Statistics Honours
Multivariate Analysis – Section A

Mr Stefan S. Britz
Department of Statistical Sciences
University of Cape Town

These notes, covering various aspects of multivariate data and the analysis thereof, are based in part on Johnson and Wichern (2007), and Rencher and Christensen (2012). The accompanying datasets and R scripts referred to in this text will be provided on Vula.

1 Introduction, Summary Statistics & Visualisation

In this section we will discuss the nature and shape of multivariate data as well as various ways of summarising this type of data, in terms of basic summary statistics and graphical displays.

1.1 Multivariate Data

Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. We are, therefore, concerned with the analysis of *multiple response variables*. For example,

1. Mass (in grams), snout-vent length (in mm) and hind limb span (in mm) measurements for lizards.

```
head(lizards)
```

```
     Mass  SVL   HLS
1   5.526 59.0 113.5
2  10.401 75.0 142.0
3   9.213 69.0 124.0
4   8.953 67.5 125.0
5   7.063 62.0 129.5
6   6.610 62.0 123.0
```

2. Air pollution data consisting of 42 measurements of 7 air-pollution variables recorded on different days.

```
head(AirPollution)
```

```
  Wind Radiation CO NO NO2 O3 HC
1    8        98  7  2  12  8  2
2    7       107  4  3   9  5  3
3    7       103  4  3   5  6  3
4   10        88  5  2   8 15  4
5    6        91  4  2   8 10  3
6    8        90  5  2  12 12  4
```

3. National Track records for men and women.

```
head(Track)
```

```
  Country  m100  m200  m400 m800 m1500 m3000 Marathon Gender
1     ARG 11.57 22.94 52.50 2.05  4.25  9.19   150.32 Female
2     AUS 11.12 22.23 48.63 1.98  4.02  8.63   143.51 Female
3     AUT 11.15 22.70 50.62 1.94  4.05  8.78   154.35 Female
4     BEL 11.14 22.48 51.45 1.97  4.08  8.82   143.05 Female
5     BER 11.46 23.05 53.30 2.07  4.29  9.81   174.18 Female
6     BRA 11.17 22.60 50.62 1.97  4.17  9.04   147.41 Female
```

Data can be represented in the form of a matrix where the rows reflect the observations and the columns reflect the variables. Typically,

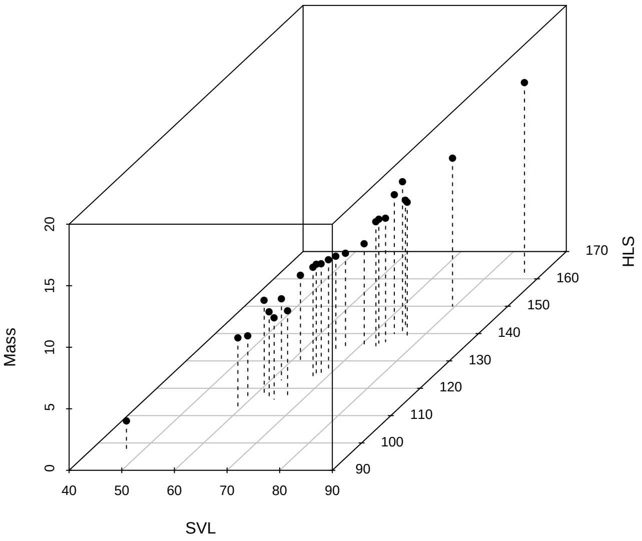
$$n = \text{Number of observations} \quad \text{and} \quad p = \text{Number of variables}$$

$$\mathbf{X}_{n\times p} = \begin{pmatrix} X_{11} & X_{12} & ... & X_{1p} \\ X_{21} & X_{22} & ... & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & ... & X_{np} \end{pmatrix}$$

So for lizard data,

$$\mathbf{X}_{25\times 3} = \begin{pmatrix} 5.526 & 59.0 & 113.5 \\ 10.401 & 75.0 & 142.0 \\ \vdots & \vdots & \vdots \\ 6.890 & 63.0 & 117.0 \end{pmatrix}$$

The data can be viewed as n points in $p-$dimensional space (we'll return to this later):



1.2 Summary Statistics

In the univariate case, where we only focus on one variable of interest, we can describe a sample's location and spread by measuring the sample mean and variance. Likewise, when working with multiple variables simultaneously, we can calculate the means and variances of each variable, as well as the covariances and correlations of all pairs of variables.

1.2.1 Mean vector

$$\bar{\boldsymbol{x}}_{p\times 1} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

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

In matrix notation, $\bar{\boldsymbol{x}} = \frac{1}{n}\boldsymbol{X'1}$, where $\mathbf{1}_{n\times 1}$ is a vector of 1's.

Example: If

$$\boldsymbol{X}_{3\times 2} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

then

$$\bar{\boldsymbol{x}} = \frac{1}{n}\boldsymbol{X'1} = \frac{1}{3}\begin{bmatrix} 4 & -1 & 3 \\ 1 & 3 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 6 \\ 9 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

1.2.2 Sample covariance

The sample covariance matrix $\boldsymbol{S}_{p\times p} = (s_{jk})$ is the matrix of sample variances and covariances of the p variables. This matrix is interchangeably referred to as the variance matrix, variance-covariance matrix or dispersion matrix.

$$\boldsymbol{S}_{p\times p} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

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

and $s_{jk} = \mathrm{Cov}(X_j, X_k) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_k), \quad \forall\ j \neq k$

In matrix notation, the **sample covariance matrix** can be expressed in terms of:

$\mathbf{1}\bar{\boldsymbol{x}}' = \frac{1}{n}\mathbf{11'}\boldsymbol{X}$ → an $n \times p$ matrix of means, and

$\boldsymbol{X} - \frac{1}{n}\mathbf{11'}\boldsymbol{X}$ → an $n \times p$ matrix of deviations such that:

$$\begin{aligned}
\boldsymbol{S} &= \frac{1}{n-1}\left[\left(\boldsymbol{X} - \frac{1}{n}\mathbf{11'}\boldsymbol{X}\right)'\left(\boldsymbol{X} - \frac{1}{n}\mathbf{11'}\boldsymbol{X}\right)\right] \\
(n-1)\boldsymbol{S} &= \left(\boldsymbol{X} - \frac{1}{n}\mathbf{11'}\boldsymbol{X}\right)'\left(\boldsymbol{X} - \frac{1}{n}\mathbf{11'}\boldsymbol{X}\right) \\
&= \boldsymbol{X}'\left(\boldsymbol{I} - \frac{1}{n}\mathbf{11'}\right)'\left(\boldsymbol{I} - \frac{1}{n}\mathbf{11'}\right)\boldsymbol{X} \\
&= \boldsymbol{X}'\left(\boldsymbol{I} - \frac{1}{n}\mathbf{11'}\right)\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{X} - \frac{1}{n}\boldsymbol{X}'\mathbf{11'}\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'
\end{aligned}$$

1.2.3 Sample correlation

The sample correlations are collected in a matrix $\mathbf{R}_{p\times p} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$, where

$$r_{jk} = \mathrm{Cor}(X_j, X_k) = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}$$

If we write $\mathbf{D} = diag(\boldsymbol{S})$ for the diagonal matrix containing the diagonal of the square matrix $\boldsymbol{S}$, then the sample correlation matrix can be expressed as

$$\mathbf{R} = [diag(\boldsymbol{S})]^{-\frac{1}{2}}\boldsymbol{S}[diag(\boldsymbol{S})]^{-\frac{1}{2}} = \mathbf{D}^{-\frac{1}{2}}\boldsymbol{S}\mathbf{D}^{-\frac{1}{2}}$$

R Examples (see `Lecture1_1.R`)

```
#Create matrix
(X <- matrix(c(4,1,-1,3,3,5), 3, 2, byrow=TRUE))
```

```
     [,1] [,2]
[1,]    4    1
[2,]   -1    3
[3,]    3    5
```

```
#Create 3x1 vector of 1s
(One <- matrix(c(1,1,1), 3, 1, byrow=TRUE))
```

```
     [,1]
[1,]    1
[2,]    1
[3,]    1
```

```
#Dimensionality
(dims <- dim(X))
```

```
[1] 3 2
```

```
(n <- dims[1])
```

```
[1] 3
```

```
#or read directly from matrix
(p <- ncol(X))
```

```
[1] 2
```

```
#Calculate mean
(X_bar <- 1/n*t(X) %*% One)
```

```
     [,1]
[1,]    2
[2,]    3
```

```
#Save matrix as a data frame
(Xdata <- as.data.frame(X))
```

```
  V1 V2
1  4  1
2 -1  3
3  3  5
```

```
#Test from data
apply(Xdata, 2, mean)
```

```
V1 V2
 2  3
```

```
#Calculate covariance matrix
#(1)
S <- t(X) %*% X - n*X_bar %*% t(X_bar)
(S <- S/(n - 1))
```

```
     [,1] [,2]
[1,]    7   -1
[2,]   -1    4
```

```
#(2)
(I3 <- diag(1, 3))
```

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

```
S1 <- t(X) %*% (I3-One %*% t(One)/n) %*% X
(S1 <- S1/(n - 1))
```

```
     [,1] [,2]
[1,]    7   -1
[2,]   -1    4
```

```
#Test from data
var(Xdata)
```

```
   V1 V2
V1  7 -1
V2 -1  4
```

```
cov(Xdata)
```

```
    V1  V2
V1   7  -1
V2  -1   4
```

```
apply(Xdata, 2, var)
```

```
V1  V2
7   4
```

```
#Calculate correlation matrix
(D <- diag(diag(S)))
```

```
     [,1]  [,2]
[1,]    7     0
[2,]    0     4
```

```
(R <- solve(D)^(1/2) %*% S %*% solve(D)^(1/2))
```

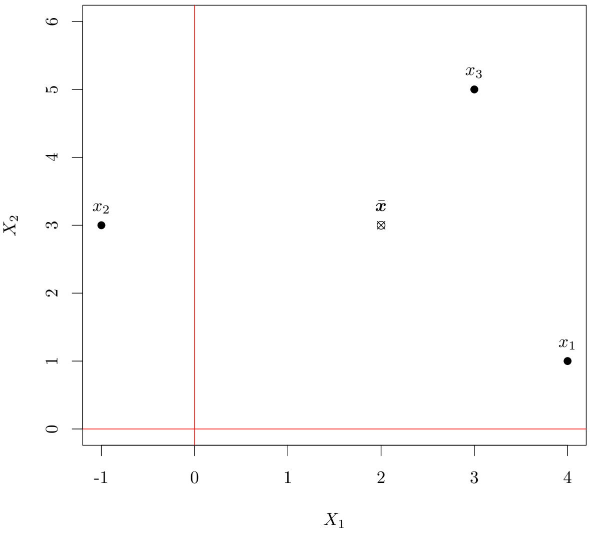
```
           [,1]       [,2]
[1,]  1.0000000 -0.1889822
[2,] -0.1889822  1.0000000
```

```
#Test from data
cor(Xdata)
```

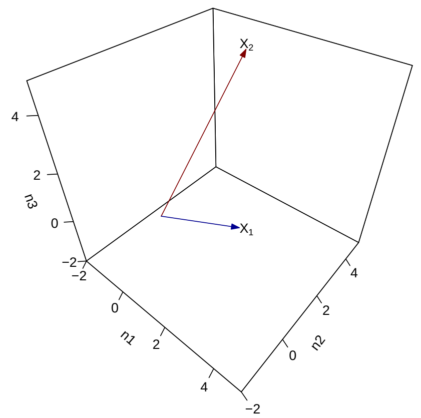
```
           V1         V2
V1  1.0000000 -0.1889822
V2 -0.1889822  1.0000000
```

1.3 Graphical Interpretation of a Matrix

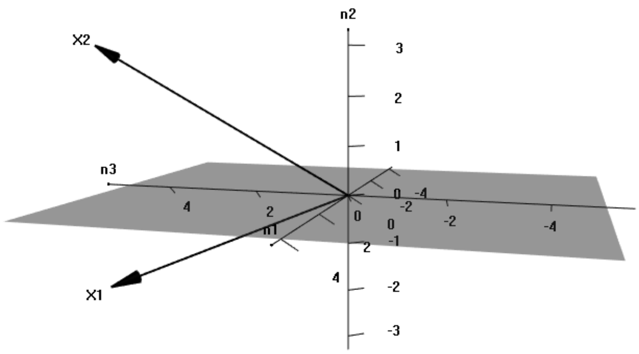
An $n \times p$ matrix $\boldsymbol{X}$ can either be viewed as n observations in $p-$dimensional space, or as p vectors in $n-$dimensional space. Continuing with the earlier example where $\boldsymbol{X}_{3\times 2} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$, we can first display $\boldsymbol{X}$ as 3 observations of 2–dimensional data points. This is referred to as a scatterplot:



This allows us to visually discern relationships between the variables by looking at the patterns of the observations. Note that $\bar{\boldsymbol{x}}$ is the balance point (centre of gravity) of the scatterplot. Alternatively, we can represent each variable as a vector, where the coordinates of each vector point are given by the n measurements on that variable:



Using plot3D::arrows3D



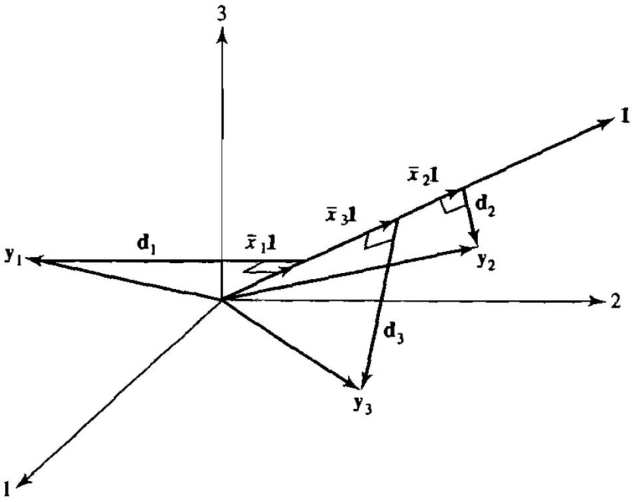
Using matlib::arrows3d & rgl

Even though we cannot visualise data in this way for $n > 3$, the geometrical relationships and associated statistical concepts remain valid regardless of dimensionality. Since three vectors of any dimension can only span a 3-dimensional space (likewise two vectors span a plane), we can select two or three variables of interest and obtain a view that preserves both the lengths of the vectors and the angles between them.

We note that the expressions for the covariance matrix depend on the deviation of the observed values from the variable means. We can define and calculate so-called "deviation vectors":

$$\boldsymbol{d}_i = \boldsymbol{x}_i - \bar{x}_i \mathbf{1} = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix}$$

Elements of the covariance matrix can then be calculated using these deviation vectors. The graphical display and interpretation of the deviation vectors is also informative, since they can now be viewed as a perpendicular projection onto the equal angular vector of ones, $\mathbf{1}$.



We can also note that the squared length of the deviation vectors is proportional to their variance:

$$L^2_{\boldsymbol{d}_i} = \boldsymbol{d}_i'\boldsymbol{d}_i = \sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2$$

Therefore, longer vectors, corresponding to further angular deviation from **1**, represent more variability. Similarly, the inner product between two deviation vectors is proportional to their covariance:

$$\boldsymbol{d}_i'\boldsymbol{d}_k = \sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

Combining these findings, the deviation vectors of a matrix $\boldsymbol{X}$ can be used to calculate elements of its covariance matrix.

For our previous example,

```
#Using deviation vectors
cent_X <- scale(X, center = T, scale = F)
(d1 <- matrix(cent_X[,1]))
```

```
     [,1]
[1,]    2
[2,]   -3
[3,]    1
```

```
(d2 <- matrix(cent_X[,2]))
```

```
     [,1]
[1,]   -2
[2,]    0
[3,]    2
```

```
#s_11 = d1'd1/(n - 1)
S[1,1]
```

```
[1] 7
```

```
t(d1)%*%d1/(n - 1)
```

```
       [,1]
[1,]     7
```

```
#s_22 = d2'd2/(n - 1)
S[2,2]
```

```
[1] 4
```

```
t(d2)%*%d2/(n - 1)
```

```
       [,1]
[1,]     4
```

```
#s_12 = d1'd2/(n - 1)
S[1, 2]
```

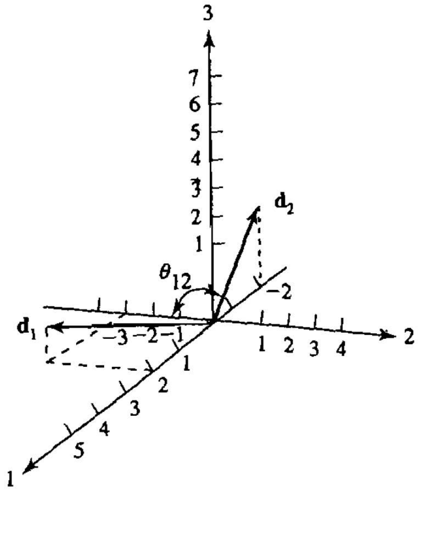
```
[1] -1
```

```
t(d1)%*%d2/(n - 1)
```

```
       [,1]
[1,]    -1
```

Finally, using the fact that the cosine of the angle between two vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be expressed as $\cos(\theta) = \dfrac{\boldsymbol{x}_1'\boldsymbol{x}_2}{L_{\boldsymbol{x}_1}L_{\boldsymbol{x}_2}}$ (show this!), we can show that the cosine of the angle between two deviation vectors is equal to the correlation coefficient between the corresponding vectors:

$$
\begin{aligned}
\cos(\theta) &= \frac{\boldsymbol{d}_1'\boldsymbol{d}_2}{L_{\boldsymbol{d}_1}L_{\boldsymbol{d}_2}} \\
&= \frac{\sum_{j=1}^n (x_{j1}-\bar{x}_1)(x_{j2}-\bar{x}_2)}{\sqrt{\sum_{j=1}^n (x_{j1}-\bar{x}_1)^2}\sqrt{\sum_{j=1}^n (x_{j2}-\bar{x}_2)^2}} \\
&= \frac{\frac{1}{n-1}\sum_{j=1}^n (x_{j1}-\bar{x}_1)(x_{j2}-\bar{x}_2)}{\sqrt{\frac{1}{n-1}\sum_{j=1}^n (x_{j1}-\bar{x}_1)^2}\sqrt{\frac{1}{n-1}\sum_{j=1}^n (x_{j2}-\bar{x}_2)^2}} \\
&= \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} \\
&= r_{x_1x_2}
\end{aligned}
$$



Summarising this information, we have:

1. The projection of a column $\boldsymbol{x}_i$ of the data matrix $\boldsymbol{X}$ into the equal angular vector $\mathbf{1}$ is the vector $\bar{x}_i\mathbf{1}$ with length $\sqrt{n}|\bar{x}_i|$, i.e., the i^{th} sample mean is related to the projection of $\boldsymbol{x}_i$ on $\mathbf{1}$.
2. The information comprising $\boldsymbol{S}$ is obtained from the deviation vectors $\boldsymbol{d}_i = \boldsymbol{x}_i - \bar{x}_i\mathbf{1}$. The square of the length of $\boldsymbol{d}_i$ is $(n_i - 1)s_{ii}$ and the inner product between $\boldsymbol{d}_i$ and $\boldsymbol{d}_k$ is $(n-1)s_{ik}$.
3. The sample correlation coefficient r_{ik} is the cosine of the angle between $\boldsymbol{d}_i$ and $\boldsymbol{d}_k$.

Homework exercise 1.1

Johnson & Wichern exercise 3.1.

Given the data matrix

$$\mathbf{X} = \begin{bmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{bmatrix}$$

1. Graph the scatterplot in $p = 2$ dimensions. Locate the sample mean on your diagram.
2. Sketch the $n = 3$ dimensional representation of the data, and plot the two deviation vectors.
3. Sketch the deviation vectors in (2) emanating from the origin. Calculate the lengths of these vectors and the cosine of the angle between them. Relate these quantities to $\boldsymbol{S}$ and $\boldsymbol{R}$.

1.4 Random Vectors

Let $\boldsymbol{X}_{p\times 1}$ be a random vector whose components are random variables $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$ where each random variable X_j has its own marginal probability distribution $f_{X_j}(x)$. Then we can define the *population* mean vector and the *population* covariance and correlation matrices:

Expected value of $\boldsymbol{X}$

$$\boldsymbol{\mu}_{p\times 1} = \mathrm{E}(\boldsymbol{X}) = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

where

$$\mu_j = \mathrm{E}(X_j) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x f_{X_j}(x)\, dx \\ \displaystyle\sum_x x \Pr(X_j = x) \end{cases}$$

Covariance matrix of $\boldsymbol{X}$

The covariance matrix $\boldsymbol{\Sigma}_{p\times p}$ is symmetric, positive semi-definite, and contains the pairwise covariances:

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\right]$$

with

$$\sigma_{jj} = \mathrm{Var}(X_j) = \mathrm{E}[(X_j-\mu_j)^2]$$

and

$$\sigma_{jk} = \mathrm{Cov}(X_j, X_k) = \mathrm{E}[(X_j-\mu_j)(X_k-\mu_k)]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Correlation matrix of $\boldsymbol{X}$

Similarly, the correlation matrix $\boldsymbol{P}_{p\times p}$ contains the pairwise correlations:

$$\boldsymbol{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

which can be written as

$$\boldsymbol{P} = [diag(\boldsymbol{\Sigma})]^{-\frac{1}{2}}\boldsymbol{\Sigma}[diag(\boldsymbol{\Sigma})]^{-\frac{1}{2}}$$

1.4.1 Moments of the sample mean and sample covariance matrix

We consider the rows of the $\boldsymbol{X}$ matrix to represent a random sample from the joint distribution of the p variables with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Therefore, $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are random variables! We'll look at their sampling distributions in detail later, for now we'll just focus on their first central moments:

$\mathrm{E}(\bar{\boldsymbol{X}}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\bar{\boldsymbol{X}}) = \dfrac{1}{n}\boldsymbol{\Sigma}$

$\mathrm{E}(\boldsymbol{S}) = \boldsymbol{\Sigma}$

Therefore, $\boldsymbol{S} = \dfrac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$ is an unbiased estimator of $\boldsymbol{\Sigma}$.

1.4.2 Linear combinations of variables

Let $\boldsymbol{X}_{p\times 1}$ be a random vector with $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Now consider the following two linear combinations of the variables contained in $\boldsymbol{X}$:

$\boldsymbol{b}'\boldsymbol{X} = b_1X_1 + b_2X_2 + \ldots + b_pX_p$ and $\boldsymbol{c}'\boldsymbol{X} = c_1X_1 + c_2X_2 + \ldots + c_pX_p$.

Then

1. the mean of $\boldsymbol{b}'\boldsymbol{X}$ is $\mathrm{E}(\boldsymbol{b}'\boldsymbol{X}) = \boldsymbol{b}'\boldsymbol{\mu}$. Likewise $\mathrm{E}(\boldsymbol{c}'\boldsymbol{X}) = \boldsymbol{c}'\boldsymbol{\mu}$.
2. the variance of $\boldsymbol{b}'\boldsymbol{X}$ is $\mathrm{Var}(\boldsymbol{b}'\boldsymbol{X}) = \boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{b}$. Likewise $\mathrm{Var}(\boldsymbol{c}'\boldsymbol{X}) = \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}$.
3. the covariance of $\boldsymbol{b}'\boldsymbol{X}$ and $\boldsymbol{c}'\boldsymbol{X}$ is $\mathrm{Cov}(\boldsymbol{b}'\boldsymbol{X}, \boldsymbol{c}'\boldsymbol{X}) = \boldsymbol{b}'\boldsymbol{\Sigma}\boldsymbol{c}$.

Homework Exercise 1.2

Johnson & Wichern exercise 3.18

Energy consumption by state from the major sources, x_1 =petroleum, x_2 =natural gas, x_3 = hydroelectric power and x_4 = nuclear electric power is recorded in quadrillions (10^{15}). The resulting mean and covariance matrix are

$$\bar{\boldsymbol{x}} = \begin{bmatrix} 0.766 \\ 0.508 \\ 0.438 \\ 0.161 \end{bmatrix}$$

and

$$\boldsymbol{S} = \begin{bmatrix} 0.856 & 0.635 & 0.173 & 0.096 \\ 0.635 & 0.568 & 0.128 & 0.067 \\ 0.173 & 0.128 & 0.171 & 0.039 \\ 0.096 & 0.067 & 0.039 & 0.043 \end{bmatrix}$$

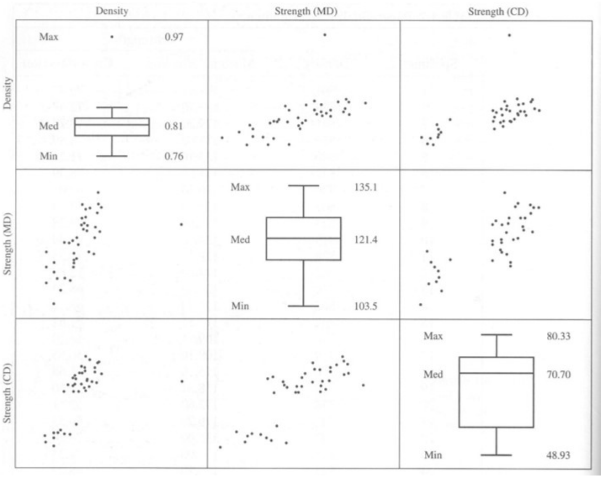
1. Determine the sample mean and variance of a state's total energy consumption for these major sources.
2. Determine the sample mean and variance of the excess of petroleum consumption over natural gas consumption. Also find the sample covariance of this variable with the total variable in part 1.

1.5 Exploratory Plots for Multivariate Data

The starting point of exploring a multivariate dataset is to view the scatterplots of all the pairwise combinations of variables, referred to as a scatterplot matrix. It also useful to view some summary of the distribution of each variable on the diagonal, in the form of a boxplot, density plot, etc. This can be done in several different ways in R; for the following examples, see `Lecture1_2.R`.

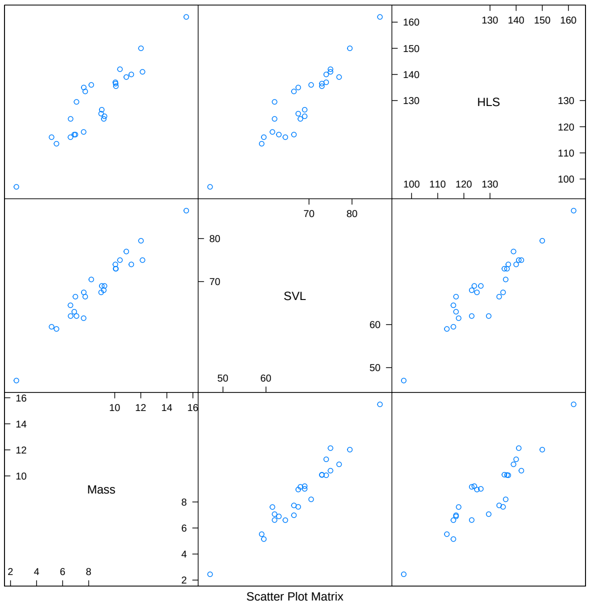
Scatterplot matrix

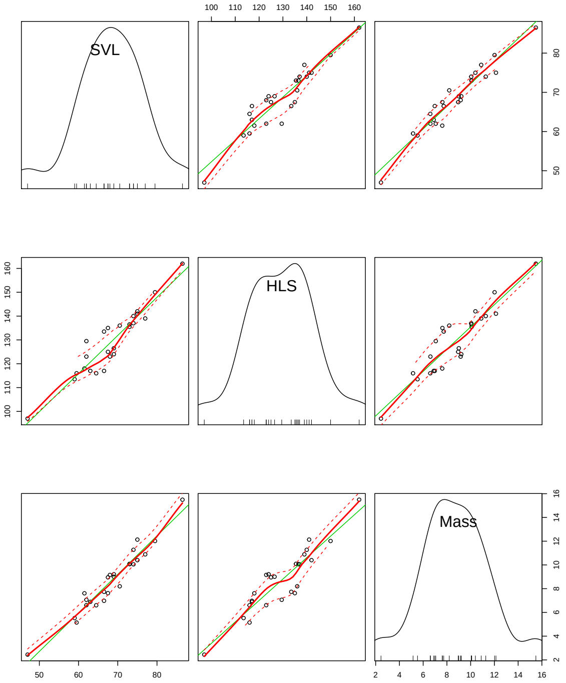
Johnson & Wichern Figure 1.5 (p. 16):



Lizards dataset:

```
library(lattice)
splom(lizards)
library(car)
scatterplotMatrix(˜ SVL + HLS + Mass, data = lizard, regLine = list(col = ’green’, lwd = 1), col = ’black’,
smooth = list(col.smooth = ’red’, lty.smooth = 1, col.spread = ’red’, lty.spread = 2))
```

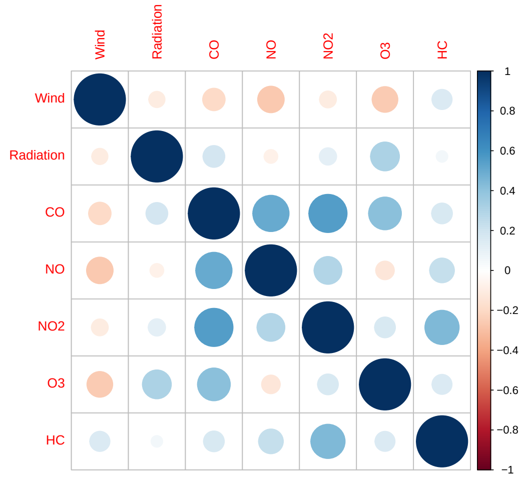


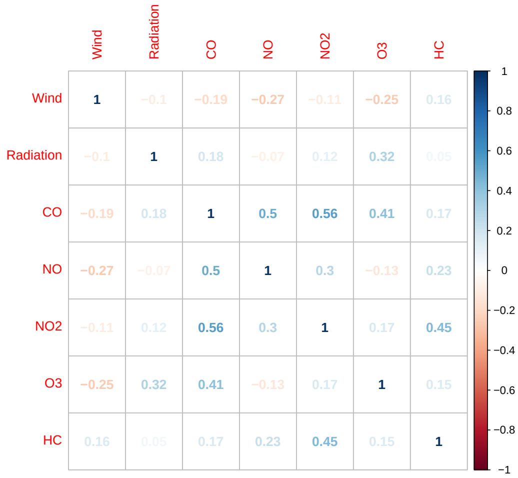


Heat map of covariance matrix

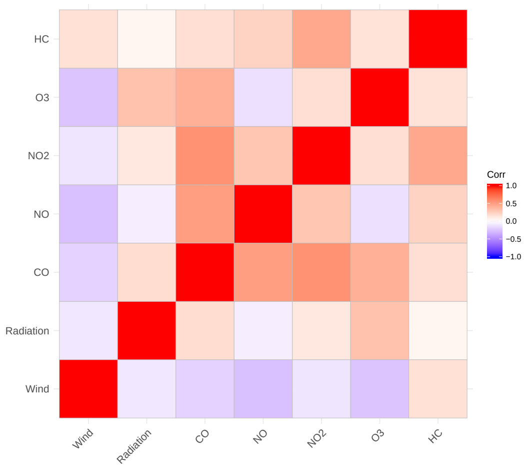
We are going to be interested in the associations among the multiple variables, as captured by the covariance or correlation matrices. This information can again be displayed in several different ways visually, although the prevailing principle is to illustrate the strength and direction of the bivariate relationships via colour coding.

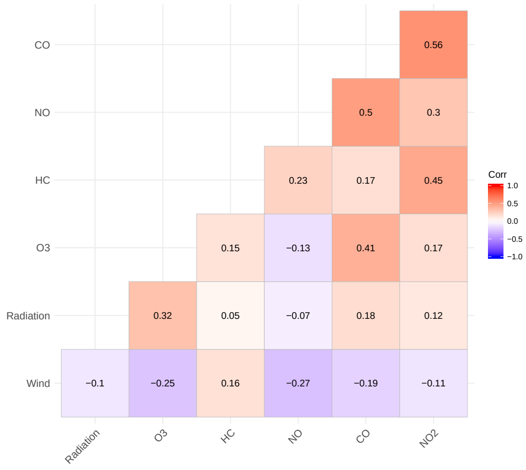
```
library(corrplot)
corrplot(AirPCor, method = 'circle')
corrplot(AirPCor, method = 'number')
```





```
library(ggcorrplot)
ggcorrplot(AirPCor)
ggcorrplot(AirPCor, hc.order = TRUE, type = "lower", lab = TRUE)
```





There are many different R libraries and functions for data visualization. Two R libraries that are particularly useful for multivariate data are the `lattice` library and the `ggplot2` library. A good book on the use of lattice is Seepayan Sarkar, “Lattice. Multivariate Visualization with R”, Springer, 2008.

There are also good blogs on the web that give you the ggplot2 functions that are equivalent to the lattice functions, for example, https://learnr.wordpress.com/2009/06/28/ggplot2-version-of-figures-in-lattice-multivariate-data-visualization-with-r-part-1/.

An excellent package for 3D visualisation is `RGL`.

Homework Exercise 1.3

Use R to visualize the data on National Track records for Men in a creative manner. Prepare a presentation of your visualization that shows both the graphs and the R-code that generated the graphs, as well as an interpretation of what the graphs are telling you about the variables and the associations in the data.

So far we have concentrated on pairwise associations of the many variables or on at most 3-dimensional displays. But how do we visualize multivariate data with many more variables, or equivalently, much higher dimensions? This is going to be a topic of much discussion going forward in this course.

2 Decomposition of Matrices

In the matrix methods pre-course you will have learnt about

1. Eigendecomposition
2. Spectral decomposition
3. Singular value decomposition (SVD)

These concepts will be important going forward, especially SVD as a method for dimension reduction. All these decompositions are inter-related, and are based on the notions of eigenvalues and eigenvectors. We will now consider a brief overview of the most pertinent aspects of eigenvalues and eigenvectors.

2.1 Eigenvectors and Eigenvalues Recap

Given a square matrix, $\boldsymbol{A}_{k\times k}$, we define a vector-scalar pair $\boldsymbol{x} \neq \boldsymbol{0}$ and λ such that

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \tag{1}$$

$\boldsymbol{x}$ is said to be an eigenvector of $\boldsymbol{A}$ and λ its corresponding eigenvalue.

There can be at most k unique eigenvalue-eigenvector pairs:

$$\lambda_1, \boldsymbol{e}_1 \quad \lambda_2, \boldsymbol{e}_2 \quad \ldots \quad \lambda_k, \boldsymbol{e}_k \quad ,$$

although a matrix can have repeated eigenvalues (and therefore repeated corresponding eigenvectors). Note that these eigenvectors aren't unique and are usually normalised such that their lengths are equal to 1.

Finding a solution to $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ is equivalent to solving $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = 0$, which only has a non-trivial solution if $|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$. This is referred to as the characteristic equation.

Geometrically, an eigenvector is one of which the span remains unchanged by the transformation of $\boldsymbol{A}$, whilst the eigenvalue is the factor by which vectors lying on this span are stretched during the transformation.

Spatial interpretation examples

Consider a sample covariance matrix, $\boldsymbol{S}_{k\times k}$. The eigenvalue-eigenvector pairs of $\boldsymbol{S}$ allow us to represent the magnitude and direction respectively of the k main vectors of variation present in the sample.

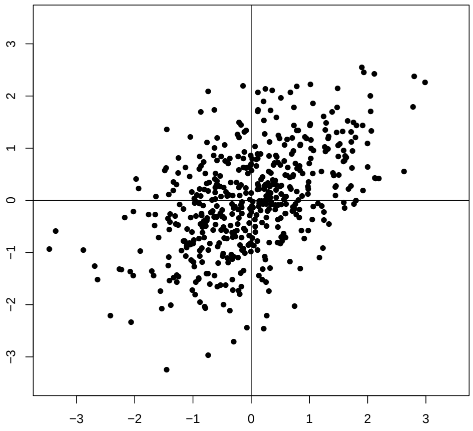
For example, consider a sample of 500 values from a bivariate distribution, shown below, with sample covariance matrix, $\boldsymbol{S} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

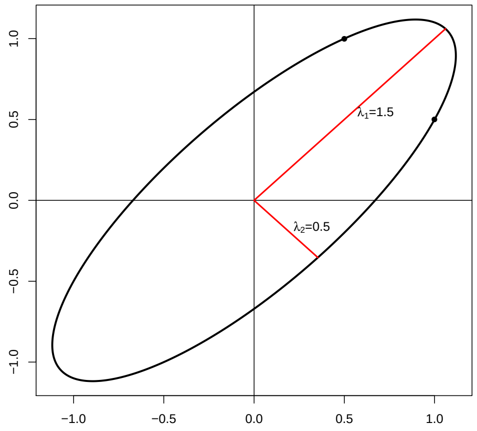
The eigenvalue and normalised eigenvector pairs for $\boldsymbol{S}$ are

$$\lambda_1 = 1.5 \quad \boldsymbol{e}_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \quad \text{and} \quad \lambda_2 = 0.5 \quad \boldsymbol{e}_2 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

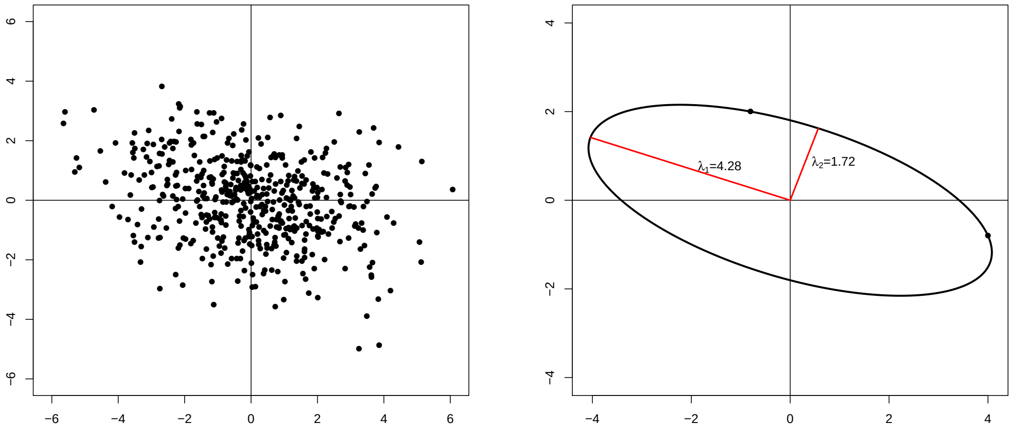
It is left as an exercise to confirm this.

We can now create two vectors, of which the directions are determined by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ respectively, and the lengths are equal to λ_1 and λ_2 respectively. An ellipse with the largest of these vectors as the primary axis and the smallest vector as the secondary axis will then capture the shape of the bivariate distribution. Note that since the data have already been centred, the ellipse goes through the points (s_{11}, s_{12}) and (s_{21}, s_{22}):





Below is the illustration for another example with $\boldsymbol{S} = \begin{bmatrix} 4 & -0.8 \\ -0.8 & 2 \end{bmatrix}$, yielding $\lambda_1 = 4.28$; $\boldsymbol{e}_1 = \begin{bmatrix} -0.94 \\ 0.33 \end{bmatrix}$ and $\lambda_2 = 1.72$; $\boldsymbol{e}_2 = \begin{bmatrix} -0.33 \\ -0.94 \end{bmatrix}$.



This principle can of course be extended to higher dimensions, with k eigenvalue-eigenvector pairs then forming a k-dimensional hyperellipsoid, at which point we lose the ability to inspect the prominent sources of variation visually. However, matrix decomposition techniques allow us to find a low-rank matrix approximations of matrices of interest.

2.2 Matrix Decomposition

Before we can define Singular Value Decomposition, we first need to discuss eigendecomposition and spectral decomposition.

2.2.1 Eigendecomposition

Let $\boldsymbol{A}$ be a $k \times k$, symmetric, positive definite matrix with $\lambda_i > 0 \quad \forall\ i$ and $\boldsymbol{e}_i,\ i = 1, \ldots, k$ indicating the k eigenvalue-eigenvector pairs.

If we arrange $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k$ into the matrix $\boldsymbol{U}$, and define $\boldsymbol{D} = diag(\lambda_i)$, then the eigendecomposition of $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \boldsymbol{UDU'}.$$

It can be shown that since $\boldsymbol{A}$ is symmetric, $\boldsymbol{U}$ is orthogonal. Therefore, $\boldsymbol{UU'} = \boldsymbol{I}$ and $\boldsymbol{U}' = \boldsymbol{U}^{-1}$. We can then write the above as

$$\boldsymbol{AU} = \boldsymbol{UD}$$

which corresponds to (1).

2.2.2 Spectral decomposition

Again define $\boldsymbol{A}$ as above. A different way of writing the eigendecomposition is by expressing $\boldsymbol{A}$ as the sum of k rank 1 matrices derived from the eigenvalue-eigenvector pairs. This is referred to as spectral decomposition:

$$\boldsymbol{A} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2' + \ldots + \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k' = \sum_{i=1}^{k} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i'$$

Square-root matrix

As a brief side-note, the spectral decomposition also helps us define a *square-root* matrix, as used in the definition of the correlation matrix, by allowing us to express the inverse of a square matrix in terms of its eigenvalues and eigenvectors. First, we note that $(\boldsymbol{UDU'})(\boldsymbol{UD^{-1}U'}) = \boldsymbol{I}$, which means

$$\boldsymbol{A}^{-1} = \boldsymbol{UD^{-1}U'} = \sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i'$$

Next, let $\boldsymbol{D}^{\frac{1}{2}}$ denote the diagonal matrix $diag(\sqrt{\lambda_i})$. The square-root matrix of a positive definite matrix $\boldsymbol{A}$ can then be defined as:

$$\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{UD^{\frac{1}{2}}U'} = \sum_{i=1}^{k} \sqrt{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i'$$

with the following properties:

1. $\boldsymbol{A}^{\frac{1}{2}}$ is symmetric: $\left(\boldsymbol{A}^{\frac{1}{2}}\right)' = \boldsymbol{A}^{\frac{1}{2}}$
2. $\boldsymbol{A}^{\frac{1}{2}} \boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{A}$
3. $\left(\boldsymbol{A}^{\frac{1}{2}}\right)^{-1} = \boldsymbol{A}^{-\frac{1}{2}} = \sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}} \boldsymbol{e}_i \boldsymbol{e}_i' = \boldsymbol{UD^{-\frac{1}{2}}U'}$, where $\boldsymbol{D}^{-\frac{1}{2}} = diag\left(\frac{1}{\sqrt{\lambda_i}}\right)$
4. $\boldsymbol{A}^{\frac{1}{2}} \boldsymbol{A}^{-\frac{1}{2}} = \boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{I}$, and $\boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{A}^{-\frac{1}{2}} = \boldsymbol{A}^{-1}$

2.2.3 Singular Value Decomposition

Consider $\boldsymbol{A}$, an $m \times k$ matrix of real numbers. There exists an $m \times m$ orthogonal matrix $\boldsymbol{U}$ and a $k \times k$ orthogonal matrix $\boldsymbol{V}$ such that

$$\boldsymbol{A} = \boldsymbol{U \Lambda V'}$$

where

- $\boldsymbol{U}$ contains the eigenvectors of $\boldsymbol{AA'}$
- $\boldsymbol{V}$ contains the eigenvectors of $\boldsymbol{A'A}$
- $\boldsymbol{\Lambda}$ contains $diag(\sqrt{\lambda_i})$, where the λ_i are the descending non-zero eigenvalues of $\boldsymbol{A'A}$ (or $\boldsymbol{AA'}$). If $m > k$, then either the square root of the diagonal matrix of eigenvalues of $(\boldsymbol{A'A})_{k \times k}$ needs to be padded with $m - k$ rows of zeros; or the last $m - k$ columns of the square root of the diagonal matrix of eigenvalues of $(\boldsymbol{AA'})_{m \times m}$ must be omitted.

One practical problem of applying SVD is that eigenvectors are defined in an arbitrary directional sense, i.e. if $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then so is $-\boldsymbol{x}$. One needs to ensure the directions of the eigenvectors in $\boldsymbol{U}$ and $\boldsymbol{V}$ are consistently defined.

If $\boldsymbol{A}$ has rank r, then there exists r positive constants $\lambda_1, \ldots, \lambda_r$, r mutually orthogonal $m \times 1$ vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ and r mutually orthogonal $k \times 1$ vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ such that

$$\boldsymbol{A} = \sum_{i=1}^{r} \sqrt{\lambda_i} \boldsymbol{u}_i \boldsymbol{v}_i'$$

similar to the spectral decomposition theorem. Note that $r \leq \min(m, k)$.

Example

Let $\boldsymbol{A} = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$

Then

$$\boldsymbol{A}\boldsymbol{A}' = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$$

To find the eigenvalues, we solve the characteristic equation:

$$\begin{aligned} |\boldsymbol{A}\boldsymbol{A}' - \lambda \boldsymbol{I}| &= \begin{vmatrix} 11-\lambda & 1 \\ 1 & 11-\lambda \end{vmatrix} \\ &= (11-\lambda)^2 - 1 \\ &= \lambda^2 - 22\lambda + 120 \\ &= (\lambda - 12)(\lambda - 10) = 0 \end{aligned}$$

$\therefore \lambda_1 = 12$ and $\lambda_2 = 10$

For $\lambda_1 = 12$:

$$\begin{gathered} \boldsymbol{A}\boldsymbol{A}'\boldsymbol{x} = \lambda_1 \boldsymbol{x} \\ \begin{bmatrix} 11x_1 + x_2 \\ x_1 + 11x_2 \end{bmatrix} = \begin{bmatrix} 12x_1 \\ 12x_2 \end{bmatrix} \end{gathered}$$

$\therefore x_1 = x_2$ such that $\boldsymbol{u}_1' = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ is a normalised eigenvector of $\lambda_1 = 12$. We can show likewise that $\boldsymbol{u}_2' = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$ is a normalised eigenvector of $\lambda_2 = 10$.

To find the columns of $\boldsymbol{V}$, consider

$$\boldsymbol{A}'\boldsymbol{A} = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}$$

such that

$$\begin{aligned} |\boldsymbol{A}'\boldsymbol{A} - \lambda \boldsymbol{I}| &= (10-\lambda)\left[(10-\lambda)(2-\lambda) - 16\right] + 2\left[-2(10-\lambda)\right] \\ &= -\lambda^3 + 22\lambda^2 - 120\lambda \\ &= -\lambda(\lambda - 12)(\lambda - 10) = 0 \end{aligned}$$

yielding eigenvalues $\lambda_1 = 12$, $\lambda_2 = 10$ and $\lambda_3 = 0$. Note that, as expected, the non-zero eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$ are the same as those of $\boldsymbol{A}\boldsymbol{A}'$.

Solving the sets of linear equations given by $\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ for each λ, we can determine the eigenvectors $\boldsymbol{v}_1' = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}$, $\boldsymbol{v}_2' = \begin{bmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & 0 \end{bmatrix}$ and $\boldsymbol{v}_3' = \begin{bmatrix} \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & -\frac{5}{\sqrt{30}} \end{bmatrix}$.

The SVD of $\boldsymbol{A}$ is therefore given by

$$\boldsymbol{A} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & -\frac{5}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

which can also be expressed as

$$\boldsymbol{A} = \sqrt{12} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} + \sqrt{10} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & 0 \end{bmatrix} = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

i.e. the sum of $r = 2$ matrices, both of rank 1.

R Illustration (see `Lecture2.R`)

```
#Create matrix from class example
A <- matrix(c(3, -1, 1, 3, 1, 1), nrow = 2)

#Creating the decomposition from scratch
(U <- eigen(A %*% t(A))$vectors)
(D <- diag(sqrt(eigen(A %*% t(A))$values)))
(V <- zapsmall(eigen(t(A) %*% A)$vectors))
```

```
> U
          [,1]       [,2]
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068

> D
         [,1]     [,2]
[1,] 3.464102 0.000000
[2,] 0.000000 3.162278

> V
           [,1]       [,2]       [,3]
[1,] -0.4082483  0.8944272  0.1825742
[2,] -0.8164966 -0.4472136  0.3651484
[3,] -0.4082483  0.0000000 -0.9128709
```

```
#Note that here we need the first position in each eigenvector to have the same sign
U[,2] <- -U[,2]
V[,1] <- -V[,1]
```

```
> U
          [,1]       [,2]
[1,] 0.7071068  0.7071068
[2,] 0.7071068 -0.7071068

> V
          [,1]       [,2]       [,3]
[1,] 0.4082483  0.8944272  0.1825742
[2,] 0.8164966 -0.4472136  0.3651484
[3,] 0.4082483  0.0000000 -0.9128709
```

```
#Checking the result
U %*% cbind(D, 0) %*% t(V)
```

```
     [,1] [,2] [,3]
[1,]    3    1    1
[2,]   -1    3    1
```

```
#Using the SVD function
(svd_func <- svd(A))
```

```
$d
[1] 3.464102 3.162278

$u
           [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,] -0.7071068  0.7071068

$v
           [,1]          [,2]
[1,] -0.4082483 -8.944272e-01
[2,] -0.8164966  4.472136e-01
[3,] -0.4082483  5.273559e-16
```

```
#Checking the result again
svd_func$u %*% diag(svd_func$d) %*% t(svd_func$v)
```

```
     [,1] [,2] [,3]
[1,]    3    1    1
[2,]   -1    3    1
```

Homework Exercise 2.1

Johnson & Wichern exercise 2.22

Using the matrix

$$\boldsymbol{A} = \begin{bmatrix} 4 & 8 & 8 \\ 3 & 6 & -9 \end{bmatrix}$$

a) Calculate $\boldsymbol{A}'\boldsymbol{A}$ and obtain its eigenvalues and eigenvectors.

b) Calculate $\boldsymbol{A}\boldsymbol{A}'$ and obtain its eigenvalues and eigenvectors. Check that the non-zero eigenvalues are the same as those in part a).

c) Obtain the singular value decomposition of $\boldsymbol{A}$.

2.3 Using SVD for lower-dimensional approximation

If $\boldsymbol{U\Lambda V'}$ is the SVD of $\boldsymbol{A}_{m\times k}$ where $m \geq k$, and $s < k = rank(\boldsymbol{A})$, then

$$\boldsymbol{B} = \sum_{i=1}^{s} \sqrt{\lambda_i}\boldsymbol{u_i v_i'}$$

is the rank s least squares approximation to $\boldsymbol{A}$ that results in the best approximation among all matrices of rank $\leq s$.

PROOF:

We use $\boldsymbol{UU'} = \boldsymbol{I}_m$ and $\boldsymbol{VV'} = \boldsymbol{I}_k$ to write the sum of squares as

$$\begin{aligned}
tr[(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}-\boldsymbol{B})'] &= tr[\boldsymbol{UU'}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{VV'}(\boldsymbol{A}-\boldsymbol{B})'] \\
&= tr[\boldsymbol{U'}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{VV'}(\boldsymbol{A}-\boldsymbol{B})'\boldsymbol{U}] \\
&= tr[\boldsymbol{U'}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{V}(\boldsymbol{U'}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{V})'] \\
&= tr[(\underbrace{\boldsymbol{U'AV}}_{\boldsymbol{\Lambda}} - \underbrace{\boldsymbol{U'BV}}_{\boldsymbol{C}})(\boldsymbol{U'AV} - \boldsymbol{U'BV})'] \\
&= tr[(\boldsymbol{\Lambda}-\boldsymbol{C})(\boldsymbol{\Lambda}-\boldsymbol{C})'] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{k}(\sqrt{\lambda_{ij}} - c_{ij})^2 \\
&= \sum_{i=1}^{m}(\sqrt{\lambda_i} - c_{ii})^2 + \underbrace{\sum\sum}_{i\neq j} c_{ij}^2
\end{aligned}$$

which will be a minimum when $c_{ij} = 0$ for all $i \neq j$ and $c_{ii} = \sqrt{\lambda_i}$ for the s largest singular values with the other $c_{ii} = 0$.

$$\therefore \boldsymbol{UBV'} = \boldsymbol{\Lambda}_s \quad \text{or} \quad \boldsymbol{B} = \sum_{i=1}^{s}\sqrt{\lambda_i}\boldsymbol{u}_i\boldsymbol{v}_i'$$

We will return to these decompositions when discussing Principal Components Analysis (PCA). Next, however, we will discuss the Multivariate Normal distribution.

3 The Multivariate Normal Distribution

The multivariate normal distribution, which will form the basis of multivariate maximum likelihood estimation and inference later on, is an extension of the univariate Normal (Gaussian) distribution.

3.1 Univariate Normal

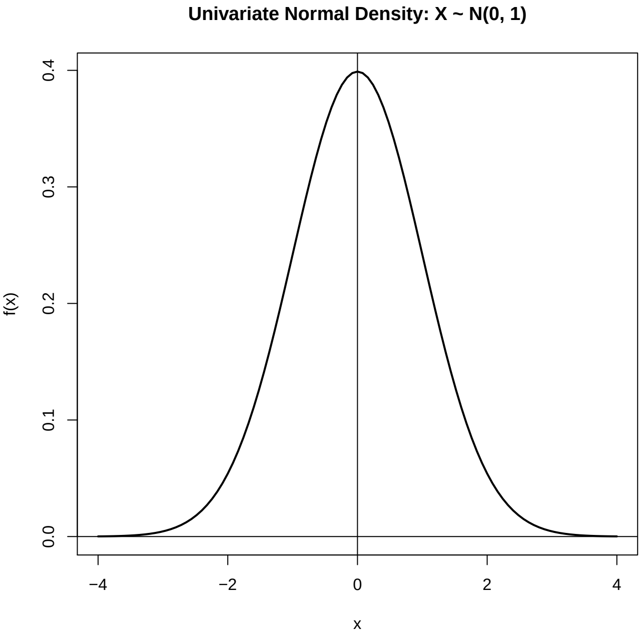
The univariate normal probability density function for a random variable $X \sim N(\mu, \sigma)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ -\infty < x < \infty$$

where

- $\dfrac{1}{\sigma\sqrt{2\pi}}$ is a normalizing constant
- $\dfrac{(x-\mu)^2}{\sigma^2} = \underbrace{(x-\mu)\sigma^{-2}(x-\mu)}_{\text{square of distance from } x \text{ to } \mu \text{ in std.dev. units}}$

The pdf of a standard normal distribution, with $\mu = 0$ and $\sigma = 1$, is shown here:



3.2 Multivariate Normal

Consider now $\boldsymbol{X}_{p\times 1}$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The multivariate generalized distance is

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

and the multivariate version of the normalizing constant is

$$(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}.$$

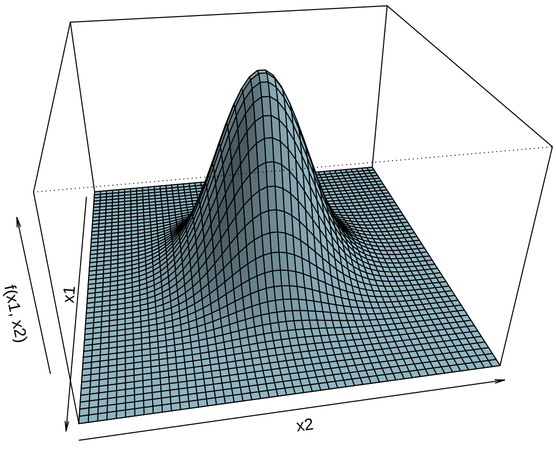
Thus the multivariate normal density function is

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

and we write $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

3.2.1 Bivariate normal Density

Consider now the case where $p = 2$, referred to as the bivariate normal distribution. We can firstly visualise the joint distribution of X_1 and X_2. For examples of several different ways of producing bivariate density plots in R, see `Lecture3.R`.



Let us explore the joint distribution of X_1 and X_2, by writing out the parameters:

- $\mu_1 = \mathrm{E}(X_1)$, $\mu_2 = \mathrm{E}(X_2)$,
- $\sigma_{11} = \mathrm{Var}(X_1)$, $\sigma_{22} = \mathrm{Var}(X_2)$, $\sigma_{12} = \sigma_{21} = \mathrm{Cov}(X_1, X_2)$
- $\rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} = \mathrm{Cor}(X_1, X_2)$

Using $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$, we can calculate

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} = \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix},$$

since $\sigma_{12}^2 = \sigma_{11}\sigma_{22}\rho_{12}^2$. Therefore, we have

$$\begin{aligned}
&(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \\
&= \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)} \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\
&= \frac{\sigma_{22}(x_1 - \mu_1)^2 + \sigma_{11}(x_2 - \mu_2)^2 - 2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)} \\
&= \frac{1}{1 - \rho_{12}^2} \times \left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho_{12} \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right) \right]
\end{aligned}$$

We can write the normalizing constant as

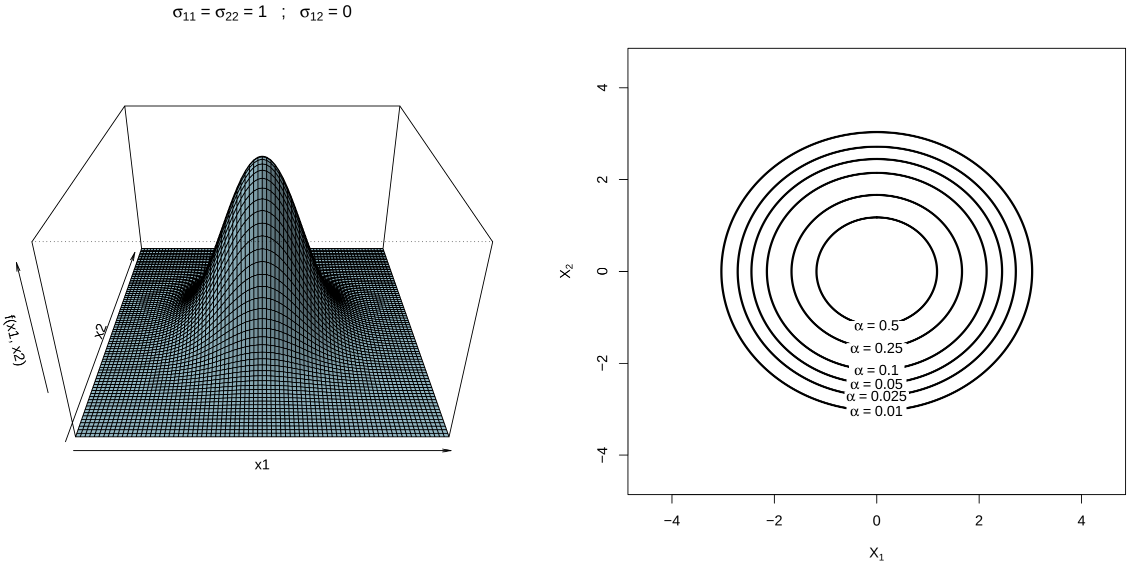
$$\frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{1/2}} = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}}$$

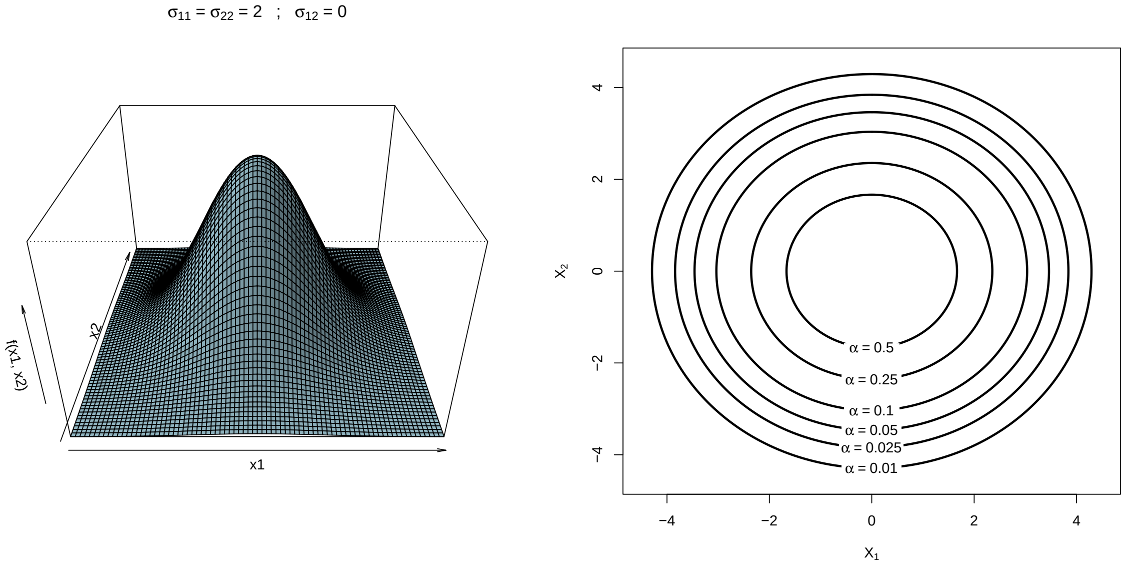
Thus

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \times e^{-\frac{1}{2(1-\rho_{12}^2)} \times \left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]}$$

If X_1 and X_2 are uncorrelated, $\rho_{12} = 0$ and $f(x_1, x_2) = f(x_1) \times f(x_2)$.

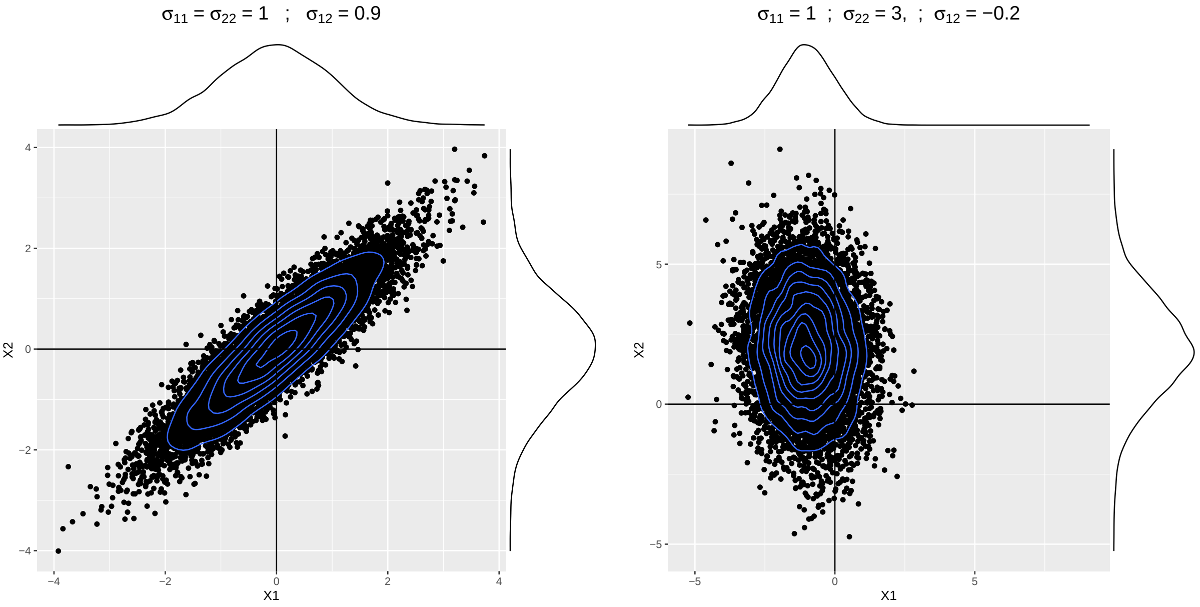
The shape of a bivariate distribution and the relationship between the variables can also be explored via a contour plot. Here we see the densities and corresponding contour lines for the distributions $N_2(\mathbf{0}, \boldsymbol{I})$ and $N_2(\mathbf{0}, 2\boldsymbol{I})$ respectively.





We can also draw contours around data simulated from bivariate normal distributions. In the following figures, 10,000 data points have been randomly drawn from the distributions $N_2\left(\mathbf{0}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$

and $N_2\left(\begin{bmatrix}-1\\2\end{bmatrix}, \begin{bmatrix}1 & -0.2\\-0.2 & 3\end{bmatrix}\right)$ respectively. Note the marginal normal distributions on both axes.



For the code corresponding to these contour plots, as well as more of ways of displaying contours, see `Lecture3.R`.

Homework exercise 3.1

Johnson & Wichern exercise 4.2

Consider a bivariate normal population with

$$\mu_1 = 0,\ \mu_2 = 2,\ \sigma_{11} = 2,\ \sigma_{22} = 1,\ \rho_{12} = 0.5$$

1. Write out the bivariate normal density.
2. Write out the squared generalized distance expression $(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ as a function of x_1 and x_2.

3.2.2 Graphical interpretation of bivariate normal density

Considering these contour plots, we see that the multivariate normal density is constant on surfaces where the square of the distance from $\boldsymbol{x}$ to the mean, $\boldsymbol{\mu}$ is constant. These constant probability contours are defined as all $\boldsymbol{x}$ such that $(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c^2$. The axes of each ellipsoid of constant density are in the direction of the eigenvectors of $\boldsymbol{\Sigma}^{-1}$ and their lengths are proportional to the reciprocals of the square roots of the eigenvalues of $\boldsymbol{\Sigma}^{-1}$. The following result allows us to work with the original covariance matrix, instead of $\boldsymbol{\Sigma}^{-1}$:

If $\boldsymbol{\Sigma}$ is positive definite such that $\boldsymbol{\Sigma}^{-1}$ exists, then $\boldsymbol{\Sigma}\boldsymbol{e} = \lambda\boldsymbol{e}$ implies $\boldsymbol{\Sigma}^{-1}\boldsymbol{e} = \dfrac{1}{\lambda}\boldsymbol{e}$, so $(\lambda,\ \boldsymbol{e})$ is an eigenvalue-eigenvector pair for $\boldsymbol{\Sigma}$ corresponding to the pair $\left(\dfrac{1}{\lambda},\ \boldsymbol{e}\right)$ for $\boldsymbol{\Sigma}^{-1}$.

Therefore, the axes are $\pm c\sqrt{\lambda_i}\boldsymbol{e}_i$, where the eigendecomposition of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$ for $i = 1, 2, \ldots, p$. We will return to this property and consider the value of c shortly, after consider the case where $p = 2$.

Bivariate normal density example

Assume $\sigma_{11} = \sigma_{22}$. Thus the characteristic equation $|\boldsymbol{\Sigma} - \lambda \boldsymbol{I}| = 0$ becomes

$$\begin{aligned} 0 &= \begin{vmatrix} \sigma_{11} - \lambda & \sigma_{12} \\ \sigma_{12} & \sigma_{11-\lambda} \end{vmatrix} \\ &= (\sigma_{11} - \lambda)^2 - \sigma_{12}^2 \\ &= (\lambda - \sigma_{11} - \sigma_{12})(\lambda - \sigma_{11} + \sigma_{12}) \end{aligned}$$

$\therefore$ the eigenvalues are $\lambda_1 = \sigma_{11} + \sigma_{12}$ and $\lambda_2 = \sigma_{11} - \sigma_{12}$.

The first eigenvector, $\boldsymbol{e}_1$, is then determined from

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = (\sigma_{11} + \sigma_{12}) \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

which gives us

$$\begin{aligned} \sigma_{11}e_1 + \sigma_{12}e_2 &= (\sigma_{11} + \sigma_{12})e_1 \\ \sigma_{12}e_1 + \sigma_{11}e_2 &= (\sigma_{11} + \sigma_{12})e_2 \end{aligned}$$

$\implies e_1 = e_2$.

After normalization we have that the first eigenvalue-eigenvector pair is

$$\lambda_1 = \sigma_{11} + \sigma_{12}, \quad \boldsymbol{e}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

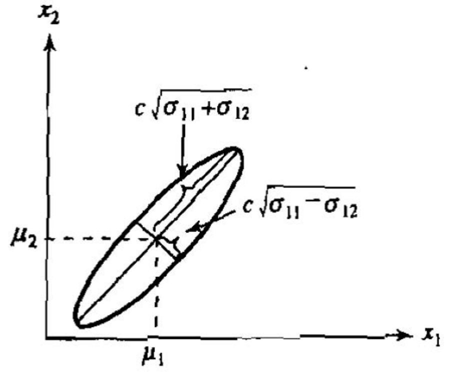
Likewise, the second pair is

$$\lambda_2 = \sigma_{11} - \sigma_{12}, \quad \boldsymbol{e}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

In this example, because $\sigma_{11} = \sigma_{22}$, the eigenvector $\boldsymbol{e}_1$ associated with λ_1 will lie on a 45° line through the point $\boldsymbol{\mu}$, regardless of the size of ρ_{12}. Since $\boldsymbol{\Sigma}$ is symmetric, the eigenvectors will be orthogonal such that $\boldsymbol{e}_2$ will lie perpendicular to $\boldsymbol{e}_1$.

If the two variables are positively correlated – i.e. both σ_{12} and ρ_{12} are larger than zero – then $\lambda_1 > \lambda_2$; for negative correlation we will have $\lambda_2 > \lambda_1$. Because the axes of the constant-density ellipse are are given by $c\sqrt{\lambda_1}\boldsymbol{e}_1$ and $c\sqrt{\lambda_2}\boldsymbol{e}_2$, and the eigenvectors have been normalized, the major axis of the ellipse will correspond to the largest eigenvalue, which indicates the direction of the association between the variables.

The figure below illustrates the bivariate case where $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} > 0$.


x_2
$c\sqrt{\sigma_{11}+\sigma_{12}}$
$c\sqrt{\sigma_{11}-\sigma_{12}}$
μ_2
μ_1
x_1

3.3 The Probability Content of the Ellipsoids of Constant Density

Let $\boldsymbol{X}$ be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| > 0$. Then

1. $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ is distributed as χ^2_p.
2. The $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution assigns a probability of $1 - \alpha$ to the solid ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \le \chi^2_p(\alpha)\},$$

where $\chi^2_p(\alpha)$ denotes the upper (100α)th percentile of the χ^2_p distribution. That is,

$$\Pr\left[(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \le \chi^2_p(\alpha)\right] = 1 - \alpha.$$

Homework exercise 3.2

1. Prove the above by using the spectral decomposition of the covariance matrix.
2. Determine (and sketch) the constant-density contour that contains 90% of the probability for the examples in exercise 3.1.

3.4 Properties of the Multivariate Normal Distribution

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We then have the following useful properties:

3.4.1 Linear combinations of the components of X are normally distributed

Result 1

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then ANY linear combination

$$\boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \ldots + a_pX_p \sim N(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$$

and vice-versa, if $\boldsymbol{a}'\boldsymbol{X} \sim N(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$ for every $\boldsymbol{a}$, then $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For example:

Let $\boldsymbol{a}' = [1, 0, \ldots, 0]$, then

$$a'X = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = X_1$$

$$\boldsymbol{a}'\boldsymbol{\mu} = \mu_1$$

$$\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \sigma_{11}$$

$$\rightarrow \boldsymbol{a}'\boldsymbol{X} \sim N(\mu_1, \sigma_{11})$$

Also, $\boldsymbol{X} + \boldsymbol{d} \sim N_p(\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{\Sigma})$ where $\boldsymbol{d}$ is a vector of constants.

Homework exercise 3.3

Johnson & Wichern exercise 4.4(a)

Given $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}' = [2,\ -3,\ 1]$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$,

find the distribution of $3X_1 - 2X_2 + X_3$.

Result 2

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then q linear combinations,

$$\underbrace{\boldsymbol{A}}_{q\times p}\underbrace{\boldsymbol{X}}_{p\times 1} = \begin{bmatrix} a_{11}X_1 + \ldots + a_{1p}X_p \\ a_{21}X_1 + \ldots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + \ldots + a_{qp}X_p \end{bmatrix} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}').$$

For example:

If $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$, then

$$\boldsymbol{AX} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \end{bmatrix}$$

$$\boldsymbol{A}\mu = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \end{bmatrix}$$

$$\boldsymbol{A\Sigma A}' = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

3.4.2 All subsets of the components of X have a (multivariate) normal distribution

If we partition $\boldsymbol{X}$ as

$$\boldsymbol{X}_{p\times 1} = \left[\begin{array}{c} \underbrace{\boldsymbol{X}_1}_{q\times 1} \\ \hline \underbrace{\boldsymbol{X}_2}_{(p-q)\times 1} \end{array}\right]$$

then

$$\boldsymbol{\mu}_{p\times 1} = \left[\begin{array}{c} \underbrace{\boldsymbol{\mu}_1}_{q\times 1} \\ \hline \underbrace{\boldsymbol{\mu}_2}_{(p-q)\times 1} \end{array}\right]$$

and

$$\boldsymbol{\Sigma}_{p\times p} = \left[\begin{array}{c|c} \underbrace{\boldsymbol{\Sigma}_{11}}_{q\times q} & \underbrace{\boldsymbol{\Sigma}_{12}}_{q\times (p-q)} \\ \hline \underbrace{\boldsymbol{\Sigma}_{21}}_{(p-q)\times q} & \underbrace{\boldsymbol{\Sigma}_{22}}_{(p-q)\times (p-q)} \end{array}\right]$$

yielding $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{(p-q)}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

The result follows from defining $\boldsymbol{A} = \left[\begin{array}{c|c} \underbrace{\boldsymbol{I}}_{q\times q} & \underbrace{\boldsymbol{0}}_{q\times (p-q)} \end{array}\right]$ and applying Result 2.

For example,

If $\boldsymbol{X} \sim N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ find the distribution of $\begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$.

Let $\boldsymbol{X}_1 = \begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$, $\boldsymbol{\mu}_1 = \begin{bmatrix} \mu_2 \\ \mu_4 \end{bmatrix}$, $\boldsymbol{\Sigma}_{11} = \begin{bmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{bmatrix}$

and then partition as follows:

$$\boldsymbol{X} = \left[\begin{array}{c} X_2 \\ X_4 \\ \hline X_1 \\ X_3 \\ X_5 \end{array}\right], \ \boldsymbol{\mu} = \left[\begin{array}{c} \mu_2 \\ \mu_4 \\ \hline \mu_1 \\ \mu_3 \\ \mu_5 \end{array}\right], \ \boldsymbol{\Sigma} = \left[\begin{array}{cc|ccc} \sigma_{22} & \sigma_{24} & \sigma_{12} & \sigma_{23} & \sigma_{25} \\ \sigma_{24} & \sigma_{44} & \sigma_{14} & \sigma_{34} & \sigma_{45} \\ \hline \sigma_{12} & \sigma_{14} & \sigma_{11} & \sigma_{13} & \sigma_{15} \\ \sigma_{23} & \sigma_{34} & \sigma_{13} & \sigma_{33} & \sigma_{35} \\ \sigma_{25} & \sigma_{45} & \sigma_{15} & \sigma_{35} & \sigma_{55} \end{array}\right]$$

$$\implies \boldsymbol{X}_1 = \begin{bmatrix} X_2 \\ X_4 \end{bmatrix} \sim N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) = N_2\left(\begin{bmatrix} \mu_2 \\ \mu_4 \end{bmatrix}, \begin{bmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{bmatrix}\right).$$

3.4.3 Zero covariance implies that the corresponding components are independently distributed

There are 3 results to consider:

1. If $\underbrace{\boldsymbol{X}_1}_{q_1\times 1}$ and $\underbrace{\boldsymbol{X}_2}_{q_2\times 1}$ are independent, then $\mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \underbrace{\boldsymbol{0}}_{q_1\times q_2}$.

2. If $\left[\begin{array}{c}\boldsymbol{X}_1\\ \hline \boldsymbol{X}_2\end{array}\right] \sim N_{q_1+q_2}\left(\left[\begin{array}{c}\boldsymbol{\mu}_1\\ \hline \boldsymbol{\mu}_2\end{array}\right], \left[\begin{array}{c|c}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{array}\right]\right)$ then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

3. If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent and are distributed as $N_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $N_{q_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, respectively, then $\left[\begin{array}{c}\boldsymbol{X}_1\\ \hline \boldsymbol{X}_2\end{array}\right] \sim N_{q_1+q_2}\left(\left[\begin{array}{c}\boldsymbol{\mu}_1\\ \hline \boldsymbol{\mu}_2\end{array}\right], \left[\begin{array}{c|c}\boldsymbol{\Sigma}_{11} & \mathbf{0}\\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{array}\right]\right)$.

For example:

If we let $\underbrace{\boldsymbol{X}}_{3\times 1} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \begin{bmatrix}4 & 1 & 0\\ 1 & 3 & 0\\ 0 & 0 & 2\end{bmatrix}$, then X_1 and X_2 are NOT independent, since $\sigma_{12} = 1$.

BUT if we partition $\boldsymbol{X}$ as follows,

$\boldsymbol{X} = \left[\begin{array}{c}X_1\\ X_2\\ \hline X_3\end{array}\right]$, $\boldsymbol{\Sigma} = \left[\begin{array}{cc|c}4 & 1 & 0\\ 1 & 3 & 0\\ \hline 0 & 0 & 2\end{array}\right]$ $\rightarrow \boldsymbol{X}_1 = \begin{bmatrix}X_1\\ X_2\end{bmatrix}$ and X_3 have covariance matrix $\boldsymbol{\Sigma}_{12} = \begin{bmatrix}0\\ 0\end{bmatrix}$ and so $\begin{bmatrix}X_1\\ X_2\end{bmatrix}$ and X_3 are independent.

Homework exercise 3.4

Johnson & Wichern exercise 4.3

Let $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}' = [-3, 1, 4]$ and $\boldsymbol{\Sigma} = \begin{bmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{bmatrix}$.

Which of the following pairs of random variables are independent? Explain.

1. X_1 and X_2
2. X_2 and X_3
3. (X_1, X_2) and X_3
4. $\dfrac{X_1 + X_2}{2}$ and X_3

3.4.4 The conditional distributions of the components are (multivariate) normally distributed

If $\boldsymbol{X} = \left[\begin{array}{c}\underbrace{\boldsymbol{X}_1}_{q\times 1}\\ \hline \underbrace{\boldsymbol{X}_2}_{(p-q)\times 1}\end{array}\right] \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \left[\begin{array}{c}\boldsymbol{\mu}_1\\ \hline \boldsymbol{\mu}_2\end{array}\right]$, and $\boldsymbol{\Sigma} = \left[\begin{array}{c|c}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{array}\right]$, then the conditional distribution of $\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean

$$\mathrm{E}(\boldsymbol{X}_1|\boldsymbol{X}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)$$

and covariance matrix

$$\mathrm{Cov}(\boldsymbol{X}_1|\boldsymbol{X}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

PROOF:

If we let $\boldsymbol{A}_{p\times p} = \left[\begin{array}{c|c} \underbrace{\boldsymbol{I}}_{q\times q} & \underbrace{-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}}_{q\times(p-q)} \\ \hline \underbrace{\boldsymbol{0}}_{(p-q)\times q} & \underbrace{\boldsymbol{I}}_{(p-q)\times(p-q)} \end{array}\right]$, then $\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{A\Sigma A}')$ with

$\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{A}\left[\begin{array}{c} \boldsymbol{X}_1-\boldsymbol{\mu}_1 \\ \hline \boldsymbol{X}_2-\boldsymbol{\mu}_2 \end{array}\right] = \left[\begin{array}{c} \boldsymbol{X}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2-\boldsymbol{\mu}_2) \\ \hline \boldsymbol{X}_2-\boldsymbol{\mu}_2 \end{array}\right]$ and

$\boldsymbol{A\Sigma A}' = \left[\begin{array}{c|c} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \hline \boldsymbol{0} & \boldsymbol{I} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{0} \\ \hline (-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1})' & \boldsymbol{I} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{array}\right].$

$\implies \boldsymbol{X}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)$ and $\boldsymbol{X}_2-\boldsymbol{\mu}_2$ have zero covariance and are thus independent.

This implies that the conditional distribution of $\boldsymbol{X}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)|\boldsymbol{X}_2-\boldsymbol{\mu}_2$ is the same as the unconditional distribution

$\boldsymbol{X}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2-\boldsymbol{\mu}_2) \sim N_q(\boldsymbol{0}, \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$.

Therefore, the random vector $\boldsymbol{X}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)$ when we have substituted $\boldsymbol{X}_2=\boldsymbol{x}_2$ is also normally distributed.

Now if $\boldsymbol{X}_2=\boldsymbol{x}_2$, then $\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)$ is a constant.

$\implies \boldsymbol{X}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)|\boldsymbol{X}_2=\boldsymbol{x}_2 \sim N_q(\boldsymbol{0}, \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$

and

$\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$.

For the bivariate normal density this simplifies to

$X_1|X_2 \sim N\left(\mu_1+\frac{\sigma_{12}}{\sigma_{22}}(x_2-\mu_2), \sigma_{11}-\frac{\sigma_{12}^2}{\sigma_{22}}\right)$

Homework exercise 3.5

Johnson & Wichern 4.5(a)

Consider again the bivariate normal population with $\mu_1=0, \mu_2=2, \sigma_{11}=2, \sigma_{22}=1, \rho_{12}=0.5$.

Specify the conditional distribution of $X_1|X_2=x_2$.

Summary of the previous property

1. All conditional distributions are multivariate normal.

2. The conditional mean is of the form

$$\begin{matrix} \mu_1 + \beta_{1,q+1}(x_{q+1} - \mu_{q+1}) + \ldots + \beta_{1,p}(x_p - \mu_p) \\ \vdots \\ \mu_q + \beta_{q,q+1}(x_{q+1} - \mu_{q+1}) + \ldots + \beta_{q,p}(x_p - \mu_p) \end{matrix} ,$$

where the $\beta's$ are defined by

$$\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} = \begin{bmatrix} \beta_{1,q+1} & \beta_{1,q+2} & \cdots & \beta_{1,p} \\ \beta_{2,q+1} & \beta_{2,q+2} & \cdots & \beta_{2,p} \\ & & \vdots & \\ \beta_{q,q+1} & \beta_{q,q+2} & \cdots & \beta_{q,p} \end{bmatrix}$$

3. The conditional covariance $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ does not depend upon the values of the conditioning variables.

3.5 Linear combinations of random variable vectors

Consider $\underbrace{\boldsymbol{V}_1}_{p\times 1} = c_1\boldsymbol{X}_1 + c_2\boldsymbol{X}_2 + \ldots + c_n\boldsymbol{X}_n = \underbrace{\begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2 & \ldots & \boldsymbol{X}_n\end{bmatrix}}_{(p\times n)}\underbrace{\boldsymbol{c}}_{n\times 1}$,

where the $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are mutually independent with each $\boldsymbol{X}_j \sim N_p(\boldsymbol{\mu}_j, \mathbf{\Sigma})$.

Then $\boldsymbol{V}_1 \sim N_p\left(\sum_{j=1}^{n} c_j\boldsymbol{\mu}_j, \left(\sum_{j=1}^{n} c_j^2\right)\mathbf{\Sigma}\right)$.

Also, $\boldsymbol{V}_1$ and $\boldsymbol{V}_2 = b_1\boldsymbol{X}_1 + b_2\boldsymbol{X}_2 + \ldots + b_n\boldsymbol{X}_n$ are jointly multivariate normal with covariance matrix

$$\begin{bmatrix} \left(\sum_{j=1}^{n} c_j^2\right)\mathbf{\Sigma} & (\boldsymbol{b}'\boldsymbol{c})\mathbf{\Sigma} \\ (\boldsymbol{b}'\boldsymbol{c})\mathbf{\Sigma} & \left(\sum_{j=1}^{n} b_j^2\right)\mathbf{\Sigma} \end{bmatrix}$$

So $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are independent if $\boldsymbol{b}'\boldsymbol{c} = \sum_{j=1}^{n} c_j b_j = 0$.

Homework exercise 3.6

Johnson & Wichern exercises 4.16 and 4.17

Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ and $\boldsymbol{X}_4$ be independent $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ random vectors.

1. Find the marginal distributions of

$$\boldsymbol{V}_1 = \frac{1}{4}\boldsymbol{X}_1 - \frac{1}{4}\boldsymbol{X}_2 + \frac{1}{4}\boldsymbol{X}_3 - \frac{1}{4}\boldsymbol{X}_4$$

and

$$\boldsymbol{V}_2 = \frac{1}{4}\boldsymbol{X}_1 + \frac{1}{4}\boldsymbol{X}_2 - \frac{1}{4}\boldsymbol{X}_3 - \frac{1}{4}\boldsymbol{X}_4$$

2. Find the joint density of $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$.

Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, $\boldsymbol{X}_4$ and $\boldsymbol{X}_5$ be independent and identically distributed random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

3. Find the mean vector and covariance matrices for

$$\frac{1}{5}\boldsymbol{X}_1 + \frac{1}{5}\boldsymbol{X}_2 + \frac{1}{5}\boldsymbol{X}_3 + \frac{1}{5}\boldsymbol{X}_4 + \frac{1}{5}\boldsymbol{X}_5$$

and

$$\boldsymbol{X}_1 - \boldsymbol{X}_2 + \boldsymbol{X}_3 - \boldsymbol{X}_4 + \boldsymbol{X}_5$$

in terms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Also, obtain the covariance between these two linear combinations.

4 Maximum Likelihood Estimation & Sampling Distributions

In this section we will focus on the parameters of the multivariate normal distribution. Specifically, the maximum likelihood estimates of the parameters and the sampling distributions of the corresponding statistics will be defined. Finally, we will discuss methods of testing for multivariate normality.

4.1 The Multivariate Normal Density and Likelihood

Consider a random sample, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from the multivariate normal population $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the $\boldsymbol{X}_j$'s mutually independent. The joint density of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is given by

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{j=1}^{n} f(\boldsymbol{x}_i) = \left(\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \right)^n \exp\left[-\frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}) \right]$$

This joint density, written as a function of the variables, regards the parameters as fixed, albeit unknown, constants. However, when observations are made, i.e. we are given values for $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, then we can consider this expression to be a function of the parameters, referred to as the *likelihood.* We are therefore trying to ascertain how likely specific values of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (viewed as variable) are, given our fixed observations.

$$\therefore L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \right)^n \exp\left[-\frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}) \right]$$

In order to rewrite this likelihood into a more convenient form going forward, we will define a matrix $\boldsymbol{A}$ as the sums of squares and cross products, which is proportional to the sample covariance matrix:

$$\boldsymbol{A} = (n-1)\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}' = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$$

Also note the following properties of traces:

- $a = tr(a)$ for any scalar a
- $tr(\boldsymbol{BC}) = tr(\boldsymbol{CB})$ if $\boldsymbol{BC}$ and $\boldsymbol{CB}$ are conformable
- $\implies \boldsymbol{x}'\boldsymbol{Bx} = tr(\boldsymbol{x}'\boldsymbol{Bx}) = tr(\boldsymbol{Bxx}')$ for $\boldsymbol{B}_{k\times k}$ and $\boldsymbol{x}_{k\times 1}$

The summation in the exponent of the likelihood function can then be written as

$$\begin{aligned}
\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}) &= \sum_{j=1}^{n} tr\left[(\boldsymbol{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})\right] \\
&= \sum_{j=1}^{n} tr\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})(\boldsymbol{x}_j-\boldsymbol{\mu})'\right] \\
&= tr\left[\sum_{j=1}^{n}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})(\boldsymbol{x}_j-\boldsymbol{\mu})'\right] \\
&= tr\left[\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})(\boldsymbol{x}_j-\boldsymbol{\mu})'\right] \\
&= tr\left[\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}\left[(\boldsymbol{x}_j-\bar{\boldsymbol{x}})+(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\right]\left[(\boldsymbol{x}_j-\bar{\boldsymbol{x}})+(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\right]'\right] \\
&= tr\left[\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})' + n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\right] \\
&= tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right] + tr\left[n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\right] \\
&= tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right] + n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})
\end{aligned}$$

Note that the cross-product terms vanish. For example,

$$\begin{aligned}
\sum_{j=1}^{n}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})' &= (\bar{\boldsymbol{x}}-\boldsymbol{\mu})\left[\sum_{j=1}^{n}\boldsymbol{x}_j' - \sum_{j=1}^{n}\bar{\boldsymbol{x}}'\right] \\
&= (\bar{\boldsymbol{x}}-\boldsymbol{\mu})\left[n\bar{\boldsymbol{x}}' - n\bar{\boldsymbol{x}}'\right] \\
&= \boldsymbol{0}
\end{aligned}$$

We can now write the joint likelihood function as

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\exp\left[-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right] - \frac{n}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\right]$$

4.2 Maximum Likelihood Estimates

We will now find the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximise the likelihood given above.

4.2.1 MLE of μ

We consider $\boldsymbol{\Sigma}$ as fixed and maximise

$$\log[L(\boldsymbol{\mu},\boldsymbol{\Sigma})] = l(\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}] - \frac{n}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$$

with respect to $\boldsymbol{\mu}$.

By inspection:

$$\begin{aligned} l &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}] - \frac{n}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \\ &\leq -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}] \end{aligned}$$

with equality when $\frac{n}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) = 0$, i.e., when

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}.$$

By differentiation:

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\mu}} &= -\frac{n}{2}2\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(-1) && \text{(See Theorem A.1)} \\ n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\hat{\boldsymbol{\mu}}) &= \mathbf{0} \\ \hat{\boldsymbol{\mu}} &= \bar{\boldsymbol{x}} \end{aligned}$$

Showing that $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ is indeed a maximum and not a minimum or saddle point:

$$\frac{\partial^2 l}{\partial \boldsymbol{\mu}\boldsymbol{\mu}'} = \frac{\partial}{\partial \boldsymbol{\mu}'}\{n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}} - n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\} = -n\boldsymbol{\Sigma}^{-1}$$

which is negative definite. Note that $\boldsymbol{\Sigma}^{-1}$ is positive definite since $\boldsymbol{\Sigma}$ is positive definite and symmetric.

4.2.2 MLE of Σ

Let $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$. We first need to find the stationary point of $l(\hat{\boldsymbol{\mu}},\boldsymbol{\Sigma})$ through differentiation with respect to $\boldsymbol{\Sigma}^{-1}$, then show that the stationary point is a maximum. Note that we require Theorem A.2 & Theorem A.3 to perform the differentiation.

$$\begin{aligned} l(\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}},\boldsymbol{\Sigma}) &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}] \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{A}] \\ \frac{\partial l}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{n}{2}\left[\left(\boldsymbol{\Sigma}^{-1}\right)^{-1}\right]' - \frac{1}{2}\boldsymbol{A}' \end{aligned}$$

Setting $\dfrac{\partial l}{\partial \mathbf{\Sigma}^{-1}} = \mathbf{0}$:

$$\begin{aligned}
\frac{n}{2}\hat{\mathbf{\Sigma}} - \frac{1}{2}\boldsymbol{A} &= \mathbf{0} \\
n\hat{\mathbf{\Sigma}} &= \boldsymbol{A} \\
\hat{\mathbf{\Sigma}} &= \frac{1}{n}\boldsymbol{A} = \frac{n-1}{n}\boldsymbol{S}
\end{aligned}$$

To show that this stationary point is a maximum, taking the second order derivative with respect to $\mathbf{\Sigma}^{-1}$ is complicated. Therefore, the likelihood is reparameterised.

$$\begin{aligned}
L(\hat{\boldsymbol{\mu}}, \mathbf{\Sigma}) &= (2\pi)^{-\frac{np}{2}} |\mathbf{\Sigma}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr[\mathbf{\Sigma}^{-1}\boldsymbol{A}]\right] \\
&= (2\pi)^{-\frac{np}{2}} |\mathbf{\Sigma}^{-1}|^{\frac{n}{2}} \left|\frac{1}{n}\boldsymbol{A}\right|^{\frac{n}{2}} \left|\frac{1}{n}\boldsymbol{A}\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr[\mathbf{\Sigma}^{-1}\boldsymbol{A}]\right] \\
&= (2\pi)^{-\frac{np}{2}} \left|\frac{1}{n}\boldsymbol{A}\mathbf{\Sigma}^{-1}\right|^{\frac{n}{2}} \left|\frac{1}{n}\boldsymbol{A}\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr[\boldsymbol{A}\mathbf{\Sigma}^{-1}]\right]
\end{aligned}$$

Let $\mathbf{\Delta}_{p\times p} = \boldsymbol{A}\mathbf{\Sigma}^{-1}$ with the i^{th} eigenvalue-eigenvector pair of $\mathbf{\Delta}$ given by $\mathbf{\Delta}\boldsymbol{b} = \lambda\boldsymbol{b}$. Then

$$\begin{aligned}
L(\hat{\boldsymbol{\mu}}, \mathbf{\Sigma}) &= (2\pi)^{-\frac{np}{2}} \left(\frac{1}{n}\right)^{\frac{np}{2}} |\mathbf{\Delta}|^{\frac{n}{2}} \left|\frac{1}{n}\boldsymbol{A}\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} tr(\mathbf{\Delta})\right] \\
&= (2n\pi)^{-\frac{np}{2}} \left(\prod_{i=1}^{p} \lambda_i^{\frac{n}{2}}\right) \left|n\boldsymbol{A}^{-1}\right|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{p}\lambda_i\right] \\
&= (2n\pi)^{-\frac{np}{2}} \left|n\boldsymbol{A}^{-1}\right|^{\frac{n}{2}} \left(\prod_{i=1}^{p} \lambda_i^{\frac{n}{2}} e^{-\frac{1}{2}\lambda_i}\right)
\end{aligned}$$

with the corresponding log-likelihood given by

$$l(\hat{\boldsymbol{\mu}}, \mathbf{\Sigma}) = -\frac{np}{2}\log(2n\pi) + \frac{n}{2}\log\left|n\boldsymbol{A}^{-1}\right| + \sum_{i=1}^{p}\left(\frac{n}{2}\log(\lambda_i) - \frac{1}{2}\lambda_i\right)$$

To maximise $l(\hat{\boldsymbol{\mu}}, \mathbf{\Sigma})$ with respect to $\mathbf{\Sigma}$, we need to maximise with respect to $\boldsymbol{A}^{-1}\mathbf{\Delta} = \boldsymbol{A}^{\mathbf{-1}}\boldsymbol{B}\mathbf{\Lambda}\boldsymbol{B'}$. We have effectively reparameterised $l(\hat{\boldsymbol{\mu}}, \mathbf{\Sigma})$ as $l(\hat{\boldsymbol{\mu}}, \mathbf{\Lambda})$, since $\boldsymbol{A}$ is effectively a constant (only a function of the observed data), and the orthogonal matrix $\boldsymbol{B}$ does not appear in the likelihood.

Since $\lambda_i,\ i = 1, \ldots, p$ appears separately in the summation terms, each summation term can be maximised separately.

$$\frac{\partial}{\partial\lambda_i}\left(\frac{n}{2}\log(\lambda_i) - \frac{1}{2}\lambda_i\right) = \frac{n}{2\lambda_i} - \frac{1}{2}$$

Setting $\dfrac{\partial l}{\partial \lambda_i} = 0$:

$$\begin{aligned} \frac{n}{2\hat{\lambda}_i} - \frac{1}{2} &= 0 \\ \frac{n}{\hat{\lambda}_i} &= 1 \\ \hat{\lambda}_i &= n \end{aligned}$$

To show that each term is a maximum when $\hat{\lambda}_i = n$:

$$\begin{aligned} \frac{\partial^2}{\partial \lambda_i^2}\left(\frac{n}{2}\log(\lambda_i) - \frac{1}{2}\lambda_i\right) &= \frac{\partial}{\partial \lambda_i}\left(\frac{n}{2}\lambda_i^{-1} - \frac{1}{2}\right) \\ &= -\frac{n}{2}\lambda_i^{-2} \\ \frac{\partial^2}{\partial \lambda_i^2}\left(\frac{n}{2}\log(\lambda_i) - \frac{1}{2}\lambda_i\right)_{\hat{\lambda}_i = n} &= -\frac{n}{2}\left(\frac{1}{n^2}\right) \\ &= -\frac{1}{2n} < 0 \end{aligned}$$

Therefore, $\hat{\lambda}_i = n, \ i = 1, \ldots, n$ is a maximum, which confirms that

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{A}$$

is the maximum likelihood estimate of $\boldsymbol{\Sigma}$. Substituting the MLE's into the likelihood, we have

$$\begin{aligned} L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) &= (2\pi)^{-\frac{np}{2}}|\hat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}tr\left[\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{A}\right] - \frac{n}{2}(\bar{\boldsymbol{x}} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{x}} - \hat{\boldsymbol{\mu}})\right] \\ &= (2\pi)^{-\frac{np}{2}}\left|\frac{1}{n}\boldsymbol{A}\right|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}tr\left[n\boldsymbol{A}^{-1}\boldsymbol{A}\right]\right] \\ &= (2\pi)^{-\frac{np}{2}}\left|\frac{1}{n}\boldsymbol{A}\right|^{-\frac{n}{2}} e^{-\frac{np}{2}} \end{aligned}$$

Homework exercise 4.1

Johnson & Wichern exercise 4.18

Find the maximum likelihood estimates of $\boldsymbol{\mu}_{2\times1}$ and $\boldsymbol{\Sigma}_{2\times2}$ based on the random sample

$$\begin{bmatrix} 3 & 6 \\ 4 & 4 \\ 5 & 7 \\ 4 & 7 \end{bmatrix}$$

from a bivariate normal population.

4.3 Sampling Distribution of $\bar{X}$ and S

Before we look at the distributions of $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$, a reminder of the univariate case (which will of course be a special instance of the multivariate normal when $p = 1$):

Univariate sampling distributions

If $X_1, X_2, \ldots, X_n$ are i.i.d random observation from $X \sim N(\mu, \sigma^2)$, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Multivariate sampling distributions

Now suppose we observe $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, which are i.i.d random vectors drawn from $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For the sample mean we have

$$\bar{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$$

For the sample variance, note that the univariate case can be expressed as

$$\begin{aligned}(n-1)s^2 &\sim \sigma^2\chi^2_{n-1}\\ &= \sigma^2\left(Z_1^2 + \ldots + Z_{n-1}^2\right)\\ &= (\sigma Z_1)^2 + \ldots + (\sigma Z_{n-1})^2\end{aligned}$$

with each $\sigma Z_i = X_i - \mu \sim N(0, \sigma^2)$.

This form is suitably generalized to the basic sampling distribution of the covariance matrix, namely the **Wishart distribution**. This distribution, which is the multivariate analogue of the χ^2-distribution, can be defined as the sum of independent products of multivariate random vectors.

If we define $\boldsymbol{Y}_i = \boldsymbol{X}_i - \boldsymbol{\mu} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and let $\boldsymbol{Y}$ represent the $p \times n$ matrix constructed from the n independent observations of $\boldsymbol{X}_i$, then

$$\boldsymbol{Y}\boldsymbol{Y}' = \sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})' \sim W_p(n, \boldsymbol{\Sigma})$$

We say that $\boldsymbol{Y}\boldsymbol{Y}'$ is p-dimensional Wishart distributed with n degrees of freedom. Similar to the univariate case, when $\bar{\boldsymbol{X}}$ is substituted for $\boldsymbol{\mu}$, the distribution remains Wishart, but with one less degree of freedom:

$$\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' = \boldsymbol{A} = (n-1)\boldsymbol{S} \sim W_p(n-1, \boldsymbol{\Sigma})$$

$$\boldsymbol{S} \sim W_p\left(n-1, \frac{1}{n-1}\boldsymbol{\Sigma}\right)$$

4.3.1 Properties of the Wishart distribution

Given $\boldsymbol{A} \sim W_p(n, \boldsymbol{\Sigma})$, we note the following properties:

- The Wishart distribution is a generalization of the χ^2-distribution, where $W_1(n, \sigma^2) = \sigma^2\chi^2_n$.
- If $\boldsymbol{c}_{p\times 1}$ is a non-zero constant vector, then $\dfrac{\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}} \sim \chi^2_n$.
- The density only exists if $n > p$.
- The density itself, although not of particular use to us, is given by

$$f(\boldsymbol{A}) = \frac{|\boldsymbol{A}|^{\frac{n-p-2}{2}} \exp\left[-\frac{1}{2}tr(\boldsymbol{A}\boldsymbol{\Sigma}^{-1})\right]}{2^{\frac{p(n-1)}{2}} \pi^{\frac{p(p-1)}{4}} |\boldsymbol{\Sigma}|^{\frac{n-1}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{n-i}{2}\right)}$$

  for $\boldsymbol{A}$ positive definite.
- $tr(\boldsymbol{A}) \sim \chi^2_{np}$.
- $\mathrm{E}(\boldsymbol{A}) = n\boldsymbol{\Sigma}$.

4.3.2 Summary of Sampling distribution results

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ be a random sample from $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

1. $\bar{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu}, \dfrac{1}{n}\boldsymbol{\Sigma}\right)$
2. $(n-1)\boldsymbol{S} \sim W_p(n-1, \boldsymbol{\Sigma})$
3. $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent.
4. $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are *sufficient statistics* for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The importance of the last point is that all the information about $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the data matrix $\boldsymbol{X}$ is contained in $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ respectively. Note that this is not generally true for non-normal populations, in which case techniques that depend solely on $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ may be ignoring useful sample information. We will discuss testing for multivariate normality shortly, after considering asymptotic behaviour.

4.3.3 Large Sample behaviour of $\bar{X}$ and S

If $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ are independent observations from some population with mean $\boldsymbol{\mu}$ and finite, nonsingular covariance $\boldsymbol{\Sigma}$, then

$$\bar{\boldsymbol{X}} \dot{\sim} N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right) \text{ or, equivalently, } \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \dot{\sim} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$$

and

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \dot{\sim} \chi^2_p$$

for large $n - p$.

Homework exercise 4.2

Johnson & Wichern exercises 4.19 & 4.21 (combined)

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{60}$ be a random sample of size $n = 60$ from an $N_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. Specify each of the distributions completely (indicate if the distribution is approximate):

1. $\bar{\boldsymbol{X}}$ and $\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$
2. $(\boldsymbol{X}_1 - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu})$
3. $(n-1)\boldsymbol{S}$
4. $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$
5. $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$

4.4 Testing the Assumption of Normality

Before we can move on to inference on the parameters of the multivariate normal distribution, we need to acknowledge the assumption of multivariate normality. If this assumption is false, it may impact on the quality of the inference. To determine whether the observations $\boldsymbol{X}_j$ appear to violate the assumption that they jointly came from a multivariate normal population, we address the following questions:

1. Do the marginal distributions of the elements of $\boldsymbol{X}$ appear to be normal?
2. Do the scatter plots of the pairs of observations on different characteristics give the elliptical appearance expected from normal populations?
3. Are there any "wild" observations that should be checked?

Our assessment will be confined to looking at one or two dimensions at a time.

4.4.1 Univariate assessment

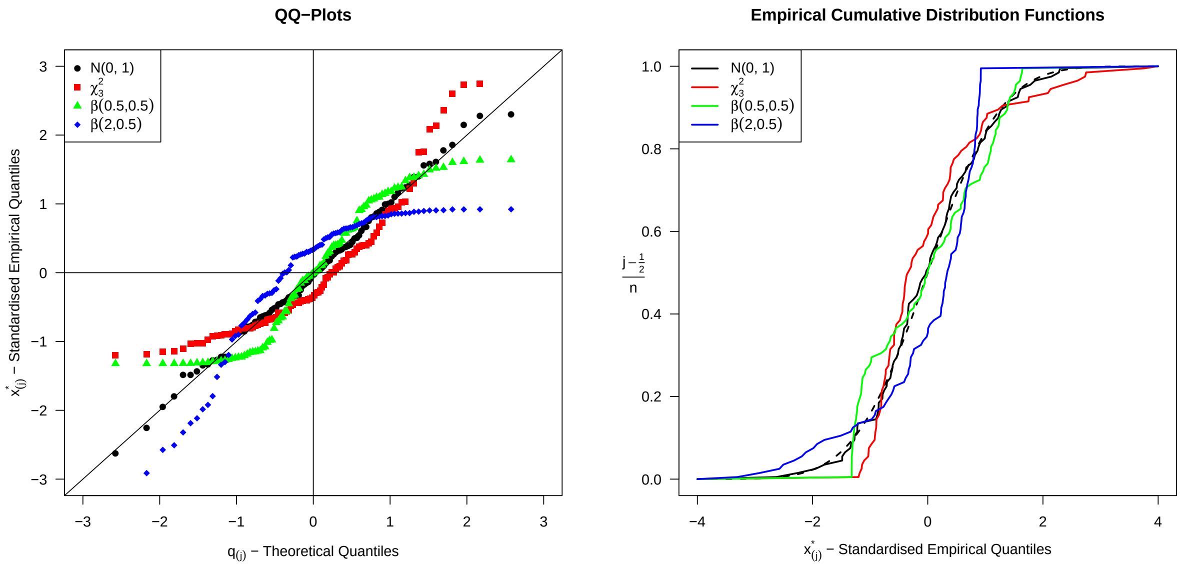
QQ-plots provide a quick, visual way of assessing how closely the distribution of observed data matches some theoretical distribution. In testing normality, we can plot the sample quantile versus the quantile one would expect to observe if the observations were actually normally distributed.

If we order the n observations such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, then if the $x_{(j)}$ are distinct, exactly j observations will be less than or equal to $x_{(j)}$. The proportion of the sample at or to the left of $x_{(j)}$, i.e. $\frac{j}{n}$, is often approximated by $\frac{j-\frac{1}{2}}{n}$ for analytical convenience.

The quantiles for a standard normal distribution are defined as those values $q_{(j)}$ such that

$$\Pr[Z \leq q_{(j)}] = \int_{-\infty}^{q_{(j)}} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}\, dz = p_{(j)} = \frac{j-\frac{1}{2}}{n}$$

Under the assumption of normality, a plot of $(q_{(j)}, x^*_{(j)})$ for all j should be a straight line through the origin, where $x^*_{(j)}$ denotes the standardised values of $\boldsymbol{x}$. Below on the left we see an example of a QQ-plot illustrating the relationship between these theoretical quantiles, $q_{(j)}$, and the standardised observed quantiles $x^*_{(j)}$ for sets of random values drawn from a few different distributions. The corresponding empirical cumulative distribution functions (CDFs) are given on the right, with the dashed line representing the theoretical standard normal CDF.



To formally test the linearity, we can calculate the correlation coefficient for the QQ plot, defined as

$$r_q = \frac{\sum_{j=1}^n (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^n (x_{(j)} - \bar{x})^2}\sqrt{\sum_{j=1}^n (q_{(j)} - \bar{q})^2}}$$

and compare it to a table of critical values, given in Appendix B.

Many other formal tests for univariate normality exist, such as Shapiro-Wilk, Anderson-Darling, Jarque-Bera and Lilliefors test, to name but a few.

4.4.2 Multivariate assessment

We can now compare the contours of constant density from observed data with the ellipsoid as defined in chapter 3.3, where a set of bivariate outcomes $\boldsymbol{x}$ such that

$$(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_p(\alpha)$$

has probability $1 - \alpha$. Typically we calculate the above for $\alpha = 0.5$ and calculate the proportion of points for which the squared distance is less than $\chi^2_p(0.5)$. If this deviates from 50%, it is evidence against the assumption of normality.

We can also construct a chi-square plot based on the assumption that, given underlying normality, the squared distances, $d^2_j = (\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$ should behave like chi-square random variables. To construct these plots,

1. Calculate $d_j^2 = (\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$.
2. Order the squared distances from smallest to largest as $d^2_{(1)} \leq d^2_{(2)} \leq \ldots \leq d^2_{(n)}$.
3. Graph the pairs $\left(q_{c,p}\left(\frac{j-\frac{1}{2}}{n}\right), d^2_{(j)}\right)$ where $q_{c,p}\left(\frac{j-\frac{1}{2}}{n}\right) = \chi^2_p\left(\frac{n-j+\frac{1}{2}}{n}\right)$.

If the variables are multivariate normal distributed, the plot should be a straight line through the origin.

Example

Example 4.14 on page 186 of Johnson & Wichern (also see `Lecture4.R`).

```
T4_3 <- read.table(’T4-3.dat’)
head(T4_3)
```

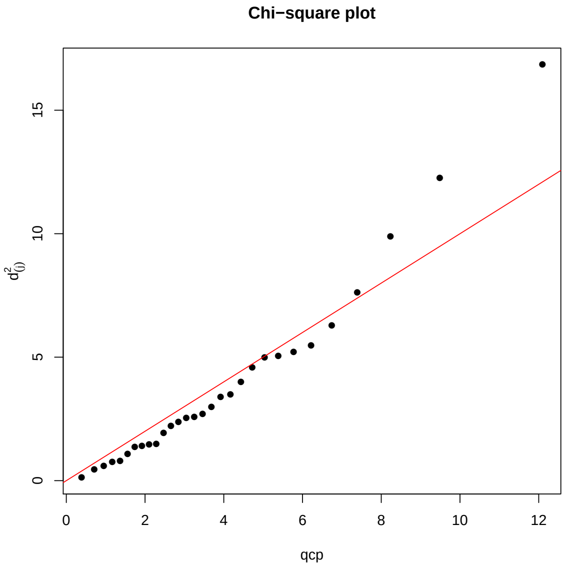
```
    V1   V2   V3   V4   V5
1 1889 1651 1561 1778 0.60
2 2403 2048 2087 2197 5.48
3 2119 1700 1815 2222 7.62
4 1645 1627 1110 1533 5.21
5 1976 1916 1614 1883 1.40
6 1712 1712 1439 1546 2.22
```

```
stiffness <- T4_3[, -5]
S <- cov(stiffness)
n <- nrow(stiffness)

#1. Calculate squared distances
d2 <- mahalanobis(stiffness, colMeans(stiffness), cov = S) #Compare with T4_3[, 5]

#2. Order the squared distances
d2_ord <- sort(d2)

#3. Graph the pairs
qcp <- qchisq((1:n - 0.5)/n, ncol(stiffness))
plot(qcp, d2_ord, main="Chi-square plot", pch = 16, ylab = ’’)
title(ylab = expression(d[(j)]^2), line = 2)
abline(a = 0, b = 1, col = ’red’)
```



Once again there are many formal tests one can apply, such as Mardia's test (a multivariate extension of skewness and kurtosis measures), the Henze-Zirkler test and Royston's test (a multivariate extension of Shapiro-Wilk). However, as with univariate normality tests, some procedures are more appropriate under certain conditions than others, and they can yield contradictory conclusions.

It is also important to note the following crucial drawbacks of all measures of fit: With small samples, only severe deviations will indicate lack of fit, whilst very large samples will invariably produce statistically significant lack of fit.

Homework exercise 4.3

Johnson & Wichern exercises 4.28 & 4.29 (combined)

Consider the air pollution data (made available in a .csv file of the same name).

1. Construct a QQ-plot for the solar radiation measurements and carry out a test for normality based on the correlation coefficient r_q. You are not prescribed to use a specific α-value; what can you report based on Table B.1?

Now examine the pairs $X_5 = \text{NO}_2$ and $X_6 = \text{O}_3$ for bivariate normality.

2. Calculate the distances $(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}}),\ j = 1, 2, \ldots, 42$, where $\boldsymbol{x}_j' = [x_{j5}, x_{j6}]$.
3. Determine the proportion of observations $\boldsymbol{x}_j' = [x_{j5}, x_{j6}],\ j = 1, 2, \ldots, 42$ falling within the approximate 50% probability contour of a bivariate normal distribution.
4. Construct a chi-square plot for the ordered distances in Part 3.

5 Inference about a Mean Vector

Now that we have covered the key aspects of the multivariate normal distribution, the distributions of its sample statistics, and testing for multivariate normality, we will next focus on inference on the mean vector, $\boldsymbol{\mu}$.

Therefore, in this section we'll be discussing ways of testing whether a given vector of values is a plausible mean of a multivariate normal distribution, given a set of observations. Finally, we will consider different approaches towards defining multivariate confidence regions and intervals for mean vectors.

5.1 Univariate Intervals

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a normal population with unknown mean μ and unknown variance σ^2. We can test the univariate hypothesis, $H_0\colon \mu = \mu_0$ vs. $H_1\colon \mu \neq \mu_0$ using a t-statistic:

$$t = \frac{\bar{X} - \mu_0}{{}^{s}\!/\!\sqrt{n}} \sim t_{n-1}$$

where $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ and $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$.

We will reject H_0 when $|t|$ is large, which is equivalent to rejecting H_0 when

$$t^2 = \frac{(\bar{X} - \mu_0)^2}{s^2/n} = n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0)$$

is large.

For an observed $\bar{X}$ and s^2, reject H_0 in favour of H_1 at a significance level of α if

$$n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0) > t^2_{n-1}\,(\alpha/2)$$

where $t_{n-1}\,(\alpha/2)$ denotes the upper $100\,(\alpha/2)^{th}$ percentile of t_{n-1}.

Note that there is actually a whole range of plausible values in support of $H_0\colon \mu = \mu_0$ since

$\{$Do not reject $H_0\colon \mu = \mu_0$ at level $\alpha\}$ or $\left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| \leq t_{n-1}\,(\alpha/2)$

is equivalent to

$\{\mu_0$ lies in the $100(1-\alpha)\%$ confidence interval $\bar{x} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}\}$ or

$$\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}$$

Since $\bar{X}$ and s are random variables, the confidence interval is a random interval. The probability that the interval contains μ is $1-\alpha$, which is the same as saying that among large numbers of such independent intervals, approximately $100(1-\alpha)\%$ of them will contain μ.

5.2 Hotelling's T^2 statistic

Now let us consider random observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ drawn from a population $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown mean and variance. The multivariate analogue of the t^2 statistic to test $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ is referred to as [1]Hotelling's T^2, which is defined as

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0),$$

where $\bar{\boldsymbol{X}}_{p\times 1} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{X}_j$, $\boldsymbol{S}_{p\times p} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'$, and $\boldsymbol{\mu}_0\colon p \times 1 = \begin{bmatrix}\mu_{10}\\ \mu_{20}\\ \vdots\\ \mu_{p0}\end{bmatrix}$.

It can be shown that this statistic is proportional to the F-distribution with p and $n-p$ degrees of freedom, such that

$$T^2 \sim \frac{(n-1)p}{n-p}F_{p,n-p}$$

[1] Harold Hotelling also developed and named PCA in the 1930s.

$\implies \Pr\left[T^2 > \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right] = \alpha$ where $F_{p,n-p}(\alpha)$ refers to the upper $(100\alpha)\%$ of the $F_{p,n-p}$ distribution.

Therefore, we reject H_0 if the observed value of $\frac{n-p}{(n-1)p}T^2$ is greater than $F_{p,n-p}(\alpha)$. Likewise, we can calculate the p-value of an observed test statistic, say T^*, as $\Pr(T^2 \geq T^*)$. This measure of the "weight of evidence" against H_0 is more insightful than a binary classification at some arbitrary level of significance α.

Note that we can write T^2 as

$$\begin{aligned}
T^2 &= \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \left(\frac{\sum_{j=1}^n(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'}{n-1}\right)^{-1} \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \\
&= (\text{MVN random vector})' \left(\frac{\text{Wishart random matrix}}{df}\right)^{-1} (\text{MVN random vector}) \\
&= N_p(\mathbf{0}, \boldsymbol{\Sigma})' \left[\frac{1}{n-1}W_p(n-1, \boldsymbol{\Sigma})\right]^{-1} N_p(\mathbf{0}, \boldsymbol{\Sigma})
\end{aligned}$$

Since N_p and W_p are independent, this product of the marginal Normal and Wishart distributions will yield their joint density. It is from this that the F-distribution of T^2 can be derived.

Example

Example 5.2 on page 214 of Johnson & Wichern (also see `Lecture5.R`).

Perspiration from 20 similar participants were analysed, with three components measured: X_1 = sweat rate, X_2 = sodium content, and X_3 = potassium content.

We will first read in the data and calculate the sample statistics $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$.

```r
sweat <- read.table('T5-1.dat')

head(sweat)
```

```
    V1   V2   V3
1  3.7 48.5  9.3
2  5.7 65.1  8.0
3  3.8 47.2 10.9
4  3.2 53.2 12.0
5  3.1 55.5  9.7
6  4.6 36.1  7.9
```

```r
n <- nrow(sweat); p <- ncol(sweat)

(xbar <- matrix(apply(sweat, 2, mean), 3))
(S <- cov(sweat))
```

```
       [,1]
[1,]  4.640
[2,] 45.400
[3,]  9.965

            V1       V2        V3
V1   2.879368  10.0100 -1.809053
V2  10.010000 199.7884 -5.640000
V3  -1.809053  -5.6400  3.627658
```

Suppose we want to test the hypothesis $H_0\colon \boldsymbol{\mu} = \begin{bmatrix} 4 \\ 50 \\ 10 \end{bmatrix}$ vs $H_1\colon \boldsymbol{\mu} \neq \begin{bmatrix} 4 \\ 50 \\ 10 \end{bmatrix}$. Using this vector of values for $\boldsymbol{\mu}_0$, we can calculate T^2 and its associated p-value.

```
(mu0 <- matrix(c(4, 50, 10), 3))
```

```
     [,1]
[1,]    4
[2,]   50
[3,]   10
```

```
# Calculate Hotteling's T^2
(T2 <- n*t(xbar - mu0)%*%solve(S)%*%(xbar - mu0))
```

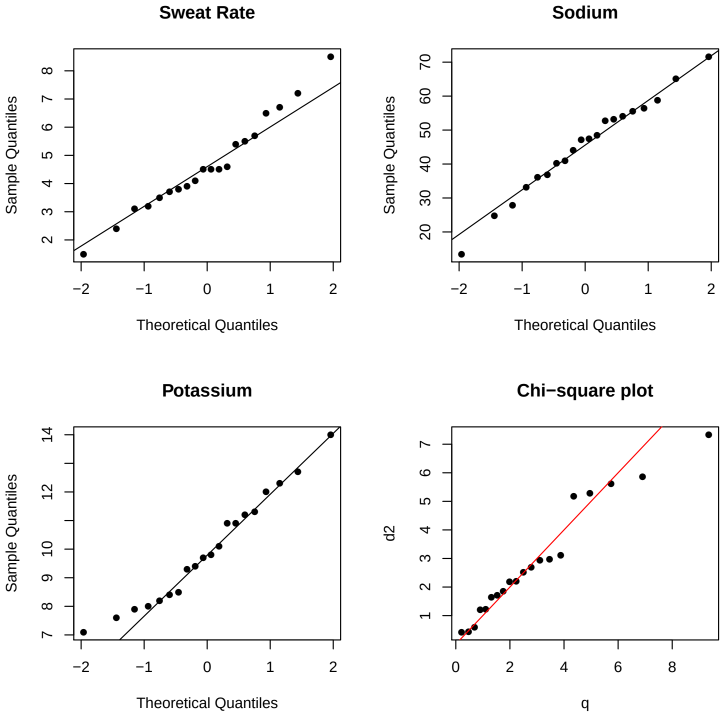
```
         [,1]
[1,] 9.738773
```

```
# Calculate p-value
1 - pf(T2*(n-p)/((n-1)*p), p, n-p)
```

```
       [,1]
[1,] 0.0649
```

Here we see that the p-value $= 0.0649$. Therefore, this particular sample is quite unlikely to have occurred, if the underlying assumption of H_0 were true. Therefore, we have reasonably strong evidence against $H_0\colon \boldsymbol{\mu} = \begin{bmatrix} 4 \\ 50 \\ 10 \end{bmatrix}$, and we will reject this hypothesis at any significance level $\alpha \geq 6.5\%$.

We will now briefly check the underlying assumption of normality by investigating the QQ-plots for each of the three variables, as well as the Chi-square plot:



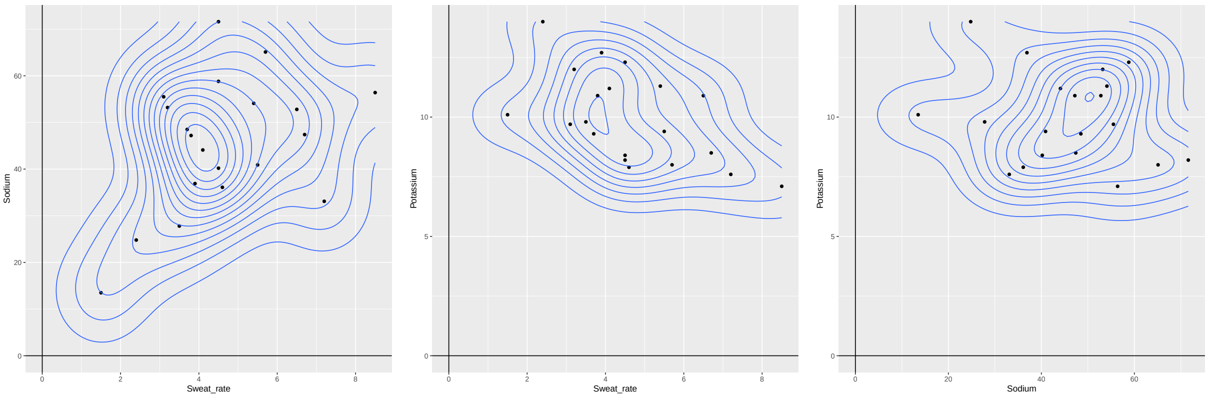
None of the variables seem to indicate severe deviation from normality. Although the chi-square plot is slightly less conclusive for the larger deviations, we cannot discount multivariate normality when considering the small sample size.

This is further confirmed by considering the proportion of the squared distances lying below the 50% quantile of the χ_p^2-distribution. If the variables were multivariate normal distributed, we would expect half of the values to be below $\chi_3^2(0.5) = 2.366$, which is indeed exactly what we observe:

```r
mean(d2 < qchisq(0.5, p))
```

```
[1] 0.5
```

When looking at the bivariate plots, we don't observe severe deviation from ellipse-shaped contours, when considering the small sample size of $n = 20$.



5.2.1 Invariance of the T^2 statistic

Let $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X} + \boldsymbol{d}$, where $\boldsymbol{C}_{p\times p}$ is non-singular and $\boldsymbol{d}_{p\times 1}$ is a vector of constants. Testing

$$H_0\colon \boldsymbol{\mu}_Y = \boldsymbol{C}\boldsymbol{\mu}_0 + \boldsymbol{d} \;\; vs \;\; H_1\colon \boldsymbol{\mu}_Y \neq \boldsymbol{C}\boldsymbol{\mu}_0 + \boldsymbol{d}$$

is equivalent to testing

$$H_0\colon \boldsymbol{\mu}_X = \boldsymbol{\mu}_0 \;\; vs \;\; H_1\colon \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_0$$

To show this, we first note that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_X, \boldsymbol{\Sigma})$, then $\boldsymbol{Y} \sim N_p(\boldsymbol{C}\boldsymbol{\mu}_X + \boldsymbol{d}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')$ such that $\bar{\boldsymbol{Y}} = \boldsymbol{C}\bar{\boldsymbol{X}} + \boldsymbol{d}$ and $\boldsymbol{S}_Y = \boldsymbol{C}\boldsymbol{S}_X\boldsymbol{C}'$.

Now

$$\begin{aligned}
T_Y^2 &= n(\bar{\boldsymbol{Y}} - \boldsymbol{C}\boldsymbol{\mu}_0 - \boldsymbol{d})'\boldsymbol{S}_Y^{-1}(\bar{\boldsymbol{Y}} - \boldsymbol{C}\boldsymbol{\mu}_0 - \boldsymbol{d}) \\
&= n(\boldsymbol{C}\bar{\boldsymbol{X}} + \boldsymbol{d} - \boldsymbol{C}\boldsymbol{\mu}_0 - \boldsymbol{d})'(\boldsymbol{C}\boldsymbol{S}_X\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{\boldsymbol{X}} + \boldsymbol{d} - \boldsymbol{C}\boldsymbol{\mu}_0 - \boldsymbol{d}) \\
&= n[\boldsymbol{C}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)]'(\boldsymbol{C}')^{-1}\boldsymbol{S}_X^{-1}\boldsymbol{C}^{-1}[\boldsymbol{C}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)] \\
&= n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{C}'(\boldsymbol{C}')^{-1}\boldsymbol{S}_X^{-1}\boldsymbol{C}^{-1}\boldsymbol{C}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \\
&= n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{S}_X^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \\
&= T_X^2
\end{aligned}$$

This shows that Hotelling's T^2 statistic is invariant under changes in the unit of measurement for $\boldsymbol{X}$ of the form $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X} + \boldsymbol{d}$.

5.3 Hotelling's T^2 statistic and Likelihood Ratio tests

The likelihood ratio test is based on comparing the maximum of the likelihood when the parameter space is restricted under H_0 to the maximum of the likelihood when parameters are allowed to take on any value in the parameter space.

A likelihood ratio test of $H_0\colon \theta \in \Theta_0$ rejects H_0 in favour of $H_1\colon \theta \notin \Theta_0$ if

$$\Lambda = \frac{\max\limits_{\theta\in\Theta_0} L(\theta)}{\max\limits_{\theta\in\Theta} L(\theta)} < c$$

for a suitably chosen constant c. It can be shown that for large sample sizes n, the value $-2\log(\Lambda)$ is approximately distributed as $\chi^2_{\nu-\nu_0}$, where $\nu = \dim(\Theta)$ and $\nu_0 = \dim(\Theta_0)$.

Consider the hypothesis $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against the alternative $H_1\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ for $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Now the unrestricted parameter space is

$$\Theta = \{-\infty < \mu_1 < \infty, \ldots, -\infty < \mu_p < \infty, \quad \sigma_{11}, \sigma_{12}, \ldots \sigma_{1p}, \sigma_{22}, \ldots, \sigma_{2p}, \ldots, \sigma_{pp}\colon \boldsymbol{\Sigma} > 0\}$$

$$\nu = \dim(\Theta) = p + \frac{p(p+1)}{2}$$

whilst the parameter space under $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is

$$\Theta = \{\mu_1 = \mu_{10}, \ldots, \mu_p = \mu_{p0}, \quad \sigma_{11}, \sigma_{12}, \ldots \sigma_{1p}, \sigma_{22}, \ldots, \sigma_{2p}, \ldots, \sigma_{pp}\colon \boldsymbol{\Sigma} > 0\}$$

$$\nu_0 = \dim(\Theta_0) = \frac{p(p+1)}{2}$$

In section 4.2.2 we showed that the maximum of the multivariate normal likelihood as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are varied over their possible values is given by

$$\max_{\theta\in\Theta} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}} \left|\frac{1}{n}\boldsymbol{A}\right|^{-\frac{n}{2}} e^{-\frac{np}{2}}$$

Under H_0, we wish to find the most likely value of $\boldsymbol{\Sigma}$ that, with $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ fixed, would have led to the observed data. This is equivalent to maximizing

$$L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)\right]$$

After taking logs we have

$$\begin{aligned} l(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_0) \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{j=1}^{n} tr\left[(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}\right] \end{aligned}$$

Differentiating and setting equal to zero:

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{n}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)' = 0$$

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_0 &= \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)' \\
&= \frac{1}{n}\sum_{j=1}^{n}[(\boldsymbol{x}_j - \bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)][(\boldsymbol{x}_j - \bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)]' \\
&= \frac{1}{n}\boldsymbol{A} + (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'
\end{aligned}$$

It can be shown, similar to what was done in section 4.2.2, that this value of $\hat{\boldsymbol{\Sigma}}_0$ indeed yields a maximum for $L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$.

Now we can write

$$\begin{aligned}
\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\hat{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_0) &= \sum_{j=1}^{n} tr\left[(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\hat{\boldsymbol{\Sigma}}_0^{-1}\right] \\
&= tr\left[\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\hat{\boldsymbol{\Sigma}}_0^{-1}\right] \\
&= tr\left[n\sum_{j=1}^{n}\left[\frac{1}{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\right]\hat{\boldsymbol{\Sigma}}_0^{-1}\right] \\
&= tr\left[n\hat{\boldsymbol{\Sigma}}_0\hat{\boldsymbol{\Sigma}}_0^{-1}\right] = tr(n\boldsymbol{I}_p) = np
\end{aligned}$$

so that

$$\max_{\theta\in\Theta_0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}}\left|\hat{\boldsymbol{\Sigma}}_0\right|^{-\frac{n}{2}} e^{-\frac{np}{2}}$$

Therefore, the likelihood ratio statistic for $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is

$$\begin{aligned}
\Lambda &= \frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)} \\
&= \frac{(2\pi)^{-\frac{np}{2}}\left|\hat{\boldsymbol{\Sigma}}_0\right|^{-\frac{n}{2}} e^{-\frac{np}{2}}}{(2\pi)^{-\frac{np}{2}}\left|\frac{1}{n}\boldsymbol{A}\right|^{-\frac{n}{2}} e^{-\frac{np}{2}}} \\
&= \left(\frac{|\frac{1}{n}\boldsymbol{A}|}{|\hat{\boldsymbol{\Sigma}}_0|}\right)^{\frac{n}{2}} \\
&= \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|}\right)^{\frac{n}{2}}
\end{aligned}$$

If the observed value of this likelihood ratio is too small, $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is unlikely to be true and so we will reject H_0 if

$$\Lambda = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|}\right)^{\frac{n}{2}} = \left(\frac{|\boldsymbol{A}|}{|\boldsymbol{A} + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'|}\right)^{\frac{n}{2}} < c_\alpha$$

where c_α is the lower $(100\alpha)^{\text{th}}$ percentile of the distribution of Λ.

Wilk's Lambda

We now define Wilk's lambda as

$$\Lambda^{\frac{2}{n}} = \frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|}$$

5.3.1 The link between Wilk's lambda and Hotelling's T^2

If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population, then the test based on T^2 is equivalent to the likelihood ratio test of $H_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ because

$$\Lambda^{\frac{2}{n}} = \left(1 + \frac{T^2}{n-1}\right)^{-1}$$

To prove this, we start by constructing a matrix

$$\boldsymbol{B}_{(p+1)\times(p+1)} = \begin{bmatrix} \boldsymbol{A} & \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \\ \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' & -1 \end{bmatrix}$$

Now, applying Theorem A.4 in Appendix A,

$$\begin{aligned}
|\boldsymbol{B}_{22}||\boldsymbol{B}_{11} - \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21}| &= |\boldsymbol{B}_{11}||\boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}| \\
|-1||\boldsymbol{A} - \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(-1)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\sqrt{n}| &= |\boldsymbol{A}||-1 - \sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{A}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\sqrt{n}| \\
|-1||\boldsymbol{A} + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'| &= |\boldsymbol{A}||-1 - n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{A}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)| \\
|-1||\boldsymbol{A} + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'| &= |-1|\left(1 + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\frac{1}{n-1}\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right)|\boldsymbol{A}| \\
\left|n\left[\frac{1}{n}\boldsymbol{A} + (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\right]\right| &= \left(1 + \frac{n}{n-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right)|\boldsymbol{A}| \\
|n\hat{\boldsymbol{\Sigma}}_0| &= \left(1 + \frac{1}{n-1}T^2\right)|\boldsymbol{A}| \\
\frac{|n\hat{\boldsymbol{\Sigma}}_0|}{|n\hat{\boldsymbol{\Sigma}}|} &= \left(1 + \frac{1}{n-1}T^2\right) \\
\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|} &= \left(1 + \frac{1}{n-1}T^2\right)^{-1} \\
\Lambda^{\frac{2}{n}} &= \left(1 + \frac{1}{n-1}T^2\right)^{-1}
\end{aligned}$$

What this shows is that T^2 can be computed from two determinants,

$$\begin{aligned}
T^2 &= \frac{(n-1)|\hat{\boldsymbol{\Sigma}}_0|)}{|\hat{\boldsymbol{\Sigma}}|} - (n-1) \\
&= \frac{(n-1)|\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'|}{|\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'|} - (n-1)
\end{aligned}$$

which does not depend on the inverse of $\boldsymbol{S}$.

5.4 Multivariate Confidence Regions

In the univariate case, we use confidence intervals – an interval around some unknown parameter – to convey information regarding the location of that parameter. Now if $\boldsymbol{\theta}$ is a vector of unknown parameters, a **confidence region** is a region of likely values for $\boldsymbol{\theta}$. The region, $R(\boldsymbol{X})$, is determined by the data, where $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]'$ is the data matrix.

$R(\boldsymbol{X})$ is said to be the $100(1-\alpha)\%$ “confidence region” if, before the sample is selected, the probability that $R(\boldsymbol{X})$ will cover the true $\boldsymbol{\theta}$ is $1-\alpha$.

For the mean, $\boldsymbol{\mu}$,

$$\Pr\left[n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right] = 1-\alpha$$

which implies that $\bar{\boldsymbol{X}}$ will be within $\left[\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right]^{\frac{1}{2}}$ of $\boldsymbol{\mu}$ with probability $1-\alpha$, if distance is defined in terms of $n\boldsymbol{S}^{-1}$.

For a particular sample, we can compute $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ and then

$$n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

will define the region $R(\boldsymbol{X})$. This region will be a p-dimensional ellipsoid centred at $\bar{\boldsymbol{x}}$ and will present the $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$.

When $p \ge 4$ we cannot graph the joint confidence region for $\boldsymbol{\mu}$, however, we can calculate the axes of the confidence ellipsoid and their relative lengths similarly to how we did it previously – by finding the eigenvalues (λ_i) and the eigenvectors ($\boldsymbol{e}_i$) of $\boldsymbol{S}$.

The direction and length of the axes of $n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \le c^2 = \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$ are determined by moving $\frac{\sqrt{\lambda_i}c_i}{\sqrt{n}} = \sqrt{\lambda_i}\sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)}$ units along the eigenvectors, $\boldsymbol{e}_i$.

Beginning at the centre $\bar{\boldsymbol{x}}$, the axes of the confidence ellipsoids are $\pm\sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}\boldsymbol{e}_i$ where $\boldsymbol{S}\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i,\ i = 1, \ldots, p$.

Thus the ratios of the λ_i will help us identify the relative amounts of elongation along pairs of axes.

Example

Example 5.3 on page 221 of Johnson & Wichern (also see `Lecture5.R`).

Radiation measurements were recorded (and transformed) for 42 microwave ovens – one measurement when the door is open, and one when it is closed.

```r
door_closed <- read.table(’T4-1.DAT’)
door_open <- read.table(’T4-5.DAT’)
mwave <- cbind(door_closed^0.25, door_open^0.25)
colnames(mwave) <- c(’x1’, ’x2’)
head(mwave)
```

```
         x1        x2
1 0.6223330 0.7400828
2 0.5477226 0.5477226
3 0.6513556 0.7400828
4 0.5623413 0.5623413
5 0.4728708 0.5623413
6 0.5885662 0.5885662
```

Now suppose we want to test whether $\boldsymbol{\mu}_0 = [0.562 \quad 0.589]'$ is within the 95% confidence region determined from the sample. This is equivalent to testing $H_0\colon \boldsymbol{\mu}_0 = [0.562 \quad 0.589]'$ against $H_1\colon \boldsymbol{\mu}_0 \neq [0.562 \quad 0.589]'$ at a significance level of $\alpha = 5\%$.

```
n <- nrow(mwave)
p <- ncol(mwave)
S <- cov(mwave)
xbar <- apply(mwave, 2 ,mean)

mu0 <- matrix(c(0.562, 0.589),2)

distance <- n*t(xbar - mu0)%*%solve(S)%*%(xbar - mu0)

alpha <- 0.05
c2 <- ((n-1)*p)/(n-p)*qf(1-alpha, p, n-p)

distance
c2
```

```
       [,1]
[1,] 1.2573

[1] 6.62504
```

Here we see that the standardised squared distance from the sample mean to the hypothesised mean is smaller than the distance to the constant density contour, which is exactly equivalent to saying that the test statistic is smaller than the critical value.

Therefore, the vector $\boldsymbol{\mu}_0$ will lie within the 2-dimensional ellipse describing the 95% confidence region for $\boldsymbol{\mu}$. The lengths of the ellipse's axes are once again determined from the eigenvalues of $\boldsymbol{S}$.

```
eigen <- eigen(S)

(axis1length <- sqrt(eigen$values[1])*sqrt(c2/n))
(axis2length <- sqrt(eigen$values[2])*sqrt(c2/n))
```

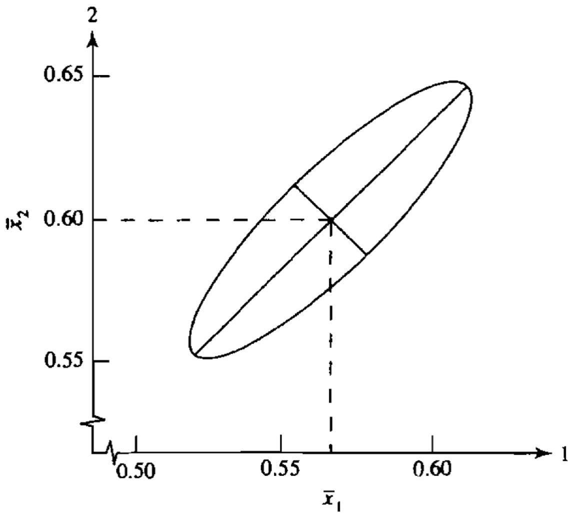
```
[1] 0.06424195
[1] 0.02075877
```

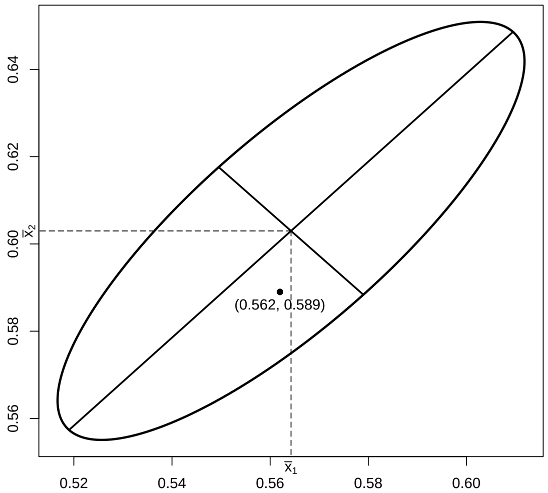
We can see that the major axis of the ellipse is 3.1 times longer than the minor axis.

```
> (axisratio <- sqrt(eigen$values[1])/sqrt(eigen$values[2]))
```

```
[1] 3.094689
```

Below we see the resulting confidence region ellipsoid, with Figure 5.1 from Johnson & Wichern on the left (using rounded values), and a recreation using exact values in R on the right. Note that the point $\boldsymbol{\mu}_0 = [0.562 \quad 0.589]'$ lies well within the bounds of the confidence region, as expected.





5.5 Simultaneous Confidence Intervals

If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then, simultaneously for all $\boldsymbol{a}$, the interval

$$\left(\boldsymbol{a}'\bar{\boldsymbol{X}} - \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}\ ;\ \boldsymbol{a}'\bar{\boldsymbol{X}} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}\right)$$

will contain $\boldsymbol{a}'\boldsymbol{\mu}$ with probability $1-\alpha$.

These are referred to as T^2 intervals since the coverage probability is determined by the distribution of T^2.

By choosing $\boldsymbol{a}' = [1, 0, \ldots, 0]$, $\boldsymbol{a}' = [0, 1, \ldots, 0]$, and so on, we obtain p confidence intervals of the form

$$\begin{aligned}
\bar{x}_1 - \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{11}}{n}} \leq & \mu_1 \leq \bar{x}_1 + \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{11}}{n}} \\
\bar{x}_2 - \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} \leq & \mu_2 \leq \bar{x}_2 + \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{22}}{n}} \\
& \vdots \\
\bar{x}_p - \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}} \leq & \mu_p \leq \bar{x}_p + \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{pp}}{n}}
\end{aligned}$$

that all hold simultaneously with probability $1-\alpha$.

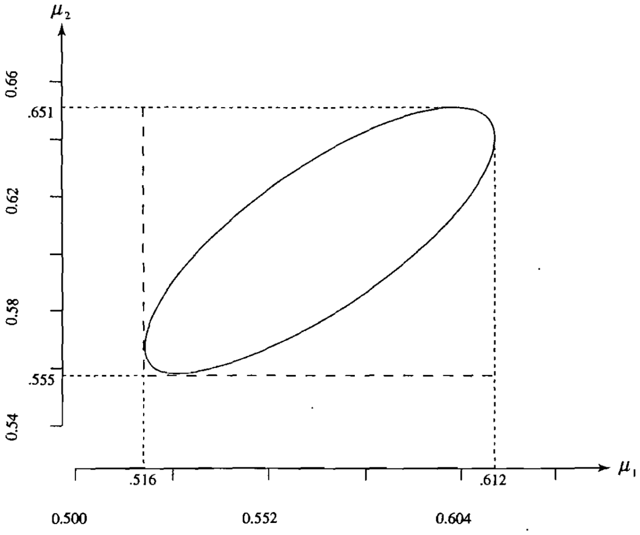
Example (continued)

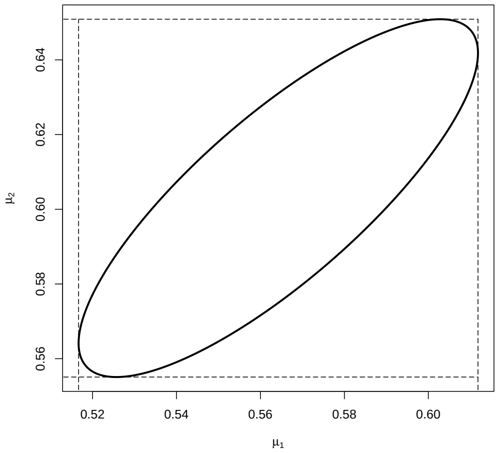
Let us calculate the simultaneous T^2 confidence intervals for the microwave radiation data in the previous example.

```r
(round(xbar[1] - sqrt(c2)*sqrt(S[1,1]/n), 3))
(round(xbar[1] + sqrt(c2)*sqrt(S[1,1]/n), 3))
(round(xbar[2] - sqrt(c2)*sqrt(S[2,2]/n), 3))
(round(xbar[2] + sqrt(c2)*sqrt(S[2,2]/n), 3))
```

```
x1
0.517
x1
0.612
x2
0.555
x2
0.651
```

Below we have a graphical illustration of the simultaneous intervals, again with Figure 5.2 from Johnson & Wichern on the left, and a recreation in R on the right. From this we can see that the simultaneous T^2 confidence intervals define the edges of the confidence region in each dimension, which can also be seen as the projection of the ellipse onto the axes of the component means.





5.5.1 Simultaneous versus one-at-a-time intervals

The form of one-at-a-time intervals, which ignores the covariance structure of the p variables, is

$$\bar{x}_i - t_{n-1}(\alpha/2)\sqrt{\frac{s_{ii}}{n}} \leq \mu_i \leq \bar{x}_i + t_{n-1}(\alpha/2)\sqrt{\frac{s_{ii}}{n}}$$

as opposed to the simultaneous interval

$$\bar{x}_i - \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{ii}}{n}} \leq \mu_i \leq \bar{x}_i + \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{ii}}{n}}.$$

Prior to sampling, the i^{th} individual interval has probability $1-\alpha$ of covering μ_i but we cannot say that ALL intervals will contain their respective μ_i with probability $1-\alpha$. Why not?

Consider the special case where $\mathbf{\Sigma} = diag(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})$, which implies that the p variables are independent. Thus

$$\Pr(\text{all t-intervals contain the } \mu_i's) = (1-\alpha)(1-\alpha)\ldots(1-\alpha) = (1-\alpha)^p < (1-\alpha) \ \forall \ \alpha > 0, p \geq 2$$

Therefore we need wider intervals and hence the different multipliers of $\sqrt{\frac{s_{ii}}{n}}$.

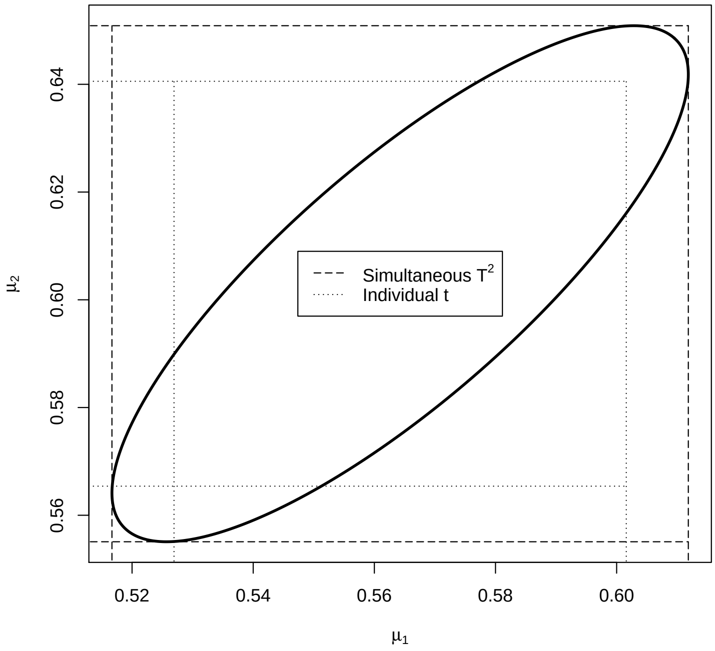
For example, for $1-\alpha = 0.95, n = 15, p = 4$,

$$\sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} = \sqrt{\frac{4 \times 14}{11} 3.36} = 4.14$$

versus

$$t_{n-1}(0.025) = 2.145.$$

In conclusion, the individual t-intervals are too precise, since they do not consider the covariance, whilst the simultaneous T^2-intervals are too wide if applied only to the p component means.



This leads us to considering the Bonferroni method.

5.6 Bonferroni Intervals

Let's assume we have m hypotheses to test or effects to estimate which can all be summarised through linear combinations of the form $\boldsymbol{a}_i'\boldsymbol{\mu},\ i = 1, \ldots, m$.

Let C_i denote a confidence statement about the value of $\boldsymbol{a}_i'\boldsymbol{\mu}$ with

$$\Pr[C_i \text{ true}] = 1 - \alpha_i,\ i = 1, \ldots, m.$$

Then the Bonferroni inequality leads to

$$\begin{aligned}\Pr[\text{all } C_i \text{ true}] &= 1 - \Pr[\text{at least one } C_i \text{ false}] \\ &\geq 1 - \sum_{i=1}^{m} \Pr(C_i \text{ false}) \\ &= 1 - \sum_{i=1}^{m} (1 - \Pr(C_i \text{ true})) \\ &= 1 - (\alpha_1 + \alpha_2 + \ldots + \alpha_m)\end{aligned}$$

Now let $\alpha_i = \dfrac{\alpha}{m}$ for the intervals $\bar{x} \pm t_{n-1}(\alpha_i/2)\sqrt{\dfrac{s_{ii}}{n}}$. Since

$$\Pr\left[\bar{X}_i \pm t_{n-1}(\alpha/2m)\sqrt{\frac{s_{ii}}{n}} \text{ contains } \mu_i\right] = 1 - \frac{\alpha}{m}$$

we have

$$\Pr\left[\bar{X}_i \pm t_{n-1}(\alpha/2m)\sqrt{\frac{s_{ii}}{n}} \text{ contains all } \mu_i\right] \geq 1 - \alpha$$

So for p means, the Bonferroni confidence intervals will be of the form

$$\bar{x}_i - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}} \leq \mu_i \leq \bar{x}_i + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}.$$

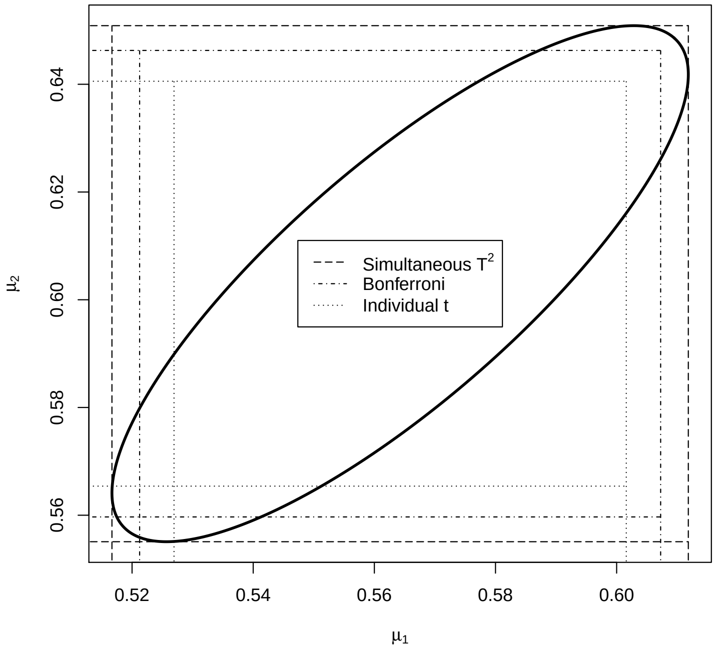
Example (continued)

Calculate the Bonferroni confidence intervals for the microwave radiation data:

```
(round(xbar[1] - qt(1-alpha/(2*p), n-1)*sqrt(S[1,1]/n), 3))
(round(xbar[1] + qt(1-alpha/(2*p), n-1)*sqrt(S[1,1]/n), 3))
(round(xbar[2] - qt(1-alpha/(2*p), n-1)*sqrt(S[2,2]/n), 3))
(round(xbar[2] + qt(1-alpha/(2*p), n-1)*sqrt(S[2,2]/n), 3))
```

```
x1
0.521
x1
0.607
x2
0.56
x2
0.646
```

The figure below illustrates that the Bonferroni intervals lie between the simultaneous T^2 and individual t-intervals.



5.7 Large Sample Inferences about the Population Mean

Without the assumption of a normal population, when $n >> p$ we base the confidence intervals on the χ^2-distribution such that.

$$\Pr\left[n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \leq \chi_p^2(\alpha)\right] \approx 1-\alpha$$

The intervals for linear combinations of the X variables are of the form

$$\boldsymbol{a}'\bar{\boldsymbol{X}} \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}{n}}$$

which leads to simultaneous confidence intervals for μ_i of the form

$$\bar{x}_i \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{s_{ii}}{n}}.$$

Homework exercise 5.1

Johnson & Wichern exercise 5.9

Harry Roberts, a naturalist for the Alaska Fish and Game department, studied grizzly bears with the goal of maintaining a healthy population. Measurements of $n = 61$ bears provided the summary statistics (sample mean and sample covariance matrix) given in the file `J&WEx5.9.RData`. The variables, in order corresponding to the summary statistics, are:

Weight (kg)	Body length (cm)	Neck (cm)	Girth (cm)	Head length (cm)	Head width (cm)

a) Obtain the large sample 95% simultaneous confidence intervals for the six population mean body measurements.

b) Obtain the large sample 95% simultaneous confidence ellipse for mean weight and mean girth.

c) Obtain the 95% Bonferroni confidence intervals for the six means in part a).

d) Refer to part b). Construct the 95% Bonferroni confidence rectangle for the mean weight and mean girth using $m = 6$. Compare this rectangle with the confidence ellipse in part b).

e) Obtain the 95% Bonferroni confidence interval for

$$\text{mean head width} - \text{mean head length}$$

using $m = 6 + 1 = 7$ to allow for this statement as well as statements about each individual mean.

5.8 Repeated measures

Let us return to the linear combinations of means we used to construct simultaneous T^2 confidence intervals in section 5.5. This approach can be extended to situations where q *repeated measures* are taken of some normally distributed variable (X) over time for n respondents, and we want to test whether the $\mu's$ at each time point are equal, or whether they change over time.

The data would be arranged as $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1 & \cdots & \boldsymbol{X}_n\end{bmatrix}$, where each $\boldsymbol{X}_j = \begin{bmatrix}X_{1j}\\X_{2j}\\\vdots\\X_{qj}\end{bmatrix}$ such that X_{ij} is the i^{th} measurement on the j^{th} respondent.

Now, testing the hypothesis $H_0\colon \mu_1 = \mu_2 = \cdots = \mu_q$ against the alternative that at least one μ_j differs, is equivalent to testing

$$H_0\colon \mu_1 - \mu_2 = \mu_1 - \mu_3 = \cdots = \mu_1 - \mu_q = 0$$

or

$$H_0\colon \mu_1 - \mu_2 = \mu_2 - \mu_3 = \cdots = \mu_{q-1} - \mu_q = 0$$

We can construct a contrast matrix containing contrast vectors $\boldsymbol{a}$ in the rows such that

$$\boldsymbol{C}_{(q-1)\times q} = \begin{bmatrix}1 & -1 & 0 & \ldots & 0\\1 & 0 & -1 & \ldots & 0\\\vdots & \vdots & \vdots & \ddots & \vdots\\1 & 0 & 0 & \ldots & -1\end{bmatrix} \quad \text{or } \boldsymbol{C}_{(q-1)\times q} = \begin{bmatrix}1 & -1 & 0 & \ldots & 0\\0 & 1 & -1 & \ldots & 0\\\vdots & \vdots & \vdots & \ddots & \vdots\\0 & 0 & 0 & \ldots & -1\end{bmatrix}$$

The hypotheses above can now be expressed as

$$H_0\colon \boldsymbol{C\mu} = \boldsymbol{0} \quad vs \quad H_1\colon \boldsymbol{C\mu} \neq \boldsymbol{0}$$

for either of these contrast matrices $\boldsymbol{C}$.

The vector of differences is constructed as

$$\boldsymbol{Y} = \boldsymbol{CX} \sim N_{q-1}(\boldsymbol{C\mu}, \boldsymbol{C\Sigma C'})$$

such that the resulting Hotelling's T^2 statistic is

$$T^2 = n(\boldsymbol{C\bar{X}})'(\boldsymbol{CSC'})^{-1}(\boldsymbol{C\bar{X}}) \sim \frac{(n-1)(q-1)}{(n-q+1)}F_{q-1,n-q+1}$$

Example

Example 6.10.1 on page 227 of Rencher (also see `Lecture 5.R`).

Dental measurement were taken for 11 girls and 16 boys at two-year intervals from age 8 to 14. For the purposes of illustration, we will only investigate whether there is a significant change in the mean measurements for the boys.

```
dental <- read.table(’T6_16_DENTAL.DAT’)
head(dental, 12)
```

```
   V1   V2   V3   V4   V5
1   1 21.0 20.0 21.5 23.0
2   1 21.0 21.5 24.0 25.5
3   1 20.5 24.0 24.5 26.0
4   1 23.5 24.5 25.0 26.5
5   1 21.5 23.0 22.5 23.5
6   1 20.0 21.0 21.0 22.5
7   1 21.5 22.5 23.0 25.0
8   1 23.0 23.0 23.5 24.0
9   1 20.0 21.0 22.0 21.5
10  1 16.5 19.0 19.0 19.5
11  1 24.5 25.0 28.0 28.0
12  2 26.0 25.0 29.0 31.0
```

```
boys <- dental[dental$V1 == 2, -1] #Extract boys’ data
xbar <- apply(boys, 2, mean)
S <- var(boys)
n <- nrow(boys)
q <- ncol(boys)

#Contrast matrix 1
(C1 <- cbind(1, diag(-1, q-1)))
```

```
     [,1] [,2] [,3] [,4]
[1,]    1   -1    0    0
[2,]    1    0   -1    0
[3,]    1    0    0   -1
```

Note that here we use the contrast matrix on the left as defined above. One can show that the results are exactly the same when using the other form.

```
#Test statistic and p-value
(T2 <- n*t(C1%*%xbar) %*% solve(C1%*%S%*%t(C1)) %*% (C1%*%xbar))
1 - pf(T2*(n-q+1)/((n-1)*(q-1)), q-1, n-q+1)
```

```
         [,1]
[1,] 77.95695

             [,1]
[1,] 1.998828e-05
```

With a p-value of 0.00002, we reject the null hypothesis of equal means and conclude that boys’ mean measurements differ at one or more of the ages. To investigate at which measurement time(s) the mean differs, we could conduct pairwise tests. We will return to this idea, combined with the Bonferroni approach, in the next chapter.

Homework exercise 5.2

Johnson & Wichern exercise 5.10

Refer to the bear growth data given in the file `T1-4.txt`, containing the weight (kg) and length (cm) of 7 female grizzly bears at ages 2, 3, 4 and 5 years respectively.

a) Obtain the 95% T^2 simultaneous confidence intervals for the four population means for length.

b) Refer to part a). Obtain the 95% T^2 simultaneous confidence intervals for the three successive yearly increases in mean length.

c) Obtain the 95% T^2 confidence ellipse for the mean increase in length from 2 to 3 years and the mean increase in length from 4 to 5 years.

d) Refer to parts a) and b). Construct the 95% Bonferroni confidence intervals for the set consisting of the four mean lengths and three successive yearly increases in mean length.

e) Refer to parts c) and d). Compare the 95% Bonferroni confidence rectangle for the mean increase in length from 2 to 3 years and the mean increase in length from 4 to 5 years with the confidence ellipse produced by the T^2-procedure.

6 MANOVA

Multivariate Analysis of Variance (MANOVA) refers to the comparison of groups with respect to many variables and is an extension of univariate analysis of variance (ANOVA), which focuses on just one response variable. As before, we are considering p jointly distributed variables. However, instead of just taking one sample of size n from some population, we will now be drawing samples of potentially different sizes from g populations (or groups), with the goal of comparing the mean vectors of each population.

These random samples can be arranged as

$$\begin{array}{rl} \text{Population 1:} & \boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1} \\ \text{Population 2:} & \boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2} \\ \vdots & \qquad \vdots \\ \text{Population g:} & \boldsymbol{X}_{g1}, \boldsymbol{X}_{g2}, \ldots, \boldsymbol{X}_{gn_g} \end{array}$$

Specifically, MANOVA investigates

1. whether the population mean vectors are the same, and
2. if not, which mean components differ.

There are some assumptions:

1. If $\boldsymbol{X}_{l1}, \boldsymbol{X}_{l2}, \ldots, \boldsymbol{X}_{ln_l}$ are a random sample of size n_l from a population with mean $\boldsymbol{\mu}_l, l = 1, \ldots g$, then random samples from different populations are *independent.*
2. All populations have a common covariance matrix, $\boldsymbol{\Sigma}$.
3. Each population is multivariate normal.

Before we define the MANOVA model, we will first give an overview of the univariate case when $p = 1$, i.e. ANOVA.

6.1 Univariate Analysis of Variance

Assume we are dealing with only one response variable, X_{lj} and wish to test

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_g$$

where $\mu_l = \mu + (\mu_l - \mu) = \mu + \tau_l$. We are interested in the deviations τ_l associated with the l^{th} population (treatment). The null hypothesis can be stated in terms of τ_l as

$$H_0\colon \tau_1 = \tau_2 = \cdots = \tau_g = 0.$$

The response $X_{lj} \sim N(\mu + \tau_l, \sigma^2)$ can be expressed as

$$X_{lj} = \underbrace{\mu}_{\text{overall mean}} + \underbrace{\tau_l}_{\text{treatment effect}} + \underbrace{e_{lj}}_{\text{random error}}$$

where $e_{lj} \sim N(0, \sigma^2)$.

To ensure uniqueness of parameters, use the constraint $\displaystyle\sum_{l=1}^{g} n_l \tau_l = 0$.

The ANOVA is based on the following decomposition of observations:

$$\begin{array}{lclclcl}
x_{lj} & = & \bar{x} & + & (\bar{x}_l - \bar{x}) & + & (x_{lj} - \bar{x}_l) \\
(x_{lj} - \bar{x})^2 & = & (\bar{x}_l - \bar{x})^2 & + & (x_{lj} - \bar{x}_l)^2 & + & 2(\bar{x}_l - \bar{x})(x_{lj} - \bar{x}_l) \\
\displaystyle\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x})^2 & = & \displaystyle\sum_{l=1}^{g} n_l(\bar{x}_l - \bar{x})^2 & + & \displaystyle\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x}_l)^2 & + & 0 \\
SS_{tot} & = & SS_{tr} & + & SS_{res} & & \\
\text{Total (Corrected) } SS & = & SS \text{ Between samples} & + & SS \text{ Within samples} & &
\end{array}$$

This is normally summarised in an ANOVA table for comparing univariate means

Source of Variation	Sum of Squares (SS)	Degrees of Freedom
Treatment	$SS_{tr} = \displaystyle\sum_{l=1}^{g} n_l(\bar{x}_l - \bar{x})^2$	$g - 1$
Residual (Error)	$SS_{res} = \displaystyle\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x}_l)^2$	$\displaystyle\sum_{l=1}^{g} n_l - g$
Total (corrected for the mean)	$SS_{tot} = \displaystyle\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x})^2$	$\displaystyle\sum_{l=1}^{g} n_l - 1$

Now to test

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0$$

at a given significance level α, we use the test statistic

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/(\sum_{l=1}^{g} n_l - g)}$$

and reject if $F > F_{g-1, \sum n_l - g}(\alpha)$.

Example

Example 6.7 on page 298 of Johnson & Wichern. Also see `Lecture6.R`, where all the components are manually calculated from the data and the results compared to that from the `aov()` function below.

$$\begin{array}{ll} \text{Population 1:} & 9, 6, 9 \\ \text{Population 2:} & 0, 2 \\ \text{Population 3:} & 3, 1, 2 \end{array}$$

```r
#Form data vectors
x1 <- matrix(c(9, 6, 9), 3)
x2 <- matrix(c(0, 2), 2)
x3 <- matrix(c(3, 1, 2), 3)

#Define n's
nl <- c(length(x1), length(x2), length(x3))

#Using aov function
X <- c(x1, x2, x3)
group <- c(rep(1, nl[1]), rep(2, nl[2]), rep(3, nl[3]))
group <- as.factor(group)
fit <- aov(X ~ group)
summary(fit)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
group         2     78      39    19.5 0.00435
Residuals     5     10       2
```

The p-value is small enough to conclude that at least one mean is significantly different from the others at a significance level of $\alpha = 0.5\%$.

6.2 MANOVA model for comparing g population mean vectors

We will now apply the same reasoning and methodology to mean *vectors* corresponding to vectors of observations. For some p-dimensional sets of observations, the model can now be expressed as

$$\boldsymbol{X}_{lj} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{e}_{lj}, j = 1, 2, \dots, n_l,\ \ l = 1, 2, \dots, g$$

where $\boldsymbol{e}_{lj} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and with $\displaystyle\sum_{l=1}^{g} n_l \boldsymbol{\tau}_l = \boldsymbol{0}$.

This implies that we have the following assumptions for this model:

- Each component of the observation vector $\boldsymbol{X}_{lj}$ satisfies the univariate model.
- The errors for the components of $\boldsymbol{X}_{lj}$ are correlated.
- The covariance matrix $\boldsymbol{\Sigma}$ is the same for all populations.

The multivariate decomposition of total variance is as follows:

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}})(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}})'}_{\text{Total (corrected for the mean) SSP}} = \underbrace{\sum_{l=1}^{g} n_l(\bar{\boldsymbol{x}}_l-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_l-\bar{\boldsymbol{x}})'}_{\text{Treatment (\underline{B}etween) SSP}} + \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}}_l)'}_{\text{Residual (\underline{W}ithin) SSP}}$$

where SSP = sum of squares and cross products.

Also note that $\boldsymbol{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)' = (n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2 + \ldots + (n_g - 1)\boldsymbol{S}_g$, where $\boldsymbol{S}_l$ = sample covariance matrix for the l^{th} sample. This can be viewed as a generalization of $(n_1 + n_2 - 2)S_p$ for the two-sample case.

This leads to the MANOVA table:

Source of Variation	Matrix of SSP's	Degrees of Freedom
Treatment	$\boldsymbol{B} = \sum_{l=1}^{g} n_l(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})'$	$g - 1$
Residual	$\boldsymbol{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)'$	$\sum_{l=1}^{g} n_l - g$
Total (corrected)	$\boldsymbol{B} + \boldsymbol{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})'$	$\sum_{l=1}^{g} n_l - 1$

To test H_0: $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \boldsymbol{\tau}_g = 0$, use Wilk's lambda,

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{W} + \boldsymbol{B}|} = \frac{|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}}_l)'|}{|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{lj} - \bar{\boldsymbol{x}})'|}$$

and reject if Λ^* is too small.

As we saw in section 5.3.1, this quantity is related to the likelihood ratio criterion. The distribution of Λ^* is tabulated, but for a few special cases we have the following exact distributions:

No. of variables	No. of groups	Sampling distribution for MVN data
$p = 1$	$g \geq 2$	$\left(\frac{\sum n_l - g}{g - 1}\right)\left(\frac{1 - \Lambda^*}{\Lambda^*}\right) \sim F_{g-1,\sum n_l - g}$
$p = 2$	$g \geq 2$	$\left(\frac{\sum n_l - g - 1}{g - 1}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(\sum n_l - g - 1)}$
$p \geq 1$	$g = 2$	$\left(\frac{\sum n_l - p - 1}{p}\right)\left(\frac{1 - \Lambda^*}{\Lambda^*}\right) \sim F_{p,\sum n_l - p - 1}$
$p \geq 1$	$g = 3$	$\left(\frac{\sum n_l - p - 2}{p}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(\sum n_l - p - 2)}$

When $n = \sum n_l$ is large,

$$-\left(n - 1 - \frac{(p+g)}{2}\right)\ln(\Lambda^*) \mathrel{\dot\sim} \chi^2_{p(g-1)}$$

and we will reject H_0 if $-\left(n - 1 - \frac{(p+g)}{2}\right)\ln\left(\frac{|\boldsymbol{W}|}{|\boldsymbol{W} + \boldsymbol{B}|}\right) > \chi^2_{p(g-1)}(\alpha)$.

Example

Example 6.9 on page 304 of Johnson & Wichern. Again see `Lecture6.R`, where all the components are manually calculated from the data and the results compared to that from the `manova()` function shown below.

$$
\begin{aligned}
&\text{Population 1:} \quad \begin{bmatrix}9\\3\end{bmatrix}, \begin{bmatrix}6\\2\end{bmatrix}, \begin{bmatrix}9\\7\end{bmatrix}\\
&\text{Population 2:} \quad \begin{bmatrix}0\\4\end{bmatrix}, \begin{bmatrix}2\\0\end{bmatrix}\\
&\text{Population 3:} \quad \begin{bmatrix}3\\8\end{bmatrix}, \begin{bmatrix}1\\9\end{bmatrix}, \begin{bmatrix}2\\7\end{bmatrix}
\end{aligned}
$$

```
#Form data matrices
X1 <- matrix(c(9,6,9,3,2,7), 3)
X2 <- matrix(c(0,2,4,0), 2)
X3 <- matrix(c(3,1,2,8,9,7), 3)

#Define n’s and p
nl <- c(nrow(X1), nrow(X2), nrow(X3))
p <- ncol(X1)

#Using manova function
X <- rbind(X1, X2, X3)
group <- c(rep(1, nl[1]), rep(2, nl[2]), rep(3, nl[3]))
group <- as.factor(group)
fit <- manova(X ~ group)
summary(fit, ’Wilks’)
```

```
          Df    Wilks approx F num Df den Df   Pr(>F)
group      2 0.038455   8.1989      4      8 0.006234
Residuals  5
```

Again we can conclude that at least one mean vector significantly differs from the others ($p < 0.01$). We can also extract the univariate ANOVA’s:

```
summary.aov(fit)
```

```
Response 1 :
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2     78      39    19.5 0.004353
Residuals    5     10       2
---

Response 2 :
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2     48    24.0       5 0.06415
Residuals    5     24     4.8
```

Here we specified the “Wilks” test statistic, which is one of four options in R.

```
summary(fit, ’?’)
```

```
Error in match.arg(test) :
’arg’ should be one of "Pillai", "Wilks", "Hotelling-Lawley", "Roy"
```

6.2.1 Other MANOVA test statistics

In the above example we notice other options of test statistics for the MANOVA model. Before considering variations on Wilk's lambda, we first note that it can also be expressed as a function of the eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_s$ of $\boldsymbol{W}^{-1}\boldsymbol{B}$, where $s = \min(p, g-1)$, such that

$$\Lambda^* = \prod_{i=1}^{s} \left(\frac{1}{1+\lambda_i} \right)$$

Although Wilk's lambda is primarily used for MANOVA significance tests, the following test statistics – all of which are based on the above ordered eigenvalues and can be linked back to Hotelling's T^2 – have also been developed:

Roy's largest root:

$$\theta = \lambda_1$$

Pillai's V criterion / Pillai's trace:

$$V^{(s)} = tr\left[(\boldsymbol{W} + \boldsymbol{B})^{-1}\boldsymbol{B} \right] = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$$

The distributions of these two statistics are tabulated.

Lawley-Hotelling trace

$$U^{(s)} = tr\left(\boldsymbol{W}^{-1}\boldsymbol{B} \right) = \sum_{i=1}^{s} \lambda_i$$

which is approximately $\chi^2_{p(g-1)}$ distributed for large n.

When $H_0\colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$ is true, all the mean vectors are at the same point. Therefore, all four MANOVA test statistics have the same Type I error rate, α; that is, all have the same probability of rejection when H_0 is true. However, when H_0 is false, the four tests have different probabilities of rejection.

Historically, Wilks' lambda has played the dominant role in significance tests in MANOVA because it was the first to be derived and has well-known χ^2- and F-approximations. It can also be partitioned in certain useful ways. However, it is not always the most powerful among the four tests, where power is measured as the probability of rejecting H_0 when it is false.

Generally, if group sizes are equal, the tests are sufficiently robust with respect to heterogeneity of covariance matrices so that we need not worry. If the n_i's are unequal and we have heteroscedasticity, then the α-level of the MANOVA test may be affected as follows:

If the larger variances and covariances are associated with the larger samples, the true α-level is reduced and the tests become conservative. On the other hand, if the larger variances and covariances come from the smaller samples, α is inflated, and the tests become liberal. Box's M-test can be used to test for homogeneity of covariance matrices.

6.2.2 Testing for equality of covariance matrices

In the MANOVA model above, it was assumed that all $\boldsymbol{e}_{lj} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and independent with the same covariance matrix $\boldsymbol{\Sigma}$ for all g groups. To validate this assumption, we use Box's M test to test the hypothesis

$$H_0\colon \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g$$

The test statistic is derived with the likelihood ratio testing method, with likelihood ratio test statistic

$$\Lambda = \prod_{i=1}^{g} \left(\frac{|\boldsymbol{S}_i|}{|\boldsymbol{S}_p|} \right)^{\frac{n_i - 1}{2}}$$

where

$$\boldsymbol{S}_p = \frac{\sum_{i=1}^{g} (n_i - 1) \boldsymbol{S}_i}{\sum_{i=1}^{g} (n_i - 1)}$$

By calculating $-2 \ln \Lambda$, we derive the test statistic for Box's M-Test:

$$M = \left(\sum_{i=1}^{g} (n_i - 1) \right) \ln |\boldsymbol{S}_p| - \left(\sum_{i=1}^{g} (n_i - 1) \ln |\boldsymbol{S}_i| \right)$$

The test statistic, after applying appropriate scaling, has the following approximate sampling distribution:

$$\left[1 - \left(\sum_{i=1}^{g} \frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{g} (n_i - 1)} \right) \frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \right] M \sim \chi^2_{\frac{p(p+1)(g-1)}{2}}$$

We do note that this test is sensitive to some forms of non-normality. For example, it is sensitive to kurtosis for which the MANOVA tests are rather robust. Thus the M-test may signal covariance heterogeneity in some cases where it is not damaging to the MANOVA tests. Hence we may not wish to automatically rule out standard MANOVA tests if the M-test leads to rejection of H_0.

6.3 Simultaneous Confidence Intervals

If we reject the hypothesis of equal mean vectors, we then want to find out which groups are different. This can be investigated by creating simultaneous confidence intervals for $\boldsymbol{\tau}_k - \boldsymbol{\tau}_l$ using the Bonferroni approach.

Since $\boldsymbol{\tau}_k$ is estimated by $\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}$,

$$\hat{\tau}_{ki} - \hat{\tau}_{li} = \bar{x}_{ki} - \bar{x}_{li},$$

which denotes the difference between two independent sample means, for $i = 1, \ldots, p$.

The variance of this contrast is $\text{Var}(\hat{\tau}_{ki} - \hat{\tau}_{li}) = \text{Var}(\bar{X}_{ki} - \bar{X}_{li}) = \left(\frac{1}{n_k} + \frac{1}{n_l} \right) \sigma_{ii}$, which can be estimated by

$$\left(\frac{1}{n_k} + \frac{1}{n_l} \right) \frac{w_{ii}}{n - g}$$

where w_{ii} is the i^{th} diagonal element of $\boldsymbol{W}$, and $n = \sum_{j=1}^{g} n_j$.

To apportion the error rate over the numerous confidence intervals, we have p variables and $\frac{g(g-1)}{2}$ pairwise differences, yielding $m = \frac{pg(g-1)}{2}$ simultaneous confidence statements. This results in a confidence interval for $\tau_{ki} - \tau_{li}$ of

$$\bar{x}_{ki} - \bar{x}_{li} \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\frac{w_{ii}}{n-g}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}$$

for all components $i = 1, \ldots, p$ and all differences $l < k = 1, \ldots, g$.

Example

Example 6.11 on page 309 of Johnson & Wichern. A study compared different kinds of nursing homes – private (271), non-profit(138) and government(107) – with respect to cost of (1) nursing, (2) dietary labour, (3) maintenance, and (4) housekeeping. Therefore, we have $n = 516$, $p = 4$, and $g = 3$.

The sample means and covariance matrices for each group are provided in the file `J&WEx6.11.RData`.

```
load('J&WEx6.11.RData')
nl <- c(271, 138, 107) ; p <- 4 ; g <- 3
n <- sum(nl)

#SSP Within
W <- (nl[1] - 1)*S1 + (nl[2] - 1)*S2 + (nl[3] - 1)*S3

#SSP Between
Xbar <- (nl[1]*X1_bar + nl[2]*X2_bar + nl[3]*X3_bar)/n
B <- nl[1]*(X1_bar - Xbar)%*%t(X1_bar - Xbar) +
nl[2]*(X2_bar - Xbar)%*%t(X2_bar - Xbar) +
nl[3]*(X3_bar - Xbar)%*%t(X3_bar - Xbar)

#Wilk's Lambda
Lam <- det(W)/(det(B+W))

#p = 2 and g = 3 (special case)
#Calculate test statistic
(F <- (n - p - 2)/p*(1-sqrt(Lam))/sqrt(Lam))
```

```
[1] 18.48786
```

```
> #Calculate p-value
> (pval <- 1 - pf(F, 2*p, 2*(n - p - 2)))
```

```
[1] 0
```

We observe a statistically highly significant difference. Note that we have enough data to have also applied the large sample chi-square approximation.

```
(c <- -(n - 1 - (p+g)/2)*log(Lam))
1 - pchisq(c, p*(g-1))
```

```
[1] 138.5215
[1] 0
```

Now that we know there is a difference in the mean vectors, we can determine where these difference are by calculating the simultaneous Bonferroni confidence intervals, starting with the difference between private and non-profit homes:

```
> #Differences between treatment effects
> tau_12 <- X1_bar - X2_bar
> tau_13 <- X1_bar - X3_bar
> tau_23 <- X2_bar - X3_bar

> #set alpha
> alpha <- 0.05

> #Calculate simultaneous Bonferroni CI’s
> options(digits = 2)
> CI_12 <- cbind(tau_12, tau_12) +
qt(1 - alpha/(p*g*(g-1)), n - g)*sqrt(diag(W)/(n-g)*(1/nl[1] + 1/nl[2]))%*%t(c(-1, 1))
> CI_12
```

```
       [,1]   [,2]
[1,] -0.281  0.079
[2,] -0.154 -0.078
[3,] -0.058 -0.026
[4,] -0.092 -0.024
```

We note that three of the four variables have a negative mean difference that is significant at $\alpha = 5\%$. Next we will consider the difference between private and government homes:

```
> CI_13 <- cbind(tau_13, tau_13) +
qt(1 - alpha/(p*g*(g-1)), n - g)*sqrt(diag(W)/(n-g)*(1/nl[1] + 1/nl[3]))%*%t(c(-1, 1))
> CI_13
```

```
       [,1]     [,2]
[1,] -0.403 -0.01068
[2,] -0.083  0.00054
[3,] -0.060 -0.02596
[4,] -0.061  0.01450
```

And finally the difference between non-profit and government homes:

```
> CI_23 <- cbind(tau_23, tau_23) +
qt(1 - alpha/(p*g*(g-1)), n - g)*sqrt(diag(W)/(n-g)*(1/nl[2] + 1/nl[3]))%*%t(c(-1, 1))
> CI_23
```

```
        [,1]  [,2]
[1,] -0.3275 0.115
[2,]  0.0281 0.122
[3,] -0.0202 0.018
[4,] -0.0073 0.077
```

Let’s create a summary of which intervals contain zero, in which case the difference is not significant.

```
> CI_contains_zero <- matrix(NA, p, 3)
> colnames(CI_contains_zero) <- c(’Private-Nonprofit’, ’Private-Government’, ’Nonprofit-Government’)
> CI_contains_zero[,1] <- (CI_12[,1] * CI_12[,2]) < 0
> CI_contains_zero[,2] <- (CI_13[,1] * CI_13[,2]) < 0
> CI_contains_zero[,3] <- (CI_23[,1] * CI_23[,2]) < 0

> CI_contains_zero
```

```
     Private-Nonprofit Private-Government Nonprofit-Government
[1,]              TRUE              FALSE                 TRUE
[2,]             FALSE               TRUE                FALSE
[3,]             FALSE              FALSE                 TRUE
[4,]             FALSE               TRUE                 TRUE
```

Based on the simultaneous confidence intervals derived using the Bonferroni method, we can see that the biggest differences in treatment effects were observed between private and non-profit nursing homes. The means of dietary labour, maintenance and housekeeping were simultaneously significant at $\alpha = 5\%$.

Homework exercise 6.1

Johnson & Wichern exercise 6.20 (adapted)

The tail lengths in mm (X_1) and wing length in mm (X_2) for 45 male hook-billed kites are given in `T6-11.dat`. Similar measurements for female hook-billed kites are given in `T5-12.dat`

a) Plot the male hook-billed data as a scatter diagram. Do any observations seem suspicious?

b) Test for the equality of mean vectors for male and female hook-billed kites.

c) Calculate simultaneous Bonferroni confidence intervals for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

6.4 Two-way Designs

Thus far we have only considered data grouped according to one factor. When including another fixed effect, this is referred to as a two-way design. We will once again start by recapping the univariate case before extending it to multiple variables.

6.4.1 Univariate two-way ANOVA

Suppose we have factor 1 with g levels, and factor 2 with b levels. Consider now n observations from each of the gb combinations of levels, all of which we will assume are independent.

Let X_{lkr} denote the r^{th} observation at level l of factor 1 and level k of factor 2. Then the univariate two-way model is given by

$$X_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + e_{lkr}, \ l = 1, \dots, g, \ k = 1, \dots, b, \ r = 1, \dots, n$$

with

$$\sum_{l=1}^{g} \tau_l = \sum_{k=1}^{b} \beta_k = \sum_{l=1}^{g} \gamma_{lk} = \sum_{k=1}^{b} \gamma_{lk} = 0$$

where

- μ represents an overall mean
- τ_l represents the fixed effect of factor 1
- β_k represents the fixed effect of factor 2
- γ_{lk} represents the interaction between factors 1 and 2
- $e_{lkr} \sim N(0, \sigma^2)$ are independent random variables

The model parameters can be estimated using

$$x_{lkr} = \bar{x} + (\bar{x}_{l\cdot} - \bar{x}) + (\bar{x}_{\cdot k} - \bar{x}) + (\bar{x}_{lk} - \bar{x}_{l\cdot} - \bar{x}_{\cdot k} + \bar{x}) + (x_{lkr} - \bar{x}_{lk})$$

Squaring and summing gives

$$\underbrace{\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{lkr}-\bar{x})^2}_{SS_{Tot}} =$$

$$\underbrace{\sum_{l=1}^{g} bn(\bar{x}_{l\cdot}-\bar{x})^2}_{SS_{fac1}} + \underbrace{\sum_{k=1}^{b} gn(\bar{x}_{\cdot k}-\bar{x})^2}_{SS_{fac2}} + \underbrace{\sum_{l=1}^{g}\sum_{k=1}^{b} n(\bar{x}_{lk}-\bar{x}_{l\cdot}-\bar{x}_{\cdot k}+\bar{x})^2}_{SS_{int}} + \underbrace{\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{lkr}-\bar{x}_{lk})^2}_{SS_{res}}$$

Source of Variation	Sum of Squares	Degrees of Freedom
Factor 1	$\sum_{l=1}^{g} bn(\bar{x}_{l\cdot}-\bar{x})^2$	$g-1$
Factor 2	$\sum_{k=1}^{b} gn(\bar{x}_{\cdot k}-\bar{x})^2$	$b-1$
Interaction	$\sum_{l=1}^{g}\sum_{k=1}^{b} n(\bar{x}_{lk}-\bar{x}_{l\cdot}-\bar{x}_{\cdot k}+\bar{x})^2$	$(g-1)(b-1)$
Residual (Error)	$\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{lkr}-\bar{x}_{lk})^2$	$gb(n-1)$
Total (corrected for the mean)	$\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{lkr}-\bar{x})^2$	$gbn-1$

6.4.2 Two-way MANOVA model

Consider now a vector response with p components, and once again two factors with g and b levels respectively. The two-way MANOVA model is then defined as

$$\boldsymbol{X}_{lkr} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{lk} + \boldsymbol{e}_{lkr}$$

with associated sum of squares and cross products,

$$\begin{aligned}\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(\boldsymbol{x}_{lkr}-\bar{\boldsymbol{x}})(\boldsymbol{x}_{lkr}-\bar{\boldsymbol{x}})' &= \sum_{l=1}^{g} bn(\bar{\boldsymbol{x}}_{l\cdot}-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{l\cdot}-\bar{\boldsymbol{x}})' \\ &+ \sum_{k=1}^{b} gn(\bar{\boldsymbol{x}}_{\cdot k}-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{\cdot k}-\bar{\boldsymbol{x}})' \\ &+ \sum_{l=1}^{g}\sum_{k=1}^{b}(\bar{\boldsymbol{x}}_{lk}-\bar{\boldsymbol{x}}_{l\cdot}-\bar{\boldsymbol{x}}_{\cdot k}+\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{lk}-\bar{\boldsymbol{x}}_{l\cdot}-\bar{\boldsymbol{x}}_{\cdot k}+\bar{\boldsymbol{x}})' \\ &+ \sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(\boldsymbol{x}_{lkr}-\bar{\boldsymbol{x}}_{lk})(\boldsymbol{x}_{lkr}-\bar{\boldsymbol{x}}_{lk})'\end{aligned}$$

Source of Variation	Sum of Squares and cross-Products	Degrees of Freedom
Factor 1	$\sum_{l=1}^{g} bn(\bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}})'$	$g-1$
Factor 2	$\sum_{k=1}^{b} gn(\bar{\boldsymbol{x}}_{\cdot k} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{\cdot k} - \bar{\boldsymbol{x}})'$	$b-1$
Interaction	$\sum_{l=1}^{g}\sum_{k=1}^{b} n(\bar{\boldsymbol{x}}_{lk} - \bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}}_{\cdot k} + \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{lk} - \bar{\boldsymbol{x}}_{l\cdot} - \bar{\boldsymbol{x}}_{\cdot k} + \bar{\boldsymbol{x}})'$	$(g-1)(b-1)$
Residual (Error)	$\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(\boldsymbol{x}_{lkr} - \bar{\boldsymbol{x}}_{lk})(\boldsymbol{x}_{lkr} - \bar{\boldsymbol{x}}_{lk})'$	$gb(n-1)$
Total (corrected)	$\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(\boldsymbol{x}_{lkr} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{lkr} - \bar{\boldsymbol{x}})'$	$gbn-1$

We can use the following statistics, with the specified distributions, to test for significance of the individual factors and their interaction:

Factor 1

Hypothesis:

$$H_0\colon \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \boldsymbol{\tau}_g = \mathbf{0} \quad \text{(No factor 1 effect)}$$

vs

$$H_1\colon \text{At least one } \boldsymbol{\tau}_l \neq \mathbf{0}$$

Test statistic:

$$\Lambda^* = \frac{|SSP_{res}|}{|SSP_{fac1} + SSP_{res}|}$$

Distribution:

$$-\left[gb(n-1) - \frac{p+1-(g-1)}{2}\right] \ln \Lambda^* \sim \chi^2_{(g-1)p}$$

Factor 2

Hypothesis:

$$H_0\colon \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_b = \mathbf{0} \quad \text{(No factor 2 effect)}$$

vs

$$H_1\colon \text{At least one } \boldsymbol{\beta}_k \neq \mathbf{0}$$

Test statistic:

$$\Lambda^* = \frac{|SSP_{res}|}{|SSP_{fac2} + SSP_{res}|}$$

Distribution:

$$-\left[gb(n-1) - \frac{p+1-(b-1)}{2}\right] \ln \Lambda^* \sim \chi^2_{(b-1)p}$$

Interaction

Hypothesis:

$$H_0\colon \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \ldots = \boldsymbol{\gamma}_{gb} = \mathbf{0} \quad \text{(No interaction effects)}$$

vs

$$H_1\colon \text{At least one } \boldsymbol{\gamma}_{lk} \neq \mathbf{0}$$

Test statistic:

$$\Lambda^* = \frac{|SSP_{res}|}{|SSP_{int} + SSP_{res}|}$$

Distribution:

$$-\left[gb(n-1) - \frac{p+1-(g-1)(b-1)}{2}\right] \ln \Lambda^* \sim \chi^2_{(g-1)(b-1)p}$$

The interaction test is usually conducted first. If interaction effects are significant, the factor effects do not have a clear interpretation. In this case it is wise to rather proceed with p univariate two-way ANOVA's to investigate whether interaction is present between some responses but not others.

If the interaction effects are very small, we focus on the contrasts in the factor 1 and factor 2 effects. Simultaneous confidence intervals for the i^{th} component of the factor 1 contrast, relating to the i^{th} variable, can be constructed using the Bonferroni method:

$$\tau_{li} - \tau_{mi} \text{ belongs to } (\bar{x}_{li\cdot} - \bar{x}_{mi\cdot}) \pm t_v\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\frac{E_{ii}}{v}\frac{2}{bn}}$$

where $v = gb(n-1)$ and E_{ii} is the i^{th} diagonal element of $\boldsymbol{E} = SSP_{res}$.

Likewise, for the factor 2 contrasts:

$$\beta_{ki} - \beta_{qi} \text{ belongs to } (\bar{x}_{ki\cdot} - \bar{x}_{qi\cdot}) \pm t_v\left(\frac{\alpha}{pb(b-1)}\right)\sqrt{\frac{E_{ii}}{v}\frac{2}{gn}}$$

Example

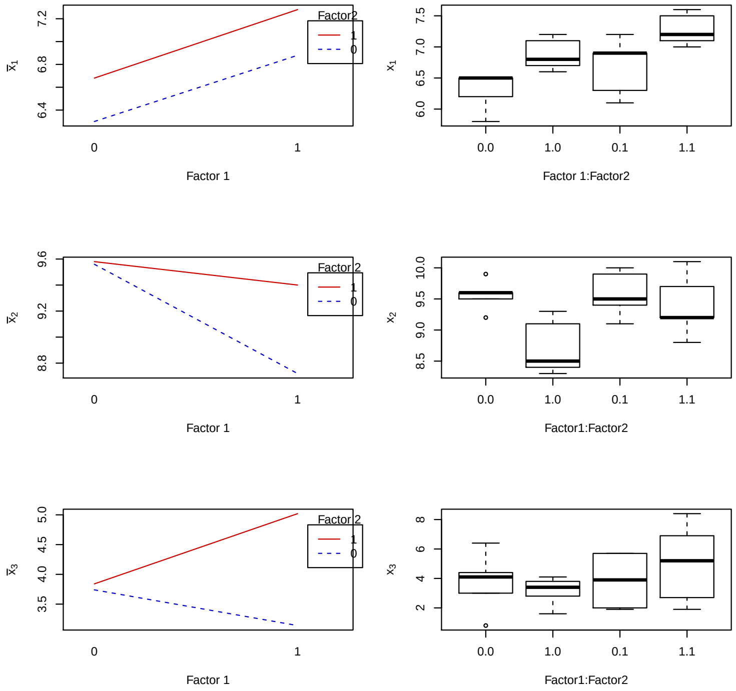
Example 6.13 on page 318 of Johnson & Wichern. We have three variables measuring the quality of plastic film: X_1 = tear resistance, X_2 =gloss, and X_3 =opacity. $n = 5$ observations were made for each combination of 2 factors, each with two levels: Factor 1 = Change in rate of extrusion, and Factor 2 = Amount of additive. The data are provided in `T6-4.dat` (also see `Lecture6.R`).

```
> plastic <- read.table('T6-4.dat')
> attach(plastic)
> X <- cbind(V3, V4, V5)

> #Fit two-way MANOVA
> summary(manova(X ~ V1*V2), test="Wilks")
```

```
          Df   Wilks approx F num Df den Df   Pr(>F)
V1         1 0.38186   7.5543      3     14 0.003034
V2         1 0.52303   4.2556      3     14 0.024745
V1:V2      1 0.77711   1.3385      3     14 0.301782
Residuals 16
```

We do not observe significant interaction. Note that both factor effects appear to be significant, which could be further explored. Below are the interaction plots for each variable separately, illustrating that although there is some interaction, it is not statistically significant.



Homework exercise 6.2

Johnson & Wichern exercise 6.31

Peanuts are an important crop in parts of the southern US. In an effort to develop improved plants, crop scientists routinely compare varieties with respect to several variables. The data for one two-factor experiment are given in `T6-17.dat`. Three varieties were grown at two locations. Three variables were measured:

X_1 = Yield (plot weight)
X_2 = Sound mature kernels (weight in grams – max 250g)
X_3 = Seed size (weight in grams of 100 seeds)

a) Perform a two-factor MANOVA, testing for variety effect, location effect, and variety-location interaction.

b) Analyse the residuals from part a). Do the usual MANOVA assumptions appear to be satisfied?

References

Johnson, R.A. and D.W. Wichern (2007). *Applied Multivariate Statistical Analysis, 6th ed.* Pearson Prentice Hall.

Rencher, A.C. and W.F. Christensen (2012). *Methods of Multivariate Analysis, 3rd ed.* John Wiley & Sons, Inc.

Appendix A
Matrix Algebra: Differentiation

Differentiation with respect to a vector $\boldsymbol{x}_{p\times 1}$

$$\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(x_1, \dots, x_p) \\ \vdots \\ \frac{\partial}{\partial x_p} f(x_1, \dots, x_p) \end{bmatrix}$$

Differentiation with respect to a matrix $\boldsymbol{X}_{n\times p}$

$$\frac{\partial}{\partial \boldsymbol{X}} f(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial}{\partial x_{11}} f(x_{11}, \dots, x_{np}) & \dots & \frac{\partial}{\partial x_{1p}} f(x_{11}, \dots, x_{np}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_{n1}} f(x_{11}, \dots, x_{np}) & \dots & \frac{\partial}{\partial x_{np}} f(x_{11}, \dots, x_{np}) \end{bmatrix}$$

The proofs for the following theorems are not required for this course.

Theorem A.1

If $\boldsymbol{A}_{p\times p}$ is a positive definite symmetric matrix, then

$$\frac{\partial \{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$$

Theorem A.2

If $\boldsymbol{A}$ is a non-singular matrix, then

$$\frac{\partial \{\log |\boldsymbol{A}|\}}{\partial \boldsymbol{A}} = \left(\boldsymbol{A}^{-1}\right)'$$

Theorem A.3

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices of dimensions such that $\boldsymbol{AB}$ is conformable, then

$$\frac{\partial \{tr(\boldsymbol{AB})\}}{\partial \boldsymbol{B}} = \boldsymbol{A}'$$

Theorem A.4

For any matrix $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}$,

$$|\boldsymbol{B}| = |\boldsymbol{B}_{22}||\boldsymbol{B}_{11} - \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21}| = |\boldsymbol{B}_{11}||\boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}|$$

Appendix B
Tables

Table B.1

Critical Points for Q-Q Plot Correlation Coefficient Test for Normality

Sample Size (n)	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
5	0.8299	0.8788	0.9032
10	0.8801	0.9198	0.9351
15	0.9126	0.9389	0.9503
20	0.9269	0.9508	0.9604
25	0.9410	0.9591	0.9665
30	0.9479	0.9652	0.9715
35	0.9538	0.9682	0.9740
40	0.9599	0.9726	0.9771
45	0.9632	0.9749	0.9792
50	0.9671	0.9768	0.9809
55	0.9695	0.9787	0.9822
60	0.9720	0.9801	0.9836
75	0.9771	0.9838	0.9866
100	0.9822	0.9873	0.9895
150	0.9879	0.9913	0.9928
200	0.9905	0.9931	0.9942
300	0.9935	0.9953	0.9960

Source: Table 4.2 Johnson and Wichern (2007, p. 181)