 holds.

Suppose now that H_j holds, where $j \geq 2$. Noting that the left-hand sides of the following equations are scalar and using the inductive hypothesis and the constraints (71), we have

$$\begin{array}{rclll} \mathbf{a}_k'\,\mathbf{S}_{12}\,\mathbf{b}_j = \mathbf{b}_j'\,\mathbf{S}_{21}\,\mathbf{a}_k & = & r_k\,\mathbf{b}_j'\,\mathbf{S}_{22}\,\mathbf{b}_k = 0 & (k < j) & (80) \\ \mathbf{b}_k'\,\mathbf{S}_{21}\,\mathbf{a}_j = \mathbf{a}_j'\,\mathbf{S}_{12}\,\mathbf{b}_k & = & r_k\,\mathbf{a}_j'\,\mathbf{S}_{11}\,\mathbf{a}_k = 0 & (k < j). & (81) \end{array}$$

Premultiplying (74) by $\mathbf{a}_k'$ and (75) by $\mathbf{b}_k'$, where $k < j$, gives

$$\begin{array}{rclll} 0 & = & 2r_j\,\mathbf{a}_k'\,\mathbf{S}_{12}\,\mathbf{b}_j - 2\lambda\,\mathbf{a}_k'\,\mathbf{S}_{11}\,\mathbf{a}_j - \sum_{i=1}^{j-1} \mu_i\,\mathbf{a}_k'\,\mathbf{S}_{11}\,\mathbf{a}_i & & \\ & = & -\mu_k & (k = 1, \ldots, j-1) & \text{by (80), (71) or its transpose, and (70)} \\ 0 & = & 2r_j\,\mathbf{b}_k'\,\mathbf{S}_{21}\,\mathbf{a}_j - 2\theta\,\mathbf{b}_k'\,\mathbf{S}_{22}\,\mathbf{b}_j - \sum_{i=1}^{j-1} \gamma_i\,\mathbf{b}_k'\,\mathbf{S}_{22}\,\mathbf{b}_i & & \\ & = & -\gamma_k & (k = 1, \ldots, j-1) & \text{by (81), (71) or its transpose, and (70)} \end{array}$$

Hence the sums in (74) and (75) are equal to zero, so the results (76) to (79) again all hold. Since we are maximising $(\mathbf{a}_j'\,\mathbf{S}_{12}\,\mathbf{b}_j)^2 = r_j^2 = \lambda$ subject to the given conditions, we want the largest permissible eigenvalue. Hence λ must be the j^{th} largest eigenvalue λ_j. Thus H_{j+1} holds, completing the proof.

Notes

(i) Use of the correlation matrix $\mathbf{R}$ gives rise to the same canonical correlations $r_1, \ldots, r_s$ as are obtained from the variance matrix $\mathbf{S}$, since the eigenvalues of $\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$ are the same as those of $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$.

(ii) In the above proposition, we used the condition (71) which encapsulated the constraint (b) that U_j is uncorrelated with U_i for $i < j$, and V_j is uncorrelated with V_i for $i < j$. As a by-product of the proposition, equations (80) and (81) offer an added bonus, namely that U_j is uncorrelated with V_i for $i < j$, and V_j is uncorrelated with U_i for $i < j$.

It follows that the canonical variates $U_1, \ldots, U_s, V_1, \ldots, V_s$ have correlation matrix

$$\begin{pmatrix} \mathbf{I}_s & \tilde{\mathbf{R}} \\ \tilde{\mathbf{R}} & \mathbf{I}_s \end{pmatrix}, \quad \text{where } \tilde{\mathbf{R}} = \text{diag}(r_1, \ldots, r_s).$$

(iii) Although $\mathbf{a}_i'\mathbf{S}_{11}\mathbf{a}_j = 0$ for $i \neq j$, the vectors $\mathbf{a}_i$ and $\mathbf{a}_j$ are not usually orthogonal (i.e. it is not true in general that $\mathbf{a}_i'\mathbf{a}_j = 0$).

(iv) If $p = 1$ then $r_1 = r_{1\cdot 2\ldots(q+1)}$, where $r_{1.2\ldots(q+1)}$ is the multiple correlation coefficient.

10.2 Using R for canonical correlation analysis

Canonical correlation can be carried out using the function `canocor` in the `calibrate` package.

Example (MKB, pp. 121 & 287)

We have previously considered parts of a data set on the head sizes of the first and second sons in a sample of 25 families. In fact, measurements (in millimetres) were made of both $\mathbf{X}_1$, the head length and the head breadth of the first son, and $\mathbf{X}_2$, the head length and the head breadth of the second son. Canonical correlation analysis was used to explore the relationship between $\mathbf{X}_1$ and $\mathbf{X}_2$. Using R, we first mimic the manual calculation:

```
> X
   FirstHL FirstHB SecondHL SecondHB
1      191     155      179      145
2      195     149      201      152
. . .
24     197     167      200      158
25     190     163      187      150
```

The correlation matrix of the complete data is:-

```
> R<-cor(X)
> R
           FirstHL   FirstHB  SecondHL  SecondHB
FirstHL  1.0000000 0.7345555 0.7107518 0.7039807
FirstHB  0.7345555 1.0000000 0.6931573 0.7085504
SecondHL 0.7107518 0.6931573 1.0000000 0.8392519
SecondHB 0.7039807 0.7085504 0.8392519 1.0000000
```

The required sub-matrices are:-

```
> R11<-R[1:2,1:2]
> R11
          FirstHL   FirstHB
FirstHL 1.0000000 0.7345555
FirstHB 0.7345555 1.0000000
> R22<-R[3:4,3:4]
> R22
         SecondHL SecondHB
SecondHL 1.000000 0.839252
SecondHB 0.839252 1.000000
> R12<-R[1:2,3:4]
> R12
         SecondHL  SecondHB
FirstHL 0.7107518 0.7039807
FirstHB 0.6931573 0.7085504
```

Hence:-

```
> IR11<-solve(R11)
> IR11
          FirstHL   FirstHB
FirstHL  2.171891 -1.595375
FirstHB -1.595375  2.171891
> IR22<-solve(R22)
> IR22
          SecondHL  SecondHB
SecondHL  3.382307 -2.838608
SecondHB -2.838608  3.382307
> M1<- IR11 %*%  R12  %*%  IR22  %*%  t(R12)
> M1
          FirstHL   FirstHB
FirstHL 0.3225003 0.3168319
FirstHB 0.3018705 0.3021324
> E1<-eigen(M1)
> E1
$values
[1] 0.621744734 0.002887956
$vectors
          [,1]       [,2]
[1,] 0.7269968 -0.7040109
[2,] 0.6866408  0.7101892
> cancorrs<-sqrt(E1$values)
> cancorrs
[1] 0.7885079 0.0537397
> M3<- IR22  %*%  t(R12)  %*%  IR11  %*%  R12
> M3
          SecondHL  SecondHB
SecondHL 0.3013980 0.3002082
SecondHB 0.3185347 0.3232347
> E3<-eigen(M3)
> E3
$values
[1] 0.621744734 0.002887956
$vectors
           [,1]       [,2]
[1,] -0.6837994 -0.7091095
[2,] -0.7296700  0.7050984
```

Alternatively, using the function canocor:-

```
> install.packages("calibrate")
  . . .
> library(calibrate)
. . .
> X1<-X[ ,1:2]
> X2<-X[ ,3:4]
> canocor(X1,X2)
```

The output includes:-

```
$ccor
          [,1]      [,2]
[1,] 0.7885079 0.0000000
[2,] 0.0000000 0.0537397

$A
           [,1]      [,2]
[1,] -0.5521896 -1.366374
[2,] -0.5215372  1.378365

$B
           [,1]      [,2]
[1,] -0.5044484 -1.768570
[2,] -0.5382877  1.758566
```

It may be verified that, when normalised, these eigenvectors agree with those produced by the previous approach. The normalised eigenvectors $\mathbf{a}_1 = (0.727,\ 0.687)'$ and $\mathbf{b}_1 = (-0.684,\ -0.730)'$ can be interpreted as measuring size or "girth".

The normalised eigenvectors $\mathbf{a}_2 = (-0.704,\ 0.710)'$ and $\mathbf{b}_2 = (-0.709,\ 0.705)'$ can be interpreted as contrasting head length and head breadth.
i.e. measuring shape.

Notice, however, that the canonical correlations are $r_1 = 0.79$ and $r_2 = 0.05$, indicating that whilst size is strongly correlated between brothers, the head shape of first and second brothers is almost uncorrelated.

Reference for Chapter 10

M: Canonical correlation §5.6.
