# An Introduction to Statistics

**Vol. 2**

**Bjarne Kjær Ersbøll and Knut Conradsen**

**7. edition - Preliminary version in English**

**Kgs. Lyngby 2007**

IMM

Printed by IMM

# Preface

This is the 7 edition of the textbook for course 02409, Multivariate Statistics. The first edition where the main parts (corresponding to the course curriculum) were translated to English was edition 6. Compared to that a large number of corrections have been made.

Errors and suggestions for corrections are very welcome.

Knut Conradsen and Bjarne Kjær Ersbøll (`be@imm.dtu.dk`)

# Contents

# Chapter 1
# Summary of linear algebra

This chapter contains a summary of linear algebra  with special emphasis on its use in statistics. The chapter is not intended to be an introduction to the subject. Rather it is a summary of an already known subject. Therefor we will not give very many examples within the areas typically covered in algebra and geometry courses. However, we will give more examples and sometimes proofs within areas which usually do not receive much attention in all-round courses, but which do enjoy significant use within algebra in statistics.

In recent years one has started to involve concepts like dual vector space in the theory of multidimensional normal analysis. Despite the advantages this might bring the author has chosen not to follow this line. Therefore the subject is not covered in this summary.

In the course of analysis of multidimensional statistical problems one often needs to invert non-regular matrices. For instance this is the case if one considers a problem given on a true sub-space of the considered $n$-dimensional vector-space. Instead of just considering the relevant sub-space, many (= most) authors prefer giving partly algebraic solutions by introducing the so-called pseudo-inverse of a non-regular matrix. In order to ease the reading of other literature (e.g. journals) we will introduce this concept and try to visualize it geometrically.

We note that use of pseudo-inverse matrices gives a very convenient way to solve many matrix equations in an algorithmic form.

## 1.1 Vector space

We start by giving an overview of the definition and elementary properties in the fundamental concept of a linear vector space.

### 1.1.1 Definition of a vector space

A **vector space (on the real numbers)** is a set $V$ with a composition rule + in the set $V \times V \to V$ which is called **vector addition** and a composition rule $\cdot$ in $R \times V \to V$ called **scalar multiplication**, which obey

i) $\forall \boldsymbol{u}, \boldsymbol{v} \in V: \quad \boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$ ( commutative law for vector addition)

ii) $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x} \in V: \quad \boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{x}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{x}$ (associative law for vector addition)

iii) $\exists \mathbf{0} \in V \forall \boldsymbol{u} \in V: \quad \boldsymbol{u} + \mathbf{0} = \boldsymbol{u}$ ( existence of a neutral element)

iv) $\forall \boldsymbol{u} \in V \exists - \boldsymbol{u} \in V: \ \boldsymbol{u} + (-\boldsymbol{u}) = \mathbf{0}$ ( existence on an inverse element)

v) $\forall \lambda \in R \forall \boldsymbol{u}, \boldsymbol{v} \in V: \quad \lambda(\boldsymbol{u} + \boldsymbol{v}) = \lambda \boldsymbol{u} + \lambda \boldsymbol{v}$ ( distributive law for scalar multiplication)

vi) $\forall \lambda_1, \lambda_2 \in R \forall \boldsymbol{u} \in V: \quad (\lambda_1 + \lambda_2)\boldsymbol{u} = \lambda_1 \boldsymbol{u} + \lambda_2 \boldsymbol{u}$ ( distributive law for scalar multiplication)

vii) $\forall \lambda_1, \lambda_2 \in R \forall \boldsymbol{u} \in V: \quad (\lambda_1 \lambda_2)\boldsymbol{u} = \lambda_1(\lambda_2 \boldsymbol{u})$ ( associative law for scalar multiplication)

viii) $\forall \boldsymbol{u} \in V: \quad 1\boldsymbol{u} = \boldsymbol{u}$

**EXAMPLE 1.1.** It is readily shown that all ordered $n$-tuples

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

of real numbers constitute a vector space, if the compositions are defined by element, i.e.

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

and

$$\lambda \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{bmatrix}$$

This vector space is denoted $R^n$ ♦

A vector space $U$ which is subset of a vector space $V$ is called a **subspace** in $V$. On the other hand, if we consider vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in V$, we can define

$$\mathrm{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$$

as the smallest subspace of $V$, which contains $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$. It is easily shown that

$$\mathrm{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\} = \{\sum_{i=1}^{k} \alpha_i \boldsymbol{v}_i | \alpha_i \in R, \quad i = 1, \ldots, k\}.$$

A vector of the form $\sum \alpha_i \boldsymbol{v}_i$ is called a linear combination of the vectors $\boldsymbol{v}_i$, $i = 1, \ldots, k$. The above result can then be expressed such that $\mathrm{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ precisely consists of all linear combinations of the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. Generally we define

$$\mathrm{span}(U_1, \ldots, U_p)$$

where $U_i \subseteq V$, as the smallest subspace of $V$, which contains all $U_i$, $i = 1, \ldots, p$.

A side-subspace is a set of the form

$$\boldsymbol{v} + U = \{\boldsymbol{v} + \boldsymbol{u} | \boldsymbol{u} \in U\},$$

where $U$ is a sub-space in $V$.

The situation is sketched in fig. 1.1.

Vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are said to be linearly independent if the relation

$$\alpha_1 \boldsymbol{v}_1 + \cdots + \alpha_n \boldsymbol{v}_n = 0$$

implies that

$$\alpha_1 = \cdots = \alpha_n = 0$$

In the opposite case they are said to be linearly dependent and at least one of them can be expressed as a linear combination of the other two.
A basis for the vector space $V$ is a set of linearly independent vectors which span all of $V$. Any vector can be expressed unambiguously as a linear combination of vectors in a basis. The number of elements in different basises of a vector space is always the same. If this number is finite it is called the dimension of the vector space and it is written $\dim(V)$.

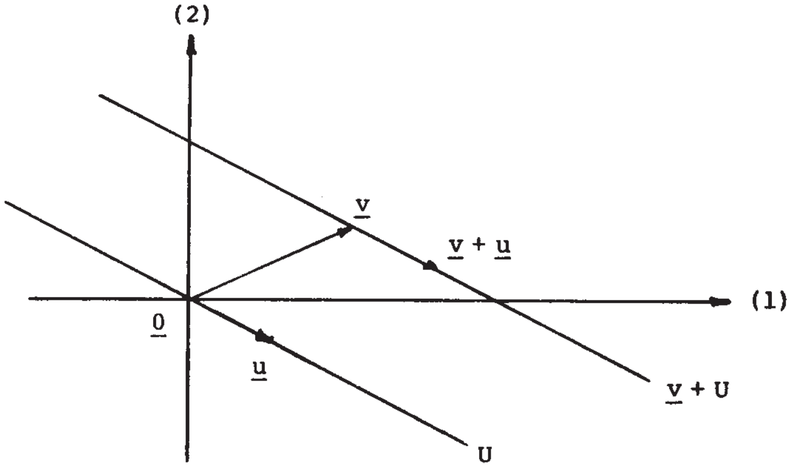


Figure 1.1: Sub-space and corresponding side-subspace in $R^2$.

**EXAMPLE 1.2.** $R^n$ has the basis

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

and is therefore $n$-dimensional ♦

In an expression like

$$\boldsymbol{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i$$

where $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a basis for $V$, we call the set $\alpha_1, \ldots, \alpha_n$ $\boldsymbol{v}$'s coordinates with respect to the basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$.

### 1.1.2 Direct sum of vector spaces

Let $V$ be a vector space (of finite dimension) and let $U_1, \ldots, U_k$ be sub-spaces of $V$. We then say that $V$ is the direct sum of the sub-spaces $U_1, \ldots, U_k$, and we write

$$V = U_1 \oplus \cdots \oplus U_k = \begin{matrix} k \\ \oplus \\ i = 1 \end{matrix} \quad U_i,$$

if an arbitrary vector $\boldsymbol{v} \in V$ in exactly one way can be expressed like

$$\boldsymbol{v} = \boldsymbol{u}_1 + \cdots + \boldsymbol{u}_k, \quad \boldsymbol{u}_1 \in U_1, \ldots, \boldsymbol{u}_k \in U_k \tag{1.1}$$

This condition is equivalent to that for vectors $\boldsymbol{u}_i \in U_i$ the following holds true

$$\boldsymbol{u}_1 + \cdots + \boldsymbol{u}_k = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{u}_1 = \cdots = \boldsymbol{u}_k = 0.$$

This is again equivalent to

$$\dim(\operatorname{span}(U_1, \ldots, U_k)) = \sum_{i=1}^{k} \dim U_i = \dim V$$

Finally, this is equivalent to that all unions of some of the $U_i$'s are $\mathbf{0}$. Of course, it is a general condition that $\operatorname{span}(U_1, \ldots, U_k) = V$, i.e. that it is at all possible to find an expression like 1.1. It is the unambiguousity of 1.1 which implies that we may call the ”sum” direct.

We sketch some examples below in fig. 1.2

If $V$ is partitioned into a direct sum

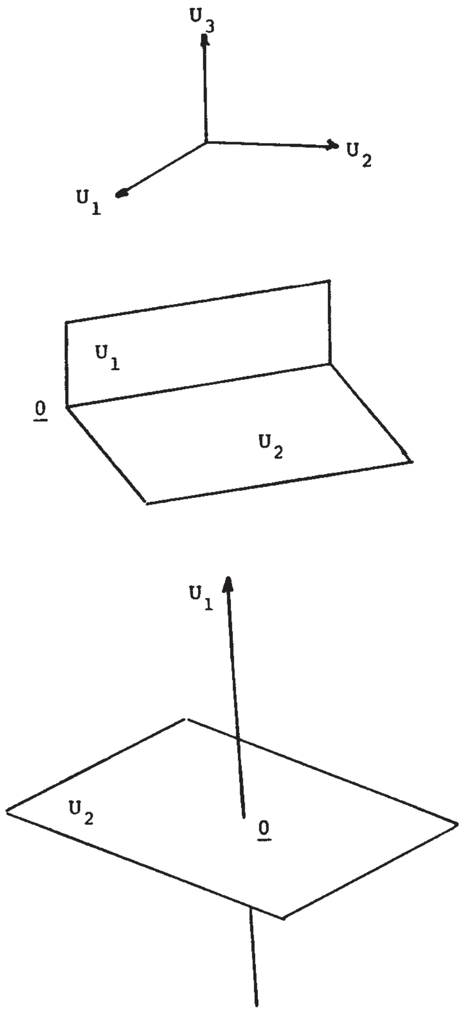
$$V = U_1 \oplus \cdots \oplus U_k$$

then we call any arbitrary vector $\boldsymbol{v}$'s component in $U_i$ for $\boldsymbol{v}$'s projection onto $U_i$ (by the direction determined by $U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_k$) and we denote it $p_i(\boldsymbol{v})$

The projection $p_i$ is idempotent, i.e. $p_i \circ p_i(\boldsymbol{v}) = p_i(\boldsymbol{v}), \forall \boldsymbol{v}$ where f $\circ$ g denotes the combination of f and g.

## 1.2 Linear projections and matrices

We start with a section on linear projections.



$U_1 \oplus U_2 \oplus U_3 = R^3$ The sum is direct because for instance $\dim U_1 + \dim U_2 + \dim U_3 = 3$

$R^3$ is not a direct sum of $U_1$ $U_2$; because $\dim U_1 + \dim U_2 = 4$

Her $U_1 \oplus U_2 = R^3$ because for instance $U_1$ and $U_2$ besides spanning $R^3$ also satisfy $U_1 \cap U_2 = \mathbf{0}$

Figure 1.2:

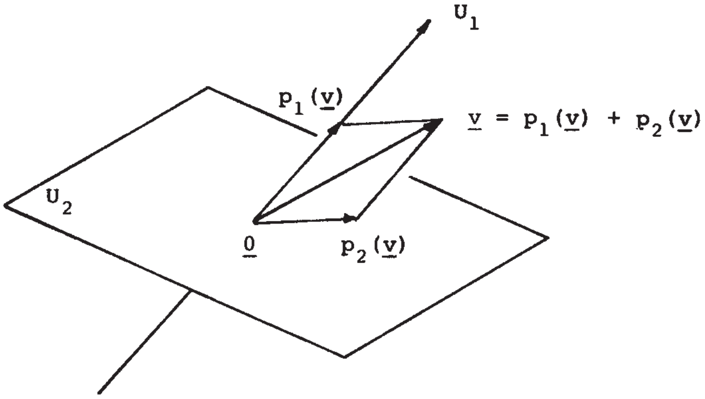


Figure 1.3: Projection of a vector.

### 1.2.1 Linear projections

A projection $A : U \to V$, where $U$ and $V$ are vector spaces are said to be linear if

$$\forall\lambda_1, \lambda_2 \in R\,\forall\boldsymbol{u}_1, \boldsymbol{u}_2 \in U : A(\lambda_1\boldsymbol{u}_1 + \lambda_2\boldsymbol{u}_2) = \\ \lambda_1 A(\boldsymbol{u}_1) + \lambda_2 A(\boldsymbol{u}_2)$$

**EXAMPLE 1.3.** A projection $A : R \to R$ is linear if its graph is a straight line through (0,0). If the graph is a straight line which does not pass through (0,0) we say the projection is affine.

By the null-space $N(A)$ of a linear projection $A : U \to V$ we mean the sub-space

$$A^{-1}(\mathbf{0}) = \{\boldsymbol{u}|A(\boldsymbol{u}) = \mathbf{0}\}$$

The following formula holds connecting the dimension of image space and null-space

$$\dim N(A) + \dim A(U) = \dim U$$

In particular we have

$$\dim A(U) \leq \dim U$$

with equality if $A$ is injective (i.e. unambiguous). If $A$ is bijective we readily see that $\dim U = \dim V$. We say that such a projection is an isomorphism and that $U$ and

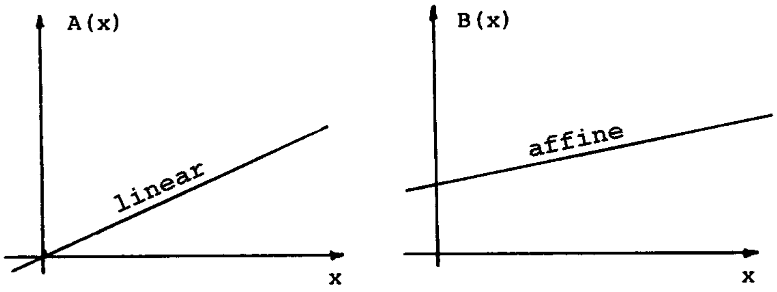


Figure 1.4: Graphs for a linear and an affine projection $R \rightarrow R$.

$V$ are isomorphic. It can be shown that any $\boldsymbol{n}$-dimensional (real) vector space is isomorphic with $\boldsymbol{R^n}$. In the following we will therefore often identify an $\boldsymbol{n}$-dimensional vector space with $\boldsymbol{R^n}$. ♦

It can be shown that the projections mentioned in the previous section are linear projections.

### 1.2.2 Matrices

By a matrix $\mathbf{A}$ we understand a rectangular table of numbers like

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}.$$

We will often use the abbreviated notation

$$\mathbf{A} = (a_{ij}).$$

More specifically we call $\mathbf{A}$ an $m \times n$ matrix because there are $m$ rows and $n$ columns. If $m = 1$ then the matrix can be called a row-vector and if $n = 1$ it can be called column-vector.

The matrix one gets by interchanging rows and columns is called the transposed matrix of $\mathbf{A}$ and we denote it by $\mathbf{A}'$, i.e.

$$\mathbf{A}' = \begin{bmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{mn} \end{bmatrix}$$

An $m \times n$ matrix is square if $n = m$. A square matrix for which $\mathbf{A} = \mathbf{A}'$ is call a symmetric matrix. The elements $a_{ii}$, $i = 1, \ldots, n$ are called the diagonal elements.

An especially important matrix is the identity matrix of order $n$

$$\mathbf{I}_n = \mathbf{I} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

A matrix which has zeroes off the diagonal is called a diagonal matrix. We use the notation

$$\boldsymbol{\Delta} = \text{diag}(\delta_1, \ldots, \delta_n) = \begin{bmatrix} \delta_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \delta_n \end{bmatrix}.$$

For given $n \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$ one defines the matrix sum

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1m} + b_{1m} \\ \vdots & & \vdots \\ a_{n1} + b_{n1} & \cdots & a_{nm} + b_{nm} \end{bmatrix}.$$

Scalar multiplication is defined by

$$c\mathbf{A} = \begin{bmatrix} ca_{11} & \cdots & ca_{1m} \\ \vdots & & \vdots \\ ca_{n1} & \cdots & ca_{nm} \end{bmatrix},$$

i.e. element-wise multiplication.

For an $m \times n$ matrix $\mathbf{C}$ and an $n \times p$ matrix $\mathbf{D}$ we define the matrix product $\mathbf{P} = \mathbf{C}\,\mathbf{D}$ by having that $\mathbf{P}$ is a $m \times p$ matrix with the $(i, j)$'th element

$$p_{ij} = \sum_{k=1}^{n} c_{ik} d_{kj}$$

We note that the matrix product is not commutative, i.e. that $\mathbf{C}\,\mathbf{D}$ generally does not equal $\mathbf{D}\,\mathbf{C}$.

For transposition we have the following rules

$$\begin{aligned}(\mathbf{A}+\mathbf{B})' &= \mathbf{A}'+\mathbf{B}' \\ (c\mathbf{A})' &= c\mathbf{A}' \\ (\mathbf{C}\,\mathbf{D})' &= \mathbf{D}'\mathbf{C}'\end{aligned}$$

### 1.2.3 Linear projections using matrix-formulation

It can be shown that for any linear projection $A : R^n \rightarrow R^m$ there is a corresponding $m \times n$ **matrix** $\mathbf{A}$, such that

$$\forall \boldsymbol{x} \in R^n : A(\boldsymbol{x}) = \mathbf{A}\,\boldsymbol{x}$$

Conversely an $A$ defined by this relation is a linear projection. $\mathbf{A}$ is easily found as the matrix which as columns has the coordinates of the projection of the unit vectors in $R^n$. E.g. we have

$$\mathbf{A}\,\boldsymbol{e}_2 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} = \boldsymbol{a}_2$$

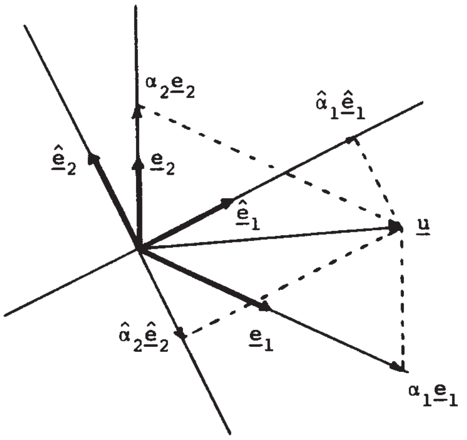
If we also have a linear projection $B : R^m \rightarrow R^k$ with corresponding matrix $\mathbf{B}$ $(k \times m)$, then we have that $B \circ A \leftrightarrow \mathbf{B}\,\mathbf{A}$ i.e.

$$\forall \boldsymbol{x} \in R^n (B \circ A(\boldsymbol{x}) = B(A(\boldsymbol{x})) = \mathbf{B}\,\mathbf{A}\,\boldsymbol{x})$$

Here we note, that an $n \times n$ matrix $\mathbf{A}$ is said to be regular if the corresponding linear projection is bijective. This is equivalent with the existence of an inverse matrix, i.e. a matrix $\mathbf{A}^{-1}$, which satisfies

$$\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix of order $n$.


$\alpha_2 \underline{e}_2$
$\hat{\alpha}_1 \underline{\hat{e}}_1$
$\underline{\hat{e}}_2$
$\underline{e}_2$
$\underline{\hat{e}}_1$
$\underline{u}$
$\hat{\alpha}_2 \underline{\hat{e}}_2$
$\underline{e}_1$
$\alpha_1 \underline{e}_1$

$$, \boldsymbol{u} = \begin{cases} \alpha_1 \boldsymbol{e}_1 + \alpha_2 \boldsymbol{e}_2 = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \\ \hat{\alpha}_1 \hat{\boldsymbol{e}}_1 + \hat{\alpha}_2 \hat{\boldsymbol{e}}_2 = \begin{bmatrix} \hat{\boldsymbol{e}}_1 & \hat{\boldsymbol{e}}_2 \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \end{cases}$$

Figure 1.5: Sketch of the coordinate transformation problem.

A square matrix which corresponds to an idempotent projection is itself called idempotent. It is readily seen that a matrix $\mathbf{A}$ is idempotent if and only if

$$\mathbf{A}\,\mathbf{A} = \mathbf{A}$$

We note that if an idempotent matrix is regular, then is equals the identity matrix, i.e. the corresponding projection is the identity.

### 1.2.4 Coordinate transformation

In this section we give formulas for the matrix formulation of a linear projection from one basis-set to another.

We first consider the change of coordinates going from one coordinate system to another. Normally, we choose not to distinguish between a vector $\boldsymbol{u}$ and its set of coordinates. This gives a simple notation and does not lead to confusion. However, when several coordinate systems are involved we do need to be able to make this distinction. In $R^n$ we consider two coordinate systems $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ and $(\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_n)$ Tre coordinates of a vector $\boldsymbol{u}$ in each of the two coordinate systems is denoted respectively $(\alpha_1, \ldots, \alpha_n)'$ and $(\hat{\alpha}_1, \ldots, \hat{\alpha}_n)'$, cf. figure 1.5.

Let the ”new” system $(\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_n)$ be given by

$$(\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_n) = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)\mathbf{S}$$

i.e.

$$\hat{\boldsymbol{e}}_i = s_{1i}\boldsymbol{e}_1 + \cdots + s_{ni}\boldsymbol{e}_i, \qquad i = 1, \ldots, n.$$

The columns in the $\mathbf{S}$-matrix are thus equal to the ”new” systems ”old” coordinates. $\mathbf{S}$ is called the coordinate transformation matrix.

**REMARK 1.1.** However, many references use the expression coordinate transformation matrix about the matrix $\mathbf{S}^{-1}$. It is therefore important to be sure which matrix one is talking about.

Since

$$(\boldsymbol{e}_1 \cdots \boldsymbol{e}_n)\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = (\hat{\boldsymbol{e}}_1 \cdots \hat{\boldsymbol{e}}_n)\begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{pmatrix},$$

(cf. fig. 1.5), the connection between a vectors ”old” and ”new” coordinates becomes

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \mathbf{S}\begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{bmatrix} \Longleftrightarrow \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{bmatrix} = \mathbf{S}^{-1}\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}.$$

▼

We now consider a linear projection $A : R^n \rightarrow R^m$, and let $A$’s matrix formulation w.r.t. the bases $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ and $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m)$ be

$$\boldsymbol{\beta} = \mathbf{A}\,\boldsymbol{\alpha}$$

and the formulation w.r.t. the bases $(\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_n) = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)\mathbf{S}$ and $(\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_m) = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m)\mathbf{T}$ be

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{A}}\hat{\boldsymbol{\alpha}}$$

Then we have

$$\hat{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\,\mathbf{T},$$

which is readily found by use of the rules of coordinate transformation on the coordinates.

If we are concerned with projections $R^n \rightarrow R^n$ and we use the same coordinate transformation, then we get the relation

$$\hat{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\,\mathbf{S}.$$

The matrices $\mathbf{A}$ and $\hat{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\,\mathbf{S}$ are then called similar matrices.

### 1.2.5 Rank of a matrix

By rank of a linear projection $A : R^n \rightarrow R^m$ we mean the dimension of the image space, i.e.

$$\mathrm{rg}(A) = \dim A(R^n).$$

By rank of a matrix $A$ we mean the rank of the corresponding linear projection.

We see that $\mathrm{rg}(\mathbf{A})$ exactly equals the number of linearly independent column vectors in $\mathbf{A}$. Trivially we therefore have

$$\mathrm{rg}(\mathbf{A}) \leq n.$$

If we introduce the transposed matrix $\mathbf{A}'$ it is easily shown that $\mathrm{rg}(\mathbf{A}) = \mathrm{rg}(\mathbf{A}')$i.e. we have

$$\mathrm{rg}(\mathbf{A}) \leq \min(m, n).$$

If $\mathbf{A}$ and $\mathbf{B}$ are two $m \times n$ matrices, then

$$\mathrm{rg}(\mathbf{A} + \mathbf{B}) \leq \mathrm{rg}(\mathbf{A}) + \mathrm{rg}(\mathbf{B}).$$

This relation is obvious when one remembers that for the corresponding projections $A$ and $B$ we have $(A + B)\ (R^n) \subseteq A(R^n) \cup B(R^n)$.

If $\mathbf{A}$ is an $(m \times n)$-matrix and $\mathbf{B}$ is an $(k \times m)$-matrix we have

$$\mathrm{rg}(\mathbf{BA}) \leq \mathrm{rg}(\mathbf{A}).$$

If $\mathbf{B}$ is regular $(m \times m)$ we have

$$\mathrm{rg}(\mathbf{BA}) = \mathrm{rg}(\mathbf{A}).$$

These relations are immediate consequences of the relation $\dim B(A(R^n)) \leq \dim(A(R^n))$, where we have equality if $B$ is injective. There are of course analogue relations for an $(n \times p)$-matrix $\mathbf{C}$:

$$\mathrm{rg}(\mathbf{A}\,\mathbf{C}) \leq \mathrm{rg}(\mathbf{A})$$

with equality if $\mathbf{C}$ is a regular $(n \times n)$-matrix. From these we can deduce for regular $\mathbf{B}$ and $\mathbf{C}$ that

$$\mathrm{rg}(\mathbf{B}\,\mathbf{A}\,\mathbf{C}) = \mathrm{rg}(\mathbf{A}).$$

Finally we mention that an $(n \times n)$-matrix $\mathbf{A}$ is regular if $\mathrm{rg}(\mathbf{A}) = n$.

### 1.2.6 Determinant of a matrix

The abstract definition of the determinant of a square $p \times p$ matrix $\mathbf{A}$ is

$$\det(\mathbf{A}) = \sum_{\text{alle } \sigma} \pm a_{1\sigma(1)} \ldots a_{p\sigma(p)},$$

where $\sigma$ is a permutation of the numbers $1, \ldots, p$ and where we use the + sign if the permutation is even (i.e. it can be composed of an even number of neighbour swaps) and - if it is odd.

We will not go into the background of this definition. We note that the determinant represents the volume of the corresponding linear projection i.e. for an $(n \times n)$ -matrix $\mathbf{A}$

$$|\det(\mathbf{A})| = \frac{\mathrm{vol}(A(I))}{\mathrm{vol}(I)},$$

where $I$ is an $n$ -dimensional box and $A(I)$ is the image of $I$ (being an $n$ -dimensional parallelepiped) found by the corresponding projection.

The situation is sketched in 2 dimensions in fig. 1.6. For $2 \times 2$ and $3 \times 3$ matrices the definition of the determinant becomes

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = aei + bfg + cdh - gec - hfa - idb.$$

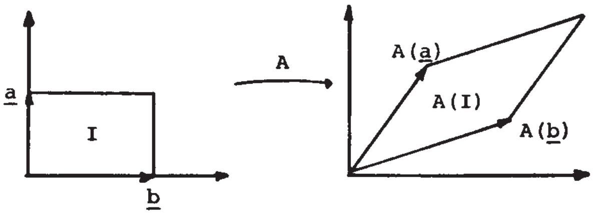


Figure 1.6: A rectangle and its image after a linear projection.

For determinants of higher order (here $n$'th order) we can develop the determinant by the $i$'th row i.e.

$$\det(\mathbf{A}) = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}\det(\mathbf{A}_{ij}),$$

where $\mathbf{A}_{ij}$ is the matrix we get after deleting the $i$'th row and the $j$'th column of $\mathbf{A}$. The number

$$A_{ij} = (-1)^{i+j}\det(\mathbf{A}_{ij})$$

is also called the element $a_{ij}$ 's cofactor. Of course an analogue procedure exists for development by columns.

When one explicitly must evaluate a determinant the following three rules are handy:

i) interchanging 2 rows (columns) in $\mathbf{A}$ multiplies $\det(\mathbf{A})$ by $-1$.

ii) multiplying a row (column) by a scalar multiplies $\det(A)$ by the scalar.

iii) adding a multiplum of a row (column) to another row (column) leaves $\det(A)$ unchanged.

When determining the rank of a matrix it can be useful to remember that the rank is the largest number $r$ for which the matrix has a determinant of the minor which different from 0 and of $r$'th order. We find as a special case that $\mathbf{A}$ is regular if and only if $\det\mathbf{A} \neq 0$. This also seems intuitively obvious when one considers the determinant being the volume. If it is 0 then the projection must in some sense "reduce the dimension".

For square matrices $\mathbf{A}$ and $\mathbf{B}$ we have

$$\text{det}(\mathbf{A}\,\mathbf{B}) = \text{det}(\mathbf{A})\det(\mathbf{B})$$

For a diagonal matrix $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ we have

$$\det(\mathbf{\Lambda}) = \lambda_1 \ldots \lambda_n$$

For a triangular matrix $\mathbf{C}$ with diagonal elements $c_1, \ldots, c_n$ we have

$$\det(\mathbf{C}) = c_1 \cdots c_n$$

By means of determinants one can directly state the inverse of a regular matrix $\mathbf{A}$. We have

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}(A_{ij})',$$

i.e. the inverse of a regular matrix $\mathbf{A}$ is the transposed of the matrix we get by substituting each element in $\mathbf{A}$ by its complement divided by $\det \mathbf{A}$. However, note that this formular is not directly applicable for the inversion of large matrices because of the large number of computations involved in the calculation of determinants.

Something similar is true for **Cramérs** theorem on solving a linear system of equations: Consider the regular matrix $\mathbf{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$. Then the solution to the equation

$$\mathbf{A}\,\boldsymbol{x} = \boldsymbol{b}$$

is given by

$$x_i = \frac{\det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{b}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n)}{\det \mathbf{A}}$$

### 1.2.7 Block-matrices

By a block-matrix we mean a matrix of the form

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \cdots & \mathbf{B}_{1n} \\ \vdots & & \vdots \\ \mathbf{B}_{m1} & \cdots & \mathbf{B}_{mn} \end{bmatrix}$$

where the blocks $\mathbf{B}_{ij}$ are matrices of order $m_i \times n_j$.

When adding and multiplying one can use the usual rules of calculation for matrices and just consider the blocks as elements. For instance we find

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{S} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\,\mathbf{R} + \mathbf{B}\,\mathbf{S} \\ \mathbf{C}\,\mathbf{R} + \mathbf{D}\,\mathbf{S} \end{bmatrix},$$

under the obvious condition that the involved products exist etc.

First we give a result on determinants of the ”triangular” matrix.

**THEOREM 1.1.** Let the square matrix $\mathbf{A}$ be partitioned into block-matrices

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}$$

where $\mathbf{B}$ and $\mathbf{D}$ er kvadratiske og are square and $\mathbf{0}$ is a matrix only containing 0’s. Then we have

$$\det(\mathbf{A}) = \det(\mathbf{B}) \det(\mathbf{D})$$

▲

**PROOF 1.1.** We have that

$$\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

where the $\mathbf{I}$ ’s are identity-matrices, not necessarily of same order. If one develops the first matrix by its 1st row we see that it has the same determinant as the matrix one gets by deleting the first row and column. By repeating this until the remaining minor is $\mathbf{D}$, we see that

$$\det \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} = \det(\mathbf{D})$$

Analogously we find that the last matrix has the determinant $\det \mathbf{B}$ and the result follows. ■

The following theorem expands this result.

**THEOREM 1.2.** Let the matrix $\mathbf{\Sigma}$ be partitioned into block matrices

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

Then we have

$$\det(\mathbf{\Sigma}) = \det(\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}) \det(\mathbf{\Sigma}_{22}),$$

under the condition that $\mathbf{\Sigma}_{22}$ is regular. ▲

**PROOF 1.2.** Since

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{12} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

the result follows immediately from the previous theorem. ■

The last theorem gives a useful result on inversion of matrices which are partitioned into block matrices.

**THEOREM 1.3.** For the symmetrical matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

we have

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{A}' \\ -\mathbf{A}\mathbf{B}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \mathbf{A}\mathbf{B}^{-1}\mathbf{A}' \end{bmatrix},$$

where

$$\begin{aligned} \mathbf{A} &= \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}^{-1} \\ \mathbf{B} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \end{aligned}$$

conditioned on the existence of the inverses involved. ▲

**PROOF 1.3.** The result follows immediately by multiplication of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$. ■

## 1.3 Pseudoinverse or generalised inverse matrix of a non-regular matrix

We consider a linear projection

$$A : E \rightarrow F$$

where $E$ is an $n$-dimensional and $F$ an $m$-dimensional (euclidian) vector space. The matrix corresponding to $A$ is usually called $\mathbf{A}$ and it has the dimensions $m \times n$. We equal the null space of $A$ to $U$ , i.e.

$$U = A^{-1}(\mathbf{0}),$$

and call its dimension $r$. The image space

$$V = A(E)$$

has dimension $s = n - r$ , cf. p. 7.

We now consider an arbitrary $s$ -dimensional space $U^* \subseteq E$, which is complementary to $U$, and an arbitrary $m-s$ dimensional subspace $V^* \subseteq F$, which is complementary to $V$.

An arbitrary vector $\boldsymbol{x} \in E$ can now be written as

$$\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{u}^*, \quad \boldsymbol{u} \in U \quad \text{og} \quad \boldsymbol{u}^* \in U^*,$$

since $\boldsymbol{u}$ and $\boldsymbol{u}^*$ are given by

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{x} - p_{U^*}(x) \\ \boldsymbol{u}^* &= p_{U^*}(x) \end{aligned}$$

Here $p_{U^*}$ denotes the projection of $E$ onto $U^*$ along the sub-space $U$. Similarly any $\boldsymbol{y} \in F$ can be written

$$\boldsymbol{y} = (\boldsymbol{y} - p_V(\boldsymbol{y})) + p_V(\boldsymbol{y}) = \boldsymbol{v}^* + \boldsymbol{v}$$

where

$$p_V : F \to V$$

is the projection of $F$ onto $V$ along $V^*$.

Since

$$A(\boldsymbol{x}) = A(\boldsymbol{u} + \boldsymbol{u}^*) = A(\boldsymbol{u}^*),$$

we see that $A$ is constant on the side-spaces

$$\boldsymbol{u}^* + U = \{\boldsymbol{u}^* + \boldsymbol{u} | \boldsymbol{u} \in U\}$$

and it follows that $A$'s restriction on $U^*$ is a bijective projection of $U^*$ onto $V$. This projection therefore has an inverse

$$B_1 : V \to U^*$$

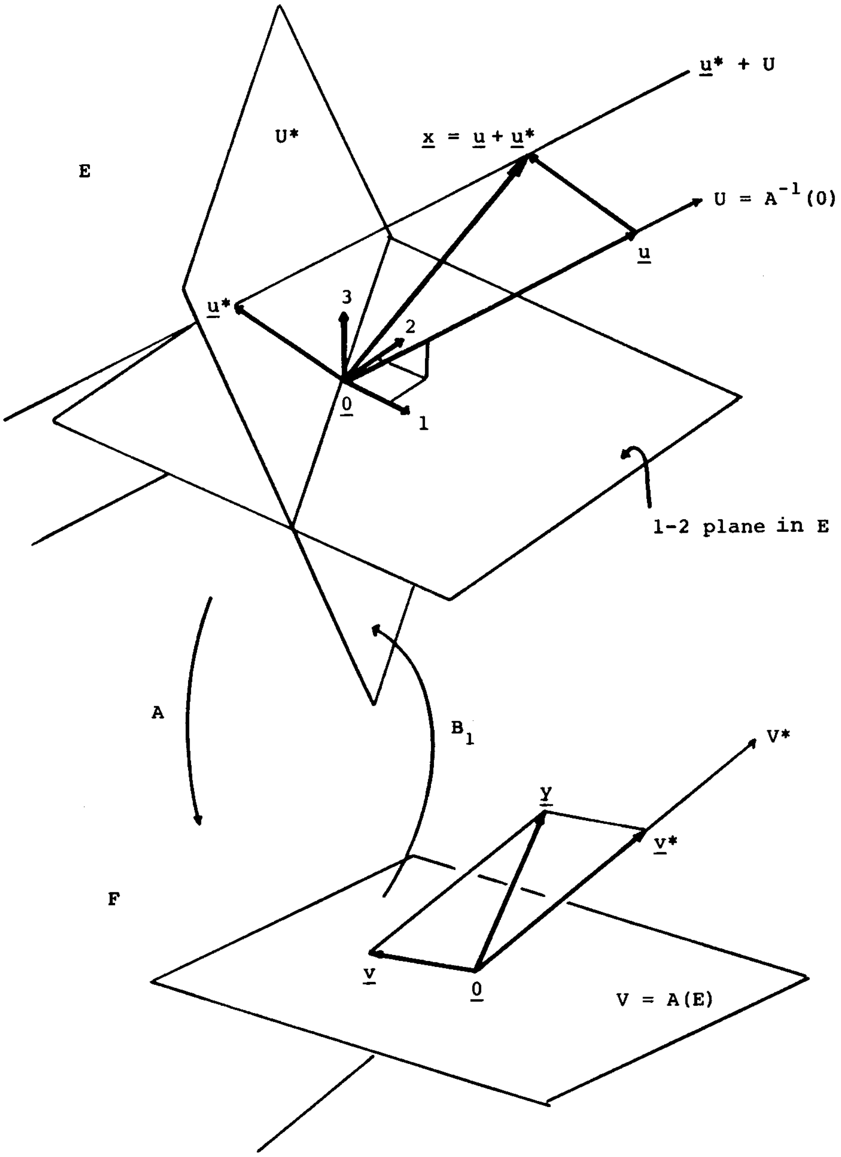


Figure 1.7: Sketch showing pseudo-inverse projection.

given by

$$B_1(\boldsymbol{v}) = \boldsymbol{u}^* \quad \Leftrightarrow \quad A(u^*) = \boldsymbol{v}$$

We are now able to formulate the definition of the pseudo-inverse projection.

**DEFINITION 1.1.** By a pseudoinverse or generalised inverse projection of the projection $A$ we mean a projection

$$B = B_1 \circ p_V : F \to E,$$

where $p_V$ and $B_1$ are as mentioned previously. ▲

**REMARK 1.2.** The pseudo-inverse is thus the combined projection onto $V$ along $V^*$ and the inverse of $A$'s restriction to $U^*$. ▼

**REMARK 1.3.** The pseudo-inverse is of course by no means unambiguous, because we get one for each choice of the sub-spaces $U^*$ and $V^*$. ▼

We can now state some obvious properties of the pseudo-inverse in the following

**THEOREM 1.4.** The pseudo-inverse $B$ of $A$ has the following properties

i) $\mathrm{rg}(B) = \mathrm{rg}(A) = s$

ii) $A \circ B = p_V : F \to V$

iii) $B \circ A = p_{U^*} : E \to U^*$

▲

It can be shown that these properties also characterise pseudo-inverse projections, because we have

**THEOREM 1.5.** Let $A : E \to F$ be linear with rank $s$. Assume that $B$ also has rank $s$, and that $A \circ B$ and $B \circ A$ both are projections of rank $s$. Then $B$ is a pseudo-inverse of $A$ as defined above. ▲

**PROOF 1.4.** Omitted (relatively simple exercise in linear algebra). ■

We now give a matrix formulation of the above mentioned definitions.

**DEFINITION 1.2.** Let $\mathbf{A}$ be an $(m \times n)$-matrix of rank $s$. An $(n \times m)$-matrix $\mathbf{B}$ of rank $s$, which satisfies

i) $\mathbf{A}\,\mathbf{B}$ idempotetn with rank $s$

ii) $\mathbf{B}\,\mathbf{A}$ idempotent with rank $s$,

is called a pseudo-inverse or a generalised inverse of $\mathbf{A}$. ▲

By means of the pseudo-inverse we can characterise the set of possible solutions of a system of linear equations. This is due to the following

**THEOREM 1.6.** Let $\mathbf{A}$ and $\mathbf{B}$ be as in definition 1.2. The general solution of the equation

$$\mathbf{A}\,\boldsymbol{x} = \mathbf{0}$$

is

$$(\mathbf{I} - \mathbf{B}\,\mathbf{A})\boldsymbol{z}, \qquad \boldsymbol{z} \in R^n,$$

and the general solution of the equation (which is assumed to be consistent)

$$\mathbf{A}\,\boldsymbol{x} = \boldsymbol{y},$$

is

$$\mathbf{B}\,\boldsymbol{y} + (\mathbf{I} - \mathbf{B}\,\mathbf{A})\boldsymbol{z}, \qquad \boldsymbol{z} \in R^n.$$

▲

**PROOF 1.5.** We first consider the homogeneous equation. A solution $\boldsymbol{x}$ is obviously a point in the null-space $N(A) = A^{-1}(\mathbf{0})$ of the linear projection corresponding to $\mathbf{A}$. The matrix $\mathbf{B}\,\mathbf{A}$ according to theorem 1.1 - corresponds precisely to the projection onto $U^*$. Therefore $\mathbf{I} - \mathbf{B}\,\mathbf{A}$ corresponds to the projection onto the null-space $U = N(A)$. Therefore, an arbitrary $\boldsymbol{x} \in N(A)$ can be written

$$\boldsymbol{x} = (\mathbf{I} - \mathbf{B}\,\mathbf{A})\boldsymbol{z}, \qquad \boldsymbol{z} \in R^n.$$

The statement regarding the homogeneous equation has now been proved.

The equation $\mathbf{A}\,\boldsymbol{x} = \boldsymbol{y}$ only has a solution (i.e. is only consistent) if $\boldsymbol{y}$ lies in the image space of $\mathbf{A}$. For such a $\boldsymbol{y}$ we have

$$\mathbf{A}\,\mathbf{B}\,\boldsymbol{y} = \boldsymbol{y},$$

according to theorem 1.4.

The result for the complete solution follows readily. ■

In order to illustrate the concept we now give

**EXAMPLE 1.4.** We consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

$\mathbf{A}$ obviously has the rank 2.

We will consider the linear projection corresponding to $\mathbf{A}$ which is

$$A : E \to F$$

where $E$ and $F$ are 3-dimensional vector spaces with bases $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ og $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3\}$. The coordinates of these bases are denoted by small $x$'s and $y$'s respectively, such that $A$ can be formulated in the coordinates

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

First we will determine the null-space

$$U = N(A) = A^{-1}(\mathbf{0})$$

for $A$. We have

$$\begin{aligned} \boldsymbol{x} \in U \quad & \Leftrightarrow \quad \mathbf{A}\,\boldsymbol{x} = \mathbf{0} \\ & \Leftrightarrow \quad x_1 + x_2 + 2x_3 = 0 \quad \wedge \quad 2x_1 + x_2 + x_3 = 0 \\ & \Leftrightarrow \quad x_1 = x_3 \quad \wedge \quad -3x_1 = x_2 \\ & \Leftrightarrow \quad \boldsymbol{x}' = x_1(1, -3, 1). \end{aligned}$$

The null-space is then

$$U = \{t \cdot \begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix} | t \in R\} = \{t \cdot \boldsymbol{u}_3 | t \in R\}$$

As complementary sub-space we choose to consider the orthogonal complement $U^*$. This has the equation

$$(1, -3, 1)\boldsymbol{x} = 0,$$

or

$$U^* = \{\boldsymbol{x} | x_1 - 3x_2 + x_3 = 0\}$$

We now consider a new basis for $E$, namely $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\}$. Coordinates in this are denoted using small $z$'s. The conversion from $z$-coordinates to $x$-coordinates is given by

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

or

$$\boldsymbol{x} = \mathbf{S}\boldsymbol{z}.$$

The columns of the $\mathbf{S}$ matrix are known to be the $\boldsymbol{u}$'s coordinates in the $\boldsymbol{e}$-system.

$A$'s image space $V$ is 2-dimensional and is spanned by $\mathbf{A}$'s columns. We can for instance choose the first two, i.e.

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad , \quad \boldsymbol{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

As complementary sub-space $V^*$ we choose $V$'s orthogonal complement. This is produced by making the cross-product of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$:

$$\boldsymbol{v}_1 \times \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \boldsymbol{v}_3$$

We now consider the new basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ for $F$. The coordinates in this are denoted using small $w$'s. The conversion from $w$-coordinates to $y$-coordinates is given by

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix},$$

or in compact notation

$$\boldsymbol{y} = \mathbf{T}\,\boldsymbol{w}.$$

We will now find coordinate expressions for $A$ in $z$- and $w$-coordinates. Since

$$\boldsymbol{y} = \mathbf{A}\,\boldsymbol{x}$$

we have

$$\mathbf{T}\,\boldsymbol{w} = \mathbf{A}\,\mathbf{S}\,\boldsymbol{z}$$

or

$$\boldsymbol{w} = \mathbf{T}^{-1}\mathbf{A}\,\mathbf{S}\,\boldsymbol{z}.$$

Now we have

$$\mathbf{T}^{-1} = \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix},$$

wherefore

$$\begin{aligned} \mathbf{T}^{-1}\mathbf{A}\,\mathbf{S} &= \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 0 \\ -3 & 11 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Since $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ spans $U^*$ and $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ spans $V$, we note that the condition

$$A : U^* \rightarrow V$$

has the coordinate expression

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ -3 & 11 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

It has the inverse projection

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{3}{22} & \frac{1}{11} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

If we consider the points as points in $E$ and $F$ - and not just as points in $U^*$ and $V$ then we get

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{3}{22} & \frac{1}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \tag{1.2}$$

The projection of $F$ onto $V$ along $V^*$ has the formulation in coordinates

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \longrightarrow \begin{bmatrix} w_1 \\ w_2 \\ 0 \end{bmatrix} \tag{1.3}$$

This is the $z - w$ coordinate formulation for the pseudo-inverse $B$ of the projection $A$. However, we want a description in $x - y$ coordinates. Since

$$\boldsymbol{z} = \mathbf{S}^{-1}\boldsymbol{x} = \mathbf{C}\,\boldsymbol{w} = \mathbf{C}\,\mathbf{T}^{-1}\boldsymbol{y}$$

we get

$$\boldsymbol{x} = \mathbf{S}\,\mathbf{C}\,\mathbf{T}^{-1}\boldsymbol{y},$$

where $\mathbf{C}$ is the matrix in formula 1.1.

We therefore have

$$\begin{aligned} \mathbf{B} &= \mathbf{S}\,\mathbf{C}\,\mathbf{T}^{-1} \\ &= \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{3}{22} & \frac{1}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= \frac{1}{22} \begin{bmatrix} -8 & 7 & 7 \\ 2 & 1 & 1 \\ 14 & -4 & -4 \end{bmatrix} \end{aligned}$$

This matrix is a pseudo-inverse of $\mathbf{A}$. ♦

As it is seen from the previous example it is rather tedious just to use the definition in order to calculate a pseudo-inverse. Often one may utilise the following

**THEOREM 1.7.** Let the $m \times n$ matrix $\mathbf{A}$ have rank $s$ and let

$$A = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix},$$

where $\mathbf{C}$ is regular with dimension $s \times s$. A (possible) pseudo-inverse of $\mathbf{A}$ is then

$$\mathbf{A}^{-} = \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where the 0-matrices have dimensions such that $\mathbf{A}^{-}$ has the dimension $n \times m$. ▲

**PROOF 1.6.** We have

$$\mathbf{A}\,\mathbf{A}^{-}\mathbf{A} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{E}\,\mathbf{C}^{-1}\mathbf{D} \end{bmatrix}.$$

Since $\mathrm{rg}(\mathbf{A}) = s$, then the last $n - s$ columns can be written as linear combinations of the first $s$ columns, i.e. there exists a matrix $\mathbf{H}$, so

$$\begin{bmatrix} \mathbf{D} \\ \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ \mathbf{E} \end{bmatrix} \mathbf{H}$$

or

$$\begin{aligned} \mathbf{D} &= \mathbf{C}\,\mathbf{H} \\ \mathbf{F} &= \mathbf{E}\,\mathbf{H} \end{aligned}$$

From this we find

$$\mathbf{F} = \mathbf{E}\,\mathbf{C}^{-1}\mathbf{D}.$$

If we insert this in the top formula we have

$$\mathbf{A}\,\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$$

By pre-multiplication with $A^{-}$ and post-multiplication with $A^{-}$ respectively, we see that $A^{-}A$ and $AA^{-}$ are idempotent. The theorem is now derived from the definition page 22. ■

We illustrate the use of the theorem in the following

**EXAMPLE 1.5.** We consider the matrix given in example 1.4

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

Since

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix},$$

we can use as pseudo-inverse:

$$\mathbf{A}^{-} = \begin{bmatrix} -1 & 1 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

♦

The advantage of using the procedure given in example 1.4 instead of the far more simple one given in example 1.5, is that one obtains a precise geometrical description of the situation.

**REMARK 1.4.** Finally, we note that the literature has a number of definitions of pseudo-inverses and generalised inverses, so it is necessary to specify exactly what the definition is. A case of special interest is the so-called **Moore-Penrose** inverse $\mathbf{A}^{+}$ of a matrix $\mathbf{A}$. It satisfies the following

i) $\mathbf{A}\,\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$

ii) $\mathbf{A}^{+}\mathbf{A}\,\mathbf{A}^{+} = \mathbf{A}^{+}$

iii) $(\mathbf{A}\,\mathbf{A}^{+})' = \mathbf{A}\,\mathbf{A}^{+}$

iv) $(\mathbf{A}^{+}\mathbf{A})' = \mathbf{A}^{+}\mathbf{A}$

It is obvious that a Moore-Penrose inverse really is a generalised inverse. The other conditions guarantee that a least squares solution of an inconsistent equation find a solution with minimal norm. We will not pursue this further here, only refer the interested reader to the literature e.g. [19]. ▼

# 1.4 Eigenvalue problems. Quadratic forms

We begin with the fundamental definitions and theorems in

## 1.4.1 Eigenvalues and eigenvectors for symmetric matrices

The definition of an eigenvector and an eigenvalue given below are valid for arbitrary square matrices. However, in the sequel we will always assume the involved matrices are symmetrical unless explicitly stated otherwise.

An eigenvalue $\lambda$ of the symmetric $n \times n$ matrix $\mathbf{A}$ is a solution to the equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

There are $n$ (real-valued) eigenvalues (some may have equal values). If $\lambda$ is an eigenvalue, then vectors $\boldsymbol{x} \neq 0,$ exist such that

$$\mathbf{A}\,\boldsymbol{x} = \lambda\boldsymbol{x},$$

i.e. vector exist such that the linear projection corresponding to $\mathbf{A}$ leads to a multiplum of its self. Such vectors are called eigenvectors corresponding to the eigenvalue $\lambda$. The number of eigenvalues different from 0 equals $\mathrm{rg}(\mathbf{A})$. An eigenvalue is to be counted as many times as its multiplicity indicates. A more interesting theorem is

**THEOREM 1.8.** If $\lambda_i$ and $\lambda_j$ are different eigenvalues, and if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the corresponding eigenvectors, then $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are orthogonal, i.e. $\boldsymbol{x}_i'\boldsymbol{x}_j = 0.$ ▲

**PROOF 1.7.** We have

$$\begin{aligned} \mathbf{A}\,\boldsymbol{x}_i &= \lambda_i\boldsymbol{x}_i \\ \mathbf{A}\,\boldsymbol{x}_j &= \lambda_j\boldsymbol{x}_j \end{aligned}$$

Here we readily find

$$\begin{aligned} \boldsymbol{x}_j'\mathbf{A}\,\boldsymbol{x}_i &= \lambda_i\boldsymbol{x}_j'\boldsymbol{x}_i \\ \boldsymbol{x}_i'\mathbf{A}\,\boldsymbol{x}_j &= \lambda_j\boldsymbol{x}_i'\boldsymbol{x}_j. \end{aligned}$$

We transpose the first relationship and get

$$\boldsymbol{x}_i'\mathbf{A}'\boldsymbol{x}_j = \lambda_i\boldsymbol{x}_i'\boldsymbol{x}_j.$$

Since $\mathbf{A}$ is symmetric this implies that

$$\lambda_i \boldsymbol{x}_i' \boldsymbol{x}_j = \lambda_j \boldsymbol{x}_i' \boldsymbol{x}_j,$$

and since $\lambda_i \neq \lambda_j$ then $\boldsymbol{x}_i' \boldsymbol{x}_j = 0$ i.e. $\boldsymbol{x}_i \perp \boldsymbol{x}_j$. ■

The result in theorem 1.8 can be supplemented with the following theorem given without proof.

**THEOREM 1.9.** If $\lambda$ is an eigenvalue with multiplicity $m$, then the set of eigenvectors corresponding to $\lambda$ forms an $m$-dimensional sub-space. This has the special implication that there exists $m$ orthogonal eigenvectors corresponding to $\lambda$. ▲

By combining these two theorems one readily sees the following

**COROLLORY 1.1.** For an arbitrary symmetric matrix $\mathbf{A}$ a basis exists for $R^n$ consisting of mutually orthogonal eigenvectors of $\mathbf{A}$.

If such a basis consisting of orthogonal eigenvectors is normed then one gets an orthonormal basis $(\boldsymbol{p}_1, \dots, \boldsymbol{p}_n)$. If we let $\mathbf{P}$ equal the $n \times n$ matrix whos columns are the coordinates of these vectors, i.e.

$$\mathbf{P} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_n)$$

we get

$$\mathbf{P}'\mathbf{P} = \mathbf{I}$$

$\mathbf{P}$ is therefore an orthogonal matrix, and

$$\mathbf{A}\,\mathbf{P} = \mathbf{P}\,\boldsymbol{\Lambda}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues for $\mathbf{A}$ (repeated corresponding to multiplicity) on the diagonal. By means of this we get the following

**THEOREM 1.10.** Let $\mathbf{A}$ be a symmetric matrix. Then an orthogonal matrix $\mathbf{P}$ exists, such that

$$\mathbf{P}'\mathbf{A}\,\mathbf{P} = \boldsymbol{\Lambda}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{A}$ 's eigenvalues on the diagonal (repeated corresponding to the multiplicity). As $\mathbf{P}$ one can choose a matrix, whos columns are orthonormed eigenvectors of $\mathbf{A}$. ▲

**PROOF 1.8.** Obvious from the above relation. ■

**THEOREM 1.11.** Let $\mathbf{A}$ be a symmetric matrix with non-negative eigenvalues. Then a regular matrix $\mathbf{B}$ exists such that

$$\mathbf{B}'\mathbf{A}\,\mathbf{B} = \mathbf{E},$$

where $\mathbf{E}$ is a diagonal matrix having 0's or 1's on the diagonal. The number of 1's equals $\mathrm{rg}(\mathbf{A})$. If $\mathbf{A}$ is of full rank then $\mathbf{E}$ becomes an identity matrix. ▲

**PROOF 1.9.** By (post-) multiplication of $\mathbf{P}$ with a diagonal matrix $\mathbf{C}$ which has the following diagonal elements

$$c_i = \begin{cases} \frac{1}{\sqrt{\lambda_i}} & \lambda_i > 0 \\ 1 & \lambda_i = 0 \end{cases},$$

we readily find the theorem with $\mathbf{B} = \mathbf{P}\,\mathbf{C}$. ■

The relation in theorem 1.10 is equivalent to

$$\mathbf{A} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}'$$

or

$$\mathbf{A} = (\boldsymbol{p}_1 \ldots \boldsymbol{p}_n) \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \boldsymbol{p}_1' \\ \vdots \\ \boldsymbol{p}_n' \end{bmatrix},$$

i.e. we have the following partitioning of the matrix

$$\mathbf{A} = \lambda_1 \boldsymbol{p}_1 \boldsymbol{p}_1' + \cdots + \lambda_n \boldsymbol{p}_n \boldsymbol{p}_n'.$$

This partitioning of the symmetrical matrix $\mathbf{A}$ is often called its spectral decomposition, since the eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ are called the spectrum of the matrix.

With the obvious definition of $\mathbf{\Lambda}^{\frac{1}{2}}$ being $\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$, we note that we can write

$$\mathbf{A} = (\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})' = \mathbf{G}\,\mathbf{G}'.$$

Here we mention that if $\mathbf{A}$ is positive definite, then there is a relation

$$\mathbf{A} = \mathbf{L}\,\mathbf{L}',$$

where $\mathbf{L}$ is a lower triangular matrix. This relation is called the Cholesky factorisation of $\mathbf{A}$ (see e.g. [21]).

Finally we have

**THEOREM 1.12.** Let $\mathbf{A}$ be a regular symmetrical matrix. Then $\mathbf{A}$ and $\mathbf{A}^{-1}$ have the same eigenvectors corresponding to reciprocal eigenvalues. ▲

**PROOF 1.10.** Let $\lambda$ be an eigenvalue of $\mathbf{A}$ and $\boldsymbol{x}$ be a corresponding eigenvector, i.e.

$$\mathbf{A}\,\boldsymbol{x} = \lambda\boldsymbol{x}.$$

Since $\mathbf{A}$ is regular then this is equivalent to

$$\mathbf{A}^{-1}\boldsymbol{x} = \frac{1}{\lambda}\boldsymbol{x},$$

which concludes the proof. ■

Finally, we note that

$$\det \mathbf{A} = \prod_i \lambda_i.$$

**EXAMPLE 1.6.** Orthogonal transformations of the plane. In order to give a geometrical understanding of the transformations which reduce a symmetrical matrix into diagonal form, we state the orthogonal transformations of the plane.

By utilising the orthogonality conditions $\mathbf{P}'\mathbf{P} = \mathbf{I}$ we readily see, that the only orthogonal $2 \times 2$-matrices are matrices of the form

$$\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \quad \text{og} \quad \begin{bmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{bmatrix}.$$

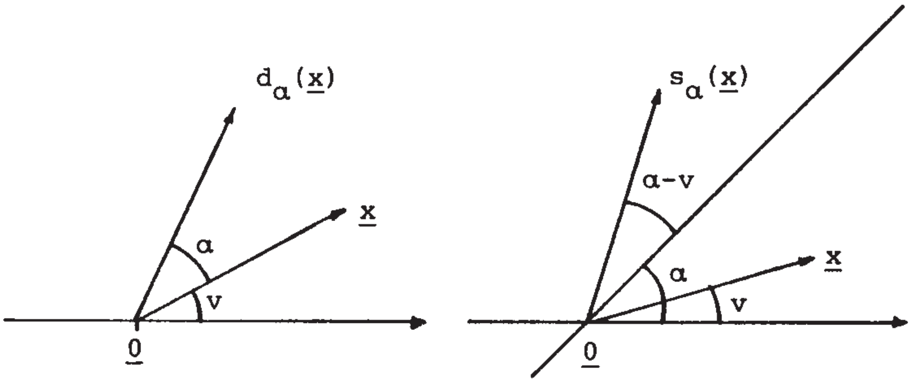


Figure 1.8: Rotation and reflection as determined by the angle $\alpha$.

We will now show that these correspond to rotations around the origin and reflections in straight lines.

We do this by determining coordinate expressions for the linear projections $d_\alpha$ and $s_\alpha$ , which respectively represent a rotation of the plane of the angle $\alpha$ and a reflection in the line having the angle $\alpha$ with the 1.st axis.

The projections are illustrated in figure 1.8. Since $\boldsymbol{x} = r(\cos v, \sin v)'$, where $r$ is equal to 1, we have

$$\begin{aligned} d_\alpha(\boldsymbol{x}) &= \begin{bmatrix} \cos(\alpha+v) \\ \sin(\alpha+v) \end{bmatrix} = \begin{bmatrix} \cos\alpha\cos v - \sin\alpha\sin v \\ \sin\alpha\cos v + \cos\alpha\sin v \end{bmatrix} \\ &= \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}. \end{aligned}$$

From this we find $\boldsymbol{d_\alpha}$ has the matrix representation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Analogously we find

$$\begin{aligned} s_\alpha(\boldsymbol{x}) &= \begin{bmatrix} \cos(2\alpha-v) \\ \sin(2\alpha-v) \end{bmatrix} = \begin{bmatrix} \cos 2\alpha\cos v + \sin 2\alpha\sin v \\ \sin 2\alpha\cos v - \cos 2\alpha\sin v \end{bmatrix} \\ &= \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}. \end{aligned}$$

so that $\boldsymbol{s_\alpha}$ has the matrix representation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This concludes the proof of the introductory statement.

It is often useful to to have the following relations between rotations and reflektions of the plane in mind

$$\begin{aligned} & s_{\frac{\pi}{4}} \circ d_\alpha = s_{\frac{\pi}{4}-\frac{\alpha}{2}} \\ & s_\alpha = s_{\frac{\pi}{4}} \circ d_{\frac{\pi}{2}-2\alpha}. \end{aligned}$$

The first relation follows from

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} = \\ \begin{bmatrix} \sin\alpha & \cos\alpha \\ \cos\alpha & -\sin\alpha \end{bmatrix} = \begin{bmatrix} \cos(\frac{\pi}{4}-\alpha) & \sin(\frac{\pi}{4}-\alpha) \\ \sin(\frac{\pi}{4}-\alpha) & -\cos(\frac{\pi}{4}-\alpha) \end{bmatrix}.$$

The last two relations are forund from the first by substituting $\alpha$ with $\frac{\pi}{2} - 2\alpha$. ♦

Part of the following section will be devoted to consider the problem of generalising the spectral decomposition of an arbitrary matrix.

### 1.4.2 Singular value decomposition of an arbitrary matrix. $Q$- and $R$-mode analysis

We first state the main result, also known as Eckart-Young's theorem.

**THEOREM 1.13.** Let $\mathbf{x}$ be an arbitrary $n \times p$ matrix of rank $r$. Then orthogonal matrices $\mathbf{U}$ $(p \times r)$ and $\mathbf{V}$ $(n \times r)$ exist, as do positive numbers $\gamma_1, \dots, \gamma_r$, such that

$$x = \mathbf{V}\,\mathbf{\Gamma}\,\mathbf{U}' = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_r] \begin{bmatrix} \gamma_1 \cdots & 0 \\ \vdots & \vdots \\ 0 \cdots & \gamma_r \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_r' \end{bmatrix} = \gamma_1 \boldsymbol{v}_1 \boldsymbol{u}_1' + \cdots + \gamma_r \boldsymbol{v}_r \boldsymbol{u}_r',$$

where $\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \dots, \gamma_r)$ and $\boldsymbol{v}_1, \dots, \boldsymbol{v}_r$ are the columns of $\mathbf{V}$ and $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ are the columns of $\mathbf{U}$. ▲

**PROOF 1.11.** Omitted. See e.g. [9]. ■

The numbers $\gamma_1, \ldots, \gamma_r$ are called $\mathbf{x}$'s singular values.

In the sequel we will investigate the relationship between $\mathbf{x}$ 's singular values and the eigenvalue problems for the symmetrical matrices $\mathbf{x}\,\mathbf{x}'$ $(n \times n)$ and $\mathbf{x}'\mathbf{x}$ $(p \times p)$.

However, first we will state

**THEOREM 1.14.** For an arbitrary (real valued) matrix $\mathbf{x}$ it holds that $\mathbf{x}'\mathbf{x}$ and $\mathbf{x}\,\mathbf{x}'$ have non-negative eigenvalues and

$$\mathrm{rg}(\mathbf{x}'\mathbf{x}) = \mathrm{rg}(\mathbf{x}\,\mathbf{x}') = \mathrm{rg}(\mathbf{x})$$

▲

**PROOF 1.12.** It suffices to prove the results for $\mathbf{x}'\mathbf{x}$. It is obvious that $\mathbf{x}'\mathbf{x}$ is symmetric, so an orthogonal matrix $\mathbf{P}$, exists such that

$$\mathbf{P}'\mathbf{x}'\mathbf{x}\,\mathbf{P} = \mathbf{\Lambda}$$

i.e.

$$(\mathbf{x}\,\mathbf{P})'(\mathbf{x}\mathbf{P}) = \mathbf{\Lambda}.$$

By letting $\mathbf{x}\,\mathbf{P} = \mathbf{B} = (b_{ij})$, we find $\mathbf{B}'\mathbf{B} = \mathbf{\Lambda}$, i.e.

$$\lambda_i = \sum_j b_{ij}^2 > 0,$$

i.e. $\mathbf{x}'\mathbf{x}$ has non-negative eigenvectors. Furthermore we see that

$$\begin{aligned} \mathrm{rg}(\mathbf{x}'\mathbf{x}) &= \text{card}(\lambda_i \neq 0) \\ &= \text{card}\{\text{columns } \boldsymbol{b}_j \text{ in } \mathbf{B} \text{ , which are} \neq \mathbf{0}\,\} \end{aligned}$$

Since $\boldsymbol{b}_i'\boldsymbol{b}_j = 0$ for $i \neq j$ (due to equation 1.1) we have

$$\mathrm{rg}(\mathbf{x}'\mathbf{x}) = \mathrm{rg}(\mathbf{B})$$

Since $\mathbf{P}$ is regular, and using a result on page 13, we find

$$\mathrm{rg}(\mathbf{B}) = \mathrm{rg}(\mathbf{x}\,\mathbf{P}) = \mathrm{rg}(\mathbf{x}).$$

■

We state a small corollary to the theorem.

**COROLLORY 1.2.** Let $\mathbf{\Sigma}$ be symmetrical and positive definite. Then for an arbitrary matrix $\mathbf{x}$ it holds that

$$\mathrm{rg}(\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}) = \mathrm{rg}(\mathbf{x}),$$

under the condition that the involved products exist.

**PROOF 1.13.** Since $\mathbf{\Sigma}^{-1}$ is also regular and positive definite, an orthogonal matrix $\mathbf{P}$ exists, such that

$$\mathbf{P}'\mathbf{\Sigma}^{-1}\mathbf{P} = \mathbf{\Lambda},$$

where $\mathbf{\Lambda}$ is a diagonal matrix. This implies

$$\mathbf{\Sigma}^{-1} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}' = \mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{P}' = \mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}}(\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})' = \mathbf{B}\,\mathbf{B}'.$$

Here $\mathbf{\Lambda}^{\frac{1}{2}}$ denotes the diagonal matrix, whos diagonal elements are the square roots of the corresponding elements of $\mathbf{\Lambda}$. It is obvious that $\mathbf{B}$ is regular. This relation is inserted and we find

$$\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} = \mathbf{x}'\mathbf{B}\,\mathbf{B}'\mathbf{x} = (\mathbf{B}'\mathbf{x})'\mathbf{B}'\mathbf{x},$$

i.e.

$$\mathrm{rg}(\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}) = \mathrm{rg}(\mathbf{B}'\mathbf{x}) = \mathrm{rg}(\mathbf{x}),$$

which concludes the proof. ■

Using the notation from theorem 1.14 we have.

**THEOREM 1.15.**

i) the matrix $\mathbf{x}\,\mathbf{x}'$ $(n \times n)$ has $r$ positive eigenvalues and $n - r$ eigenvalues equal to 0. The positive eigenvalues are $\gamma_1^2, \dots, \gamma_r^2$, where $\gamma_1, \dots, \gamma_r$ are the singular values of $\mathbf{x}$. The corresponding eigenvectors are $\boldsymbol{v}_1, \dots, \boldsymbol{v}_r$.

ii) Similarly $\mathbf{x}'\mathbf{x}$ $(p \times p)$ has $r$ positive and $(p - r)$ 0-eigenvalues. The positive eigenvalues are $\gamma_1^2, \dots, \gamma_r^2$ and the corresponding eigenvectors are $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$.

iii) The positive eigenvalues of $\mathbf{x}\,\mathbf{x}'$ and $\mathbf{x}'\mathbf{x}$ are therefore equal and the relationship between the corresponding eigenvectors is $(m = 1, \dots, r)$

$$\boldsymbol{v}_m = \frac{1}{\gamma_m}\mathbf{x}\,\boldsymbol{u}_m \quad \text{og} \quad \boldsymbol{u}_m = \frac{1}{\gamma_m}\mathbf{x}'\boldsymbol{v}_m,$$

or in a more compact notation

$$\mathbf{V} = \mathbf{x}\,\mathbf{U}\,\boldsymbol{\Gamma}^{-1} \quad \text{og} \quad \mathbf{U} = \mathbf{x}'\mathbf{V}\,\boldsymbol{\Gamma}^{-1}$$

▲

**PROOF 1.14.** Follows by use of Eckart-Young's theorem. ■

**REMARK 1.5.** Analysis of the matrix $\mathbf{x}'\mathbf{x}$ is called $\boldsymbol{R}$ -mode analysis and the analysis of $\mathbf{x}\,\mathbf{x}'$ is called $\boldsymbol{Q}$ -mode analysis. These names originate from factor analysis, cf. chapter 8. ▼

**REMARK 1.6.** The theorem implies that one can find the results for an R-mode analysis from a Q-mode analysis ad vice versa. For practical use one should therefore consider which of the matrices $\mathbf{x}'\mathbf{x}$ and $\mathbf{x}\,\mathbf{x}'$ has lowest order. ▼

### 1.4.3 Quadratic forms and positive semi-definite matrices

In this section we still consider symmetrical matrices only.

By the quadratic form corresponding to the symmetrical matrix $\mathbf{A}$ we mean the projection

$$\boldsymbol{x} \to \boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = \sum a_{ii}x_i^2 + 2\sum_{1<j} a_{ij}x_i x_j.$$

We say that a symmetrical matrix $\mathbf{A}$ is positive definite respectively positive semi-definite if the corresponding quadratic form is positive respectively non-negative for vectors different from the 0-vector, i.e. if

$$\forall \boldsymbol{x} \neq \mathbf{0} : \boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} > 0,$$

respectively

$$\forall \boldsymbol{x} \neq \mathbf{0} : \boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} \geq 0.$$

We then also say the quadratic form is positive definite respectively positive semi-definite.

We have the following

**THEOREM 1.16.** The symmetrical matrix $\mathbf{A}$ is positive definite respectively semi-definite, if all $\mathbf{A}$ 's eigenvalues are positive respectively non-negative. ▲

**PROOF 1.15.** With $\mathbf{P}$ as in theorem 1.10 we have

$$\begin{aligned} \boldsymbol{x}'\mathbf{A}\boldsymbol{x} &= \boldsymbol{x}'\mathbf{P}'\mathbf{P}\,\mathbf{A}\,\mathbf{P}\,\mathbf{P}'\boldsymbol{x} = (\mathbf{P}'\boldsymbol{x})'\boldsymbol{\Lambda}(\mathbf{P}'\boldsymbol{x}) \\ &= \boldsymbol{y}'\boldsymbol{\Lambda}\,\boldsymbol{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2. \end{aligned}$$

■

Another useful result is

**THEOREM 1.17.** A symmetrical $n \times n$ matrix $\mathbf{A}$ is positive definite if all principal minors

$$d_i = \det \begin{bmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} \end{bmatrix}, \qquad i = 1, \ldots, n,$$

are positive. ▲

**PROOF 1.16.** Omitted ■

We now state a very important theorem on extrema of quadratic forms

**THEOREM 1.18.** If we let the eigenvalues for the symmetrical matrix $\mathbf{A}$ equal $\lambda_1 \geq \cdots \geq \lambda_n$ with corresponding eigenvectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$, and we define

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}},$$

and

$$M_k = \{\boldsymbol{x} | \boldsymbol{x}'\boldsymbol{p}_i = 0, \qquad i = 1, \dots, k-1\},$$

Then it holds that

$$\begin{aligned} \sup_{\boldsymbol{x}} R(\boldsymbol{x}) &= R(\boldsymbol{p}_1) = \lambda_1, \\ \inf_{\boldsymbol{x}} R(\boldsymbol{x}) &= R(\boldsymbol{p}_n) = \lambda_n, \\ \sup_{\boldsymbol{x} \in M_k} R(\boldsymbol{x}) &= R(\boldsymbol{p}_k) = \lambda_k. \end{aligned}$$

▲

**PROOF 1.17.** An arbitrary vector $\boldsymbol{x}$ can be written

$$\boldsymbol{x} = \alpha_1 \boldsymbol{p}_1 + \cdots + \alpha_n \boldsymbol{p}_n.$$

If $\boldsymbol{p}_i'\boldsymbol{x} = 0$, $i = 1, \dots, k-1$, we find $\alpha_1 = \cdots = \alpha_{k-1} = 0$, i.e.

$$\boldsymbol{x} = \alpha_k \boldsymbol{p}_k + \cdots + \alpha_n \boldsymbol{p}_n.$$

Therefore we have

$$\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = \alpha_k^2 \lambda_k + \cdots + \alpha_n^2 \lambda_n,$$

and

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \frac{\alpha_k^2 \lambda_k + \cdots + \alpha_n^2 \lambda_n}{\alpha_k^2 + \cdots + \alpha_n^2}$$

It is obvious that this expression is maximal for

$$(\alpha_k, \dots, \alpha_n) = (\alpha_k, 0, \dots, 0),$$

where it takes the value $\lambda_k$. The result with inf is proved analogously. ■

**REMARK 1.7.** The theorem say for $k = 1$, that the unit vector, i.e. the ”direction”, for which the quadratic form takes its maximal value, is the eigenvector corresponding to the largest eigenvalue. If we only consider the quadratic form in unit vectors which are orthogonal to eigenvectors corresponding to the $k-1$ largest eigenvalues, then the theorem says that maximum is in the direction corresponding to the eigenvector which corresponds to the $k$'th largest eigenvalue. ▼

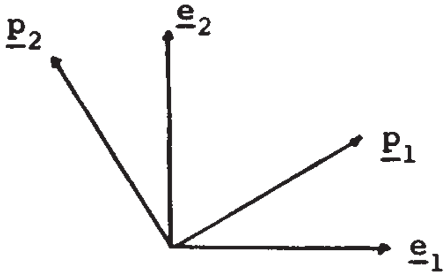


Figure 1.9: Illustration showing change of basis

**REMARK 1.8.** $R(x)$ is also called Rayleigh's coefficient. ▼

We will now describe the level curves for positive definite forms.

**THEOREM 1.19.** Let $\mathbf{A}$ be positive definite. Then the set of solutions for the equation

$$\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = c, \qquad c > 0,$$

is an ellipsoid with principle axes in the directions of the eigenvectors. The first principle axis corresponds with the smallest eigenvalue, the second to the second smallest eigenvalue etc. ▲

**PROOF 1.18.** We consider the matrix $\mathbf{P} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_n)$, whos columns are the coordinates of orthonormed eigenvectors of $\mathbf{A}$. Assuming $\boldsymbol{y} = \mathbf{P}'\boldsymbol{x}$ the following holds

$$\begin{aligned} \boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} &= \boldsymbol{y}'\boldsymbol{\Lambda}\boldsymbol{y} \\ &= \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2 \\ &= \frac{y_1^2}{(1/\sqrt{\lambda_1})^2} + \cdots + \frac{y_n^2}{(1/\sqrt{\lambda_n})^2} \end{aligned} \tag{1.4}$$

The matrix equation

$$\boldsymbol{y} = \mathbf{P}'\boldsymbol{x} \quad \Leftrightarrow \quad \boldsymbol{x} = \mathbf{P}\,\boldsymbol{y}$$

corresponds to a change of basis from the original orthonormal basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ to the orthonormal basis $\{\boldsymbol{p}_1, \dots, \boldsymbol{p}_n\}$.

This is seen by letting $S$ be a point whos $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$-coordinates are called $\boldsymbol{x}$ and whos $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\}$-coordinates are called $\boldsymbol{y}$. Then it holds that

$$x_1\boldsymbol{e}_1 + \cdots + x_n\boldsymbol{e}_n = y_1\boldsymbol{p}_1 + \cdots + y_n\boldsymbol{p}_n,$$

or

$$(\boldsymbol{e}_1 \cdots \boldsymbol{e}_n)\boldsymbol{x} = (\boldsymbol{p}_1 \cdots \boldsymbol{p}_n)\boldsymbol{y},$$

i.e.

$$\mathbf{I}\,\boldsymbol{x} = \mathbf{P}\,\boldsymbol{y},$$

where $\mathbf{I}$ is a unit matrix.

The expression in 1.4 therefore shows the equation of the set of solutions in $y$-coordinates corresponding to the coordinate system consisting of orthonormed eigenvectors. This shows that we are dealing with an ellipsoid. The rest of the theorem now follows by noting that the 1.st principle axis corresponds to the $y_i$, for which $1/\sqrt{\lambda_i}$ is maximal, i.e. for which $\lambda_i$ is minimal. ■

**REMARK 1.9.** If the matrix is only positive semi-definite then the set of solutions to the equation correspond to an elliptical cylinder. This can be seen by change of base to the base $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\}$ consisting of orthonormal eigenvectors, where we for simplicity assume that $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r$ corresponds to the eigenvalues which are different form 0. We then have

$$\begin{aligned} \boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = c \quad &\Leftrightarrow \quad \lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 + 0y_{r+1}^2 + \cdots + 0y_n^2 = c \\ &\Leftrightarrow \quad \lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 = c. \end{aligned}$$

This leads to the the statement. If we consider the restriction of the quadratic form to the subspace spanned by the eigenvectors corresponding to eigenvectors $> 0$, then the set of solutions becomes an ellipsoid. ▼

**EXAMPLE 1.7.** We consider the symmetrical positive definite matrix

$$\mathbf{A} = \begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix}.$$

The quadratic form corresponding to $\mathbf{A}$ is

$$\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = 3x_1^2 + 2x_2^2 + 2\sqrt{2}x_1x_2,$$

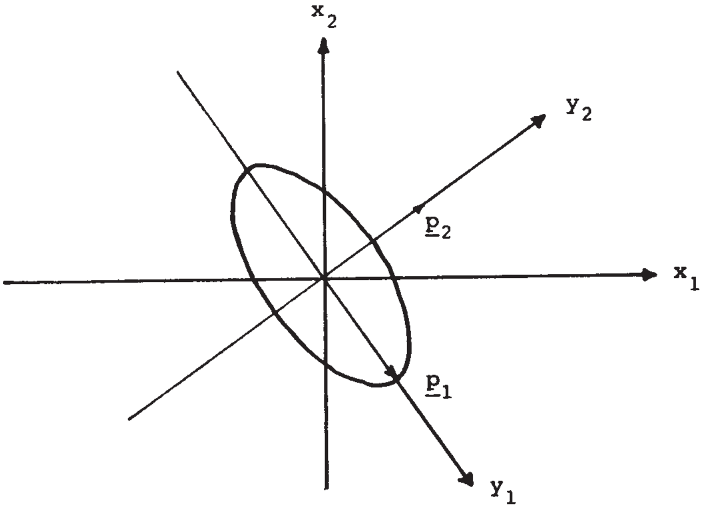


Figure 1.10: Ellipse determined by the quadratic form given in example 1.7.

so the unit ellipse corresponding to $\mathbf{A}$ is the set of solutions to the equation

$$3x_1^2 + 2x_2^2 + 2\sqrt{2}x_1x_2 = 1.$$

In order to determine the principle axes we determine $\mathbf{A}$'s eigenvalues. We find

$$\begin{aligned}(\mathbf{A} - \lambda\mathbf{I}) = 0 \quad &\Leftrightarrow \quad \lambda^2 - 5\lambda + 4 = 0\\ &\Leftrightarrow \quad \lambda = 1 \quad \vee \quad \lambda = 4.\end{aligned}$$

Eigen vectors corresponding to $\lambda = 1$ respectively $\lambda = 4$ are seen to be of the form $t(1, -\sqrt{2})$ respectively $t(1, \sqrt{2}/2)$. We norm these and get

$$\boldsymbol{p}_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{-\sqrt{6}}{3} \end{bmatrix} \quad , \quad \boldsymbol{p}_2 = \begin{bmatrix} \frac{\sqrt{6}}{3} \\ \frac{-\sqrt{3}}{3} \end{bmatrix}.$$

If we choose the base $\{\boldsymbol{p}_1, \boldsymbol{p}_2\}$, then the coordinate representation of the quadratic form becomes

$$\boldsymbol{y} \to y_1^2 + 4y_2^2,$$

The ellipse has the equation

$$\frac{y_1^2}{1^2} + \frac{y_2^2}{\frac{1}{2}^2} = 1.$$

It is illustrated in figure 1.10

Since

$$\begin{aligned}\boldsymbol{p}_1 &= \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{-\sqrt{6}}{3} \end{bmatrix} = \begin{bmatrix} 0.577 \\ -0.820 \end{bmatrix} \\ &\simeq \begin{bmatrix} \cos(-54.7^\circ) \\ \sin(-54,7^\circ) \end{bmatrix},\end{aligned}$$

the new coordinate system corresponds to a rotation of the old one with the angle $-54.7°$. ♦

### 1.4.4 The general eigenvalue problem for symmetrical matrices

For use with the theory of canonical correlations and in discriminant analysis we will need a slightly more general concept of eigenvalues than seen in the previous sections. We introduce the concept in

**DEFINITION 1.3.** Let $\mathbf{A}$ and $\mathbf{B}$ be real-valued $m \times m$ symmetrical matrices and let $\mathbf{B}$ be of full rank. A number $\lambda$, for which

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0,$$

is termed an eigenvalue of $\mathbf{A}$ w.r.t. $\mathbf{B}$. For such a $\lambda$ it is possible to find an $\boldsymbol{x} \neq 0$ such that

$$\mathbf{A}\,\boldsymbol{x} = \lambda\mathbf{B}\,\boldsymbol{x}.$$

Such a vector $\boldsymbol{x}$ is called an eigenvector for $\mathbf{A}$ w.r.t. $\mathbf{B}$. ▲

**REMARK 1.10.** The concepts given above can be traced back to eigenvalues and eigenvectors for the **non**-symmetrical matrix $\mathbf{B}^{-1}\mathbf{A}$. ▼

**THEOREM 1.20.** We consider again the situation in the definition 1.3 and further let $\mathbf{B}$ be positive definite. There are then $m$ real eigenvalues of $\mathbf{A}$ w.r.t. $\mathbf{B}$. If $\mathbf{A}$ is positive semi-definite, then these will be non-negative and if $\mathbf{A}$ is positive definite then they will be positive. ▲

**PROOF 1.19.** According to theorem 1.11 there is a matrix $\mathbf{I}$ where

$$\mathbf{T}'\mathbf{B}\,\mathbf{T} = \mathbf{I}.$$

Let

$$\mathbf{D} = \mathbf{T}'\mathbf{A}\,\mathbf{T}$$

$\mathbf{D}$ is obviously symmetrical, and since

$$\boldsymbol{x}'\mathbf{D}\,\boldsymbol{x} = (\mathbf{T}\,\boldsymbol{x})'\mathbf{A}(\mathbf{T}\,\boldsymbol{x}),$$

we see that $\mathbf{D}$ and $\mathbf{A}$ are at the same time respectively positive semi-definite and positive definite.

Now we have

$$\begin{aligned}(\mathbf{D}-\lambda\mathbf{I})\boldsymbol{v} = 0 \quad &\Leftrightarrow \quad (\mathbf{T}'\mathbf{A}\,\mathbf{T} - \lambda\mathbf{T}'\mathbf{B}\,\mathbf{T})\boldsymbol{v} = 0\\ &\Leftrightarrow \quad (\mathbf{A}-\lambda\mathbf{B})(\mathbf{T}\,\boldsymbol{v}) = 0\end{aligned}$$

From this we deduce that $\mathbf{D}$'s eigenvalues equal $\mathbf{A}$'s eigenvalues w.r.t. $\mathbf{B}$, and that the eigenvectors of $\mathbf{A}$ w.r.t. $\mathbf{B}$ are found by using the transformation $\mathbf{T}$ on $\mathbf{D}$ 's eigenvectors. The result regarding the sign of the eigenvalues follows trivially. ■

**THEOREM 1.21.** Let the situation be as above. Then a basis exists for $R^m$ consisting of eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ of $\mathbf{A}$ w.r.t. $\mathbf{B}$. These vectors can be chosen as conjugated vectors both w.r.t. $\mathbf{A}$ as well as w.r.t. $\mathbf{B}$ , i.e.

$$\boldsymbol{u}_i'\mathbf{A}\,\boldsymbol{u}_j = \boldsymbol{u}_i'\mathbf{B}\,\boldsymbol{u}_j = 0.$$

▲

**PROOF 1.20.** Follows from the proof of the above theorem and of the corollary to theorem 1.9, remembering that

$$0 = \boldsymbol{v}_i'\boldsymbol{v}_j = (\boldsymbol{v}_i'\mathbf{T}')\mathbf{T}'^{-1}\mathbf{T}^{-1}(\mathbf{T}\,\boldsymbol{v}_j) = \boldsymbol{u}_i'\mathbf{B}\,\boldsymbol{u}_j,$$

where $\boldsymbol{v}_i, \ldots, \boldsymbol{v}_m$ is an orthonormal basis for $R^m$ consisting of eigenvectors of $\mathbf{D}$.

Finally we have

$$\boldsymbol{u}_i'\mathbf{A}\,\boldsymbol{u}_j = \lambda_j\boldsymbol{u}_i'\mathbf{B}\,\boldsymbol{u}_j = 0$$

■

**THEOREM 1.22.** Let $\mathbf{A}$ be symmetrical and let $\mathbf{B}$ be positive definite. Then a regular matrix $\mathbf{R}$ exists with

$$\mathbf{R}'\mathbf{A}\,\mathbf{R} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_n),$$

and

$$\mathbf{R}'\mathbf{B}\,\mathbf{R} = \mathbf{I},$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $\mathbf{A}$ w.r.t. $\mathbf{B}$. If we term the $i$'th column in $\mathbf{R}'^{-1}$ $s_i$ then these relations can be written

$$\mathbf{A} = \lambda_1 \boldsymbol{s}_1 \boldsymbol{s}_1' + \cdots + \lambda_m \boldsymbol{s}_m \boldsymbol{s}_m',$$

and

$$\mathbf{B} = \boldsymbol{s}_1 \boldsymbol{s}_1' + \ldots + \boldsymbol{s}_m \boldsymbol{s}_m'.$$

▲

**PROOF 1.21.** From the proof of theorem 1.20 we consider the $\mathbf{D} = \mathbf{T}'\mathbf{A}\,\mathbf{T}$. Since $\mathbf{D}$ is symmetrical, according to theorem 1.10 there exists an orthogonal matrix $\mathbf{C}$ with

$$\mathbf{C}'\mathbf{D}\,\mathbf{C} = \boldsymbol{\Lambda},$$

because we have that $\mathbf{D}$'s eigenvalues are $\mathbf{A}$'s eigenvalues w.r.t. $\mathbf{B}$.

If we choose $\mathbf{R} = \mathbf{T}\,\mathbf{C}$, then we have that

$$\mathbf{R}'\mathbf{B}\,\mathbf{R} = \mathbf{C}'\mathbf{T}'\mathbf{B}\,\mathbf{T}\,\mathbf{C} = \mathbf{C}'\mathbf{C} = \mathbf{I},$$

and

$$\mathbf{R}'\mathbf{A}\,\mathbf{R} = \mathbf{C}'\mathbf{T}'\mathbf{A}\,\mathbf{T}\,\mathbf{C} = \mathbf{C}'\,\mathbf{D}\,\mathbf{C} = \boldsymbol{\Lambda}.$$

■

Finally we state an analogue of theorem 1.18 in the following

**THEOREM 1.23.** Let $\mathbf{A}$ be positive semi-definite and let $\mathbf{B}$ be positive definite. Let $\mathbf{A}$'s eigenvalues w.r.t. $\mathbf{B}$ be $\lambda_1 \geq \cdots \geq \lambda_m$ and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ denote a basis for $R^m$ consisting of the corresponding eigenvectors with $\boldsymbol{v}_i \mathbf{B}\, \boldsymbol{v}_j = 0 \quad i \neq j$. We let

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x}}{\boldsymbol{x}'\mathbf{B}\,\boldsymbol{x}}$$

and

$$M_k = \{\boldsymbol{x} | \boldsymbol{x}'\mathbf{B}\,\boldsymbol{v}_1 = \cdots = \boldsymbol{x}'\mathbf{B}\,\boldsymbol{v}_{k-1} = 0\},$$

and we then obtain

$$\begin{aligned} \sup_{x} R(\boldsymbol{x}) &= R(\boldsymbol{v}_1) = \lambda_1 \\ \inf_{x} R(\boldsymbol{x}) &= R(\boldsymbol{v}_m) = \lambda_m \\ \sup_{\boldsymbol{x} \in M_k} R(\boldsymbol{x}) &= R(\boldsymbol{v}_k) = \lambda_k. \end{aligned}$$

▲

**PROOF 1.22.** Without loss of generality the $\boldsymbol{v}_i$'s can be chosen so that $\boldsymbol{v}_i'\mathbf{B}\boldsymbol{v}_i = 1$, and since an arbitrary vector $\boldsymbol{x}$ can be written

$$\boldsymbol{x} = \alpha_1 \boldsymbol{v}_1 + \cdots + \alpha_m \boldsymbol{v}_m,$$

we find

$$R(\boldsymbol{x}) = \frac{\sum \alpha_i^2 \boldsymbol{v}_i' \mathbf{A}\, \boldsymbol{v}_i}{\sum \alpha_i^2 \boldsymbol{v}_i' \mathbf{B}\, \boldsymbol{v}_i} = \frac{\sum \lambda_i \alpha_i^2}{\sum \alpha_i^2}.$$

From this the two first statements are easily seen. If $\boldsymbol{x} \in M_k$, then $\boldsymbol{x}$ can be written

$$\boldsymbol{x} = \alpha_k \boldsymbol{v}_k + \cdots + \alpha_m \boldsymbol{v}_m,$$

and

$$R(\boldsymbol{x}) = \frac{\lambda_k \alpha_k^2 + \cdots + \lambda_m \alpha_m^2}{\alpha_m^2 + \cdots + \alpha_m^2},$$

which leads to the desired result. ■

### 1.4.5 The trace of a matrix

By the term trace of the (symmetrical) matrix $\mathbf{A}$ we mean the sum of the diagonal elements. i.e.

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

For (square) matrices $\mathbf{A}$ and $\mathbf{B}$ the following holds

$$\operatorname{tr}(\mathbf{A}\,\mathbf{B}) = \operatorname{tr}(\mathbf{B}\,\mathbf{A}). \tag{1.5}$$

Furthermore we have that the trace equals the sum of eigenvalues, i.e.

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i.$$

This follows trivially from 1.5 and theorem 1.10
For positive semi-definite matrices the trace is therefore another measure of ”size” of a matrix. If the trace is large then at least some of the eigenvalues are large. On the other hand this measure is not sensitive to if some eigenvalues might be 0, i.e. if the matrix is degenerate. The determinant is sensitive to that, since we recall

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i.$$

We note further that for an idempotent matrix $\mathbf{A}$ we have that

$$\operatorname{tr}(\mathbf{A}) = \operatorname{rg}(\mathbf{A}).$$

Further we have

$$\operatorname{tr}(\mathbf{B}\,\mathbf{B}^{-}) = \operatorname{rg}(\mathbf{B}),$$

where $\mathbf{B}^{-}$ is an arbitrary pseudo-inverse of $\mathbf{B}$.

Finally we note that for a regular matrix $\mathbf{S}$ we have that

$$\operatorname{tr}(\mathbf{S}^{-}\mathbf{B}\,\mathbf{S}) = \operatorname{tr}(\mathbf{B}).$$

### 1.4.6 Differentiation of linear form and quadratic form

Let f : $R^n \to R$. We will use the following notation for the vector of partial derivatives

$$\frac{\partial \mathrm{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial \mathrm{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \mathrm{f}(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial \mathrm{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix}.$$

The following theorem holds for differentiation of certain forms

**THEOREM 1.24.** For a symmetrical $(n \times n)$-matrix $\mathbf{A}$ and an arbitrary $n$-dimensional vector $\boldsymbol{b}$ it holds that

i) $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{b}'\boldsymbol{x}) = \boldsymbol{b}$

ii) $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{x}) = 2\boldsymbol{x}$

iii) $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x}) = 2\mathbf{A}\,\boldsymbol{x}$.

▲

**PROOF 1.23.** The proof of i) and ii) are trivial. iii) is (strangely) proved most easily by means of the definition. For an arbitrary vector $\boldsymbol{h}$ we have that

$$(\boldsymbol{x} + \boldsymbol{h})'\mathbf{A}(\boldsymbol{x} + \boldsymbol{h}) = \boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} + \boldsymbol{h}'\mathbf{A}\,\boldsymbol{h} + 2\boldsymbol{h}'\mathbf{A}\,\boldsymbol{x}$$

By choosing $\boldsymbol{h} = (0, \ldots, h, \ldots, 0)'$ we see that

$$\frac{\partial}{\partial \boldsymbol{x}_i}(\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x}) = 2\sum_{j=1}^{h} a_{ij}x_j,$$

and the result follows readily. ■

We will illustrate the use of the theorem in the following

**EXAMPLE 1.8.** We want to find the minimum of the function

$$\mathrm{g}(\boldsymbol{\theta}) = (\boldsymbol{y} - \mathbf{A}\,\boldsymbol{\theta})'\mathbf{B}(\boldsymbol{y} - \mathbf{A}\,\boldsymbol{\theta}),$$

where $\boldsymbol{y}$, $\mathbf{A}$ and $\mathbf{B}$ are given and $\mathbf{B}$ is further positive semidefinite (and symmetrical). Since $\mathrm{g}(\boldsymbol{\theta})$ is convex (a paraboloid, possibly degenerate), then the point corresponding to the minimum is found by solving the equation

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{g}(\boldsymbol{\theta}) = \mathbf{0}.$$

First we rewrite $\mathrm{g}$. We have that

$$\begin{aligned} \mathrm{g}(\boldsymbol{\theta}) &= \boldsymbol{y}'\mathbf{B}\,\boldsymbol{y} - \boldsymbol{\theta}'\mathbf{A}'\mathbf{B}\,\boldsymbol{y} + \boldsymbol{\theta}'\mathbf{A}'\mathbf{B}\,\mathbf{A}\,\boldsymbol{\theta} - \boldsymbol{y}'\mathbf{B}\,\mathbf{A}\,\boldsymbol{\theta} \\ &= \boldsymbol{y}'\mathbf{B}\,\boldsymbol{y} - 2\boldsymbol{y}'\mathbf{B}\,\mathbf{A}\,\boldsymbol{\theta} + \boldsymbol{\theta}'\mathbf{A}'\mathbf{B}\,\mathbf{A}\,\boldsymbol{\theta}. \end{aligned}$$

Here we have used that

$$\boldsymbol{\theta}'\mathbf{A}'\mathbf{B}\,\boldsymbol{y} = \boldsymbol{y}'\mathbf{B}\,\mathbf{A}\,\boldsymbol{\theta}$$

(both $1 \times 1$ matrices, i.e. a scalar, and each others transposed). From this follows that

$$\frac{\partial \mathrm{g}}{\partial \boldsymbol{\theta}} = -2\mathbf{A}'\mathbf{B}\,\boldsymbol{y} + 2\mathbf{A}'\mathbf{B}\,\mathbf{A}\,\boldsymbol{\theta},$$

and it is seen that

$$\frac{\partial \mathrm{g}}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad \leftrightarrow \quad \mathbf{A}'\mathbf{B}\,\mathbf{A}\boldsymbol{\theta} = \mathbf{A}'\mathbf{B}\,\boldsymbol{y}.$$

This equation has as mentioned always at least one root. If $\mathbf{A}'\mathbf{B}\,\mathbf{A}$ is regular then we have

$$\boldsymbol{\theta}_{\min} = (\mathbf{A}'\mathbf{B}\,\mathbf{A})^{-1}\mathbf{A}'\mathbf{B}\,\boldsymbol{y}.$$

If the matrix is singular, then we can write

$$\boldsymbol{\theta}_{\min} = (\mathbf{A}'\mathbf{B}\,\mathbf{A})^{-}\mathbf{A}'\mathbf{B}\,\boldsymbol{y},$$

where $(\mathbf{A}'\mathbf{B}\,\mathbf{A})^{-}$ denotes a pseudo-inverse of $\mathbf{A}'\mathbf{B}\,\mathbf{A}$. ♦

We are now able to find an alternative description of the principle axes in an ellipsoid, due to

**THEOREM 1.25.** Let $\mathbf{A}$ be a positive definite symmetrical matrix. The principle directions of the ellipsoid $E_c$ with the equation

$$\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = c, \qquad c > 0$$

are those directions where $\boldsymbol{x}'\boldsymbol{x}$, $\boldsymbol{x} \in E_c$, has stationary points. ▲

**PROOF 1.24.** We may assume that $x = 1$. We then need to find the stationary points for

$$\mathrm{f}(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{x}$$

with the condition that

$$\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = 1$$

We apply a Lagrange multiplier technique and define

$$\varphi(\boldsymbol{x}, \lambda) = \boldsymbol{x}'\boldsymbol{x} - \lambda(\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} - 1).$$

Be differentiation we obtain

$$\frac{\partial \varphi}{\partial \boldsymbol{x}} = 2\boldsymbol{x} - 2\lambda\mathbf{A}\,\boldsymbol{x}.$$

If this quantity is to equal $\mathbf{0}$, then

$$\boldsymbol{x} = \lambda\mathbf{A}\,\boldsymbol{x}$$

or

$$\mathbf{A}\,\boldsymbol{x} = \frac{1}{\lambda}\boldsymbol{x},$$

i.e. $\boldsymbol{x}$ must be an eigenvector. ■

## 1.5 Tensor- or Kronecker product of matrices

It is an advantage to use this product when treating the multidimensional general linear model.

**DEFINITION 1.4.** Let $\mathbf{A}$ be an $m \times n$ matrix and let $\mathbf{B}$ be a $k \times \ell$ matrix. By the term tensor - or Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ we mean the matrix

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B}) = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \tag{1.6}$$

This concept corresponds to the tensor product of linear projections, which can be stated independently of coordinate system (see e.g. [3]). It this is introduced in coordinate form then we can either use 1.6 or equivalently, $\mathbf{A} \otimes \mathbf{B} = (\mathbf{A} b_{ij})$. This only corresponds to changing the order of the coordinates, i.e. to changing row and columns in the respective matrices. ▲

We briefly give some rules of calculation for the tensor-product. These are proved trially by means of the definition.

$\mathbf{O} \otimes \mathbf{A} = \mathbf{A} \otimes \mathbf{O} = \mathbf{O}$

$(\mathbf{A}_1 + \mathbf{A}_2) \otimes \mathbf{B} = \mathbf{A}_1 \otimes \mathbf{B} + \mathbf{A}_2 \otimes \mathbf{B}$

iii) $\mathbf{A} \otimes (\mathbf{B}_1 + \mathbf{B}_2) = \mathbf{A} \otimes \mathbf{B}_1 + \mathbf{A} \otimes \mathbf{B}_2$

iv) $\alpha\mathbf{A} \otimes \beta\mathbf{B} = \alpha\beta\mathbf{A} \otimes \mathbf{B}$

v) $\mathbf{A}_1\mathbf{A}_2 \otimes \mathbf{B}_1\mathbf{B}_2 = (\mathbf{A}_1 \otimes \mathbf{B}_1)(\mathbf{A}_2 \otimes \mathbf{B}_2)$

vi) $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$, if the inverses exist

vii) $(\mathbf{A} \otimes \mathbf{B})^{-} = \mathbf{A}^{-} \otimes \mathbf{B}^{-}$

viii) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$

ix) Let $\mathbf{A}$ be symmetrical and $p \times p$, have eigenvalues $\alpha_1, \ldots, \alpha_p$ and eigenvectors $\boldsymbol{x}_i$, and let $\mathbf{B}$, be symmetrical and $q \times q$, have eigenvalues $\beta_1, \ldots, \beta_q$ and eigenvectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q$. Then $\mathbf{A} \otimes \mathbf{B}$ will have the eigenvalues $\alpha_i\beta_j$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, with corresponding eigenvectors.

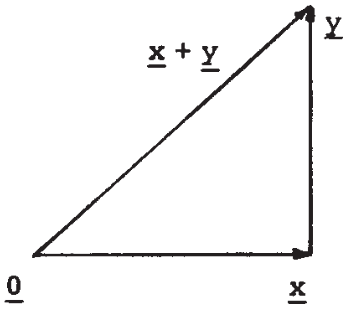
$$(\boldsymbol{x}_i \otimes \boldsymbol{y}_j \sim) \begin{bmatrix} x_{1i}\boldsymbol{y}_j \\ \vdots \\ x_{pi}\boldsymbol{y}_j \end{bmatrix}$$

x) $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^q (\det \mathbf{B})^p$

## 1.6 Inner products and norms

For $n$-dimensional vectors we note that the inner product or scalar product or dot product of $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}'\boldsymbol{y} = (x_1 \ldots x_n) \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=n}^{n} x_i y_i,$$



$$
\begin{aligned}
&\|\boldsymbol{x}+\boldsymbol{y}\|^2\\
&= (\boldsymbol{x}+\boldsymbol{y})'(\boldsymbol{x}+\boldsymbol{y})\\
&= \boldsymbol{x}'\boldsymbol{x}+\boldsymbol{x}'\boldsymbol{y}+\boldsymbol{y}'\boldsymbol{x}+\boldsymbol{y}'\boldsymbol{y}\\
&= \boldsymbol{x}'\boldsymbol{x}+\boldsymbol{y}'\boldsymbol{y}\\
&= \|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2.
\end{aligned}
$$

and we note that $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal if and only if

$$\boldsymbol{x}\cdot\boldsymbol{y} = \boldsymbol{x}'\boldsymbol{y} = 0.$$

The corresponding norm is

$$\|\boldsymbol{x}\| = (\boldsymbol{x}\cdot\boldsymbol{x})^{\frac{1}{2}} = (\boldsymbol{x}'\boldsymbol{x})^{\frac{1}{2}} = \sqrt{x_1^2+\cdots+x_n^2}$$

We note that $\|\boldsymbol{x}-\boldsymbol{y}\|$ represents the euclidian distance between the points $\boldsymbol{x}$ and $\boldsymbol{y}$.

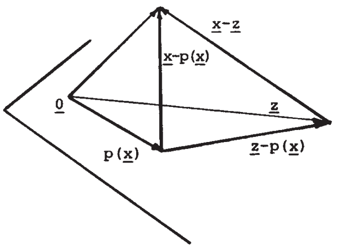
For orthogonal vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ (i.e. $\boldsymbol{x}\perp\boldsymbol{y}$) we have the pythagorean theorem

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2;$$

see figure 1.6. Further we note that the (orthogonal) projection $p(\boldsymbol{x})$ of a vector $\boldsymbol{x}$ onto the sub-space $U$ can be determined by means of the norm, since we have that $p(\boldsymbol{x})$ is given by

$$\|\boldsymbol{x}-p(\boldsymbol{x})\| = \min_{\boldsymbol{z}\in U}\|\boldsymbol{x}-\boldsymbol{z}\|$$

**PROOF 1.25.**



Due to the Pythagorean theorem we have that

$$
\begin{aligned}
&\|\boldsymbol{x}-p(\boldsymbol{x})\|^2 - \|\boldsymbol{z}-p(\boldsymbol{x})\|^2\\
&= \|\boldsymbol{x}-\boldsymbol{z}\|^2,
\end{aligned}
$$

i.e. the minimal value of

$= \|\boldsymbol{x}-\boldsymbol{z}\|^2$, and therefore of

$= \|\boldsymbol{x}-\boldsymbol{z}\|$ is achieved for

$\boldsymbol{z} = p(\boldsymbol{x})$.

■

It is now very easy to show that the validity of the above results only depend on 4 fundamental properties of the inner product. If we term the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$ by $(\boldsymbol{x}|\boldsymbol{y})$ then they are

$$\begin{aligned}
&\text{IP1}: && (\boldsymbol{x}|\boldsymbol{y}) = (\boldsymbol{y}|\boldsymbol{x})\\
&\text{IP2}: && (\boldsymbol{x}+\boldsymbol{y}|\boldsymbol{z}) = (\boldsymbol{x}|\boldsymbol{z}) + (\boldsymbol{y}|\boldsymbol{z})\\
&\text{IP3}: && (k\boldsymbol{x}|\boldsymbol{y}) = k(\boldsymbol{x}|\boldsymbol{y})\\
&\text{IP4}: && \boldsymbol{x} \neq \mathbf{0} \Rightarrow (\boldsymbol{x}|\boldsymbol{x}) > 0.
\end{aligned}$$

For an arbitrary bi-linear form $(\cdot|\cdot)$ , which satisfies the above one can define a concept of orthogonality by

$$\boldsymbol{x} \perp \boldsymbol{y} \quad \overset{d}{\Leftrightarrow} \quad (\boldsymbol{x}|\boldsymbol{y}) = 0.$$

For an arbitrary positive definite symmetrical matrix $\mathbf{A}$ we can define an inner product by

$$(\boldsymbol{x}|\boldsymbol{y})_{\mathbf{A}} = \boldsymbol{x}'\mathbf{A}\,\boldsymbol{y}.$$

It is trivial to prove that IP 1-4 are satisfied. for this inner product and the corresponding norm given by

$$\|\boldsymbol{x}\|_{\mathbf{A}} = \sqrt{(\boldsymbol{x}|\boldsymbol{x})_{\mathbf{A}}} = \sqrt{\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x}},$$

we will - whenever it does not lead to confusion - use the terms $(\boldsymbol{x}|\boldsymbol{y})$ and $\|\boldsymbol{x}\|$.

We note that the set of points with constant $\mathbf{A}$ -norm equal to 1 is the set

$$\{\boldsymbol{x}|\ \|\boldsymbol{x}\|^2 = 1\} = \{\boldsymbol{x}|\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = 1\},$$

i.e. the points on an ellipsoid.

Conversely, to any non-degenerate ellipsoid there is a corresponding positive definite matrix $\mathbf{A}$, so

$$E = \{\boldsymbol{x}|\boldsymbol{x}'\mathbf{A}\,\boldsymbol{x} = 1\} = \{\boldsymbol{x}|\ \|\boldsymbol{x}\|^2_{\mathbf{A}} = 1\}.$$

In this way we have brought about a connection between the set of possible inner products and the set of ellipsoids.

Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal (with respect to $\mathbf{A}$), if

$$\boldsymbol{x}'\mathbf{A}\,\boldsymbol{y} = 0,$$

i.e. if $\boldsymbol{x}$ and $\boldsymbol{y}$ are conjugate directions in the ellipsoid corresponding to $\mathbf{A}$.

It is also possible to introduce a concept of angle by means of the definition

$$\cos(\angle \boldsymbol{a}, \boldsymbol{b}) = \frac{(\boldsymbol{a}|\boldsymbol{b})}{\|\boldsymbol{a}\|\;\|\boldsymbol{b}\|}.$$

We now give a lemma which we will need for the theorems of independence of projections of normally distributed stochastic variables.

**LEMMA 1.1.** Let $R^n$ be partitioned in a direct sum

$$R^n = U_1 \oplus \cdots \oplus U_k$$

of $n_i$ dimensional sub-spaces. These are orthogonal w.r.t. the positive definite matrix $\boldsymbol{\Sigma}^{-1}$, i.e.

$$\boldsymbol{x} \perp \boldsymbol{y} \Leftrightarrow \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = 0.$$

For $i = 1, \ldots, k$ we let the projection $p_i$ onto $U_i$ be given by the matrix $\mathbf{C}_i$. Then

$$\mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{C}_j' = 0$$

for all $i \neq j$. Furtmermore, we have

$$\boldsymbol{\Sigma}^{-1}\mathbf{C}_i = \mathbf{C}_i'\boldsymbol{\Sigma}^{-1} = \mathbf{C}_i'\boldsymbol{\Sigma}\,\mathbf{C}_i.$$

**PROOF 1.26.** Since $p_i \circ p_i = p_i$, we have

$$\mathbf{C}_i\mathbf{C}_i = \mathbf{C}_i,$$

and since

$$p_i(\boldsymbol{x}) \perp \boldsymbol{x} - p_i(\boldsymbol{x}),$$

(cf. the illustration) we have

$$p_i(\boldsymbol{x})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - p_i(\boldsymbol{x})) = 0,$$

i.e.

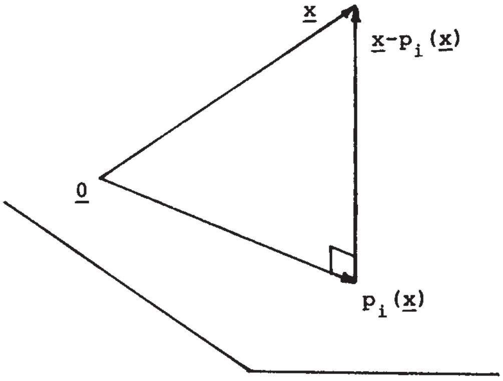
$$\boldsymbol{x}\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}[\boldsymbol{x} - \mathbf{C}_i\boldsymbol{x}] = 0.$$

This holds for all $\boldsymbol{x}$, and therefore

$$\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}(\mathbf{I} - \mathbf{C}_i) = \mathbf{0},$$

or

$$\mathbf{C}_i'\boldsymbol{\Sigma}^{-1} = \mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i.$$



The right hand side of the equation is obviously symmetrical, so that

$$\mathbf{C}_i'\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}\mathbf{C}_i.$$

By pre- and post-multiplication with $\boldsymbol{\Sigma}$ we get

$$\boldsymbol{\Sigma}\,\mathbf{C}_i' = \mathbf{C}_i\boldsymbol{\Sigma},$$

so

$$\mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{C}_i' = \mathbf{C}_i\mathbf{C}_i\boldsymbol{\Sigma} = \mathbf{C}_i\boldsymbol{\Sigma}.$$

This gives

$$\mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{C}_j' = \mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{C}_i'\mathbf{C}_j' = \mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{0} = \mathbf{0}.$$

The second-last equal sign follows from the fact that the sum is direct, so for all $\boldsymbol{x}$ it holds that

$$p_j(p_i(\boldsymbol{x})) = \mathbf{0},$$

i.e.

$$\mathbf{C}_j\mathbf{C}_i\boldsymbol{x} = \mathbf{0}.$$

Since $\boldsymbol{x}$ - as was mentioned previously - is arbitrary, then this implies

$$\mathbf{C}_j\mathbf{C}_i = \mathbf{0},$$

or

$$\mathbf{C}_i'\mathbf{C}_j' = \mathbf{0}.$$

■

# Chapter 2

# Multidimensional variables

In this chapter we start by supplementing the results on multidimensional stochastic variables, given in chapter 0, volume 1. Then we discuss the multivariate normal distribution and distributions derived from it. Finally we shortly describe the special considerations that estimation and testing give rise to.

## 2.1 Moments of multidimensional stochastic variables

We start with

### 2.1.1 The mean value

Let there be given a stochastic matrix, i.e. a matrix, where the single elements are stochastic variables:

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kn} \end{bmatrix}$$

We then define the mean value, or the expectation value, or the expected value of $\mathbf{X}$ as

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix} \mathrm{E}(X_{11}) & \cdots & \mathrm{E}(X_{1n}) \\ \vdots & & \vdots \\ \mathrm{E}(X_{k1}) & \cdots & \mathrm{E}(X_{kn}) \end{bmatrix} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \vdots & & \vdots \\ \mu_{k1} & \cdots & \mu_{kn} \end{bmatrix} = \mu.$$

**THEOREM 2.1.** Let $\mathbf{A}$ be a $k \times n$ matrix of constants. Then

$$\mathrm{E}(\mathbf{A} + \mathbf{X}) = \mathbf{A} + \mathrm{E}(\mathbf{X}).$$

This theorem follows trivially from the definition as does the following. ▲

**THEOREM 2.2.** Let $\mathbf{A}$ and $\mathbf{B}$ be constant stochastic matrices, so that $\mathbf{A}\,\mathbf{x}$ and $\mathbf{x}\,\mathbf{B}$ exist. Then

$$\begin{aligned} \mathrm{E}(\mathbf{A}\,\mathbf{X}) &= \mathbf{A}\;\mathrm{E}(\mathbf{X}) \\ \mathrm{E}(\mathbf{X}\,\mathbf{B}) &= \mathrm{E}(\mathbf{X})\mathbf{B} \end{aligned}$$

▲

Finally we have

**THEOREM 2.3.** Let $\mathbf{X}$ and $\mathbf{Y}$ be stochastic matrices of the same rank. Then

$$\mathrm{E}(\mathbf{X} + \mathbf{Y}) = \mathrm{E}(\mathbf{X}) + \mathrm{E}(\mathbf{Y}).$$

▲

**REMARK 2.1.** We have not mentioned that we of course assume, that the involved expected values exist. This is assumed here and in all the following, where these are mentioned.

▼

### 2.1.2 The variance-covariance matrix (dispersion matrix).

The generalisation of the variance of a stochastic variable is the variance-covariance matrix (or dispersion matrix) for a stochastic vector $\boldsymbol{X}$. It is defined by

$$\mathrm{D}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \mathrm{E}\{(\boldsymbol{X} - \boldsymbol{\mu})\,(\boldsymbol{X} - \boldsymbol{\mu})'\},$$

where

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}).$$

It should be noted, that $\mathrm{D}(\boldsymbol{X})$ also often is called the covariance-matrix and is then denoted $\mathrm{Cov}(\boldsymbol{X})$. However, this is a bit misleading, since it could misunderstood as the covariance between two (multidimensional) stochastic variables. Another commonly used notation is $\mathrm{V}(\boldsymbol{X})$. Furthermore, we note that

$$(\boldsymbol{X}-\boldsymbol{\mu})\,(\boldsymbol{X}-\boldsymbol{\mu})' = \begin{bmatrix} X_1-\mu_1 \\ \vdots \\ X_n-\mu_n \end{bmatrix} (X_1-\mu_1,\ldots,X_n-\mu_n) =$$

$$\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_n-\mu_n) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n-\mu_n)(X_1-\mu_1) & (X_n-\mu_n)(X_2-\mu_2) & \cdots & (X_n-\mu_n)^2 \end{bmatrix}$$

i.e. the variance-covariance matrix's $(i,j)$'th element is $\mathrm{Cov}(X_i,X_j)$, or

$$\boldsymbol{\Sigma} = \mathrm{D}(\boldsymbol{X}) = \begin{bmatrix} \mathrm{V}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{V}(X_2) & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{V}(X_n) \end{bmatrix}.$$

We will often use the following notation

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix},$$

i.e. the variances can be denoted both as $\sigma_i^2$ and as $\sigma_{ii}$. We note, that $\boldsymbol{\Sigma}$ is symmetric. More interesting is the following

**THEOREM 2.4.** The variance-covariance matrix $\boldsymbol{\Sigma}$ for a stochastic vector (i.e. a multidimensional stochastic vector) is positive semidefinite. ▲

**PROOF 2.1.** For any vector $\boldsymbol{y}$ we have

$$\begin{aligned} \boldsymbol{y}'\,\boldsymbol{\Sigma}\,\boldsymbol{y} &= \boldsymbol{y}'\,\mathrm{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\}\boldsymbol{y} \\ &= \mathrm{E}\{\boldsymbol{y}'\,(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{y}\} \\ &= \mathrm{E}\{\,[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{y}]'[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{y}]\,\} \\ &\geq 0\,, \end{aligned}$$

since the expression in the curly brackets is $\geq 0$. ■

Theorems exist which are analogous to the ones known from the one dimensional stochastic variables.

**THEOREM 2.5.** Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent. Then

$$\mathrm{D}(\boldsymbol{X}+\boldsymbol{Y}) = \mathrm{D}(\boldsymbol{X}) + \mathrm{D}(\boldsymbol{Y}).$$

Let $\boldsymbol{b}$ be a constant. Then we have

$$\mathrm{D}(\boldsymbol{b}+\boldsymbol{X}) = \mathrm{D}(\boldsymbol{X}).$$

If $\mathbf{A}$ is a constant matrix, so that $\mathbf{A}\,\boldsymbol{X}$ exists, then the following holds

$$\mathrm{D}(\mathbf{A}\,\boldsymbol{X}) = \mathbf{A}\ \mathrm{D}(\boldsymbol{X})\mathbf{A}'.$$

▲

**PROOF 2.2.** The first relation comes from

$$\begin{aligned}
\mathrm{Cov}(X_i+Y_i, X_j+Y_j) &= \mathrm{Cov}(X_i,X_j) + \mathrm{Cov}(X_i,Y_j) + \\
&\quad \mathrm{Cov}(Y_i,X_j) + \mathrm{Cov}(Y_i,Y_j) \\
&= \mathrm{Cov}(X_i,X_j) + \mathrm{Cov}(Y_i,Y_j),
\end{aligned}$$

since $\mathrm{Cov}(Y_i,X_j)=0,$ because $X_j$ and $Y_i$ are independent. The second relation is trivial. The last one comes from

$$\begin{aligned}
\mathrm{D}(\mathbf{A}\,\boldsymbol{X}) &= \mathrm{E}\{(\mathbf{A}\,\boldsymbol{X} - \mathbf{A}\,\boldsymbol{\mu})(\mathbf{A}\,\boldsymbol{X} - \mathbf{A}\,\boldsymbol{\mu})'\} \\
&= \mathrm{E}\{\mathbf{A}[\boldsymbol{X}-\boldsymbol{\mu}][\boldsymbol{X}-\boldsymbol{\mu}]'\mathbf{A}'\} \\
&= \mathbf{A}\,\mathrm{E}\{[\boldsymbol{X}-\boldsymbol{\mu}][\boldsymbol{X}-\boldsymbol{\mu}]'\}\mathbf{A}' \\
&= \mathbf{A}\,\mathrm{D}(\boldsymbol{X})\mathbf{A}' \\
&= \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}'
\end{aligned}$$

■

If we let

$$\mathbf{V} = \mathrm{diag}\left(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_n}\right) = \begin{bmatrix} \sigma_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^{-1} \end{bmatrix}$$

and we "scale" $\boldsymbol{X}$ by $\mathbf{V}$, we get

$$\mathrm{D}(\mathbf{V}\,\boldsymbol{X}) = \mathbf{V}\,\boldsymbol{\Sigma}\,\mathbf{V}' = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma_1\,\sigma_2} & \cdots & \frac{\sigma_{1n}}{\sigma_1\,\sigma_n} \\ \frac{\sigma_{12}}{\sigma_1\,\sigma_2} & 1 & \cdots & \frac{\sigma_{2n}}{\sigma_2\,\sigma_n} \\ \vdots & \vdots & & \vdots \\ \frac{\sigma_{1n}}{\sigma_1\,\sigma_n} & \frac{\sigma_{2n}}{\sigma_2\,\sigma_n} & \cdots & 1 \end{bmatrix}.$$

We note, that the elements are the correlation coefficients between $\boldsymbol{X}$'s components, which is why this matrix is also called the correlation matrix for $\boldsymbol{X}$, and we write

$$R(\boldsymbol{X}) = \begin{bmatrix} 1 & \cdots & \rho_{1n} \\ \vdots & & \vdots \\ \rho_{1n} & \cdots & 1 \end{bmatrix},$$

where

$$\rho_{ij} = \mathrm{Cor}(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{V}(X_i)\,\mathrm{V}(X_j)}}.$$

### 2.1.3 Covariance

Let there be given two stochastic variables

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \quad \text{and} \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix}$$

with mean values $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. We now define the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ as

$$\mathrm{C}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\nu})'] = \begin{bmatrix} \mathrm{Cov}(X_1, Y_1) & \cdots & \mathrm{Cov}(X_1, Y_q) \\ \vdots & & \vdots \\ \mathrm{Cov}(X_p, Y_1) & \cdots & \mathrm{Cov}(X_p, Y_q) \end{bmatrix}.$$

Then

$$\mathrm{C}(\boldsymbol{X}, \boldsymbol{X}) = \mathrm{D}(\boldsymbol{X})$$

and

$$\mathrm{C}(\boldsymbol{X}, \boldsymbol{Y}) = [\mathrm{C}(\boldsymbol{Y}, \boldsymbol{X})]'.$$

Less trivial is

**THEOREM 2.6.** Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be as above, and let $\mathbf{A}$ and $\mathbf{B}$ be $n \times p$ and $m \times q$ matrices of constants respectively. Then

$$\mathrm{C}(\mathbf{A}\,\boldsymbol{X}, \mathbf{B}\,\boldsymbol{Y}) = \mathbf{A}\,\mathrm{C}(\boldsymbol{X}, \boldsymbol{Y})\mathbf{B}'.$$

If $\boldsymbol{U}$ is a $p$-dimensional and $\boldsymbol{V}$ is a $q$-dimensional stochastic variable the following holds

$$\mathrm{C}(\boldsymbol{X} + \boldsymbol{U}, \boldsymbol{Y}) = \mathrm{C}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{C}(\boldsymbol{U}, \boldsymbol{Y})$$

$$\mathrm{C}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{V}) = \mathrm{C}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{C}(\boldsymbol{X}, \boldsymbol{V}).$$

Finally

$$\mathrm{D}(\boldsymbol{X} + \boldsymbol{U}) = \mathrm{D}(\boldsymbol{X}) + \mathrm{D}(\boldsymbol{U}) + \mathrm{C}(\boldsymbol{X}, \boldsymbol{U}) + \mathrm{C}(\boldsymbol{U}, \boldsymbol{X}).$$

▲

**PROOF 2.3.** According to the definition we have

$$\begin{aligned}
\mathrm{C}(\mathbf{A}\,\boldsymbol{X}, \mathbf{B}\,\boldsymbol{Y}) &= \mathrm{E}[(\mathbf{A}\,\boldsymbol{X} - \mathbf{A}\,\boldsymbol{\mu})(\mathbf{B}\,\boldsymbol{Y} - \mathbf{B}\,\boldsymbol{\nu})'] \\
&= \mathrm{E}[\mathbf{A}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\nu})'\mathbf{B}'] \\
&= \mathbf{A}\,\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\nu})']\mathbf{B}' \\
&= \mathbf{A}\,\mathrm{C}(\boldsymbol{X}, \boldsymbol{Y})\mathbf{B}'.
\end{aligned}$$

This proves the first statement. Similarly - if we let $E(\boldsymbol{U}) = \boldsymbol{\delta}$ -

$$\begin{aligned}
\mathrm{C}(\boldsymbol{X} + \boldsymbol{U}, \boldsymbol{Y}) &= \mathrm{E}[(\boldsymbol{X} + \boldsymbol{U} - \boldsymbol{\mu} - \boldsymbol{\delta})(\boldsymbol{Y} - \boldsymbol{\nu})'] \\
&= \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\nu})' + (\boldsymbol{U} - \boldsymbol{\delta})(\boldsymbol{Y} - \boldsymbol{\nu})'] \\
&= \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\nu})'] + \mathrm{E}[(\boldsymbol{U} - \boldsymbol{\delta})(\boldsymbol{Y} - \boldsymbol{\nu})'] \\
&= \mathrm{C}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{C}(\boldsymbol{U}, \boldsymbol{Y}),
\end{aligned}$$

and the corresponding relation with $\boldsymbol{Y} + \boldsymbol{V}$ is shown analogously. Finally we have

$$\begin{aligned}
\mathrm{D}(\boldsymbol{X} + \boldsymbol{U}) &= \mathrm{C}(\boldsymbol{X} + \boldsymbol{U}, \boldsymbol{X} + \boldsymbol{U}) \\
&= \mathrm{C}(\boldsymbol{X}, \boldsymbol{X}) + \mathrm{C}(\boldsymbol{X}, \boldsymbol{U}) + \mathrm{C}(\boldsymbol{U}, \boldsymbol{X}) + \mathrm{C}(\boldsymbol{U}, \boldsymbol{U}).
\end{aligned}$$

■

If $C(\boldsymbol{X},\boldsymbol{Y}) = \mathbf{0}$ then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be uncorrelated. This corresponds to all components of $\boldsymbol{X}$ being uncorrelated with all components of $\boldsymbol{Y}$.

Later, when we consider the multidimensional general linear model we will need the following

**THEOREM 2.7.** Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent, $p$-dimensional stochastic variables with the same variance-covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. We let

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{X}'_1 \\ \vdots \\ \boldsymbol{X}'_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{p1} \\ \vdots & & \vdots \\ X_{1n} & \cdots & X_{pn} \end{bmatrix}$$

(Note, that the variable index is the first index and the repetition index is the second). If we define

$$\text{vc}(\boldsymbol{X}) = \begin{bmatrix} X_{11} \\ \vdots \\ X_{1n} \\ \vdots \\ X_{p1} \\ \vdots \\ X_{pn} \end{bmatrix}$$

i.e. as the vector consisting of the columns in $\mathbf{X}$ (vc = vector of columns) we get

$$\text{D}(\text{vc}(\mathbf{X})) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the identity matrix of n'th order. ▲

**PROOF 2.4.** Follows trivially from the definition of a tensor-product and from the definition of the variance-covariance matrix.

■

## 2.2 The multivariate normal distribution

The multivariate normal distribution plays the same important role in the theory of multidimensional variables, as the normal distribution does in the univariate case. We start with

### 2.2.1 Definition and simple properties

Let $X_1, \ldots, X_p$ be mutually independent, N(0,1) distributed variables. We then say that

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix},$$

are standardised (normed) $p$-dimensionally normally distributed, and we write

$$\boldsymbol{X} \in \mathbf{N}(\mathbf{0}, \mathbf{I}) = \mathrm{N}_p(\mathbf{0}, \mathbf{I}),$$

where the last notation is used, if there is any doubt about the dimension.We note, that

$$\mathrm{E}(\boldsymbol{X}) = \mathbf{0}, \quad \mathrm{D}(\boldsymbol{X}) = \mathbf{I}.$$

We define the multivariate normal distribution with general parameters in

**DEFINITION 2.1.** We say that the $p$-dimensional stochastic variable $\boldsymbol{X}$ is normally distributed with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, if $\boldsymbol{X}$ has the same distribution as

$$\boldsymbol{\mu} + \mathbf{A}\,\boldsymbol{U},$$

where $\mathbf{A}$ satisfies

$$\mathbf{A}\,\mathbf{A}' = \boldsymbol{\Sigma},$$

and where $\boldsymbol{U}$ is standardised $p$-dimensional normally distributed. We write

$$\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where the last notation again is used, if there is any doubt about the dimension. ▲

**REMARK 2.2.** The definition is only valid, if one shows, that $\mathbf{A}\,\mathbf{A}' = \mathbf{B}\,\mathbf{B}'$ implies

$$\pounds(\boldsymbol{\mu} + \mathbf{A}\,\boldsymbol{U}) = \pounds(\boldsymbol{\mu} + \mathbf{B}\,\boldsymbol{V}),$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are standardised normally distributed and not necessarily of the same dimension. The relation is valid, but we will not pursue this further here. From theorem 1.10 follows that for any positive semidefinite matrix $\boldsymbol{\Sigma}$ there exists a matrix $\mathbf{A}$

with $\mathbf{A}\,\mathbf{A}' = \mathbf{\Sigma}$, so the expression $\mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ makes sense for any positively semidefinite $p \times p$ matrix $\mathbf{\Sigma}$ and any $p$-dimensional vector $\boldsymbol{\mu}$.

Trivially, we note that

$$\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \quad \Rightarrow \quad \text{i) } \mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu} \quad \wedge \quad \text{ii) } \mathrm{D}(\boldsymbol{X}) = \mathbf{\Sigma}$$

i.e. the distribution is parametrised by its mean and variance-covariance matrix. ▼

If $\mathbf{\Sigma}$ has full rank, then the distribution has the density given in

**THEOREM 2.8.** Let $\boldsymbol{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, and let $\mathrm{rg}(\mathbf{\Sigma}) = p$. Then $\boldsymbol{X}$ has the density

$$\begin{aligned} \mathrm{f}(\boldsymbol{x}) &= \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \mathbf{\Sigma}}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \\ &= \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \mathbf{\Sigma}}} \exp[-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2], \end{aligned}$$

where the norm used is the one defined by $\mathbf{\Sigma}^{-1}$ , see p. 53. ▲

**PROOF 2.5.** Let $\mathbf{U} \in \mathrm{N}_p(\mathbf{0}, \mathbf{I})$. Then $\boldsymbol{U}$ has the density

$$\begin{aligned} \mathrm{h}(\boldsymbol{u}) &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u_i^2) = \frac{1}{\sqrt{2\pi}^p} \exp(-\frac{1}{2}\sum_{i=1}^{p} u_i^2) \\ &= \frac{1}{\sqrt{2\pi}^p} \exp(-\frac{1}{2}\boldsymbol{u}'\,\boldsymbol{u}). \end{aligned}$$

We then consider the transformation from $R^p \to R^p$ given by

$$\boldsymbol{u} \to \boldsymbol{x} = \boldsymbol{\mu} + \mathbf{A}\,\boldsymbol{u}$$

where $\mathbf{A}\,\mathbf{A}' = \mathbf{\Sigma}$. From theorem 1.14 it follows that $\mathbf{A}$ is regular. We obtain

$$\boldsymbol{u} = \mathbf{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}),$$

giving

$$\begin{aligned} \boldsymbol{u}'\boldsymbol{u} &= (\boldsymbol{x}-\boldsymbol{\mu})'\mathbf{A}^{-1\prime}\mathbf{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}). \end{aligned}$$

Furthermore, since

$$\det(\mathbf{\Sigma}) = \det(\mathbf{A}\,\mathbf{A}') = \det(\mathbf{A})^2,$$

i.e.

$$\det(\mathbf{A}^{-1}) = \frac{1}{\sqrt{\det \mathbf{\Sigma}}}$$

the result follows from theorem 0.8 in volume 1. ■

We note that the inverse variance-covariance matrix $\mathbf{\Sigma}^{-1}$ is often called the precision of the normal distribution.

If $\mathbf{\Sigma}$ is not regular, then the distribution is degenerate and has no density. We then introduce the concept of the affine support in

**DEFINITION 2.2.** Let $\boldsymbol{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. By the (affine) support for $\boldsymbol{X}$ we mean the smallest (side-) sub-space of $R^p$, where $\boldsymbol{X}$ is defined with probability 1. ▲

**REMARK 2.3.** If we restrict the considerations to the affine support, then $\boldsymbol{X}$ is regularly distributed and has a density as shown in theorem 2.8. ▼

We have different possibilities of determining the support of a $p$-dimensional normal distribution. Firstly

**THEOREM 2.9.** Let $\boldsymbol{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, and let $\mathbf{A}$ be an $p \times m$ matrix, so that $\mathbf{A}\,\mathbf{A}' = \mathbf{\Sigma}$. We then let V equal $\mathbf{A}$'s projection-space, i.e.

$$V = \{\boldsymbol{v} \in R^p | \exists \boldsymbol{u} \in R^m : \boldsymbol{v} = \mathbf{A}\,\boldsymbol{u}\}.$$

Then the (affine) support for $\boldsymbol{X}$ is the (side-) sub-space

$$\boldsymbol{\mu} + V = \{\boldsymbol{\mu} + \boldsymbol{v} | \boldsymbol{v} \in V\}.$$

▲

**PROOF 2.6.** Omitted. ■

Further, we have

**THEOREM 2.10.** Let $\boldsymbol{X}$ be as in the previous theorem. Then the subspace $V$ equals the direct sum of the eigen-spaces corresponding to those eigenvalues in $\boldsymbol{\Sigma}$ which are different from 0. ▲

**PROOF 2.7.** Omitted. ■

Finally we have

**THEOREM 2.11.** Let $\boldsymbol{X}$ be as in the previous theorems. Then the subspace V equals the orthogonal complement to the null-space for $\boldsymbol{\Sigma}$, i.e.

$$V = \{\boldsymbol{v} | \boldsymbol{\Sigma}\, \boldsymbol{v} = \mathbf{0}\}^{\perp}$$

▲

**PROOF 2.8.** Omitted. ■

The three theorems are illustrated in

**EXAMPLE 2.1.** We consider

$$\boldsymbol{X} \in \mathrm{N}\left( \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix} \right) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Since

$$\det\left( \begin{bmatrix} 1 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix} \right) = 0,$$

then $\boldsymbol{X}$ is singularly distributed, and we will determine the affine support.

We first seek a matrix $\mathbf{A}$, so $\mathbf{A}\,\mathbf{A}' = \boldsymbol{\Sigma}$. To do that we first determine $\boldsymbol{\Sigma}$'s eigenvalues and (normed) eigenvectors. These are

$$\begin{aligned}\lambda_1 = 9 \quad &\wedge \quad \boldsymbol{p}_1 = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}, \\ \lambda_2 = 2 \quad &\wedge \quad \boldsymbol{p}_2 = \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}, \\ \lambda_3 = 0 \quad &\wedge \quad \boldsymbol{p}_3 = \begin{bmatrix} \frac{2\sqrt{2}}{3} \\ -\frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \end{bmatrix}.\end{aligned}$$

It now follows that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \\ \frac{2}{3} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \end{bmatrix} \begin{bmatrix} 9 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{2\sqrt{2}}{3} & -\frac{\sqrt{2}}{6} & -\frac{\sqrt{2}}{6} \end{bmatrix}$$

From this we see that we as $\mathbf{A}$-matrix can choose

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & -1 & 0 \end{bmatrix} \quad (= \begin{bmatrix} \frac{1}{3} & 0 & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \\ \frac{2}{3} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \end{bmatrix} \begin{bmatrix} \sqrt{9} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}).$$

If we regard $\mathbf{A}$ as the matrix for a linear projection $R^3 \rightarrow R^3$ we then obtain that the projection-space is

$$\begin{aligned} V &= \{\mathbf{A}\,\boldsymbol{u} | \boldsymbol{u} \in R^3\} \\ &= \{u_1 \boldsymbol{p}_1 + u_2 \boldsymbol{p}_2 | u_1 \in R \wedge u_2 \in R\}. \end{aligned}$$

It is immediately noted that this is also the direct sum of the eigen-spaces corresponding to the eigenvalues which are different from 0.

The null-space for $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}\,\boldsymbol{u} = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{u} = t \cdot \boldsymbol{p}_3.$$

This again gives the same description of V.

The affine support for $\boldsymbol{Y}$ is then the (side-) sub-space

$$\boldsymbol{\mu} + V = \left\{ \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} + u_1 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix} | u_1, u_2 \in R \right\}.$$

♦

**REMARK 2.4.** From the example the proofs of theorems 2.9-2.11 can nearly be deduced completely. ▼

We now formulate a trivial but useful theorem.

**THEOREM 2.12.** Let $\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$\mathbf{A}\,\boldsymbol{X} + \boldsymbol{b} \in \mathrm{N}(\mathbf{A}\,\boldsymbol{\mu} + \boldsymbol{b},\ \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}'),$$

where we implicitly require that the implied matrix-products etc. exist.

▲

**PROOF 2.9.** Trivial from the definition. ■

## 2.2.2 Independence and contour ellipsoids.

In this section we will give the conditions for independence of the normally distributed stochastic variables, and we will prove that the isocurves for the density functions are ellipsoids. First we have

**THEOREM 2.13.** Let

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \in \mathrm{N}\left( \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Then

$$\boldsymbol{X}_i \in \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii}),$$

and

$$\boldsymbol{X}_1, \boldsymbol{X}_2 \text{ are stochastically independent } \quad \Leftrightarrow \quad \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}' = \mathbf{0},$$

where $\mathbf{0}$ is the null matrix. ▲

**PROOF 2.10.** The first statement follows from the previous theorem. The second follows by proving that the condition $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ assures, that the distribution becomes a product distribution. ■

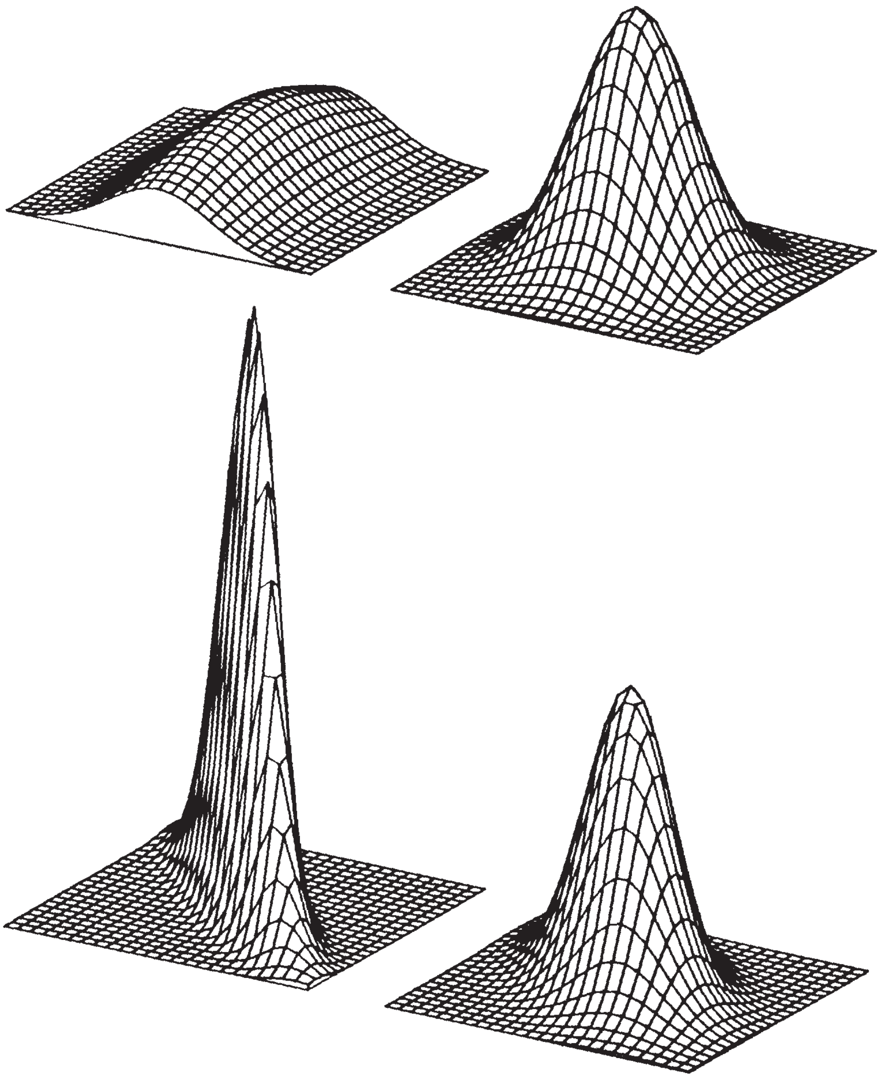


Figure 2.1: Density functions for two-dimensional normal distributions with the variance-covariance matrices
$\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$.

From the theorem follows that the components in a vector $\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are stochastically independent if $\boldsymbol{\Sigma}$ is a diagonal matrix. We will now show that independence is just a question of choosing a suitable coordinate-system.

Let $\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\boldsymbol{\Sigma}$ have the ortho-normed eigenvectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$. We now consider a coordinate system, with origo in $\boldsymbol{\mu}$ and the vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ as base-vectors. The coordinates in this system are called $\boldsymbol{y}$.

If we let

$$\mathbf{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n),$$

we have the following correspondence between the original coordinates $\boldsymbol{x}$ and the new coordinates $\boldsymbol{y}$ for any point $\in R^n$.

$$\boldsymbol{y} = \mathbf{P}'(\boldsymbol{x} - \boldsymbol{\mu}) \quad \Leftrightarrow \quad \boldsymbol{x} = \mathbf{P}\,\boldsymbol{y} + \boldsymbol{\mu},$$

cf. p. 12.

Note: The above relation is a relation between coordinates for a fixed vector viewed in two coordinate-systems.

Using this, if we let $\boldsymbol{Y}$ be the new coordinates for $\boldsymbol{X}$ we have

**THEOREM 2.14.** Let $\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\boldsymbol{Y}$ be as above. Then

$$\boldsymbol{Y} \in \mathrm{N}(\mathbf{0}, \boldsymbol{\Lambda}),$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\Sigma}$'s eigenvalues on the diagonal. ▲

**PROOF 2.11.** Follows from theorem 2.12 and theorem 1.10. ■

**REMARK 2.5.** By translating and rotating (or reflection of) the original coordinate-system we have obtained, that the variance-covariance matrix is a diagonal matrix. I.e. that the components in the stochastic vector are uncorrelated and thereby also independent. ▼

By rescaling the axes we can even obtain that the variance-covariance matrix has zeros or ones on the diagonal. Considering the base-vectors

$$c_1\boldsymbol{p}_1, \ldots, c_n\boldsymbol{p}_n,$$

where

$$c_i = \begin{cases} \frac{1}{\sqrt{\lambda_i}} & \text{if } \lambda_i > 0 \\ 1 & \text{if } \lambda_i = 0 \end{cases} ,$$

cf. p. 31, and calling the coordinates in this system $\boldsymbol{z},$ we get the equation

$$\boldsymbol{z} = \mathbf{C}'\mathbf{P}'(\boldsymbol{x} - \boldsymbol{\mu}) = (\mathbf{P}\,\mathbf{C})'(\boldsymbol{x} - \boldsymbol{\mu}),$$

where $\mathbf{C} = \mathrm{diag}(c_1, \ldots, c_n)$.

If we let the $\boldsymbol{z}$ -coordinates for $\boldsymbol{X}$ equal $\boldsymbol{Z}$ we get

$$\boldsymbol{Z} = \mathrm{N}(\mathbf{0}, \mathbf{E}),$$

where

$$\mathbf{E} = (\mathbf{P}\,\mathbf{C})'\boldsymbol{\Sigma}\,\mathbf{P}\,\mathbf{C} = \mathbf{C}'\mathbf{P}'\boldsymbol{\Sigma}\,\mathbf{P}\,\mathbf{C} = \mathbf{C}'\boldsymbol{\Lambda}\,\mathbf{C}$$

has zeros or ones on the diagonal.

The transformation into the new bases is closely related to the isocurves for the density function for the normal distribution.

As mentioned earlier the density for an $\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\begin{aligned} \mathrm{f}(\boldsymbol{x}) &= k \cdot \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})) \\ &= k \cdot \exp(-\frac{1}{2}(\|\boldsymbol{x} - \boldsymbol{\mu}\|)^2). \end{aligned}$$

Therefore we have

$$\mathrm{f}(\boldsymbol{x}) = k_1 \quad \Leftrightarrow \quad (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c,$$

where $k_1$ and $c$ are constants. Since $\boldsymbol{\Sigma}^{-1},$ is positive definite the isocurves

$$E_c = \{\boldsymbol{x} | \mathrm{f}(\boldsymbol{x}) = k_1\}$$

will be ellipsoids, cf. p. 40. From theorem 1.19 is also seen that the major axes in these ellipsoids are the eigenvectors for $\boldsymbol{\Sigma}^{-1},$ but from theorem 1.12 we note that they are also eigenvectors for $\boldsymbol{\Sigma}.$ In the new coordinates the densities become

$$\mathrm{g}(\boldsymbol{y}) = k \cdot \exp(-\frac{1}{2}\Sigma\frac{1}{\lambda_i}y_i^2),$$

where $\lambda_i$ is the $i$'th eigenvalue for $\mathbf{\Sigma}$, and

$$\mathrm{h}(\boldsymbol{z}) = k_1 \cdot \exp(-\frac{1}{2}\Sigma z_i^2).$$

The ellipsoids $E_i$ are often called contour-ellipsoids. From the above we get

**THEOREM 2.15.** Let $\mathbf{P}$ and $\mathbf{C}$ be as above. Then

$$(\boldsymbol{X}-\boldsymbol{\mu})'(\mathbf{P}\,\mathbf{C})(\mathbf{P}\,\mathbf{C})'(\boldsymbol{X}-\boldsymbol{\mu}) \in \chi^2(\mathrm{rg}\,\mathbf{\Sigma}).$$

If $\mathbf{\Sigma}$ has full rank $p$ then

$$(\boldsymbol{X}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \|\boldsymbol{X}-\boldsymbol{\mu}\|^2 \in \chi^2(p).$$

▲

**PROOF 2.12.** $$(\boldsymbol{X}-\boldsymbol{\mu})'(\mathbf{P}\,\mathbf{C})(\mathbf{P}\,\mathbf{C})'(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{Z}'\boldsymbol{Z} = \Sigma\delta_i Z_i^2,$$

where $\delta_i = 1$ if $\lambda_i \neq 0$ and equal to 0 otherwise.

Since the non-degenerate components in $\boldsymbol{Z}$ are stochastically independent and N(0,1)-distributed the result follows immediately. The last remark comes from

$$\mathbf{P}\,\mathbf{C}(\mathbf{P}\,\mathbf{C})' = \mathbf{P}\,\mathbf{C}\,\mathbf{C}'\mathbf{P}' = \mathbf{P}\,\mathbf{\Lambda}^{-1}\mathbf{P}' = \mathbf{\Sigma}^{-1}$$

■

**REMARK 2.6.** The result of the theorem is that the probability of an outcome being within the contour ellipsoid can be computed using a $\chi^2$-distribution. ▼

Examples of these concepts will be given in example 2.3, where we consider the two-dimensional normal distribution.

## 2.2.3 Conditional distributions

In this section we consider the partitioning of a stochastic variable $\boldsymbol{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, into

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}; \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}; \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

We then have

**THEOREM 2.16.** If $\boldsymbol{X}_2$ is regularly distributed, i.e. if $\boldsymbol{\Sigma}_{22}$ has full rank, then the distribution of $\boldsymbol{X}_1$ conditioned on $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is again a normal distribution, and the following holds

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\ \mathrm{D}(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

If $\boldsymbol{\Sigma}_{22}$ does not have full rank then the conditional distribution is still normal and $\boldsymbol{\Sigma}_{22}^{-1}$ in the above equations should be substituted by a generalised inverse $\boldsymbol{\Sigma}_{22}^{-}$. ▲

**PROOF 2.13.** The proof is technical and is omitted, however cf. section 2.2.5. ■

**REMARK 2.7.** It is seen that the conditional variance is independent of $\boldsymbol{x}_2$. This result is not valid for all distributions, but is special for the normal distribution. Also we see the conditional mean is an affine function of $\boldsymbol{x}_2$, cf. the discussion in section 2.3.3. ▼

We will not discuss the implications of the theorem here. Instead we refer to the examples in section 2.2.5.

### 2.2.4 Theorem of reproductivity and the central limit theorem.

Analogous to the theorem of reproductivity for the univariate normal distribution we have

**THEOREM 2.17.** (Theorem of reproductivity). Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ be independent, and let $\boldsymbol{X}_i \in \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

Then

$$\sum_{i=1}^{k} \boldsymbol{X}_i \in \mathrm{N}\left(\sum_{i=1}^{k} \boldsymbol{\mu}_i, \sum_{i=1}^{k} \boldsymbol{\Sigma}_i\right).$$

▲

**PROOF 2.14.** Omitted. ■

As in the univariate case, central limit theorems exist, i.e. sums of independent multidimensional stochastic variables are under generel assumptions asymptotically normally distributed. We state an analogue to Lindeberg-Levy's theorem.

**THEOREM 2.18.** (Central limit theorem). Let the independent and identically distributed variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \ldots$ have finite first and second moments

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}_i), \boldsymbol{\Sigma} = \mathrm{D}(\boldsymbol{X}_i).$$

Then we have - with $\bar{\boldsymbol{X}}_n = \frac{1}{n}(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n)$ - that

$$\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})$$

has an $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ -distribution as its limiting distribution, and we say that $\bar{\boldsymbol{X}}_n$ is asymptotically $\mathrm{N}(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$ distributed. ▲

**PROOF 2.15.** This and the previous theorem can be proved from the corresponding univariate theorems by first using a theorem, which characterises the multivariate distribution (a multidimensional variable is normally distributed if and only if all linear combinations of its components are (univariate normally distributed); and by using a theorem which characterises a multivariate limiting distribution as limiting distributions of linear combinations of the components (coordinates). However, this is out of the scope of this presentation and the interested reader is referred to the literature e.g. [18], section 2c.5. ■

### 2.2.5 Estimation of the parameters in a multivariate normal distribution.

We consider a number of observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n,$ which are assumed independent and identically $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributed. We assume there are more observations than the dimension indicates, i.e. that $n > p$. In this section we will give estimates of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

We introduce the notation

$$\begin{aligned}
\boldsymbol{X}_i &= \begin{bmatrix} X_{1i} \\ \vdots \\ X_{pi} \end{bmatrix} \\
\bar{\boldsymbol{X}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix} \\
\boldsymbol{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i' - \frac{n}{n-1}\bar{\boldsymbol{X}}\,\bar{\boldsymbol{X}}'.
\end{aligned}$$

If we consider the data-matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{p1} \\ \vdots & & \vdots \\ X_{1n} & \cdots & X_{pn} \end{bmatrix},$$

where the $i$'th row corresponds to the $i$'th observation, we can also write

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

With this we can now state

**THEOREM 2.19.** Let the situation be as stated above. Then the maximum likelihood estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ equal

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \bar{\boldsymbol{X}} \\
\hat{\boldsymbol{\Sigma}} &= \frac{n-1}{n}\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'.
\end{aligned}$$

$\hat{\boldsymbol{\mu}}$ is an unbiased estimate of $\boldsymbol{\mu}$, and $\mathbf{S}$ is an unbiased estimate of $\boldsymbol{\Sigma}$. ▲

**PROOF 2.16.** Proof. Omitted, see e.g. [2], chapter 3. ■

**REMARK 2.8.** Since the empirical variance-covariance matrix $\mathbf{S}$ is an unbiased estimate $\boldsymbol{\Sigma}$, and since it only differs from the maximum likelihood estimator by the factor $\frac{n}{n-1}$, we often prefer $\mathbf{S}$ as the estimate. Often one will see the notation $\hat{\boldsymbol{\Sigma}}$ used for $\mathbf{S}$. One should in each case be aware of what the expression $\hat{\boldsymbol{\Sigma}}$ precisely means.

The distribution of $\hat{\boldsymbol{\mu}}$ comes trivially from theorem 2.2.4. The following holds

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \in \mathrm{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}).$$

The distribution of $\mathbf{S}$ is more complicated. It is stated in section 2.5.

We give an example of estimating the parameters in the following section. ▼

### 2.2.6 The two-dimensional normal distribution.

We now specialise the results from before to two dimensions.

Let $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be normally distributed with $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Since

$$\det(\boldsymbol{\Sigma}) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$$

is, if $\det(\boldsymbol{\Sigma}) \neq 0$,

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix}.$$

Introducing the correlation coefficient $\rho$

$$\rho = \frac{\sigma_{12}}{\sigma_1\,\sigma_2},$$

we get

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix},$$

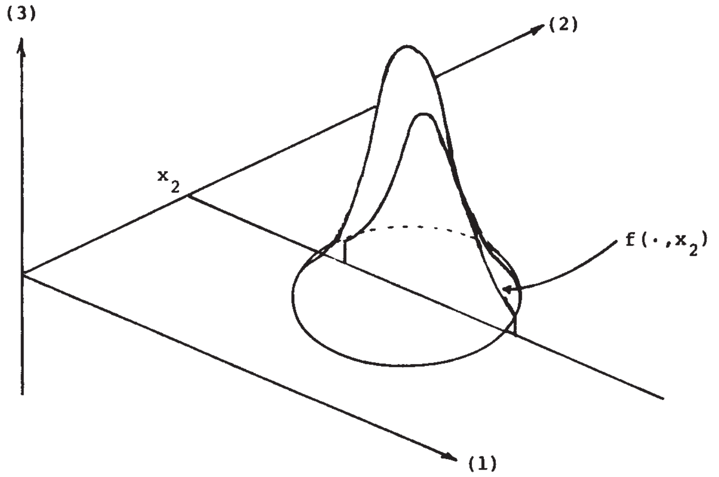


Figure 2.2: The density of a two-dimensional normal distribution.

and the density becomes

$$\begin{aligned}&\mathrm{f}(x_1,x_2) = \\ &\quad \frac{1}{2\pi}\frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2}\frac{1}{1-\rho^2}\left\{\left[\frac{x_1-\mu_1}{\sigma_1}\right]^2\right.\right. \\ &\quad \left.\left.-2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}+\left[\frac{x_2-\mu_2}{\sigma_2}\right]^2\right\}\right].\end{aligned}$$

The graph is shown in fig. 2.2 It is immediately seen that we have a product distribution i.e. that $X_1$ and $X_2$ are stochastically independent, if $\rho = 0,$ i.e. if $\mathbf{\Sigma}$ is a diagonal matrix.

The conditional distribution of $X_1$ conditioned on $X_2 = x_2$ is proportional to the intersecting curve between the plane through $(0, x_2, 0)$ parallel to the (1)-(3) plane. If we denote the density as $\mathrm{g}$ we have

$$\mathrm{g}(\cdot) = c\mathrm{f}(\cdot, x_2),$$

where $c$ is a normalisation constant. We have

$$\begin{aligned}
\mathrm{g}(x_1) &= k_1 \cdot \exp\left[-\frac{1}{2}\frac{1}{1-\rho^2}\left\{\left[\frac{x_1-\mu_1}{\sigma_1}\right]^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}\right\}\right] \\
&= k_2 \cdot \exp\left[-\frac{1}{2}\frac{1}{1-\rho^2}\left[\frac{x_1-\mu_1}{\sigma_1} - \rho\frac{x_2-\mu_2}{\sigma_2}\right]^2\right] \\
&= k_3 \cdot \exp\left[-\frac{1}{2}\frac{1}{\sigma_1^2(1-\rho^2)}\left(x_1-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right)^2\right] \\
&= k_3 \cdot \exp\left[-\frac{1}{2\gamma^2}(x_1-\xi_1)^2\right].
\end{aligned}$$

Note that no bookkeeping has been done with respect to $x_2$. It has disappeared into different constants. From the final result we note that the conditional distribution is normal and that

$$k_3 = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}},$$

and finally that

$$\mathrm{E}(X_1|X_2 = x_2) = \xi_1 = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)$$

and

$$\mathrm{V}(X_1|X_2 = x_2) = \gamma^2 = \sigma_1^2(1-\rho^2).$$

We have shown the result of theorem 2.16 for the case $n = 2$. Note, that the conditional mean depends linearly (or more correctly: affinely) upon $x_2$, and that the conditional variance is independent of $x_2$. Further we have

$$\mathrm{V}(X_1|X_2 = x_2) \leq \mathrm{V}(X_1),$$

and the squared coefficient of correlation represents the reduction in variance. i.e. the fraction of $X_1$'s variance, which can be explained by $X_2$, since

$$\rho^2 = \frac{\mathrm{V}(X_1) - \mathrm{V}(X_1|X_2 = x_2)}{\mathrm{V}(X_1)}.$$

In the following example we consider a numerical example which also involves an estimation problem.

**EXAMPLE 2.2.** In the following table corresponding values of the air's content of flying dust measured in $\frac{\mu g}{m^3}$. is shown. Two different measuring principles were used, a measure of grey-value (using a so-called OECD instrument) and a weighing principle (using a so-called High Volume Sampler). Among other things the reason for the large deviations is that the measurements using the grey value principle are sensitive to flying dust's deviation from "normal dust". In this way, a large content of calcium dust in the air could result in the measurements being systematically too small.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | I | 2 | 5 | 15 | 16 | 16 | 19 | 26 | 24 | 16 | 36 |
| | II | 2 | 12 | 4 | 21 | 41 | 14 | 31 | 29 | 31 | 8 |
| | I | 39 | 42 | 44 | 40 | 42 | 42 | 50 | 51 | 58 | 64 |
| | II | 30 | 44 | 26 | 60 | 34 | 34 | 14 | 41 | 58 | 47 |

We consider this data as being observations from independent identically distributed stochastic variables

$$\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \dots, \begin{bmatrix} X_{20} \\ Y_{20} \end{bmatrix}.$$

We will examine whether we can assume the distribution is normal with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If the distribution is normal, we find the estimates

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix} = \begin{bmatrix} \bar{X} \\ \bar{Y} \end{bmatrix} = \begin{bmatrix} 32.35 \\ 29.05 \end{bmatrix},$$

and

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{bmatrix} = \begin{bmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{bmatrix} = \begin{bmatrix} 311 & 182 \\ 182 & 279 \end{bmatrix},$$

where $\hat{\boldsymbol{\Sigma}}$ is the unbiased estimate of $\boldsymbol{\Sigma}$. Specially we have

$$S_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

We now want to check if the observations can be assumed to come from a normal distribution with parameters $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. To do that we first estimate the contour ellipses. The eigenvalues and eigenvectors for $\hat{\boldsymbol{\Sigma}}$ are

$$\hat{\lambda}_1 = 477.613 \quad \text{and} \quad \hat{\boldsymbol{p}}_1 = \begin{bmatrix} 0.736 \\ 0.678 \end{bmatrix}$$

and

$$\hat{\lambda}_2 = 112.676 \quad \text{and} \quad \hat{\boldsymbol{p}}_2 = \begin{bmatrix} -0.678 \\ 0.736 \end{bmatrix}.$$

If we choose the coordinate system with origo in $\hat{\boldsymbol{\mu}}$ and with $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ as base vectors, the contour ellipsoids have equations of the form

$$\frac{z_1^2}{\hat{\lambda}_1} + \frac{z_2^2}{\hat{\lambda}_2} = c,$$

or

$$\frac{z_1^2}{477.613} + \frac{z_2^2}{112.676} = c,$$

where the new coordinates are given by

$$\mathbf{P}\,\boldsymbol{z} = (\boldsymbol{p}_1\boldsymbol{p}_2)\boldsymbol{z} = \boldsymbol{x} - \hat{\boldsymbol{\mu}}.$$

In figure 2.2 we show the observations and 3 contour ellipses corresponding to the $c$-values $c_1 = \chi^2(2)_{0.40} = 1.02$, $c_2 = \chi^2(2)_{0.80} = 3.22$ and $c_3 = \chi^2(2)_{0.95} = 5.99$. This has the effect (see theorem 2.15) that in the normal distribution with parameters $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ we have the probabilities $40\%$, $80\%$ and $95\%$ of having the observations within the inner, the middel and the outer ellipse. For the areas between the ellipses resp. outside these, we have the probabilities $40\%$, $40\%$, $15\%$ and $5\%$. These numbers can be compared to the corresponding observed relative probabilities $40\%$, $30\%$, $30\%$ and $0\%$. The fit is - if not overwhelming - at least acceptable.

If one wants a more precise result, one can perform a $\chi^2$ -test. It would then be reasonable to divide the plane further according to the eigenvectors. In the case shown, this would result in $4 \times 4$ areas with estimated probabilities of $10\%$, $10\%$, $3.75\%$ and $1.25\%$. One can then compute the usual $\chi^2$ test-statistic:

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

and compare it with a $\chi^2(n-6)$ distribution (we have estimated 5 parameters). In the present case there are not really enough observations to perform this analysis.
The correlation coefficient is estimated at

$$\hat{\rho} = \frac{182}{\sqrt{311 \cdot 279}} = 0.62,$$

and the conditional variances are estimated at

$$\begin{aligned} \hat{\mathrm{V}}(X|Y=y) &= 311(1-\hat{\rho}^2) = 192 \\ \hat{\mathrm{V}}(Y|X=x) &= 279(1-\hat{\rho}^2) = 172. \end{aligned}$$

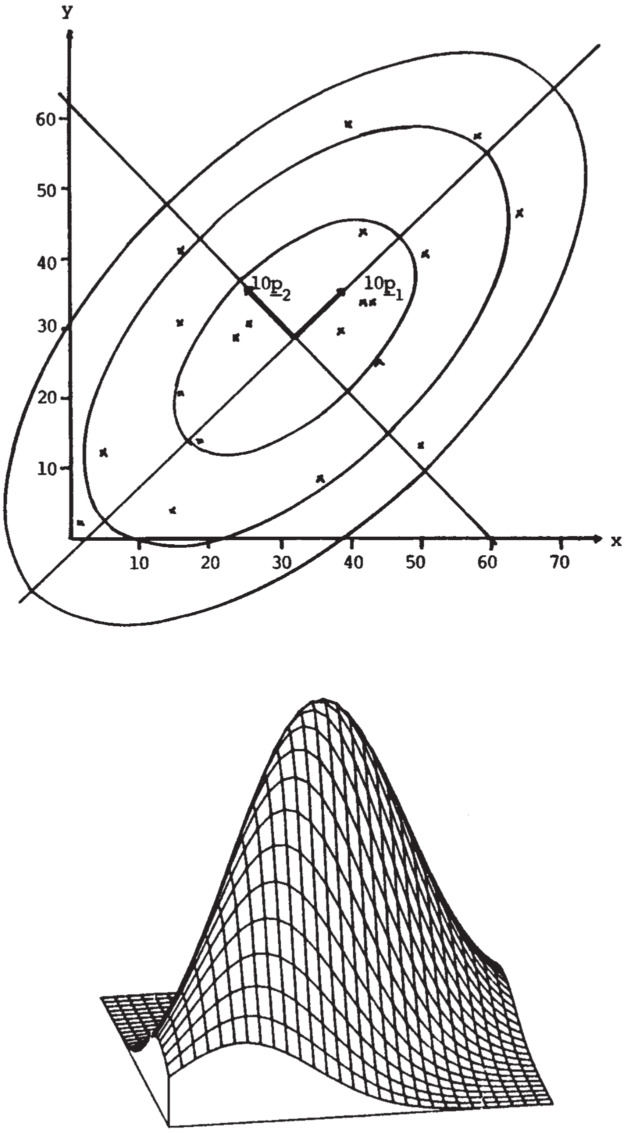


Figure 2.3: Estimated contour ellipses and estimated density function corresponding to the data in example 2.2

We see, that the conditional variances have been reduced by 38% corresponding to $\rho^2 = 0.38$. That the conditional variance of e.g. an OECD-measurement for given High Volume Sampler measurement is substantially less than the unconditional variance seems rather reasonable. We know, that the amount of flying dust measured using a High Volume Sampler is found as e.g. $2\frac{\mu g}{m^3}$, so we would not expect to get results from the OECD-instrument, which deviate grossly. This corresponds to a small conditional variance. If the result from the High Volume Sampler is unknown, then we must expect a measurement from the OECD-instrument can lie anywhere in its natural range of variation - corresponding to a larger unconditional variance. ♦

## 2.3 Correlation and regression

In this section we will discuss the meaning of parameters in a multidimensional normal distribution in greater detail. First we will try to generalise the properties of the correlation coefficient seen in the previous section.

### 2.3.1 The partial correlation coefficient.

The starting point is the formula for the conditional distributions in a multi-dimensional normal distribution. Let $\boldsymbol{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let the variables be partitioned as follows

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}; \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}; \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\boldsymbol{X}_1$ consists of the $m$ first elements in $\boldsymbol{X}$ and likewise with the others. Then the conditional dispersion of $\boldsymbol{X}_1$ for given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is, as was shown in theorem 2.16, equal to

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

By the partial correlation coefficient between $X_i$ and $X_j$, $i, j \leq m$, conditioned on (or: for given) $\boldsymbol{X}_2 = \boldsymbol{x}_2$ we will understand the correlation in the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$. It is denoted by $\rho_{ij|m+1,\ldots,p}$.

Let

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{1p} & \cdots & \sigma_p^2 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}=\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{1m} & \cdots & a_{mm} \end{bmatrix},$$

we now have

$$\rho_{ij|m+1,\ldots,n}=\frac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}}.$$

For the special case of $\boldsymbol{X}$ being three dimensional we have with

$$\boldsymbol{\Sigma}=\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix},$$

that

$$\begin{aligned}&\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\\ &=\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}-\frac{1}{\sigma_3^2}\begin{bmatrix} \rho_{13}^2\sigma_1^2\sigma_3^2 & \rho_{13}\rho_{23}\sigma_1\sigma_2\sigma_3^2 \\ \rho_{13}\rho_{23}\sigma_1\sigma_2\sigma_3^2 & \rho_{23}^2\sigma_2^2\sigma_3^2 \end{bmatrix}\\ &=\begin{bmatrix} \sigma_1^2(1-\rho_{13}^2) & \sigma_1\sigma_2(\rho_{12}-\rho_{13}\rho_{23}) \\ \sigma_1\sigma_2(\rho_{12}-\rho_{13}\rho_{23}) & \sigma_2^2(1-\rho_{23}^2) \end{bmatrix}.\end{aligned}$$

From this follows that the partial correlation coefficient between $X_1$ and $X_2$ conditioned on $X_3$ is

$$\rho_{12|3}=\frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}.$$

For a $p$-dimensional vector $\boldsymbol{X}$ we therefore find

$$\rho_{ij|k}=\frac{\rho_{ij}-\rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}. \qquad (**)$$

Since it is possible to find conditional distributions for given $X_{m+1},\ldots,X_p$ by successive conditionings we can therefore determine partial correlation coefficients of higher order by successive use of (**). E.g. we find

$$\rho_{ij|kl}=\frac{\rho_{ij|k}-\rho_{il|k}\cdot\rho_{jl|k}}{\sqrt{(1-\rho_{il|k}^2)\cdot(1-\rho_{jl|k}^2)}},$$

| | C $_3$S | C$_3$A | BLAINE | Strength 3 | Strength 28 |
|---|---|---|---|---|---|
| C$_3$S | 1 | -0.309 | 0.091 | 0.158 | 0.344 |
| C$_3$A | -0.309 | 1 | 0.192 | 0.120 | -0.166 |
| BLAINE | 0.091 | 0.192 | 1 | 0.745 | 0.320 |
| Strength 3 | 0.158 | 0.120 | 0.745 | 1 | 0.464 |
| Strength 28 | 0.344 | -0.166 | 0.320 | 0.464 | 1 |

Table 2.1: The correlation matrix for 5 cement variables.

here we have first conditioned on $X_k$ and then conditioned on $X_l$.

In section 2.2.6 we saw that the (squared) correlation coefficient is a measure of the reduction in variance if we condition on one of the variables. Since the partial correlation coefficients are just correlations in conditional distributions we can use the same interpretation here. We have e.g. that $\rho^2_{ij|kl}$ gives the fraction of $X_i$'s variance for given $X_k = x_k$ and $X_l = x_l$ which is explained by $X_j$. It should be emphasised that these interpretations are strongly dependent on the assumption of normality. For the general case the conditioned variances will depend on the values with which they are conditioned (i.e. depend on $x_k$ and $x_l$).

When estimating the partial correlations one just estimates the variance-covariance matrix and then computes the partial correlations as shown. If the estimate of the variance-covariance matrix is a maximum-likelihood estimator then the estimates of the partial correlations computed in this way will also be maximum likelihood estimates (cf. theorem 10 p. 2.28 in volume I).

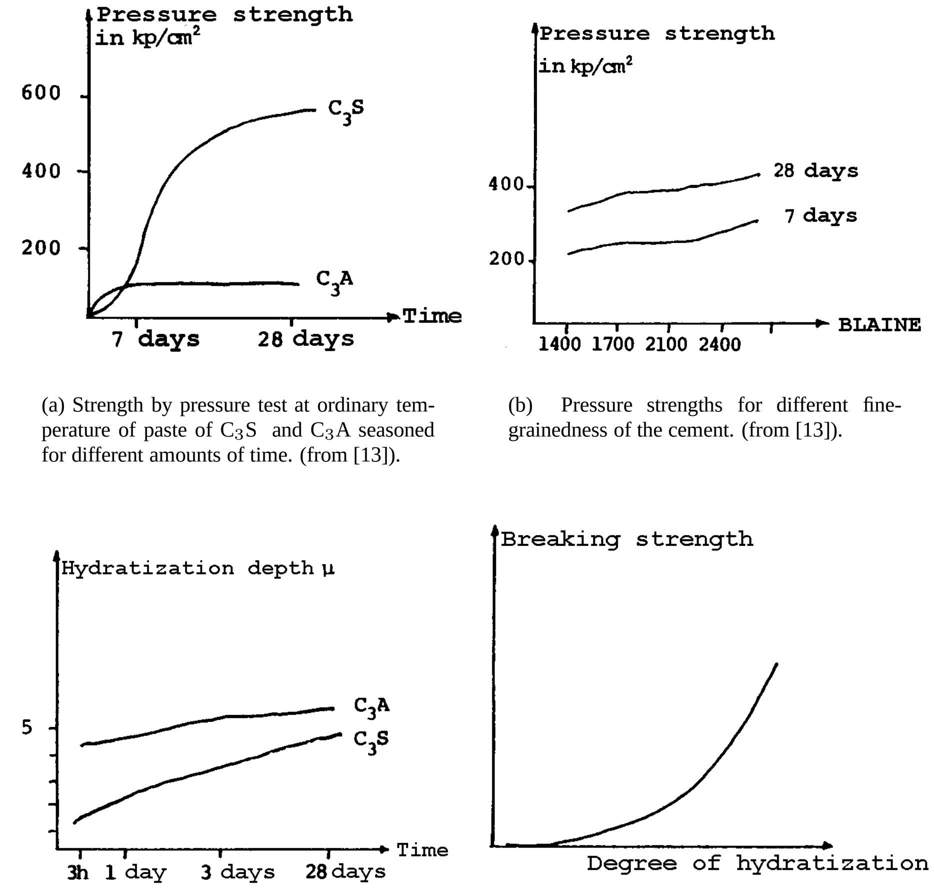
We will now illustrate the concepts in

**EXAMPLE 2.3.** (Data are from [17]).

In table 2.1 correlation coefficients between 3- and 28-day strengths for Portland Cement and the content of minerals C$_3$S (Alit, Tricalciumsilicat Ca$_3$SiO$_5$) and C$_3$A (Aluminat, Tricalciumaluminat, Ca$_3$Al$_2$O$_6$), and the degree of fine-grainedness (BLAINE) are given. The correlations are estimated using 51 corresponding observations.

It should be noted that C$_3$S constitutes about 35-60% of normal portland clinkers and C$_3$A is about 5-18% of clinker. The BLAINE is a measure of the specific surface so that a large BLAINE corresponds to a very fine-grained cement.

We will be especially interested in the relationship between C$_3$A content in clinker and the two strengths. It is commonly accepted cf. the following figure, that a large content of C$_3$A gives a larger 3-day strength which is also in correspondence with $\hat{\rho}_{\mathrm{C_3A,Strength3}} = 0.120$. The problem is that this larger 3-day strength for cement with large content of C$_3$A only depends on C$_3$A 's larger degree of hydratisation (the faster the water reacts with the cement the faster it will have greater strength. C$_3$A's far greater hydratisation after 3 days as seen from figure 2.4(c) and the degree of hydratisation and its influence on the strengths has been sketched in figure 2.4(d).



(a) Strength by pressure test at ordinary temperature of paste of $C_3S$ and $C_3A$ seasoned for different amounts of time. (from [13]).

(b) Pressure strengths for different finegrainedness of the cement. (from [13]).

(c) Degree of hydratisation for cement minerals and their dependence on time (from [13]).

(d) Relationship between degree of hydratisation and strength (from [13]).

Figure 2.4:

| | $C_3S$ | $C_3A$ | Strength 3 | Strength 28 |
|---|---|---|---|---|
| $C_3S$ | 1 | -0.333 | 0.137 | 0.333 |
| $C_3A$ | -0.333 | 1 | -0.035 | -0.246 |
| Strength 3 | 0.137 | -0.035 | 1 | 0.358 |
| Strength 28 | 0.333 | -0.246 | 0.358 | 1 |

Table 2.2: Correlation matrix for 4 cement variables conditioned on BLAINE.

If we look at the correlation matrix we also see that the content of $C_3A$ is positively correlated with the BLAINE i.e. cements with a very high content of $C_3A$ will usually be very fine-grained and as it is seen in figure 2.4(b) this should also help increase the strength.

Finally we see that the 28-day strength is slightly negatively correlated with the content of $C_3A$ This does not seem strange if we consider the temporal dependence of $C_3S$'s and $C_3A$'s as seen in e.g. in figure 2.4(a) even though the finer grain (for cement with large content of $C_3A$ ) should also be seen in the 28-day strength cf. figure 2.4(b).

In order to separate the different characteristics of $C_3A$ from the effects which arise from a $C_3A$ -rich cement seems to be easier to grind and therefore often is seen in a bit more fine-grained form. Therefore, we will estimate the conditional correlations for fixed value of BLAINE. These are seen in table 2.3. We see that the partial correlation coefficient between 3-day strength and $C_3A$ for given fine-grainedness is negative (note the unconditioned correlation coefficient was positive). This implies that we for fixed fine-grainedness must expect that cements with a high content of $C_3A$ will tend to have lower strengths. This might indicate that the large 3-day strength for cements with high content of $C_3A$ rather depends on these cements having a large BLAINE (that they are crushed somewhat easier) than that $C_3A$ hydrates quickly!

We see a corresponding effect on the correlation between $C_3A$ and 28-day strength. Here the unconditional correlation is -0.168 and the partial correlation for fixed BLAINE has become -0.246. ♦

**REMARK 2.9.** The example above shows that one has to be very cautious in the interpretation of correlation coefficients. It would be directly misleading e.g. to say that a large content of $C_3A$ assures a large 3-day strength. First of all it is not possible to conclude anything about the relation between two variables just by looking at their correlation. What you can conclude is that there seems to be a tendency that a high content of $C_3A$ and a high 3-day strength appear at the same time. The reason for this could be that they both depend on a third but unknown factor without there having to be any direct relation between the two variables. Secondly we also see that going from unconditioned to partial correlations can even give a change of sign corresponding to an effect which is the opposite of that we get by a direct analysis. The reason for this is a correlation with a 3rd factor in this case BLAINE which disturbs the picture. ▼

In many situations we would like to test if the correlation coefficient can be assumed to be 0. You can then use

**THEOREM 2.20.** Let $R = R_{ij|m+1\ldots p}$ be the empirical partial correlation coefficient between $X_i$ and $X_j$ conditioned on (or: for given) $X_{m+1,\ldots,X_p}$. It is assumed to be computed from the unbiased estimates of the variance-covariance matrix and from $n$ observations. Then

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} \in \mathrm{t}(n-2-(p-m)),$$

if $\rho_{ij|m+1,\ldots,p} = 0$. ▲

**PROOF 2.17.** Omitted. ■

**REMARK 2.10.** The number $(p-m)$ is the number of variables which are fixed (conditioned upon). The degrees of freedom are therefore equal to the number of observations minus 2 minus the number of fixed variables. The theorem is also valid if $p - m = 0$ i.e. if we have the case of an unconditional correlation coefficient. ▼

We continue example 2.3 in

**EXAMPLE 2.4.** Let us investigate whether the value of $r_{24|3}$ is significantly different from 0. We find with $r_{24|3} = R$:

$$\begin{aligned}\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} &= \frac{-0.035}{\sqrt{1-0.035^2}}\cdot\sqrt{51-2-(5-4)} \\ &= -0.243 = \mathrm{t}(48)_{40\%}.\end{aligned}$$

A hypothesis that $\rho_{24|3}$ is 0 will therefore be accepted using a test at level $\alpha$ for $\alpha < 80\%$. (Note: this is by nature a two-sided test.) ♦

If we wish to test other values of $\rho$ or to determine confidence intervals we can use

**THEOREM 2.21.** Assume the situation is as in the previous theorem. We consider the hypothesis

$$H_0 : \rho_{ij|m+1,\ldots,p} = \rho_0$$

versus

$$H_1 : \rho_{ij|m+1,\ldots,p} \neq \rho_0.$$

We let

$$Z = \frac{1}{2} \ln \frac{1 + R_{ij|m+1,\ldots,p}}{1 - R_{ij|m+1,\ldots,p}}$$

and

$$z_0 = \frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0}.$$

Under $H_0$ we will have

$$(Z - z_0) \cdot \sqrt{n - (p - m) - 3} \quad \text{approx.} \in \mathrm{N}(0, 1).$$

▲

**PROOF 2.18.** Omitted. ■

**EXAMPLE 2.5.** Let us determine a 95% confidence interval for $\rho_{24|3}$ in example 2.4. We have

$$\begin{array}{ll} \mathrm{P} & \{-1.96 < (Z - z) \cdot \sqrt{51 - (5 - 4) - 3} < 1.96\} \simeq 95\% \\ \Leftrightarrow & \mathrm{P}\{-1.96 - 6.86Z < -6.86z < 1.96 - 6.86Z\} \simeq 95\% \\ \Leftrightarrow & \mathrm{P}\{Z - 0.29 < z < Z + 0.29\} \simeq 95\%. \end{array}$$

The relationship between $z$ and $\rho_{24|3} = \rho$ is

$$z = \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} \quad \Leftrightarrow \quad \rho = \frac{e^{2z} - 1}{e^{2z} + 1}$$

The observed value of $Z$ is

$$Z = \frac{1}{2} \ln \frac{1 - 0.035}{1 + 0.035} = -0.03501.$$

The limits for $z$ become

$$[-0.3250, 0.2549].$$

The corresponding limits for $\rho_{25|4}$ are

$$\left[ \frac{e^{-0.6500} - 1}{e^{-0.6500} + 1}, \frac{e^{0.5098} - 1}{e^{0.5098} + 1} \right] = [-0.31, 0.25].$$

♦

### 2.3.2 The multiple correlation coefficient

The partial correlation coefficient is one possible generalisation of the correlation between two variables. The partial correlations are mostly intended to describe the degree of relationship (correlation, covariance) between two variables. Instead we will now consider the formula on p. 79

$$\rho^2 = \frac{\mathrm{V}(X_1) - \mathrm{V}(X_1 | X_2 = x_2)}{\mathrm{V}(X_1)},$$

This is the "degree of reduction in variation" interpretation of the (squared) correlation coefficient. This we now seek to generalise. We again consider the partition of the $p$-dimensionally normally distributed vector $\boldsymbol{X}$ i an $m$-dimensional vector $\boldsymbol{X}_1$ and a $(p-m)$-dimensional vector $\boldsymbol{X}_2$, and the resulting partitioning of the parameters i.e.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}; \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}; \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

We now define the multiple correlation coefficient between $X_i$, $i = 1, \ldots, m$ and $\boldsymbol{X}_2$ as the maximal correlation between $X_i$ and a linear combination of $\boldsymbol{X}_2$'s elements. It is denoted $\rho_{i|m+1,\ldots,p}$.

It can be shown that the optimal linear combination of $\boldsymbol{X}_2$'s elements is

$$\beta_i' \boldsymbol{X}_2 = (\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1})_i \boldsymbol{X}_2,$$

where $\beta_i'$ is the $i$'th row in the matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. This matrix appears in the expression for the conditional mean of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$. As stated before this is

$$\mathrm{E}(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) = \boldsymbol{\mu}_1 + \begin{bmatrix} \boldsymbol{\beta}_1' \\ \vdots \\ \boldsymbol{\beta}_m' \end{bmatrix} (\boldsymbol{x}_2 - \boldsymbol{\mu}_2).$$

It can also be shown that

$$\inf_{\alpha} \mathrm{V}(X_i - \boldsymbol{\alpha}' \boldsymbol{X}_2) = \mathrm{V}(X_i - \boldsymbol{\beta}_i' \boldsymbol{X}_2),$$

i.e. the considered linear combination minimises the variance of $(X_i - \boldsymbol{\alpha}'\boldsymbol{X}_2)$.

We now have the following important

**THEOREM 2.22.** We consider the situation above. Let $\boldsymbol{\sigma}_i$ be the $i$'th column in $\boldsymbol{\Sigma}_{21}$, i.e. $\boldsymbol{\sigma}_i'$ is the $i$'th row in $\boldsymbol{\Sigma}_{12}$.

Then

$$\rho_{i|m+1,\dots,p} = \frac{\sqrt{\boldsymbol{\sigma}_i' \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_i}}{\sqrt{\sigma_{ii}}}.$$

If we let

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{ii} & \sigma_i' \\ \sigma_i & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

then

$$1 - \rho^2_{i|m+1,\dots,p} = \frac{\det \boldsymbol{\Sigma}_i}{\sigma_{ii} \det \boldsymbol{\Sigma}_{22}} = \frac{\mathrm{V}(X_i|\boldsymbol{X}_2)}{\mathrm{V}(X_i)},$$

▲

**PROOF 2.19.** The proofs to the claims before the theorem are quite simple. One just has to use a Lagrange multiplier and also use that the variance-covariance matrix is positive semidefinite. What is claimed in the theorem then follows by using the formula for the conditional variance-covariance structure (p. 74) on $\boldsymbol{\Sigma}_i$ by use of the matrix formulas in section 1.2.7. ■

**REMARK 2.11.** In the theorem we have obtained a large number of characteristics for the multiple correlation coefficient and since

$$\rho^2_{i|m+1,\dots,p} = \frac{\mathrm{V}(X_i) - \mathrm{V}(X_i|\boldsymbol{X}_2)}{\mathrm{V}(X_i)},$$

we note that we have generalised the property of reduction in variance. It is important to note that we can see from the determinant formula that it is possible to compute the multiple correlation coefficient from the correlation matrix by using the same formulas valid when computing it from the variance-covariance matrix. ▼

With regard to the estimation of multiple correlation coefficients the same remark as on p. 85 regarding the estimation of partial coefficients holds.

In the next example we continue example 2.4.

**EXAMPLE 2.6.** To get an impression of to which degree the content of $C_3A$ and $C_3S$ in example 2.4 can explain the variation in e.g. 3-day strength we can compute the multiple correlation coefficient between strength day 3 and ($\mathrm{C_3S}$, and $\mathrm{C_3A}$). We find

$$1-\hat{\rho}^2_{4|12} = \frac{\det\begin{bmatrix} 1 & 0.158 & 0.120 \\ 0.158 & 1 & -0.309 \\ 0.120 & -0.309 & 1 \end{bmatrix}}{1\cdot\det\begin{bmatrix} 1 & -0.309 \\ -0.309 & 1 \end{bmatrix}}$$

where the indices of the variables correspond to those used in example 2.3. We find

$$\hat{\rho}^2_{4|12} = 1 - 0.9435 = 0.0565.$$

The data therefore indicate that only about 6% of the variation in the strength of the cement (from samples which have been collected the way these data have been collected) can be explained by variations in $\mathrm{C_3S}$- and $\mathrm{C_3A}$- content alone. ♦

If the multiple correlation coefficient is 0 (i.e. if $\sigma_i = 0$) it is not difficult to determine the distribution of $\hat{\rho}^2_{i|m+1,\ldots,p}$. We give the results in the slightly changed form in

**THEOREM 2.23.** Let $R = \hat{\rho}_{i|m+1,\ldots,p}$ be the empirical multiple correlation coefficient between $X_i$ and $\boldsymbol{X}_2 = (X_{m+1},\ldots,X_p)$ based upon $n$ observations. Then

$$\frac{R^2}{1-R^2}\cdot\frac{n-(p-m)-1}{p-m} \in \mathrm{F}(p-m, n-(p-m)-1),$$

if $\rho_{i|m+1,\ldots,p} = 0$. ▲

**PROOF 2.20.** Omitted ■

This can be used in testing the hypotheses

$$H_0 : \rho_{i|m+1,\ldots,p} = 0 \qquad \text{against} \qquad H_1 : \rho_{i|m+1,\ldots,p} \neq 0.$$

We reject the null hypothesis for large values of the test statistic. This is illustrated in

**EXAMPLE 2.7.** Consider the situation in example 2.6. We now want to examine if it can be assumed that the multiple correlation between $X_4$ and $(X_1, X_2)$ is 0. (Note that $p = 3$ and $m = 1$.) We find the statistic

$$\frac{R^2}{1-R^2}\frac{51-(3-1)-1}{3-1} = \frac{0.0565}{0.9435}\cdot\frac{48}{2} = 1.44.$$

Since

$$\mathrm{F}(2,48)_{0.90} = 2.42,$$

we will at least accept a hypothesis that $\rho_{4|12} = 0$ for any level $\alpha < 10\%$. With the available data it cannot be rejected that $\rho_{4|12} = 0$. This does not mean that it is not different from 0 (which it probably is), only that we cannot be sure using the available data because the true (but unknown) value of $\rho_{4|12}$ is probably rather small. ♦

We shall not consider tests for other values of $\rho_{i|m+1,\ldots,n}$.

## 2.3.3 Regression

We will not give any deep introduction to the so-called regression theory which must not be confused with what we in the following section will call (linear) regression analysis.

Let $\begin{bmatrix} Y \\ \boldsymbol{X} \end{bmatrix}$ be a stochastic vector. By the term regression of $Y$ on $\boldsymbol{x}$ we mean the function given by

$$\mathrm{g}(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}),$$

i.e. the conditional mean as a function of the conditioned variable.

Let $\begin{bmatrix} Y \\ \boldsymbol{X} \end{bmatrix}$ be normally distributed with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_1' \\ \boldsymbol{\sigma}_1 & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then theorem 2.16 shows that

$$\mathrm{g}(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \mu_1 + \boldsymbol{\sigma}_1'\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2),$$

i.e. the regression is linear (affine).

We now specialise to two dimensions.

Let $\begin{bmatrix} Y \\ X \end{bmatrix}$ be normally distributed with parameters

$$\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix}.$$

Then the regression of $Y$ on $X$ is given by

$$\mathrm{E}(Y|X=x)=\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x),$$

and the regression of $X$ on $Y$ is given by

$$\mathrm{E}(X|Y=y)=\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y).$$

Let us assume that we have measurements $\begin{bmatrix} Y_1 \\ X_1 \end{bmatrix},\dots,\begin{bmatrix} Y_n \\ X_n \end{bmatrix}$.

The maximum likelihood estimates for the slopes are obtained by using the maximum likelihood estimators for the parameters in the formula. Then

$$\begin{aligned}
\hat{\rho} &= \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}}=\frac{\mathrm{SP}_{xy}}{\sqrt{\mathrm{SAK}_x\mathrm{SAK}_y}},\\
\hat{\sigma_x^2} &= \frac{1}{n}\sum(X_i-\bar{X})^2,\\
\hat{\sigma_y^2} &= \frac{1}{n}\sum(Y_i-\bar{Y})^2,
\end{aligned}$$

and we see e.g. that the estimates of the slope in the expression for the regression of $Y$ on $X$ becomes

$$\hat{\rho}\frac{\hat{\sigma_y}}{\hat{\sigma_x}}=\frac{\mathrm{SP}_{xy}}{\mathrm{SAK}_x}.$$

This gives the empirical regression equation

$$\hat{\mathrm{E}}(Y|X=x)=\bar{Y}+\frac{\mathrm{SP}_{xy}}{\mathrm{SAK}_x}(x-\bar{X}),$$

i.e. precisely the same result as we obtained in the one dimensional linear regression analysis cf. section 5.2 in volume 1. However, there the assumptions were completely different since then we assumed that the values of the independent variable (here $X$, in volume 1 $t$) were deterministic values. In the present text we assume that they are observations of a normally distributed variable which is correlated with the dependent variable. Concerning the estimation it is not important which of the two models one works with but the interpretation of the results are of course dependent hereon. We now continue with example 2.8.

**EXAMPLE 2.8.** In this example we will determine the linear relations from a measurement by one of the two methods stated in example 2.2 to the other measurement.

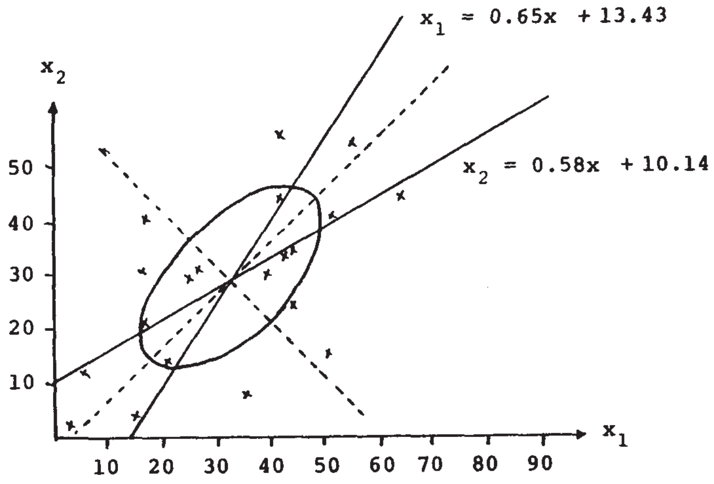


Figure 2.5:

We find the regressions

$$\begin{aligned}\hat{\mathrm{E}}(X_1|X_2=x_2) &= \bar{x_1}+\hat{\rho}\frac{s_1}{s_2}(x_2-\bar{x_2}) \\ &= 0.65x_2+13.43\end{aligned}$$

and

$$\begin{aligned}\hat{\mathrm{E}}(X_2|X_1=x_1) &= \bar{x_2}+\hat{\rho}\frac{s_2}{s_1}(x_1-\bar{x_1}) \\ &= 0.58x_1+10.14.\end{aligned}$$

These lines are shown in figure 2.5. If we wish to check if there might be some sort of relation between $X_1$ and $X_2$ we can examine the correlation coefficient. It has been found to be

$$\hat{\rho}=\frac{182}{\sqrt{311\cdot 279}}=0.617,$$

i.e.

$$\hat{\rho}^2=0.380.$$

The test statistic for a test of the hypothesis $\rho=0$ is, cf. p. 88, with $p=m=2$

$$t=\frac{0.617}{\sqrt{1-0.380}}\sqrt{20-2}=3.32>\mathrm{t}(18)_{0.995}.$$

Using a test at level $\alpha > 1\%$ we must reject the hypothesis and we assume that $\rho \neq 0$, is different from 0. I.e. we now assume there exists a linear relationship between the methods of measurements in the two cases and it is estimated by the two regressions. We can then find estimates of the errors etc. in the usual fashion.

In the figure we have also shown a contour-ellipse and its main axes. It can be shown that the first axis is the line which is obtained by minizing the orthogonal squared distance to the points. On the other hand the regression equations are found by minimizing the vertical and horizontal distances respectively. The first main axis is therefore also called the orthogonal regression. In chapter 4 we will return to this concept. ♦

## 2.4 The partition theorem

In this section we will consider a stochastic variable $\boldsymbol{x} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is regular of order $n$. We will consider the inner product defined by $\boldsymbol{\Sigma}^{-1}$ and the corresponding norm i.e.

$$(\boldsymbol{x}|\boldsymbol{y}) = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

and

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}|\boldsymbol{x})} = \sqrt{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$$

Now let the sub-spaces $U_1, \ldots, U_k$ be orthogonal (with respect to this inner product) so that

$$R = U_1 \oplus \ldots \oplus U_k.$$

We let $\dim U_i = n_i$ and call the projection onto $U_i$ for $p_i$. The corresponding projection matrix is called $\mathbf{C}_i$.

Using the notation mentioned above the following is valid

**THEOREM 2.24.** (**The partition theorem**) If we let

$$\boldsymbol{Y}_i = p_i(\boldsymbol{x} - \boldsymbol{\mu}), \qquad i = 1, \ldots, k$$

and

$$K_i = \|\boldsymbol{Y}_i\|^2 = \|p_i(\boldsymbol{x} - \boldsymbol{\mu})\|^2, \qquad i = 1, \ldots, k,$$

then

$$\boldsymbol{x}-\boldsymbol{\mu}=\sum_{i=1}^{k}\boldsymbol{Y}_i$$

and

$$\|\boldsymbol{x}-\boldsymbol{\mu}\|^2=\sum_{i=1}^{k}K_i.$$

Furthermore $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_k$ are stochastically independent and normally distributed and $K_1,\dots,K_k$ are stochastically independent and $\chi^2(n_i)$ -distributed variables. ▲

**PROOF 2.21.** We have that $\boldsymbol{Y}_i=\mathbf{C}_i(\boldsymbol{x}-\boldsymbol{\mu})$ therefore

$$\boldsymbol{Y}=\begin{bmatrix}\boldsymbol{Y}_1\\ \vdots\\ \boldsymbol{Y}_k\end{bmatrix}=\begin{bmatrix}\mathbf{C}_1\\ \vdots\\ \mathbf{C}_k\end{bmatrix}(\boldsymbol{X}-\boldsymbol{\mu}).$$

From this we obtain

$$\mathrm{D}(\boldsymbol{Y})=\begin{bmatrix}\mathbf{C}_1\\ \vdots\\ \mathbf{C}_k\end{bmatrix}\cdot\boldsymbol{\Sigma}\cdot(\mathbf{C}_1',\dots,\mathbf{C}_k')=(\mathbf{C}_i\boldsymbol{\Sigma}\mathbf{C}_j')_{(i,j)}.$$

Now for $i\neq j$ it follows from the lemma on page 54 that

$$\mathbf{C}_i\boldsymbol{\Sigma}\mathbf{C}_j'=\mathbf{0}.$$

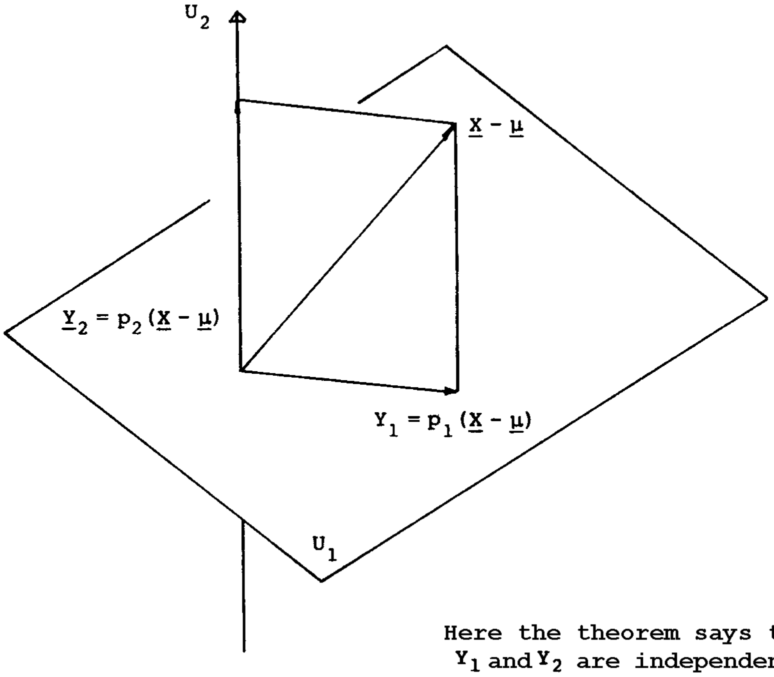
From this it follows that the components of $\boldsymbol{Y}$ are stochastically independent (because $\boldsymbol{Y}$ is normally distributed).

We must now determine the distribution of $\|p_i(\boldsymbol{X}-\boldsymbol{\mu})\|^2$. We have that $\boldsymbol{X}$ can be written

$$\boldsymbol{X}=\boldsymbol{\mu}+\mathbf{A}\boldsymbol{Z}$$

where $\boldsymbol{Z}\in\mathrm{N}(0,\mathbf{I})$ and $\mathbf{A}\,\mathbf{A}'=\boldsymbol{\Sigma}$. From this it follows that

$$\begin{aligned}\|p_i(\boldsymbol{X}-\boldsymbol{\mu})\|^2 &= \|p_i(\mathbf{A}\boldsymbol{Z})\|^2=\|\mathbf{C}_i\mathbf{A}\boldsymbol{Z}\|^2\\ &= \boldsymbol{Z}'\mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A}\boldsymbol{Z}=\boldsymbol{Z}'\mathbf{D}_i\boldsymbol{Z}.\end{aligned}$$


$U_2$
$\underline{X} - \underline{\mu}$
$\underline{Y}_2 = P_2(\underline{X} - \underline{\mu})$
$Y_1 = P_1(\underline{X} - \underline{\mu})$
$U_1$

Here the theorem says that $Y_1$ and $Y_2$ are independent and that $\| Y_1 \|^2 \in \chi^2(2)$ and $\| Y_2 \|^2 \in \chi^2(1)$ .

Figure 2.6:

Now

$$\begin{aligned}\mathbf{D}_i\mathbf{D}_i &= \mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A}\mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A}\\ &= \mathbf{A}'\mathbf{C}_i'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\,\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\,\mathbf{A}\\ &= \mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A}\\ &= \mathbf{D}_i,\end{aligned}$$

i.e. $\mathbf{D}_i$ is idempotent. In the above we have used the lemma on p. 54 repeatedly. It is obvious that $\mathrm{rg}(\mathbf{D}_i) = n_i$. Now, since

$$\begin{aligned}\mathbf{D}_i &= \mathbf{A}'\mathbf{C}_i'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{C}_i\mathbf{A}\\ &= (\mathbf{A}^{-1}\mathbf{C}_i\mathbf{A})'(\mathbf{A}^{-1}\mathbf{C}_i\mathbf{A}),\end{aligned}$$

then $\mathbf{D}_i$ is positive semidefinite (cf. theorem 1.16 p. 38) therefore there exists an orthogonal (and even orthonormal) matrix $\mathbf{P}'$ (theorem 1.10) so that

$$\mathbf{P}'\mathbf{D}_i\mathbf{P} = \boldsymbol{\Lambda}_i \quad \text{or} \quad \mathbf{D}_i = \mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}',$$

where $\boldsymbol{\Lambda}_i$ is a diagonal matrix with rank $n_i$. Since $\mathbf{D}_i$ is idempotent we obtain

$$\mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}' = \mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}'\mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}' = \mathbf{P}\,\boldsymbol{\Lambda}_i^2\mathbf{P}',$$

or $\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_i^2$. Therefore $\boldsymbol{\Lambda}_i$ has $n_i$ 1's and $n - n_i$ 0's on the diagonal. Therefore

$$\begin{aligned}\boldsymbol{Z}'\mathbf{D}_i\boldsymbol{Z} &= \boldsymbol{Z}'\mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}'\boldsymbol{Z} = (\mathbf{P}'\boldsymbol{Z})'\boldsymbol{\Lambda}_i(\mathbf{P}'\boldsymbol{Z})'\\ &= \boldsymbol{V}'\boldsymbol{\Lambda}_i\boldsymbol{V}\\ &= \underbrace{V_1^2 + \cdots + V_n^2}_{n_i \text{ components } \neq 0}.\end{aligned}$$

Since $\boldsymbol{V} \in \mathrm{N}(\mathbf{0}, \mathbf{P}'\mathbf{P}) = \mathrm{N}(\mathbf{0}, \mathbf{I})$ it is seen that

$$\boldsymbol{Z}'\mathbf{D}_i\boldsymbol{Z} = \|p_i(\boldsymbol{X} - \boldsymbol{\mu})\|^2 \in \chi^2(n_i).$$

■

**EXAMPLE 2.9.** Let $X_1, \ldots, X_n$ be independent and $\mathrm{N}(\mu, \sigma^2)$ -distributed. Then

$$\boldsymbol{X} = \begin{bmatrix} X_1\\ \vdots\\ X_n\end{bmatrix} \in \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}).$$

We consider the subspace $U_1$ given by

$$\boldsymbol{x} \in U_1 \quad \Leftrightarrow \quad x_1 = \ldots = x_n,$$

and the orthogonal subspace to $U_1$ (with respect to $\sigma^2 \mathbf{I}$) called $U_2$. (This concept of orthogonality corresponds to the usual one). Now the identity

$$\sum (x_i - y)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - y)^2,$$

shows that the projection onto $U_1$ is given by

$$p_1(\boldsymbol{x}) = \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix},$$

which means

$$p_2(\boldsymbol{x}) = \boldsymbol{x} - p_1(\boldsymbol{x}) = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}.$$

Since $\dim U_1 = 1$ and $\dim U_2 = n - 1$ we find from the partition theorem that

$$p_1(\boldsymbol{X} - \boldsymbol{\mu}) \quad \text{and} \quad \|p_2(\boldsymbol{X} - \boldsymbol{\mu})\|^2$$

are stochastically independent. $p_1(\boldsymbol{X} - \boldsymbol{\mu})$ is normally distributed and $\|p_2(\boldsymbol{X} - \boldsymbol{\mu})\|^2$ is $\chi^2(n-1)$ distributed.

Since

$$p_1(\boldsymbol{X} - \boldsymbol{\mu}) = \begin{bmatrix} \bar{X} - \mu \\ \vdots \\ \bar{X} - \mu \end{bmatrix},$$

and

$$\|p_2(\boldsymbol{X} - \boldsymbol{\mu})\|^2 = \frac{1}{\sigma^2} \sum_1 (X_i - \bar{X})^2,$$

we again find the results of the distribution of $\bar{X}$ and $(n-1)S^2 = \frac{1}{\sigma^2} \sum (X_i - \bar{X})^2$. ♦

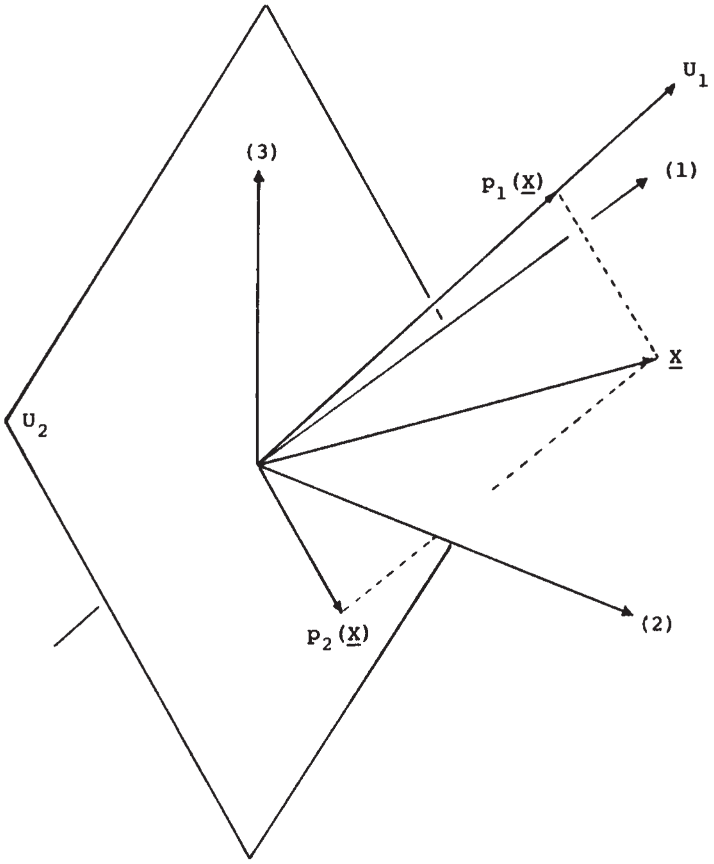


Figure 2.7:

## 2.5 The Wishart distribution and the generalised variance

In the one dimensional case a number of sample-distributions are derived from the normal distribution. The most important of these is the $\chi^2$-distribution, which corresponds to the sum of squared normally distributed data. Its multi-dimensional analog is the Wishart distribution.

We give the definition by means of the density in

**DEFINITION 2.3.** Let $\mathbf{V}$ be a continuously distributed random $p \times p$-matrix, which is symmetrical and positive semi-definite with probability 1. Then $\mathbf{V}$ is said to be

**Wishart distributed** with parameters $(n, \mathbf{\Sigma})$, $(n \geq p)$, if the density for $\mathbf{V}$ is

$$\mathrm{f}(\mathbf{v}) = c \cdot [\det(\mathbf{v})]^{\frac{1}{2}(n-p-1)} \exp(-\frac{1}{2}\,\mathrm{tr}(\mathbf{v} \cdot \mathbf{\Sigma}^{-1})),$$

for $\mathbf{v}$ positive definite and 0 otherwise. Here $\mathbf{\Sigma}$ is a positive definite $p \times p$-matrix, and $c$ is the constant given by

$$\frac{1}{c} = 2^{\frac{1}{2}np} \pi^{p(p-1)/4} (\det \mathbf{\Sigma})^{\frac{1}{2}n} \prod_{i=1}^{p} \Gamma(\frac{1}{2}(n+1-i)).$$

Abbreviated we write

$$\mathbf{V} \in \mathrm{W}(n, \mathbf{\Sigma}) = \mathrm{W}_p(n, \mathbf{\Sigma}).$$

where the first version is used whenever there is doubt about the dimension.

We now give a remark about the mean and variance of the components in a Wishart distribution

Let $\mathbf{V} = (V_{ij})$ be Wishart distributed $\mathrm{W}(n, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma_{ij})$. Then it holds that

$$\begin{aligned}
&\mathrm{E}(V_{ij}) = n\sigma_{ij} \\
&\mathrm{V}(V_{ij}) = n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}) \\
&\mathrm{Cov}(V_{ij}, V_{k\mathrm{l}}) = n(\sigma_{ik}\sigma_{j\mathrm{l}} + \sigma_{i\mathrm{l}}\sigma_{jk}).
\end{aligned}$$

▲

**PROOF 2.22.** Omitted. ■

The analogy with the $\chi^2$-distribution is seen in

**THEOREM 2.25.** Let $\boldsymbol{X}_i \in \mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma})$, $i = 1, \ldots, n$, be independent and regularly distributed. Then for $n \geq p$ it holds that

$$\mathbf{Y} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \in \mathrm{W}(n, \mathbf{\Sigma}).$$

▲

**PROOF 2.23.** Omitted. ■

**REMARK 2.12.** If $n < p$ then $\mathbf{Y}$ as it is defined in the theorem does not have a density function. However, we still choose to say, that $\boldsymbol{Y}$ is Wishart distributed with parameters $(n, \boldsymbol{\Sigma})$.

Corresponding remarks hold if $\boldsymbol{\Sigma}$ is singular. Using this convention the theorem holds without the restriction $n \le p$. ▼

A nearly trivial implication of the above now is

**THEOREM 2.26.** Let $\mathbf{V}_1, \dots, \mathbf{V}_k$ be independent random $p \times p$-matrices, which are $\mathrm{W}(n_i, \boldsymbol{\Sigma})$-distributed. Then it holds

$$\mathbf{V} = \mathbf{V}_1 + \cdots + \mathbf{V}_k \in \mathrm{W}(n_1 + \cdots + n_k, \boldsymbol{\Sigma}).$$

One of the main theorems in the theory of sampling functions of normally distributed random variables is that $\bar{X}$ and $S^2$ are independent and that $S^2$ is $\sigma^2\chi^2/f$-distributed with 1 degree of freedom less than the number of observations. This theorem has its multidimensional analog in ▲

**THEOREM 2.27.** Let $\boldsymbol{X}_i \in \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$, $i = 1, \dots, n$, be stochastically independent. We let

$$\begin{aligned} \bar{\boldsymbol{X}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i, \\ \mathbf{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'. \end{aligned}$$

Then

$$\bar{\boldsymbol{x}} \in \mathrm{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

and

$$\mathbf{S} \in \mathrm{W}(n-1, \frac{1}{n-1}\boldsymbol{\Sigma}).$$

Furthermore, $\bar{\boldsymbol{X}}$ and $\mathbf{S}$ are stochastically independent. ▲

**PROOF 2.24.** Omitted. ■

We will now consider some results on marginal distributions. We have that

**THEOREM 2.28.** Let $\mathbf{V}$ be Wishart distributed with parameters $(n, \mathbf{\Sigma})$. We consider the partitioning

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

It then holds that

$$\mathbf{V}_{ii} \in \mathrm{W}(n, \mathbf{\Sigma}_{ii}).$$

▲

Further, it holds that

**THEOREM 2.29.** We again consider the above situation. If $\mathbf{\Sigma}_{12}$ and $\mathbf{\Sigma}_{21}$ are $\mathbf{0}$-matrices, then $\mathbf{V}_{11}$ and $\mathbf{V}_{22}$ are stochastically independent. ▲

**PROOF 2.25.** for the theorems. They follow readily by considering the corresponding partitions of normally distributed vectors, which produce the Wishart distributions. ■

Since the multidimensional normal distribution can be defined independent of the coordinate system, then it is not surprising that something similar holds for the Wishart distribution. Because change form coordinates in one coordinate system to coordinates in another is performed by manipulating matrices we have the following

**THEOREM 2.30.** Let $\mathbf{V} \in \mathrm{W}_p(n, \mathbf{\Sigma})$ and let $\mathbf{A}$ be an arbitrary fixed $r \times p$-matrix. Then

$$\mathbf{A}\,\mathbf{V}\,\mathbf{A}' \in \mathrm{W}_r(n, \mathbf{A}\,\mathbf{\Sigma}\,\mathbf{A}').$$

▲

**PROOF 2.26.** As indicated above one just has to consider the normally distributed vectors which result in $V$ and then transform them. The resultat then follows readily. ■

We now conclude the chapter by introducing a different generalisation from the one-dimensional variance to the multidimensional case than the variance-covariance matrix.

**DEFINITION 2.4.** Let the $p$-dimensional vector $\boldsymbol{X}$ have the variance-covariance matrix $\boldsymbol{\Sigma}$. By the term **the generalised variance** of $\boldsymbol{X}$ we mean the determinant of the variance-covariance matrix, i.e.

$$\text{gen.var.}(\boldsymbol{X}) = \det(\boldsymbol{\Sigma}).$$

▲

**REMARK 2.13.** In section 1.2.6 we established that the determinant of a matrix corresponds to the volume relationship of the corresponding linear projection, i.e. it is a intuitively sensible measure of the ”size” of a matrix. ▼

If we have observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, then we define the **empirical generalised variance** in a straight forward way from the empirical variance-covariance matrix

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})',$$

i.e. as its determinant.

In the normal case we can establish the distribution of the empirical generalised variance., We have

**THEOREM 2.31.** Let $\boldsymbol{X}_i \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, be stochastically independent. Then the empirical generalised variance follows the same distribution as

$$\frac{\det \boldsymbol{\Sigma}}{(n-1)p} \cdot Z_1 \ldots Z_p,$$

where $Z_1, \ldots, Z_p$ are stochastically independent and $Z_i \in \chi^2(n-i)$. ▲

**PROOF 2.27.** Omitted. ■

For $p = 1$ and 2 it is possible to find the density of the empirical generalised variance. However, for larger values of $p$ this density involves integrals, which cannot readily be written as known functions, but for $n \to \infty$ we do have

**THEOREM 2.32.** Let $\mathbf{S}$ be as above (in the normal case). Then it holds that

$$\sqrt{n-1}\left(\frac{\det(\mathbf{S})}{\det(\boldsymbol{\Sigma})} - 1\right) \quad \text{asymptotically} \quad \in \mathrm{N}(0, 2p).$$

▲

**PROOF 2.28.** Omitted. ■

# Chapter 3

# The general linear model

In this chapter we will formulate a model which is a natural generalisation of the variance and regression analysis models known from introductory statistics. The theorems and definitions will to a large extent be interpreted geometrically in order to give a more intuitive understanding of problems.

## 3.1 Estimation in the general linear model

We first give a description of the model in

### 3.1.1 Formulation of the Model.

We consider an $n$ -dimensional stochastic variable $\boldsymbol{Y} \in \mathrm{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is assumed known. Consider the norm given by $\boldsymbol{\Sigma}^{-1}$ i.e.

$$\|\boldsymbol{x}\|^2 = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}.$$

The norm $(\sigma^2\boldsymbol{\Sigma})^{-1}$ defined by the inverse variance-covariance matrix is given by

$$\|\boldsymbol{x}\|^2_{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \frac{1}{\sigma^2}\|\boldsymbol{x}\|^2.$$

The two norms are seen to be proportional and they result in the same concept of orthogonality. We will now consider a number of problems in connection with the estimation and testing of the mean value $\boldsymbol{\mu}$ in cases where $\boldsymbol{\mu}$ is a known linear function of unknown parameters i.e.

$$\boldsymbol{\mu} = \mathrm{x}\boldsymbol{\theta}$$

or

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix},$$

where $\mathrm{x}$ is assumed known.

Geometrically this can be expressed such that we assume the expected value of the stochastic vector $\boldsymbol{Y}$ is contained in a subspace $M$ of $R^n$. $M$ is the image of $R^k$ corresponding to the linear projection $\mathrm{x}$. The dimension of $M$ is $\mathrm{rg}(\mathrm{x}) \leq k$. The situation is depicted in the following figure.

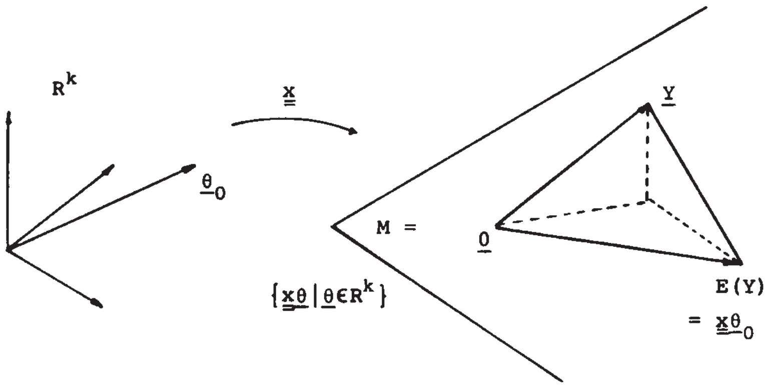

$R^k$
$\underline{\underline{x}}$
$\underline{\theta}_0$
$\underline{Y}$
M =
$\underline{0}$
$\{\underline{\underline{x}}\underline{\theta} \mid \underline{\theta} \in R^k\}$
E(Y)
$= \underline{\underline{x}}\underline{\theta}_0$

Figure 3.1: Geometrical sketch of the general linear model.

We will call such a model, where the unknown mean value $\boldsymbol{\mu}$ is a (known) linear function of the parameter $\boldsymbol{\theta}$ a (general) linear model. This is also valid without the assumption $\boldsymbol{Y}$ has to be normally distributed.

**EXAMPLE 3.1.** Consider an ordinary one-dimensional regression analysis model i.e. we have observations

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\mathrm{E}(\varepsilon_i) = 0$. This model can be written

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

or

$$\boldsymbol{Y} = \mathrm{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

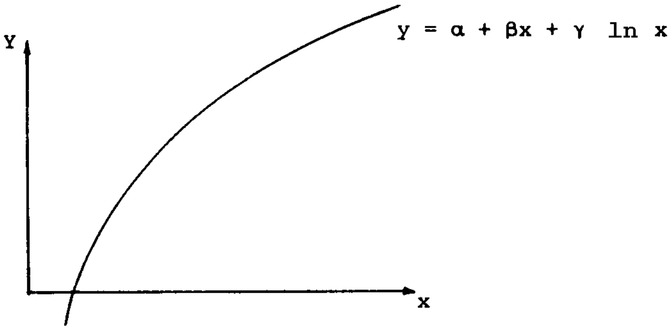
i.e. the model is linear in the meaning stated above. ♦

Another example is

**EXAMPLE 3.2.** We now consider a situation, where

$$Y_i = \alpha + \beta x_i + \gamma \ln x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

and still we have $\mathrm{E}(\varepsilon_i) = 0$. Even in this case we have a linear model which is



$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \ln x_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & \ln x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

We note that the term linear has nothing to do with $\mathrm{E}(Y|X) = \alpha + \beta\, x + \gamma \ln\, x$ being linear in the independent variable $x$, rather that $\mathrm{E}(Y|x)$ considered as a function of the unknown parameter $(\alpha, \beta, \gamma)'$ should be linear. If we had had a model such as

$$Y_i = \alpha + \beta \ln(\gamma x_i + \delta) + \varepsilon_i,$$

where $\alpha, \beta, \gamma$ and $\delta$ are the unknown parameters it would not be possible to write

$$\boldsymbol{Y} = \mathbf{x} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \boldsymbol{\varepsilon}$$

with the known $\mathbf{x}$ -matrix and we would therefore not have a linear model. ♦

### 3.1.2 Estimation in the regular case

We will first formulate the result of estimating $\boldsymbol{\theta}$ in

**THEOREM 3.1.** Let $\mathbf{x}$ and $\boldsymbol{\theta}$ be given as in the preceding section and let $\boldsymbol{Y} \in \mathrm{N}_n(\mathbf{x}\,\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is given by $\mathbf{x}\,\hat{\boldsymbol{\theta}}$ being the projection (with respect to $\boldsymbol{\Sigma}$ ) onto $M$, $\hat{\boldsymbol{\theta}}$ is a solution to the so-called normal equation(s)

$$(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

If $\mathbf{x}$ has full rank $k$, then

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

and since a linear combination of normally distributed variables $\hat{\boldsymbol{\theta}}$ is also normally distributed with parameters

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathrm{D}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}. \end{aligned}$$

It is especially noted that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$. ▲

**PROOF 3.1.** If $\boldsymbol{Y} \in \mathrm{N}(\mathbf{x}\,\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is regular then the density for $\boldsymbol{Y}$

$$\begin{aligned} \mathrm{f}(\boldsymbol{y}) &= \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sigma^n}\frac{1}{\sqrt{\det\boldsymbol{\Sigma}}}\exp[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\mathbf{x}\,\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\mathbf{x}\,\boldsymbol{\theta})] \\ &= k\cdot\frac{1}{\sigma^n}\exp[-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\mathbf{x}\,\boldsymbol{\theta}\|^2]. \end{aligned}$$

We have the likelihood function

$$\mathrm{L}(\boldsymbol{\theta}) = k\cdot\frac{1}{\sigma^n}\exp[-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\mathbf{x}\,\boldsymbol{\theta}\|^2],$$

taking the logarithm on each side gives

$$\ln \mathrm{L}(\boldsymbol{\theta}) = k_1 - \frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathrm{x}\,\boldsymbol{\theta}\|^2.$$

It is now evident that maximisation of the likelihood function is equivalent to minimisation of the squared distance between any point in $M$ and the observation i.e. equivalent to minimisation of

$$\|\boldsymbol{y} - \mathrm{x}\,\boldsymbol{\theta}\|^2 = (\boldsymbol{y} - \mathrm{x}\,\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mathrm{x}\,\boldsymbol{\theta}).$$

From the result p. 52 the value of $\mathrm{x}\,\boldsymbol{\theta},$ giving the minimum is equal to the orthogonal projection (with respect to $\boldsymbol{\Sigma}^{-1}$) of $\boldsymbol{y}$ on $M$. From example 1.8 p. 48 the optimal $\boldsymbol{\theta}$ is the solution to the equation

$$(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})\boldsymbol{\theta} = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

If $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ has full rank $k$, i.e. if $\mathbf{x}$ has rank $k$ (cf. p. 35) we therefore have

$$\boldsymbol{\theta}_{\text{opt.}} = (\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

We have now shown the first half of the theorem.

From theorem 2.2 we find that

$$\mathrm{E}(\hat{\boldsymbol{\theta}}) = (\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathrm{x}\,\boldsymbol{\theta} = \boldsymbol{\theta},$$

And from theorem 2.5 we find

$$\begin{aligned}\mathrm{D}(\hat{\boldsymbol{\theta}}) &= (\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Sigma}^{-1}(\sigma^2\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{x}(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\\ &= \sigma^2(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1},\end{aligned}$$

■

The situation is illustrated in the following figure 3.2.

**REMARK 3.1.** We note that $\boldsymbol{\theta}$ is estimated by minimising the squared distance onto $M$. $\hat{\boldsymbol{\theta}}$ is therefore also a least squares estimate of $\boldsymbol{\theta}$. If we do not have the distributional assumption we will often be able to use the estimator $\hat{\boldsymbol{\theta}}$ in theorem 3.1 as an estimate of $\boldsymbol{\theta}$. It can be shown that the least squares estimator $\hat{\boldsymbol{\theta}}$ has the least generalised variance among all the estimators that are linear functions of the observations (the so-called Gauss-Markov theorem) cf. [12]. ▼

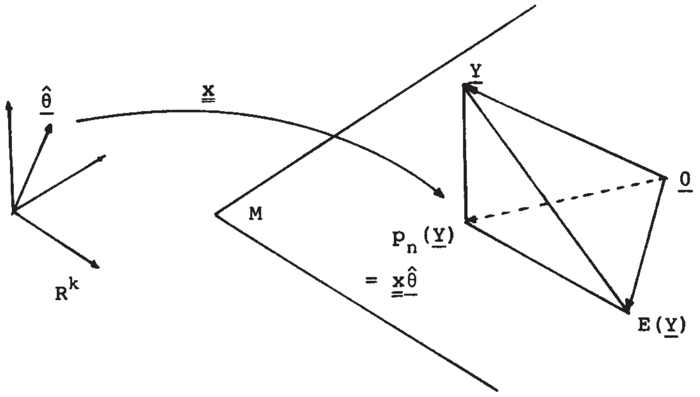

$\hat{\underline{\theta}}$
$\underline{\underline{x}}$
$\underline{Y}$
$\underline{0}$
M
$R^k$
$p_n(\underline{Y})$
$= \underline{\underline{x}}\hat{\underline{\theta}}$
$E(\underline{Y})$

Figure 3.2: Geometric sketch of the problem of estimation in the general linear model.

Since $\sigma^2$ is often unknown we will now find estimators for it. We have

**THEOREM 3.2.** Let the situation be as above. The maximum likelihood estimator of $\sigma^2$ is

$$\frac{1}{n}\|\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2.$$

The unbiased estimator of $\sigma^2$ is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - \operatorname{rg}\mathbf{x}}\|\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 \\ &= \frac{1}{n - \operatorname{rg}\mathbf{x}}(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})\end{aligned}$$

where $\mathbf{x}\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\mathrm{E}(\boldsymbol{Y})$. The following holds

$$\hat{\sigma}^2 \in \sigma^2\chi^2(n - \operatorname{rg}\mathbf{x})/(n - \operatorname{rg}\mathbf{x})$$

and $\hat{\sigma}^2$ is independent of the maximum likelihood estimator of the expected value and is therefore independent of $\hat{\boldsymbol{\theta}}$. ▲

**PROOF 3.2.** The likelihood function is

$$\mathrm{L}(\boldsymbol{\theta}, \sigma^2) = k \cdot \frac{1}{\sigma^n}\exp[-\frac{1}{2}\frac{1}{\sigma^2}\|\boldsymbol{y} - \mathbf{x}\boldsymbol{\theta}\|^2],$$

and

$$\ln \mathrm{L}(\theta, \sigma^2) = k_1 - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\frac{1}{\sigma^2}\|\boldsymbol{y} - \mathbf{x}\boldsymbol{\theta}\|^2.$$

now

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\ln \mathrm{L} &= -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\|\boldsymbol{y}-\mathbf{x}\boldsymbol{\theta}\|^2\\ &= -\frac{n}{2}\frac{1}{\sigma^4}(\sigma^2-\frac{1}{n}\|\boldsymbol{y}-\mathbf{x}\boldsymbol{\theta}\|^2).\end{aligned}$$

After differentiating with respect to $\boldsymbol{\theta}$ we get the ordinary system of normal equations. We therefore find that the maximum likelihood estimates to $(\hat{\boldsymbol{\theta}},\hat{\sigma}^2)$ for $(\boldsymbol{\theta},\sigma^2)$ are solutions for

$$\begin{aligned}&\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}\hat{\boldsymbol{\theta}}=\mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}\\ &\hat{\sigma}^2=\frac{1}{n}\|\boldsymbol{Y}-\mathbf{x}\hat{\boldsymbol{\theta}}\|^2=\frac{1}{n}(\boldsymbol{Y}-\mathbf{x}\hat{\boldsymbol{\theta}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\mathbf{x}\hat{\boldsymbol{\theta}}).\end{aligned}$$

If we consider the partitioning of $R^n$ as the direct sum of $M$ and $M^{\perp}$, where $M^{\perp}$ is the orthogonal component (with respect to $\boldsymbol{\Sigma}^{-1}$) of $M$, we get that

$$\mathrm{P}_M(\boldsymbol{Y}-\mathbf{x}\boldsymbol{\theta})=\mathbf{x}\hat{\boldsymbol{\theta}}-\mathbf{x}\boldsymbol{\theta}$$

and

$$\boldsymbol{Y}-\mathbf{x}\hat{\boldsymbol{\theta}}$$

are stochastically independent and that

$$\begin{aligned}&\|\boldsymbol{Y}-x\hat{\boldsymbol{\theta}}\|^2\in\sigma^2\chi^2(\dim M^{\perp})\\ &=\sigma^2\chi^2(n-\mathrm{rg}\,\mathbf{x}).\end{aligned}$$

From this we especially get

$$\mathrm{E}(\hat{\sigma}^2)=\frac{1}{n}(n-\mathrm{rg}\,\mathbf{x})\sigma^2,$$

i.e. the likelihood estimator of $\sigma^2$ is not unbiased. If we want an unbiased estimate we can obviously use

$$\frac{1}{n-\mathrm{rg}\,\mathbf{x}}\|\boldsymbol{Y}-\mathbf{x}\hat{\boldsymbol{\theta}}\|^2.$$

Most often we will be using the unbiased estimate of $\sigma^2$, and we will therefore use the notation $\hat{\sigma}^2$ for this. ■

**REMARK 3.2.** If $\boldsymbol{\Sigma}$ is the identity matrix then $\|\boldsymbol{y}\|^2 = \sum y_i^2$. So in this case we have

$$\hat{\sigma}^2 = \frac{1}{n - \operatorname{rg}\mathbf{x}} \sum_{i=1}^{n} (Y_i - \hat{\mathrm{E}}(Y_i))^2,$$

where $\hat{\mathrm{E}}(Y_i) = (\mathbf{x}\hat{\boldsymbol{\theta}})_i$. The quantity $Y_i - \hat{\mathrm{E}}(i)$ is equal to the $i$'th observations deviation from the estimated model, and it is called the $i$'th residual. In the case $\boldsymbol{\Sigma} = \mathbf{I}$, we have that the estimate of variances proportional to the sum of the squared residuals called $\mathrm{SS}_{\mathrm{res}}$. We will generally use this notation for the squared distance between the observation and the estimated model i.e.

$$\mathrm{SS}_{\mathrm{res}} = \|\boldsymbol{Y} - \mathbf{x}\boldsymbol{\theta}\|^2 = (\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}).$$

▼

Before we will go on we will give a small example for the purpose of illustration.

**EXAMPLE 3.3.** In the production of a certain synthetic product two raw materials A and B are mainly used. The quality of the end product can be described by a stochastic variable which is normally distributed with mean value $\mu$ and variance $\sigma^2$. The mean-value is known to depend linearly on the added amount of A and B respectively i.e.

$$\mu = x_{\mathrm{A}}\theta_{\mathrm{A}} + x_{\mathrm{B}}\theta_{\mathrm{B}},$$

where $x_{\mathrm{A}}$ is the added amount of A and $x_{\mathrm{B}}$ is the corresponding added amount of B. $\sigma^2$ is assumed to be independent of the added amount of raw-materials. For the determination of $\theta_{\mathrm{A}}$ and $\theta_{\mathrm{B}}$ three experiments were performed after the following plan.

| Experiment | Content of A | Content of B |
|---|---|---|
| 1 | 100% | 0% |
| 2 | 0% | 100% |
| 3 | 50% | 50% |

The single experiments are assumed to be stochastically independent. The simultaneous distribution of the experimental results $Y_1, Y_2, Y_3$ is then a three dimensional normal distribution with mean value

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_{\mathrm{A}} \\ \theta_{\mathrm{B}} \end{bmatrix} = \mathbf{x}\boldsymbol{\theta},$$

and variance-covariance matrix $\sigma^2\mathbf{I}$.

We have

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} \frac{5}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{5}{4} \end{bmatrix} \quad \Rightarrow \quad (\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix},$$

and

$$\mathbf{x}'\mathbf{y} = \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix},$$

giving

$$\begin{bmatrix} \hat{\theta}_{\mathrm{A}} \\ \hat{\theta}_{\mathrm{B}} \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix} \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix} = \begin{bmatrix} \frac{5}{6}y_1 - \frac{1}{6}y_2 + \frac{1}{3}y_3 \\ -\frac{1}{6}y_1 + \frac{5}{6}y_2 + \frac{1}{3}y_3 \end{bmatrix}.$$

In this case we observed

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 10.11 \\ 0.81 \\ 5.24 \end{bmatrix},$$

so that

$$\begin{bmatrix} \hat{\theta}_{\mathrm{A}} \\ \hat{\theta}_{\mathrm{B}} \end{bmatrix} = \begin{bmatrix} 10.037 \\ 0.735 \end{bmatrix}.$$

From this we easily find

$$\hat{\mathrm{E}}(\boldsymbol{Y}) = \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 10.037 \\ 0.735 \\ 5.386 \end{bmatrix},$$

and

$$\boldsymbol{Y} - \hat{\mathrm{E}}(\boldsymbol{Y}) = \boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 0.07 \\ 0.07 \\ -0.15 \end{bmatrix}.$$

This gives the residual sum of squares

$$(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = 0.07^2 + 0.07^2 + 0.15^2 = 0.0338,$$

which means that an unbiased estimate of $\sigma^2$ is

$$\frac{1}{3-2}0.0338 = 0.0338.$$

♦

### 3.1.3 The case of $\mathrm{x}'\Sigma^{-1}\mathrm{x}$ singular

If $\mathrm{rg}(\mathbf{x}) = p < k$ then $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ is singular and we cannot find an ordinary solution to the equation.

$$(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

If we can find a pseudo inverse for $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ then we can write

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

However, sometimes it is possible to use a little trick in the determination of the pseudo inverse. The reason for the singularity is that we have too many parameters. It would therefore be reasonable to restrict $\boldsymbol{\theta}$ to only vary freely in a (side-)subspace of $R^k$. One of those could e.g. be determined by $\boldsymbol{\theta}$ satisfying the linear equations (restrictions)

$$\mathbf{b}\,\boldsymbol{\theta} = \boldsymbol{c}$$

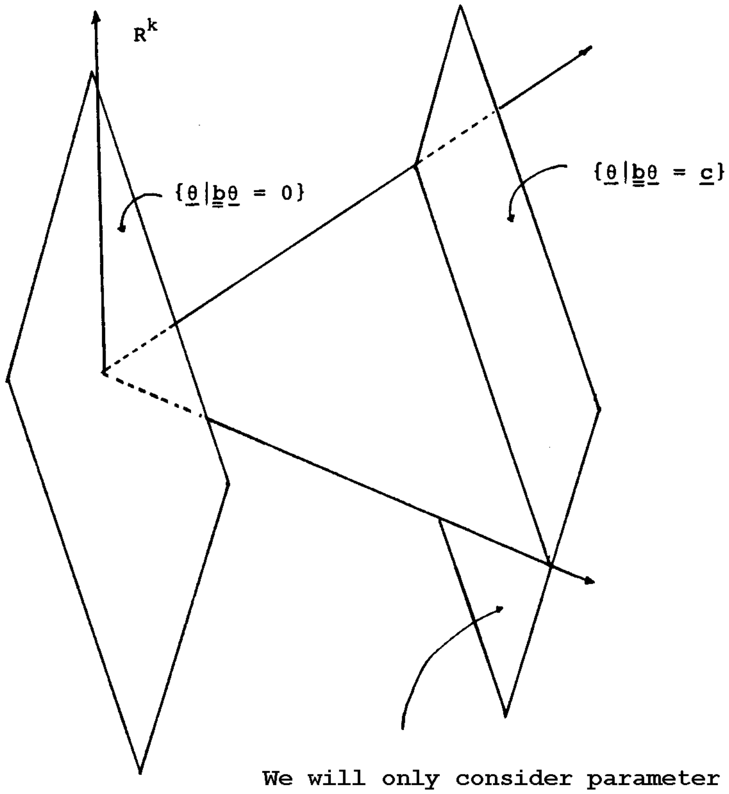
or

$$\begin{bmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}.$$

If there exist $\boldsymbol{\theta}$ s that satisfy this equation system then they span a subspace of dimension $k - \mathrm{rg}(\mathbf{b})$.

Since

$\mathrm{rg}(\mathbf{x}) = p,$ and we have $k\ \theta$ -components it would be reasonable to remove $k-p$ of these i.e. impose the restriction $k - \mathrm{rg}(\mathbf{b}) = p$ or $k = p + \mathrm{rg}(\mathbf{b})$.


$R^k$
$\{\underline{\theta} | \underline{\underline{b}}\underline{\theta} = 0\}$
$\{\underline{\theta} | \underline{\underline{b}}\underline{\theta} = \underline{c}\}$

We will only consider parameter values $\underline{\theta}$, which lie in this side-subspace in $R^k$.

Now if

$$\mathrm{rg}\begin{bmatrix}\mathbf{x}\\ \mathbf{b}\end{bmatrix}=\mathrm{rg}\begin{bmatrix}x_{11} & \cdots & x_{1k}\\ \vdots & & \vdots\\ x_{n1} & \cdots & x_{nk}\\ b_{11} & \cdots & b_{1k}\\ \vdots & & \vdots\\ b_{m1} & \cdots & b_{mk}\end{bmatrix}=k,$$

we can now consider the model

$$\begin{bmatrix}\boldsymbol{Y}\\ \mathbf{c}\end{bmatrix}=\begin{bmatrix}\mathbf{x}\\ \mathbf{b}\end{bmatrix}\boldsymbol{\theta}+\begin{bmatrix}\varepsilon\\ \mathbf{0}\end{bmatrix}.$$

We let

$$\mathbf{D}=\begin{bmatrix}\boldsymbol{\Sigma}^{-1} & \mathbf{0}_{n,m}\\ \mathbf{0}_{m,n} & \mathbf{I}_{m,m}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\Sigma}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix},$$

where the short notation should not cause confusion.

If we in the usual way compute

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \{[\mathbf{x}'\mathbf{b}']\mathbf{D}\begin{bmatrix}\mathbf{x}\\ \mathbf{b}\end{bmatrix}\}^{-1}\{[\mathbf{x}'\mathbf{b}']\mathbf{D}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{c}\end{bmatrix}\}\\ &= \{\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{b}'\mathbf{b}\}^{-1}\{\mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}+\mathbf{b}'\mathbf{c}\},\end{aligned}$$

then we have a quantity which minimises

$$\begin{aligned}\mathrm{g}(\boldsymbol{\theta}) &= \{\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{c}\end{bmatrix}-\begin{bmatrix}\mathbf{x}\\ \mathbf{b}\end{bmatrix}\boldsymbol{\theta}\}'\mathbf{D}\{\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{c}\end{bmatrix}-\begin{bmatrix}\mathbf{x}\\ \mathbf{b}\end{bmatrix}\boldsymbol{\theta}\}\\ &= \begin{bmatrix}\boldsymbol{y}-\mathbf{x}\boldsymbol{\theta}\\ \mathbf{0}\end{bmatrix}'\begin{bmatrix}\boldsymbol{\Sigma}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{y}-\mathbf{x}\boldsymbol{\theta}\\ \mathbf{0}\end{bmatrix}\\ &= (\boldsymbol{y}-\mathbf{x}\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\mathbf{x}\boldsymbol{\theta})\\ &= \|\boldsymbol{y}-\mathbf{x}\boldsymbol{\theta}\|^2.\end{aligned}$$

Since this is exactly the same quantity we must determine in order to find the ML-estimates, we therefore find that

$$\hat{\boldsymbol{\theta}}=\{\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{b}'\mathbf{b}\}^{-1}\{\mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}+\mathbf{b}'\boldsymbol{c}\}$$

really is the maximum likelihood estimator for $\boldsymbol{\theta}$. The only requirement is that we must find a matrix $\mathbf{b}$ so $\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}$ has full rank and this corresponds to restricting $\boldsymbol{\theta}$'s region of variation.

The variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ becomes

$$\mathrm{D}(\hat{\boldsymbol{\theta}}) = \sigma^2\{\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{b}'\mathbf{b}\}^{-1}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}\{\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{b}'\mathbf{b}\}^{-1}.$$

This expression is found immediately by using theorem 2.5.

As before the unbiased estimate of $\sigma^2$ is

$$\frac{1}{n - \operatorname{rg}\mathbf{x}}\|\boldsymbol{y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2$$

Here we have $n - \operatorname{rg}\mathbf{x} = n - k + \operatorname{rg}\mathbf{b}$.

First we give a little theoretical

**EXAMPLE 3.4.** Consider a very simple one-sided analysis of variance with two groups with two observations in each group. We could imagine that we were examining the effect of a catalyst on the results of some process. We therefore conduct four experiments, two with the catalyst at level A and two with the catalyser at level B. We therefore have the following observations

level A: $Y_{11}, Y_{12}$
level B: $Y_{21}, Y_{22}$

If we assume that the observations are stochastically independent and have mean values

$$\begin{aligned} \mathrm{E}(Y_{11}) &= \mathrm{E}(Y_{12}) = \theta_1 \\ \mathrm{E}(Y_{21}) &= \mathrm{E}(Y_{22}) = \theta_2, \end{aligned}$$

then we can express the model as

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{x}\,\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

We easily find that

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

and

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix},$$

which are the usual estimators. If we instead use the (commonly used) parametrisation

$$\begin{aligned} \mathrm{E}(Y_{11}) &= \mathrm{E}(Y_{12}) = \mu + \alpha_1 \\ \mathrm{E}(Y_{21}) &= \mathrm{E}(Y_{22}) = \mu + \alpha_2 \end{aligned}$$

i.e. we express the effect of a catalyst as a level plus the specific effect of that catalyst. Then we have

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{x}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

It is easily seen that $\mathbf{x}$ has rank 2 (the sum of the last two columns equals the first). We will therefore try to introduce a linear restriction between the parameters. We will try with

$$\alpha_1 + \alpha_2 = 0 \quad \text{i.e.:} \quad \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = 0.$$

We can now formally introduce the model

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ 0 \end{bmatrix},$$

or

$$\begin{bmatrix} \boldsymbol{Y} \\ 0 \end{bmatrix} = \begin{bmatrix} & \mathbf{x} & \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ 0 \end{bmatrix}.$$

We now have that

$$\begin{bmatrix} & \mathbf{x} & \\ 0 & 1 & 1 \end{bmatrix}' \begin{bmatrix} & \mathbf{x} & \\ 0 & 1 & 1 \end{bmatrix} = \mathbf{x}'\mathbf{x} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

The inverse of this matrix is

$$\begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}.$$

Now, since

$$\begin{bmatrix} & \mathbf{x} & \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} \sum y_{ij} \\ y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix},$$

we have

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha_1} \\ \hat{\alpha_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \sum y_{ij} \\ y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{y}_1 - \bar{y} \\ \bar{y}_2 - \bar{y} \end{bmatrix},$$

i.e. exactly the same estimators we are used to from a balanced one-sided analysis of variance (note: We know in beforehand that we will get these estimators. cf. p. 119). ♦

We will now give a more practical example of the estimation of parameters in the case where $\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}$ is singular.

**EXAMPLE 3.5.** In the production of enzymes one can use two principally different types of bacteria. Via its metabolism one type of bacterie liberates acid during the production (acid producer). The other produces neutral metabolic products. In order to regulate the pH-value in the substrate on which the bacterias are produced, one can add a so-called pH-buffer. It is known, that the pH-buffer itself does not have any effect on the production of the enzyme, rather it works through an interaction with the acid content and the metabolic products of the bacteria.

For a ”neutral” type of bacteria which lives on a substrate without pH-buffer the mean production of enzyme (normal production) is known. In order to estimate the above mentioned interactions one has measured the difference between the normal production and the actual production of enzyme in 7 experiments as shown below.

First we will formulate a mathematical model that can describe the above mentioned experiment.

| | | pH-buffer | |
|---|---|---|---|
| | | added | not added |
| bacteria culture | acid producer | 0,-2 | -19,-15 |
| | neutral | -6, 0,-2 | |

Table 3.1: Differences between nominal yield and actual yield under different experimental circumstances.

We have observations

$$\begin{aligned} y_{11\nu}, \quad \nu &= 1,2 \\ y_{12\nu}, \quad \nu &= 1,2 \\ y_{21\nu}, \quad \nu &= 1,2,3. \end{aligned}$$

These are assumed to have the mean values

$$\begin{aligned} \mathrm{E}(y_{11\nu}) &= \mu_1 + \theta_{11} \\ \mathrm{E}(y_{12\nu}) &= \mu_1 + \theta_{12} \\ \mathrm{E}(y_{21\nu}) &= \theta_{21}, \end{aligned}$$

where $\mu_1$ is the effect of using acid producing bacteria and $\theta_{ij}$ is the interaction between pH-buffer and bacteria culture.

Furthermore we assume that the observations are stochastically independent and we have the same but unknown variance $\sigma^2$.

We can now formulate the model as a general linear model. We have

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{213} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \end{bmatrix} + \boldsymbol{\varepsilon},$$

where the error $\boldsymbol{\varepsilon} \in \mathrm{N}_7(\mathbf{0}, \sigma^2 \mathbf{I})$.

We find

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

and

$$\mathbf{x}'\boldsymbol{y} = \begin{bmatrix} y_{1..} \\ y_{11.} \\ y_{12.} \\ y_{21.} \end{bmatrix},$$

where a dot as an index-value indicates that we have summed over the corresponding index.

Since $\mathbf{x}'\mathbf{x}$ only has the rank 3, we are unable to invert it. Instead we can find a pseudo-inverse. We use the theorem 1.7 p. 27 and get

$$(\mathbf{x}'\mathbf{x})^{-} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix},$$

so the estimates from the parameters become - with this special choice of pseudo-inverse -

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'\boldsymbol{y} = \begin{bmatrix} 0 \\ \bar{y}_{11.} \\ \bar{y}_{12.} \\ \bar{y}_{21.} \end{bmatrix},$$

where e.g.

$$\bar{y}_{21.} = \frac{1}{3}\sum_{\nu=1}^{3} y_{21\nu}.$$

Now, since

$$\mathbf{I} - (\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

we have

$$(\mathbf{I} - (\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'\mathbf{x})\boldsymbol{z} = \begin{bmatrix} z_1 \\ -z_1 \\ -z_1 \\ 0 \end{bmatrix}$$

From theorem 1.6 the complete solution to the normal equations is therefore all vectors of the form

$$\hat{\boldsymbol{\theta}} + \begin{bmatrix} t \\ -t \\ -t \\ 0 \end{bmatrix} = \begin{bmatrix} t \\ \bar{y}_{11.} - t \\ \bar{y}_{12.} - t \\ \bar{y}_{21.} \end{bmatrix}, \qquad t \in R.$$

An arbitrary maximum likelihood estimator for $\boldsymbol{\theta}$ is then of this form.

The observed value of $\hat{\theta}$ is

$$\hat{\theta}_{\text{obs}} = \begin{bmatrix} 0 \\ -1 \\ -17 \\ -2\frac{2}{3} \end{bmatrix}.$$

It is obvious that this estimator is not very satisfactory since e.g. $\hat{\mu}_1$ always will be 0. In order to get estimators which correspond to our expectations about physical reality we must impose some constraints on the parameters. It seems reasonable to demand that

$$\theta_{11} + \theta_{12} = 0,$$

i.e.

$$\begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \end{bmatrix} = 0,$$

or

$$\mathbf{b}\,\boldsymbol{\theta} = 0.$$

It is obvious that

$$\text{rg}(\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}) = 4,$$

so we can use the result from p. 119. We find

$$\begin{aligned}\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b} &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 3 & 1 & 0 \\ 2 & 1 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.\end{aligned}$$

Since

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix},$$

we find

$$(\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b})^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

We now get

$$\hat{\theta} = (\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b})^{-1}\mathbf{x}'\boldsymbol{y} = \begin{bmatrix} \bar{y}_{1\cdot\cdot} \\ \bar{y}_{11\cdot} - \bar{y}_{1\cdot\cdot} \\ \bar{y}_{12\cdot} - \bar{y}_{1\cdot\cdot} \\ \bar{y}_{21\cdot} \end{bmatrix}.$$

The observed value is

$$\begin{bmatrix} -9 \\ 8 \\ -8 \\ -2\frac{2}{3} \end{bmatrix} \left( = \begin{bmatrix} \text{acid producing effect} \\ \text{buffer \& acid interaction} \\ \text{(-buffer) \& acid interaction} \\ \text{buffer \& neutral interaction} \end{bmatrix} \right).$$

We now find the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$. We have

$$\begin{aligned} \mathrm{D}(\hat{\theta}) &= \sigma^2(\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b})^{-1}\mathbf{x}'\mathbf{x}(\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b})^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}, \end{aligned}$$

i.e. the estimators are not independent.

In order to estimate $\sigma^2$ we find the vector of residuals. Since

$$\mathrm{x}\hat{\theta} = \begin{bmatrix} \hat{\mu}_1 + \hat{\theta}_{11} \\ \hat{\mu}_1 + \hat{\theta}_{11} \\ \hat{\mu}_1 + \hat{\theta}_{12} \\ \hat{\mu}_1 + \hat{\theta}_{12} \\ \hat{\theta}_{21} \\ \hat{\theta}_{21} \\ \hat{\theta}_{21} \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -17 \\ -17 \\ -2\frac{2}{3} \\ -2\frac{2}{3} \\ -2\frac{2}{3} \end{bmatrix},$$

the vector of residuals is

$$\boldsymbol{y} - \mathrm{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 1 \\ -1 \\ -2 \\ 2 \\ -3\frac{1}{3} \\ 2\frac{2}{3} \\ \frac{2}{3} \end{bmatrix}.$$

We then find

$$\|\boldsymbol{y} - \mathrm{x}\hat{\boldsymbol{\theta}}\|^2 = (\boldsymbol{y} - \mathrm{x}\hat{\boldsymbol{\theta}})'(\boldsymbol{y} - \mathrm{x}\hat{\boldsymbol{\theta}}) = 1^2 + \cdots + (\frac{2}{3})^2 = 28\frac{2}{3}.$$

**An unbiased estimate of $\boldsymbol{\sigma^2}$** is therefore

$$s^2 = \frac{1}{7-3} \cdot 28\frac{2}{3} = 7\frac{1}{6}.$$

♦

### 3.1.4 Constrained estimation

This section is omitted.

### 3.1.5 Confidence-intervals for estimated values. Prediction-intervals

We consider the model $(n > k)$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where

$$\boldsymbol{\varepsilon} \in \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{\Sigma}).$$

Here we will denote the $Y$ 's as dependent variables and the $x$ 's as the independent variables.

As usual $\sigma^2$ is (assumed) unknown and $\mathbf{\Sigma}$ is (assumed) known. We have the estimator

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{\Sigma}^{-1}\boldsymbol{Y}$$

for $\boldsymbol{\theta}$ and $\sigma^2$ is estimated using

$$\begin{aligned} \hat{\sigma}^2 &= s^2 = \frac{1}{n-k}\|\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 \\ &= \frac{1}{n-k}(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'\mathbf{\Sigma}^{-1}(\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}). \end{aligned}$$

If we wish to predict the expected value of the new observation $Y$ of the dependent variable corresponding to the values of the independent variables:

$$(z_1, \dots, z_k) = \boldsymbol{z}'$$

it is obvious that we will use

$$Z = (z_1, \dots, z_k) \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_k \end{bmatrix} = \boldsymbol{z}'\hat{\boldsymbol{\theta}}$$

as our ”best” guess.

We have that $\mathrm{E}(Z) = \mathrm{E}(Y)$ and that

$$\begin{aligned} V(Z) &= \boldsymbol{z}' \, \mathrm{D}(\hat{\boldsymbol{\theta}}) \boldsymbol{z} \\ &= \sigma^2 \boldsymbol{z}' (\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \boldsymbol{z} \\ &= \sigma^2 c, \end{aligned}$$

where

$$c = (z_1, \ldots, z_k)(\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix}.$$

We therefore immediately have

$$\frac{Z - \mathrm{E}(Y)}{\sigma\sqrt{c}} \in \mathrm{N}(0,1),$$

and therefore also

$$\frac{Z - \mathrm{E}(Y)}{S\sqrt{c}} \in \mathrm{t}(n-k).$$

We are now able to formulate and prove

**THEOREM 3.3.** Let the situation be as above. Then the $(1-\alpha)$ -confidence interval for the expected value of a new observation $Y$ will be

$$[z - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}, \quad z + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}].$$

▲

**PROOF 3.3.** From the above considerations we immediately have

$$1 - \alpha = P\{Z - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c} \;\le\; \mathrm{E}(Y) \;\le\; Z + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}\},$$

and therefore also have the theorem. ■

Often one is more interested in a confidence interval for the new (or future) observations than for the expected value of the observations. We now consider the more general problem of determining the confidence interval for the average $\bar{Y}_q$ of $q$ observations taken at $(z_1, \dots, z_k)$. If $Y_{iq} \in \mathrm{N}(\mathrm{E}(Y), c_1\sigma^2)$, then we have that

$$\bar{Y}_q \in \mathrm{N}(\mathrm{E}(Y), \frac{c_1}{q}\sigma^2).$$

If we now assume that the new (or future) observations are independent of those we already have then

$$Z - \bar{Y}_q \in \mathrm{N}(0, \sigma^2(c + \frac{c_1}{q})),$$

i.e.

$$\frac{Z - \bar{Y}_q}{S\sqrt{c + \frac{c_1}{q}}} \in \mathrm{t}(n-k).$$

From this we can as before derive

**THEOREM 3.4.** Let us assume that $q$ new observations taken at $(z_1, \dots, z_k)$ each have a variance $c_1\sigma^2$ Furthermore, they are independent of each other and independent of the earlier observations. In that case a $(1-\alpha)$ confidence interval for the average of the $q$ observations equals the interval

$$[z - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c + \frac{c_1}{q}}, z + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c + \frac{c_1}{q}}].$$

▲

**REMARK 3.3.** The above mentioned interval is a confidence interval for an observation and not for a parameter as we are used to. One therefore often speaks of a prediction interval in order to distinguish between the two situations. ▼

**REMARK 3.4.** We see that the correspondence to the interval for $\bar{Y}_q$ instead of the interval for $\mathrm{E}(\bar{Y}_q) = \mathrm{E}(Y)$ just consists of the expression under the square root sign being larger by an amount equal to $\frac{c_1}{q}$ which is the variance of $\frac{\bar{Y}_q}{\sigma}$. ▼

**EXAMPLE 3.6.** We consider the following corresponding observations of an independent variable $x$ and a dependent variable $y$:

| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| y | 0.4 | 0.3 | 1.5 | 1.3 | 1.9 | 4.2 | 8 |

We assume that the $y$ 's originate from independent stochastic variables $Y_1, \ldots, Y_7$ which are normally distributed with mean values

$$\mathrm{E}(Y|x) = \beta x^2$$

and variances

$$\mathrm{V}(Y|0) = \sigma^2, \quad \mathrm{V}(Y|x) = x^2\sigma^2, \qquad x > 0.$$

We would now like to find a confidence interval for a new (or future) observation corresponding to $x = 10$. This observation is called $Y$, and we have

$$\begin{aligned} \mathrm{E}(Y) &= 100\beta \\ \mathrm{V}(Y) &= 100\sigma^2 . \end{aligned}$$

We now reformulate the problem in matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \\ \cdot \\ \cdot \\ \cdot \\ \\ \varepsilon_7 \end{bmatrix} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathrm{D}(\boldsymbol{\varepsilon}) = \sigma^2 \begin{bmatrix} 1 & & \cdot & \cdot & \cdot & & 0 \\ & 1 & & & & & \\ \cdot & & 4 & & & & \cdot \\ \cdot & & & 9 & & & \cdot \\ \cdot & & & & 16 & & \cdot \\ & & & & & 25 & \\ 0 & & \cdot & \cdot & \cdot & & 36 \end{bmatrix} = \sigma^2 \boldsymbol{\Sigma}.$$

We have that

$$\begin{aligned} \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} &= (0,1,4,9,16,25,36)\,\mathrm{diag}(1,1,\frac{1}{4},\ldots,\frac{1}{36}) \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 36 \end{bmatrix} \\ &= 91. \\ \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} &= 0.3 + 1.5 + 1.3 + 1.9 + 4.2 + 8.0 = 17.2. \end{aligned}$$

so

$$\hat{\beta} = \frac{17.2}{91} = 0.1890,$$

and

$$P_M(\boldsymbol{y}) = \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \cdot 0.1890 = \begin{bmatrix} 0 \\ 0.1890 \\ 0.7560 \\ 1.7010 \\ 3.0240 \\ 4.7250 \\ 6.8040 \end{bmatrix}.$$

The residuals are

$$\boldsymbol{y} - P_M(\boldsymbol{y}) = \begin{bmatrix} 0.4000 \\ 0.1110 \\ 0.7440 \\ -0.4010 \\ -1.1240 \\ -0.5250 \\ 1.1960 \end{bmatrix},$$

so

$$\begin{aligned} \|\boldsymbol{y} - P_M(\boldsymbol{y})\|^2 &= (0.4000 \cdots 1.1960) \begin{bmatrix} \frac{1}{1} & & & \\ & \frac{1}{1} & & \\ & & \ddots & \\ & & & \frac{1}{36} \end{bmatrix} \begin{bmatrix} 0.4000 \\ \vdots \\ 1.1960 \end{bmatrix} \\ &= 0.45829 \end{aligned}$$

i.e.

$$\hat{\sigma^2} = s^2 = \frac{1}{7-1} 0.45829 = 0.07638 = 0.27637^2.$$

The constants $c$ and $c_1$ are equal to

$$\begin{aligned} c &= 100 \cdot \frac{1}{91} \cdot 100 = 109.89 \\ c_1 &= 10^2 = 100. \end{aligned}$$

The prediction for $x$ = 10 is

$$z = 100\hat{\beta} = 18.90$$

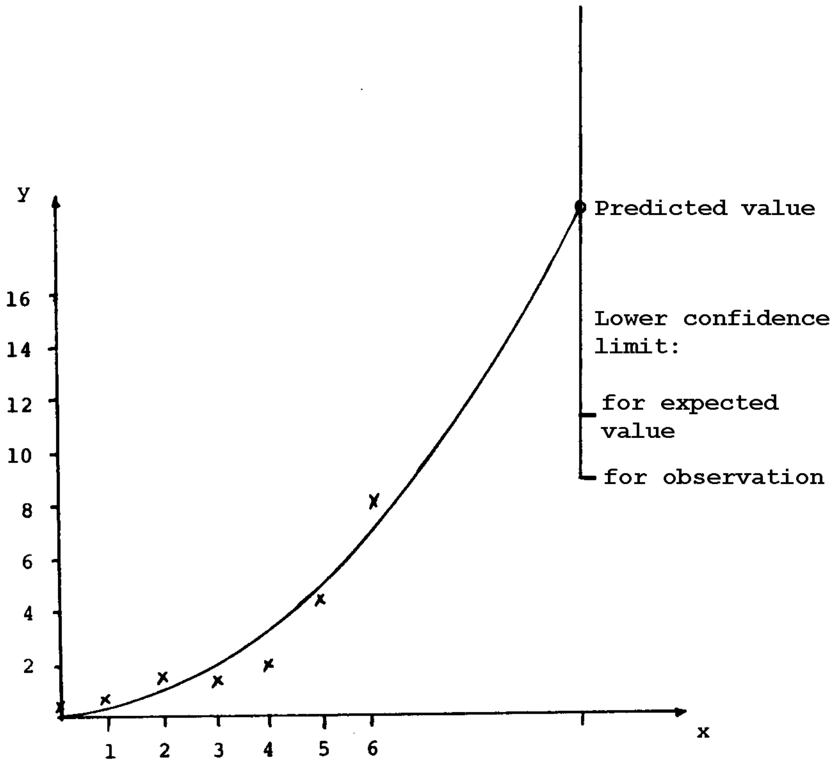
The confidence interval for the expected value at $x = 10$ is therefore given by

$$\begin{aligned} & 18.90 \pm \mathrm{t}(6)_{0.975} 0.2764\sqrt{109.89} \\ = \quad & 18.90 \pm 2.447 \cdot 0.2764\sqrt{109.89} \\ = \quad & 18.90 \pm 7.09. \end{aligned}$$

The corresponding prediction interval for the next observation is

$$\begin{aligned} & 18.90 \pm \mathrm{t}(6)_{0.975} \cdot 0.2764\sqrt{109.89 + 100}. \\ = \quad & 18.90 \pm 9.80, \end{aligned}$$

i.e. a somewhat broader interval than for the expected value. The explanation is simply that we have a variance of $10^2\sigma^2 = 100\sigma^2$ in x=10. We depict the observations and estimated polynomial in the following graph. Further the two confidence intervals are given. ♦


y
16
14
12
10
8
6
4
2
1
2
3
4
5
6
x
Predicted value
Lower confidence limit:
for expected value
for observation

## 3.2 Tests in the general linear model

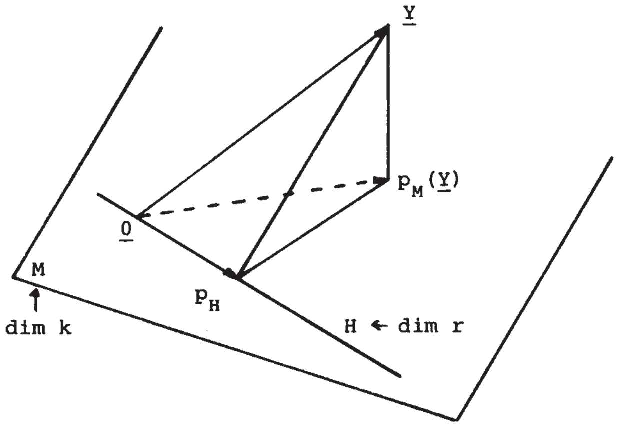
In this section we will check if the mean vector can be assumed to lie in a true sub-space of the model space and also check if the mean vector successively can be assumed to lie in sub-spaces of smaller and smaller dimensions. First

### 3.2.1 Test for a lower dimension of model space

Let $\boldsymbol{Y} \in \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is regular and known. We assume that $\boldsymbol{\mu} \in M$, is a $k$-dimensional sub-space and we will test the hypothesis

$$H_0 : \boldsymbol{\mu} \in H \quad \text{against} \quad H_1 : \boldsymbol{\mu} \in M \backslash H,$$

where $H$ is an $r$-dimensional sub-space of $M$. In the following we will consider the norm given by $\boldsymbol{\Sigma}^{-1}$ . The maximum likelihood estimator for $\boldsymbol{\mu}$ is then the projection $p_M(\boldsymbol{Y})$ onto M and if $H_0$ is true then the maximum likelihood estimator $p_M(\boldsymbol{Y})$, is $\boldsymbol{Y}$ 's projection onto $H$. The ML estimator for $\sigma^2$ in the two cases are respectively $\frac{1}{n}\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2$ and $\frac{1}{n}\|\boldsymbol{y} - p_H(\boldsymbol{y})\|^2$.



The likelihood function is

$$\begin{aligned}
\mathrm{L}(\boldsymbol{\mu}, \sigma^2) &= \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sigma^n}\frac{1}{\sqrt{\det \boldsymbol{\Sigma}}}\exp(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})) \\
&= k \cdot \sigma^{-n}\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{\mu}\|^2).
\end{aligned}$$

With this notation we have

**THEOREM 3.5.** Let the situation be as above. Then the ratio test at level $\alpha$ of

$$H_0 : \boldsymbol{\mu} \in H \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \in M \backslash H,$$

is equivalent to the test given by the critical region

$$C_\alpha = \{(y_1, \dots, y_n) | \frac{\|p_M(\boldsymbol{y}) - p_H(\boldsymbol{y})\|^2/(k-r)}{\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2/(n-k)} > \mathrm{F}(k-r, n-k)_{1-\alpha}\}.$$

▲

**PROOF 3.4.** The ratio test statistic is

$$\begin{aligned} Q &= \frac{\sup_{H_0} \mathrm{L}(\boldsymbol{\mu}, \sigma^2)}{\sup \mathrm{L}(\boldsymbol{\mu}, \sigma^2)} = \frac{\mathrm{L}(p_H(\boldsymbol{y}), \hat{\sigma^2})}{\mathrm{L}(p_M(\boldsymbol{y}), \hat{\sigma^2})} \\ &= \left[\frac{\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2}{\|\boldsymbol{y} - p_H(\boldsymbol{y})\|^2}\right]^{\frac{n}{2}} \frac{\exp(-\frac{n}{2})}{\exp(-\frac{n}{2})} = \left[\frac{\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2}{\|\boldsymbol{y} - p_H(\boldsymbol{y})\|^2}\right]^{\frac{n}{2}}. \end{aligned}$$

From this we see

$$Q < q \quad \Leftrightarrow \quad \frac{\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2}{\|\boldsymbol{y} - p_H(\boldsymbol{y})\|^2} < k_1.$$

Since we reject the hypothesis for small values of $Q$ we see that we reject when the length of the leg $\boldsymbol{Y} - p_M(\boldsymbol{Y})$ is much less than the length of the hypotenuse. From Pythagoras we have that

$$\|\boldsymbol{y} - p_H(\boldsymbol{y})\|^2 = \|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2 + \|p_H(y) - p_M(y)\|^2,$$

we see that we may just as well compare the two legs i.e. use

$$Q < q \quad \Leftrightarrow \quad \frac{\|p_M(\boldsymbol{y}) - p_H(\boldsymbol{y})\|^2/(k-r)}{\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2/(n-k)} > c. \tag{3.1}$$

Under $H_0$ both the numerator and denominator are $\sigma^2\chi^2(\mathrm{f})/\mathrm{f}$ distributed with respectively $k-r$ and $n-k$ degrees of freedom and they are furthermore independent (follows from the partition theorem). The ratio will therefore be F-distributed under $H_0$, and the theorem follows from this. The reason why we in (3.1) have divided the respective norms with the dimension of the relevant sub-space is of course that we want the test statistic to be F-distributed under $H_0$, and not just proportional to an F-distribution. ■

Q.E.D.

One usually collects the calculations in an analysis of variance table.

| Variation | SS | Degrees of freedom $=$ dimension |
|---|---|---|
| Of model from hypothesis | $\\|p_M(\boldsymbol{Y}) - p_H(\boldsymbol{Y})\\|^2$ | $k - r$ |
| Of observations from model | $\\|\boldsymbol{Y} - p_M(\boldsymbol{Y})\\|^2$ | $n - k$ |
| Of observations from hypothesis | $\\|\boldsymbol{Y} - p_H(\boldsymbol{Y})\\|^2$ | $n - r$ |

**REMARK 3.5.** Often one will be in the situation that the sub-spaces $M$ and $H$ are parameterised, i.e.

$$\begin{array}{ccccc} \boldsymbol{\mu} & \in & M & \Leftrightarrow & \exists \boldsymbol{\theta} \in R^k (\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\theta}) \\ \boldsymbol{\mu} & \in & H & \Leftrightarrow & \exists \gamma \in R^r (\boldsymbol{\mu} = \mathbf{x}_0 \gamma), \end{array}$$

where $\mathbf{x}$ and $\mathbf{x_0}$ are $n \times k$ respectively $n \times r$ (with $r \leq k$) matrices. We then have that $p_M(\boldsymbol{y}) = \mathbf{x}\hat{\boldsymbol{\theta}}$ and $p_H(\boldsymbol{y}) = \mathbf{x}_0 \hat{\gamma}$ are computed by solving the equations

$$\begin{array}{rcl} (\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} & = & \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \\ (\mathbf{x}_0'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0)\hat{\gamma} & = & \mathbf{x}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \end{array}$$

with respect to $\hat{\boldsymbol{\theta}}$ and $\hat{\gamma}$. ▼

Once again we consider the model from p. 114.

**EXAMPLE 3.7.** We have the model

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \boldsymbol{\epsilon}.$$

We observe data where $y' = (10.11,\ 0.81,\ 5.24)$. We wish to test the hypothesis

$$H_0 : \theta_2 = 0 \quad \text{versus} \quad H_1 : \theta_2 \neq 0.$$

We reformulate the hypothesis into

$$H_0 : \mathrm{E}(\boldsymbol{Y}) = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \theta_1 = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \gamma.$$

The estimator for $\gamma$ is

$$\hat{\gamma} = [(\,1 \quad 0 \quad \tfrac{1}{2}\,) \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix}]^{-1}[(\,1 \quad 0 \quad \tfrac{1}{2}\,) \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}] = \frac{4}{5}y_1 + \frac{2}{5}y_3.$$

The observed value is $\hat{\gamma} = 10.184.$ From this we have

$$\mathbf{x}_0\hat{\gamma} = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} 10.184 = \begin{bmatrix} 10.184 \\ 0 \\ 5.092 \end{bmatrix},$$

and

$$\|\boldsymbol{y} - \mathbf{x}_0\hat{\gamma}\|^2 = (\boldsymbol{y} - \mathbf{x}_0\hat{\gamma})'(\boldsymbol{y} - \mathbf{x}_0\hat{\gamma}) = 0.6835.$$

Since we had (p. 115)

$$\|\boldsymbol{y} - \mathbf{x}_0\hat{\theta}\|^2 = (\boldsymbol{y} - \mathbf{x}\hat{\theta})'(\boldsymbol{y} - \mathbf{x}\hat{\theta}) = 0.0338,$$

we get

$$\|\mathbf{x}\,\hat{\boldsymbol{\theta}} - \mathbf{x}_0\hat{\gamma}\|^2 = 0.6835 - 0.0338 = 0.6497.$$

From this the test statistic becomes

$$\frac{\|\mathbf{x}\,\hat{\boldsymbol{\theta}} - \mathbf{x}_0\hat{\boldsymbol{\gamma}}\|^2/(2-1)}{\|\boldsymbol{y} - \mathbf{x}\,\hat{\boldsymbol{\theta}}\|^2/(3-2)} = 19.22 < \mathrm{F}(1,1)_{0.90},$$

and we accept the hypothesis at least for any $\alpha <$ 10%.
Explanation of the degrees of freedom:

$$\begin{aligned} \operatorname{rg}\mathbf{x} &= \operatorname{rg}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = 2 = k \\ \operatorname{rg}\mathbf{x}_0 &= \operatorname{rg}\begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} = 1 = r \\ n &= 3. \end{aligned}$$

♦

We will now look at the continuation of example 3.5 p. 121.

**EXAMPLE 3.8.** From the formulation of the problem it seems reasonable to assume that the parameter $\theta_{21} = 0$. We will therefore test the hypothesis

$$H_0 : \theta_{21} = 0 \qquad \text{against} \qquad H_1 : \theta_{21} \neq 0.$$

The hypothesis-space $H$ is therefore given by

$$\mathrm{E}(\boldsymbol{Y}) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \end{bmatrix} = \begin{bmatrix} \mu_1 + \theta_{11} \\ \mu_1 + \theta_{11} \\ \mu_1 + \theta_{12} \\ \mu_1 + \theta_{12} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We now find

$$\mathbf{x}_1'\mathbf{x}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix},$$

and

$$\mathbf{x}_1'\boldsymbol{Y} = \begin{bmatrix} Y_{1..} \\ Y_{11.} \\ Y_{12.} \end{bmatrix}.$$

We see that $\mathbf{x}_1'\mathbf{x}_1$ is singular, and we add the linear restriction

$$\mathbf{b}\,\boldsymbol{\theta} = (\ 0 \quad 1 \quad 1\ ) \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \end{bmatrix} = \theta_{11} + \theta_{12} = 0.$$

Since

$$\mathbf{b}'\mathbf{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

we have

$$\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b}=\begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

This matrix is inverted on p. 121. We therefore find the estimator under $H_0$ as

$$\hat{\theta}_1=\begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}\begin{bmatrix} Y_{1..} \\ Y_{11.} \\ Y_{12.} \end{bmatrix}=\begin{bmatrix} \bar{Y}_{1..} \\ \bar{Y}_{11.}-\bar{Y}_1 \\ \bar{Y}_{12.}-\bar{Y}_1 \end{bmatrix}.$$

The observed value is $(-9,\ +8,\ -8)'$. The new residual vector is

$$\boldsymbol{y}-\mathbf{x}_1\hat{\boldsymbol{\theta}}_1=(1,-1,-2.+2,-6,0,-2)'.$$

The norm of this vector is 50, and the number of degrees of freedom is 7-2 = 5. We therefore find that

$$\begin{aligned} \|p_M(\boldsymbol{y})-p_H(y)\|^2 &= \|y-p_H(y)\|^2-\|y-p_M(y)\|^2 \\ &= 50-28\frac{2}{3}=21\frac{1}{3}. \end{aligned}$$

We now collect the calculations in the following analysis of variance table.

| Variation | SS | $f$ | $S^2$ | Test |
|---|---|---|---|---|
| $M-H$ | $21\frac{1}{3}$ | $3-2=1$ | $21\frac{1}{3}$ | 2.97 |
| $O-M$ | $28\frac{2}{3}$ | $7-3=4$ | $7\frac{1}{6}$ | |
| $O-H$ | 50 | $7-2=5$ | | |

Since the observed value of the test statistic $2.97 < \mathrm{F}(1,4)_{0.90}$ we will accept the hypothesis, and therefore assume that $H_0$ is true. ♦

### 3.2.2 Successive testing in the general linear model.

In this section we will illustrate the test procedure one should follow, when one successively wants to investigate if the mean vector for ones observations lies in sub-spaces $H_i$ with

$$H_0 \supseteq H_1 \supseteq H_2 \supseteq \cdots \supseteq H_m, \qquad m \leq k.$$

We will start by considering the following numbers from the yield of penicillin fermentation using two different types of sugar namely: lactose and cane sugar, at the concentrations 2%, 4%, 6% and 8% (in g./100 ml.).

| | | Factor B: concentration | | | |
|---|---|---|---|---|---|
| | | 2% | 4% | 6% | 8% |
| Factor A: | Lactose | 0.606 | 0.660 | 0.984 | 0.908 |
| | Cane sugar | 0.761 | 0.933 | 1.072 | 0.979 |

The numbers are from [5] p. 314. The yield has been expressed by the logarithm of the weight of the mycelium after one week of growth.

We are now interested in investigating the two factors A's and B's influence on the yield. We assume that the observations are stochastic independent and normally distributed. They are called

$$L : Y_{11}, Y_{12}, Y_{13}, Y_{14}$$

and

$$R : Y_{21}, Y_{22}, Y_{23}, Y_{24}$$

further we will assume that

$$\mathrm{E}(Y_{ij}) = \alpha_i' + \beta_i' x_j' + \gamma_i' x_j^2$$

where $x_j'$ gives the $j$'th sugar concentration. We will perform change in scale of the sugar concentration

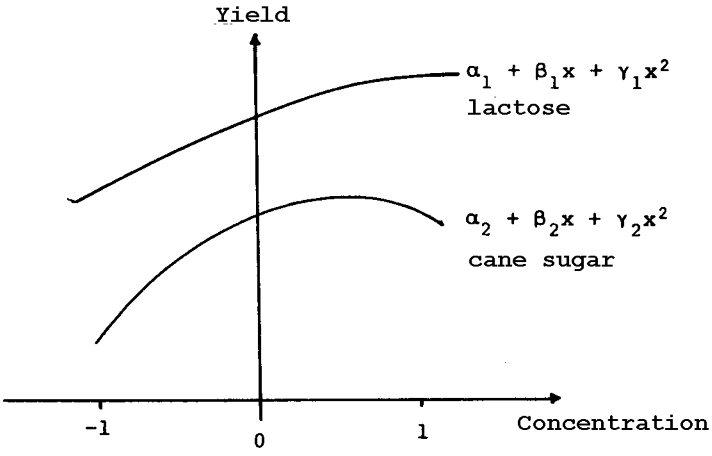
| | |
|---|---|
| $2\%$ | $-3$ |
| $4\%$ | $-1$ |
| $6\%$ | $1$ |
| $8\%$ | $3,$ |

or more stringently define $x$ by

$$x_j = \frac{x_j' - 5\%}{1\%}.$$

We then get the following expression for the mean values

$$\mathrm{E}(Y_{ij}) = \alpha_i + \beta_i x_j + \gamma_i x_j^2.$$



We are assuming that the yield within the given limits can be expressed as polynomials of second degree.

One could now e.g. successively investigate

1) if $\gamma_1 = \gamma_2 = 0$, i.e. if a description by affine functions is sufficient

2) if that is accepted then if $\beta_1 = \beta_2 = \beta$, i.e. if the marginal effect by increasing the concentration is the same for the two types of sugar

3) if that is accepted then if $\alpha_1 = \alpha_2 = \alpha$, i.e. if the two types of sugar are equal with respect to the yield and if this is accepted

4) then if $\beta = 0$, i.e. if the concentration has any influence at all

i) We first write the model in matrix form

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \\ \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix},$$

or

$$\boldsymbol{Y} = \mathrm{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

We find

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ -3 & -1 & 1 & 3 & 0 & 0 & 0 & 0 \\ 9 & 1 & 1 & 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -3 & -1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 9 & 1 & 1 & 9 \end{bmatrix} \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 0 & 20 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 & 0 & 0 \\ 20 & 0 & 164 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 20 \\ 0 & 0 & 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 20 & 0 & 164 \end{bmatrix}.$$

Since

$$\begin{bmatrix} 4 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{41}{64} & 0 & -\frac{5}{64} \\ 0 & \frac{1}{20} & 0 \\ -\frac{5}{64} & 0 & \frac{1}{64} \end{bmatrix},$$

then

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} \frac{41}{64} & 0 & -\frac{5}{64} & 0 & 0 & 0 \\ 0 & \frac{1}{20} & 0 & 0 & 0 & 0 \\ -\frac{5}{64} & 0 & \frac{1}{64} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{41}{64} & 0 & -\frac{5}{64} \\ 0 & 0 & 0 & 0 & \frac{1}{20} & 0 \\ 0 & 0 & 0 & -\frac{5}{64} & 0 & \frac{1}{64} \end{bmatrix}.$$

From this we see that

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} -\frac{1}{16}y_{11} + \frac{9}{16}y_{12} + \frac{9}{16}y_{13} - \frac{1}{16}y_{14} \\ -\frac{3}{20}y_{11} - \frac{1}{20}y_{12} + \frac{1}{20}y_{13} + \frac{3}{20}y_{14} \\ \frac{1}{16}y_{11} - \frac{1}{16}y_{12} - \frac{1}{16}y_{13} + \frac{1}{16}y_{14} \\ -\frac{1}{16}y_{21} + \frac{9}{16}y_{22} + \frac{9}{16}y_{23} - \frac{1}{16}y_{24} \\ \frac{3}{20}y_{21} - \frac{1}{20}y_{22} - \frac{1}{20}y_{23} + \frac{3}{20}y_{24} \\ \frac{1}{16}y_{21} - \frac{1}{16}y_{22} - \frac{1}{16}y_{23} + \frac{1}{16}y_{24} \end{bmatrix} = \begin{bmatrix} 0.830 \\ 0.062 \\ -0.008 \\ 1.019 \\ 0.040 \\ -0.017 \end{bmatrix}.$$

The model corresponds to a 6-dimensional sub-space $M$ in $R^8$ $(\operatorname{rg}\mathbf{x} = 6),$ and since we are using the norm corresponding to $\mathbf{I}$ we have that the projection onto

$M$ is

$$p_M(\boldsymbol{y}) = \boldsymbol{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} 0.830 \\ 0.062 \\ -0.008 \\ 1.019 \\ 0.040 \\ -0.017 \end{bmatrix} = \begin{bmatrix} 0.572 \\ 0.760 \\ 0.884 \\ 0.944 \\ 0.746 \\ 0.962 \\ 1.042 \\ 0.986 \end{bmatrix}.$$

We therefore have the residuals

$$\boldsymbol{y} - p_M(\boldsymbol{y}) = \begin{bmatrix} 0.034 \\ -0.100 \\ 0.100 \\ -0.036 \\ 0.015 \\ -0.029 \\ 0.030 \\ -0.007 \end{bmatrix}.$$

The squared length of this vector is

$$\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2 = 0.034^2 + \cdots + (-0.007)^2 = 0.024467.$$

As an estimate of $\sigma^2$ we can therefore use

$$\hat{\sigma}^2 = \frac{1}{8-6} 0.024467 = 0.0122335.$$

ii) If the hypothesis $\boldsymbol{\mu} \in H_1$, i.e. $\gamma_1 = \gamma_2 = 0$, or

$$\boldsymbol{y} = \begin{bmatrix} 1 & -3 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{x}_1 \boldsymbol{\delta}_1 + \boldsymbol{\varepsilon}_1,$$

is true, then we get the estimates

$$\hat{\boldsymbol{\delta}}_1 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\boldsymbol{y} = \begin{bmatrix} \frac{1}{4}y_{11} + \frac{1}{4}y_{12} + \frac{1}{4}y_{13} + \frac{1}{4}y_{14} \\ -\frac{3}{20}y_{11} - \frac{1}{20}y_{12} + \frac{1}{20}y_{13} + \frac{3}{20}y_{14} \\ \frac{1}{4}y_{21} + \frac{1}{4}y_{22} + \frac{1}{4}y_{23} + \frac{1}{4}y_{24} \\ -\frac{3}{20}y_{21} - \frac{1}{20}y_{22} + \frac{1}{20}y_{23} + \frac{3}{20}y_{24} \end{bmatrix} = \begin{bmatrix} 0.790 \\ 0.062 \\ 0.936 \\ 0.040 \end{bmatrix}$$

The residuals are

$$\boldsymbol{y} - p_{H_1}(\boldsymbol{y}) = \boldsymbol{y} - \mathrm{x}_1\hat{\boldsymbol{\delta}}_1 = \begin{bmatrix} 0.002 \\ -0.068 \\ 0.132 \\ -0.068 \\ -0.055 \\ 0.037 \\ 0.096 \\ -0.077 \end{bmatrix}.$$

The squared length of this vector is

$$\|\boldsymbol{y} - p_{H_1}(\boldsymbol{y})\|^2 = 0.002^2 + \cdots + (-0.077)^2 = 0.046215.$$

iii) If $\boldsymbol{\mu} \in H_2$, d.v.s. $\beta_1 = \beta_2 = \beta$, the model becomes

$$\boldsymbol{y} = \begin{bmatrix} 1 & 0 & -3 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & 3 \\ 0 & 1 & -3 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \boldsymbol{\varepsilon}_2 = \boldsymbol{x}_2\boldsymbol{\delta}_2 + \boldsymbol{\varepsilon}_2.$$

The estimates become

$$\hat{\boldsymbol{\delta}}_2 = (\mathrm{x}_2'\mathrm{x}_2)^{-1}\mathrm{x}_2'\boldsymbol{y} = \begin{bmatrix} 0.790 \\ 0.936 \\ 0.051 \end{bmatrix},$$

and the residuals

$$\boldsymbol{y} - p_{H_2}(\boldsymbol{y}) = \begin{bmatrix} -0.031 \\ -0.079 \\ 0.143 \\ -0.035 \\ -0.022 \\ 0.048 \\ 0.085 \\ -0.110 \end{bmatrix}.$$

The squared norm of the residual vector is

$$\|\boldsymbol{y} - p_{H_2}(\boldsymbol{y})\|^2 = (-0.031)^2 + \cdots + (-0.110)^2 = 0.050989.$$

iv) If $\boldsymbol{\mu} \in H_3$, i.e. $\beta_1 = \beta_2 = \beta$ and $\alpha_1 = \alpha_2 = \alpha$, then the model is

$$\boldsymbol{y} = \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \\ 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \boldsymbol{\varepsilon}_3 = \mathbf{x}_3 \boldsymbol{\delta}_3 + \boldsymbol{\varepsilon}_3$$

We find

$$\hat{\boldsymbol{\delta}}_3 = (\mathbf{x}_3' \mathbf{x}_3)^{-1} \mathbf{x}_3' \boldsymbol{y} = \begin{bmatrix} 0.863 \\ 0.051 \end{bmatrix},$$

and

$$\boldsymbol{y} - p_{H_3}(\boldsymbol{y}) = \begin{bmatrix} -0.104 \\ -0.152 \\ 0.070 \\ -0.108 \\ 0.051 \\ 0.121 \\ 0.158 \\ -0.037 \end{bmatrix},$$

giving

$$\|\boldsymbol{y} - p_{H_3}(\boldsymbol{y})\|^2 = 0.094059.$$

v) Finally we consider the case $\boldsymbol{\mu} \in H_4$, i.e. $\beta = 0$, or

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \alpha = \mathbf{x}_4 \boldsymbol{\delta}_4 + \boldsymbol{\varepsilon}_4.$$

We find

$$\hat{\boldsymbol{\delta}}_4 = \hat{\alpha} = (\mathbf{x}_4' \mathbf{x}_4)^{-1} \mathbf{x}_4' \boldsymbol{y}' = 0.863,$$

giving

$$\boldsymbol{y} - p_{H_4}(\boldsymbol{y}) = \begin{bmatrix} -0.250 \\ -0.203 \\ 0.121 \\ 0.045 \\ -0.102 \\ 0.070 \\ 0.209 \\ 0.116 \end{bmatrix},$$

and

$$\|\boldsymbol{y} - p_{H_4}(\boldsymbol{y})\|^2 = 0.196365.$$

Since we let $\mathrm{rg}(\mathbf{x}_i) = r_i$ and $\mathrm{rg}(\mathbf{x}) = k$ we can summarise the testing procedure in an analysis of variance table such as

| Variation | SS | Degrees of freedom = dimension |
|---|---|---|
| $H_4 - H_3$ | $\|p_{H_4}(\boldsymbol{y}) - p_{H_3}(\boldsymbol{y})\|^2$ | $r_3 - r_4 = 2 - 1 = 1$ |
| $H_3 - H_2$ | $\|p_{H_3}(\boldsymbol{y}) - p_{H_2}(\boldsymbol{y})\|^2$ | $r_2 - r_3 = 3 - 2 = 1$ |
| $H_2 - H_1$ | $\|p_{H_2}(\boldsymbol{y}) - p_{H_1}(\boldsymbol{y})\|^2$ | $r_1 - r_2 = 4 - 3 = 1$ |
| $H_1 - M$ | $\|p_{H_1}(\boldsymbol{y}) - p_M(\boldsymbol{y})\|^2$ | $k - r_1 = 6 - 4 = 2$ |
| $M - \text{obs.}$ | $\|p_M(\boldsymbol{y}) - \boldsymbol{y}\|^2$ | $n - k = 8 - 6 = 2$ |
| $H_4 - \text{obs.}$ | $\|p_{H_4}(\boldsymbol{y}) - \boldsymbol{y}\|^2$ | $n - r_4 = 8 - 1 = 7$ |

This table is a simple extension of the table on p. 136. We can use the partition theorem and get, under the different hypotheses, that the sum of squares are independent and distributed as $\sigma^2\chi^2$ with the respective degrees of freedom.

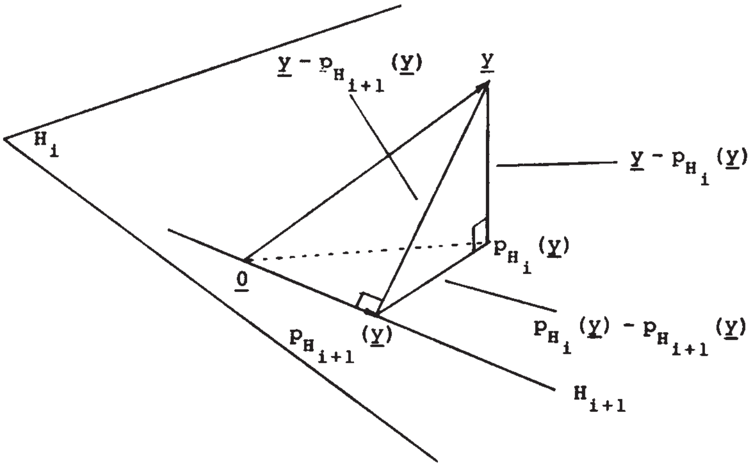
If a hypothesis $H_i$ is accepted then the test statistic for the test of $H_{i+1}$ becomes

$$\frac{\|p_{H_i}(\boldsymbol{y}) - p_{H_{i+1}}(\boldsymbol{y})\|^2/(r_i - r_{i+1})}{\|p_{H_i}(\boldsymbol{y}) - \boldsymbol{y}\|^2/(n - r_i)}.$$

Under the hypothesis this measure is $\mathrm{F}(r_i - r_{i+1}, n - r_i)$ distributed (according to the partition theorem) and - still following the theory from the previous section - we reject for large values of $Z$ i.e. for

$$Z > \mathrm{F}(r_i - r_{i+1}, n - r_i)_{1-\alpha}.$$

Before we start testing it would be appropriate to give some computational formulas. We consider the transition from $H_i$ to $H_{i+1} \subset H_i$.



Using Pythagoras' theorem we now see that there are two alternative ways of computation for

$$z = \|p_{H_{i+1}}(\boldsymbol{y}) - p_{H_i}(\boldsymbol{y})\|^2,$$

they are

$$z = \|p_{H_i}(\boldsymbol{y})\|^2 - \|p_{H_{i+1}}(\boldsymbol{y})\|^2 \tag{3.2}$$

and

$$z = \|\boldsymbol{y} - p_{H_{i+1}}(\boldsymbol{y})\|^2 - \|\boldsymbol{y} - p_{H_i}(\boldsymbol{y})\|^2. \tag{3.3}$$

Of these the first must be preferred from numerical reasons but if one has computed the residuals sum of squares anyhow it seems to be easier to use ( reffor:3.3)).

The analysis of variance table in our case becomes

| Variation | SS | $f$ | Test statistic |
|---|---|---|---|
| $H_4 - H_3$ | 0.102306 | 1 | $\frac{0.102306/1}{0.094059/6} = 5.44$ |
| $H_3 - H_2$ | 0.043070 | 1 | $\frac{0.043070/1}{0.050981/5} = 4.22$ |
| $H_2 - H_1$ | 0.004774 | 1 | $\frac{0.004774/1}{0.046215/4} = 0.41$ |
| $H_1 - M$ | 0.021748 | 2 | $\frac{0.021748/2}{0.024467/2} = 0.89$ |
| $M - \text{obs}$ | 0.024467 | 2 | |
| $\text{Obs} - H_4$ | 0.196365 | 7 | |

Since

$$4.22 \simeq \mathrm{F}(1,5)_{0.91},$$

and

$$5.44 \simeq \mathrm{F}(1,6)_{0.94},$$

we will not by testing at say, level $\alpha = 5\%$ - reject any of the hypothesis $H_1, H_2, H_3$ or $H_4$.

**NOTE 1.** We will of course not test e.g. $H_2$, if we had rejected $H_1$, since $H_2$ is a sub-hypothesis of $H_1$.

The conclusion is therefore that we (until new investigations reject this) will continue to work with the model that the yield $Y$ by penicillin fermentation is independent of type of sugar and the concentration (2% $\leq$ concentration $\leq$ 8%) at which the fermentation takes place. We have with

$$\mathrm{E}(Y) = \alpha \quad \text{and} \quad \mathrm{V}(Y) = \sigma^2,$$

that

$$\hat{\alpha} = 0.863,$$

and

$$\hat{\sigma}^2 = \frac{0.196365}{7} = 0.028052 \simeq 0.17^2.$$

Finally

$$\mathrm{V}(\hat{\alpha}) = \frac{\sigma^2}{8} \simeq \frac{\hat{\sigma}^2}{8} = 0.0035 \simeq 0.059^2.$$

# Chapter 4
# Regression analysis

In this chapter we will give an overview on regression analysis. Most of it is a special case of the general linear model but since a number of uses are often concerned with regression situations we will try to describe the results in this language.

There is a small section on orthogonal regression (not to be confused with regression by orthogonal polynomials). From a statistical point of view this is more related to the section on principle components and factor analysis and considering ways of computation we also refer to that chapter. However, from a curve-fitting point of view we have found it sensible to mention the concept in the present chapter too.

## 4.1 Linear regression analysis

In this section linear regression analysis will be analysed by means of the theory for the general linear model. We start with

### 4.1.1 Notation and model.

In the ordinary regression analysis we work with the model

$$\mathrm{E}(Y) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k,$$

where the $x$'s are known variables and the $\beta$'s (and $\alpha$) are unknown parameters. If we have given $n$ observations of $Y$ we could more precisely write the model as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

or

$$\boldsymbol{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

We assume as usual that

$$\mathrm{D}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is known and $\sigma^2$ is (usually) unknown.

The estimators are found in the usual way by solving the normal equations

$$\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}\boldsymbol{\beta} = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

or if $\boldsymbol{\Sigma} = \mathbf{I}$

$$\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}} = \mathbf{x}'\boldsymbol{Y}.$$

In the first case we talk of a weighted regression analysis.

Before we go on it is probably appropriate once again to stress what is meant by the word linear in the term linear regression analysis.

As in the ordinary linear model the meaning is that we have linearity in the parameters. We can easily do regression by e.g. time and the logarithm of the time. The model will then just be

$$\mathrm{E}(Y) = \alpha + \beta_1 t + \beta_2 \ln t,$$

cf. example 3.2.

With $n$ observations the model in matrix form becomes

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \ln t_1 \\ \vdots & \vdots & \vdots \\ 1 & t_n & \ln t_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

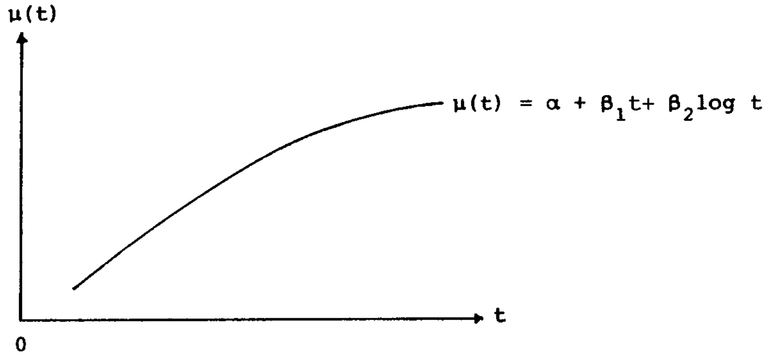


Figure 4.1:

Another banality that could be useful to stress is that one can force the regression surface through 0 by deleting the $\alpha$ and first column in the $\mathbf{x}$ -matrix i.e. use the model

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & & \vdots \\ x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

It can be useful to note that you can use the following trick if you wish the regression surface to go through 0. We assume that $\mathbf{\Sigma} = \mathbf{I}$.

We consider the observations $Y_1, \ldots, Y_n$ and the corresponding values of the independent variables $1, x_{i1}, \ldots, x_{ik}$, $i = 1, \ldots, n$. if we add $-Y_1, \ldots, -Y_n$ and $1, -x_{i1}, \ldots, -x_{ik}$, $i = 1, \ldots, n$ and write down the usual model we get

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \\ -Y_1 \\ \vdots \\ -Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \\ 1 & -x_{11} & \cdots & -x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & -x_{1n} & \cdots & -x_{kn} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon,$$

or more compactly

$$\begin{bmatrix} \boldsymbol{Y} \\ -\boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mathbf{1} & -\mathbf{x} \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} + \varepsilon,$$

where we - compared to the notation on p. 150 - have used a slightly different definition of the $\mathbf{x}$ matrix and $\boldsymbol{\beta}$.

The normal equations become

$$\begin{bmatrix} \mathbf{1}' & \mathbf{1}' \\ \mathbf{x}' & -\mathbf{x}' \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mathbf{1} & -\mathbf{x} \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{1}' & \mathbf{1}' \\ \mathbf{x}' & -\mathbf{x}' \end{bmatrix} \begin{bmatrix} \boldsymbol{Y} \\ -\boldsymbol{Y} \end{bmatrix},$$

or

$$\begin{bmatrix} 2n & 0 \\ 0 & 2\mathbf{x}'\mathbf{x} \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} 0 \\ 2\mathbf{x}'\boldsymbol{Y} \end{bmatrix}.$$

If we write out the equations we get

$$\begin{aligned} 2n\alpha &= 0 \\ 2\mathbf{x}'\mathbf{x}\boldsymbol{\beta} &= 2\mathbf{x}'\boldsymbol{Y}, \end{aligned}$$

or

$$\begin{aligned} \alpha &= 0 \\ \mathbf{x}'\mathbf{x}\boldsymbol{\beta} &= \mathbf{x}'\boldsymbol{Y}. \end{aligned}$$

In other words in this way we have found the estimators of the coefficients to a regression surface which has been forced through $\mathbf{0}$.

The reason why the above is useful is that a number of standard programmes cannot force the surface through $\mathbf{0}$. Using the above mentioned trick the problem can be circumvented.

The output from such a programme should be interpreted cautiously since all the sums of squares are twice their correct size. E.g. the residual sums of squares will be computed as

$$\begin{aligned} &\left( \begin{bmatrix} \boldsymbol{Y} \\ -\boldsymbol{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{x}\hat{\boldsymbol{\beta}} \\ -\mathbf{x}\hat{\boldsymbol{\beta}} \end{bmatrix} \right)' \left( \begin{bmatrix} \boldsymbol{Y} \\ -\boldsymbol{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{x}\hat{\boldsymbol{\beta}} \\ -\mathbf{x}\hat{\boldsymbol{\beta}} \end{bmatrix} \right) \\ = &\ ([\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\beta}}]', [-\boldsymbol{Y} + \mathbf{x}\hat{\boldsymbol{\beta}}]') \begin{bmatrix} \boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\beta}} \\ -\boldsymbol{Y} + \mathbf{x}\hat{\boldsymbol{\beta}} \end{bmatrix} \\ = &\ 2[\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\beta}}]'[\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\beta}}], \end{aligned}$$

i.e. twice the correct residual sum of squares. The mentioned degrees of freedom will not be correct either. We have to write up the ordinary linear model and find the correct degrees of freedom by considering the dimensions.

### 4.1.2 Correlation and regression.

In theorem 2.23 p. 92 a result was stated, which can be used for a test if the multiple correlation coefficient between normally distributed variables is 0. We will now show that this result corresponds to a certain test in a regression model.

We will assume that we have the usual model p. 149 and we assume that $\mathbf{\Sigma} = \mathbf{I}$.

Without any problems we can use the theory from chapter 3 to test different hypothesis about the parameters $\alpha, \beta_1, \ldots, \beta_k$.

By formal calculations we can estimate the multiple correlation coefficient between $Y$ and $x_1, \ldots, x_k$ using expressions mentioned in section 2.3.2.

It can be shown that we get

$$R^2 = \frac{\|\boldsymbol{Y} - p_0(\boldsymbol{Y})\|^2 - \|\boldsymbol{Y} - p_M(\boldsymbol{Y})\|^2}{\|\boldsymbol{Y} - p_0(\boldsymbol{Y})\|^2},$$

where

$$p_0(\boldsymbol{Y}) = \begin{bmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix} \qquad (= \mathbf{x} \cdot \hat{\boldsymbol{\beta}}),$$

and

$$p_M(\boldsymbol{Y}) = \mathbf{x}\hat{\boldsymbol{\beta}} = \hat{\mathrm{E}}(\boldsymbol{Y}).$$

These results are not very surprising. We remember that the multiple correlation coefficient could be found as the linear combination of $\boldsymbol{X}$ which minimises the variance of $(Y - \boldsymbol{\alpha}'\boldsymbol{X})$ and this corresponds exactly to writing the condition for least squares estimates.

If we let

$$\mathrm{SS}_{\mathrm{tot}} = \|\boldsymbol{Y} - p_0(\boldsymbol{Y})\|^2 = \sum_i (Y_i - \bar{Y})^2,$$

and

$$\mathrm{SS}_{\mathrm{res}} = \|\boldsymbol{Y} - \mathbf{x}\hat{\boldsymbol{\beta}}\|^2 = \sum_i (Y_i - \hat{\mathrm{E}}(Y_i))^2,$$

we can write

$$R^2 = \frac{\mathrm{SS}_{\mathrm{tot}} - \mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}},$$

i.e. the squared multiple correlation coefficient can also be expressed as the part of the total variation in the $Y$'s which are explained using the independent variables.

A corresponding re-interpretation of the partial correlations is of course also possible.

Furthermore, we see that if we formally write the test on p. 92 for $\rho_{Y|x_1,\ldots,x_k} = 0$ we get

$$\begin{aligned}
\frac{R^2}{1-R^2}\frac{n-k-1}{k} &= \frac{\|\boldsymbol{Y}-p_0(\boldsymbol{Y})\|^2 - \|\boldsymbol{Y}-p_M(\boldsymbol{Y})\|^2}{\|\boldsymbol{Y}-p_M(\boldsymbol{Y})\|^2}\frac{n-k-1}{k} \\
&= \frac{\|p_M(\boldsymbol{Y})-p_0(\boldsymbol{Y})\|^2/k}{\|\boldsymbol{Y}-P_M(\boldsymbol{Y})\|^2/(n-k-1)} \\
&= \frac{(\mathrm{SS}_{\mathrm{tot}}-\mathrm{SS}_{\mathrm{res}})/k}{\mathrm{SS}_{\mathrm{res}}/(n-k-1)}
\end{aligned}$$

From the normal theory (p. 135) this is exactly the test statistic for the hypothesis

$$\begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and the distribution of the test statistic is a $\mathrm{F}(k, n-k-1)$-distribution - exactly the same as we found on p. 135.

For testing it is from the numerical point of view therefore of no importance if we choose to consider the $x$'s as observations of a $k$-dimensional normally distributed stochastic variable or as fixed deterministic variables.

This issue can therefore be separated from the assumptions we will consider in the next section.

### 4.1.3 Analysis of assumptions.

If we for corresponding $x$-values

$$x_{1i}, \ldots, x_{pi}$$

have more observations of $Y$, it would be possible to compute the usual tests for distributional type (histograms, quantile diagrams, $\chi^2$ -tests, etc.) and for the homogeneity of variances (Bartlett's test and others). Finally we could also do run tests for randomness etc. etc.

However, the situation is often that we very seldom have (more than maybe a couple) of repetitions for different values of the independent variable. It is therefore not possible to do these types of checks of the assumptions. Instead we consider the residuals

$$\mathrm{E}_i = Y_i - \hat{\mathrm{E}}(Y_i) = Y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}.$$

If the model is valid these will be approximately independent and $\mathrm{N}(0, \sigma^2)$ distributed.

If one depicts the residuals in different ways and thereby sees something which does not look (or could not be) observations of independently $\mathrm{N}(0, \sigma^2)$ -distributed stochastic variables then we have an indication that there is something wrong with the model.

Most often we would probably start with a usual analysis of the distribution of the residuals i.e. do run-tests, draw histograms, quantile diagrams etc.

Afterwards we could depict the residuals against different quantities (time, independent variables, etc.). We show the following 4 sketches to illustrate often seen residual plots.
We will now give a short description of what the reason for plots of this kind could be. First we note that 1 always is acceptable (however, cf. p. 157).

i) Plot of residuals against time

2 The variance increases with time. Perform a weighted analysis.

3 Lack terms of the form $\beta$·time

4 Lack terms of the form $\beta_1 \cdot \text{time} + \beta_2 \cdot \text{time}^2$

ii) Plot of residuals against $\hat{\mathrm{E}}(Y_i)$

2 The variance increases with $\mathrm{E}(Y_i)$. Perform a weighted analysis or transform the $Y$'s (e.g. with the logarithm or equivalent)

3 Lack constant term (the regression is possibly erroneously forced through 0). Error in the analysis.

4 Bad model. Try with a transformation of the $Y$'s.

iii) Plot against independent variable $\boldsymbol{x_i}$

2 The variance grows with $x_i$. Perform a weighted analysis or transform the Y's.

3 Error in the computations

4 Lacks quadratic term in $x_i$

The above is not meant to be an exhaustive description of how to analyse residual plots but may be considered as an indication of how such an analysis could be done.

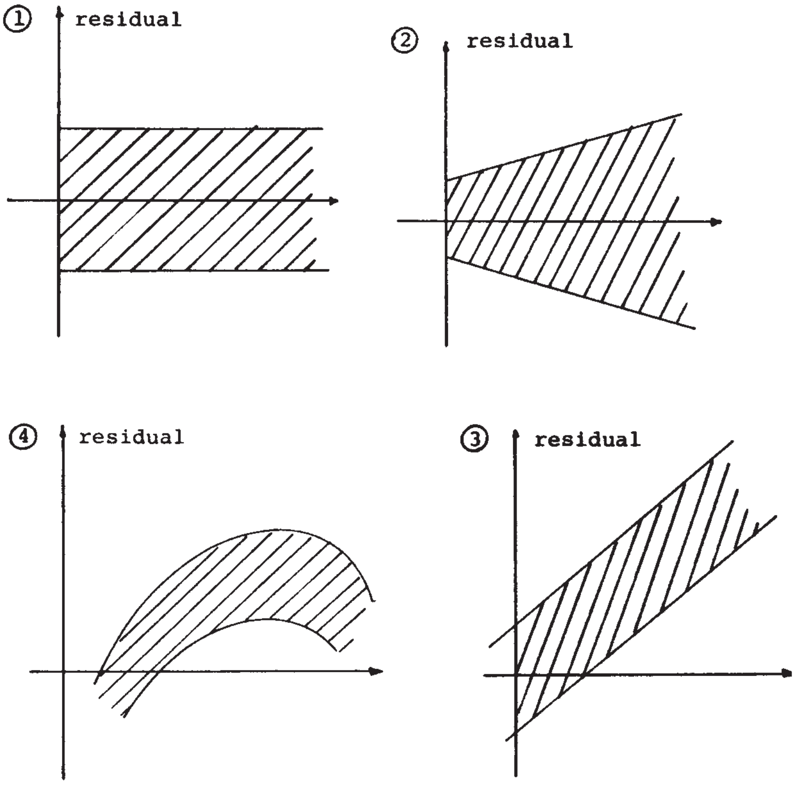


Figure 4.2: Residual plots.

**REMARK 4.1.** One often sees residual plots of the type residual versus dependent variable i.e.

$$Y_i - \hat{\mathrm{E}}(Y_i) \quad \text{against} \quad Y_i,$$

and people are often surprised that the picture is as displayed in 3). However, there is nothing abnormal in this. It can be shown that

$$\mathrm{Cor}(Y_i, Y_i - \hat{\mathrm{E}}(Y_i)) = 1 - R^2,$$

i.e. they are positively correlated. If the multiple correlation coefficient is anything less than 1 we would therefore get a picture as 3. Only if the regression surface goes through all points i.e. $R^2 = 1$, then we will have a picture as in 1.

In practise we will often have our residual plot printed on printer listings. Then the plots might look as shown on p. 158. The 4 plots have been taken from [20] p. 14-15 in appendix C.

When interpreting these plots we should remember that there are not always an equal numbers of observations for each value of the independent variable.

This is e.g. the case in the plot which depicts the residual against variable 10.

There are 7 observations corresponding to $x_{10} \sim 0.2704\,\mathrm{E}\,04$ and 35 observations corresponding to $x_{10} \sim 0.7126\,\mathrm{E}\,03$. The range of variation for the residuals is approximately the same in the two cases. If the residuals corresponding to the 2 values of $x_{10}$ had the same variance we would, however, expect the range of variation for the one with many observations to be the largest.

In other words if one has most observations around the centre of gravity for an independent variable a residual plot should rather be elliptical than of the form 1 to be satisfactory. ▼

### 4.1.4 On "Influence Statistics"

When judging the quality of a regression analysis one often consider the following two possibilities:

1) Check if deviations from the model look random.

2) Check the effect of single observations on the parameter estimates etc.

Considerations regarding 1) are given in section 4.1.3 above. Here we will briefly consider 2).

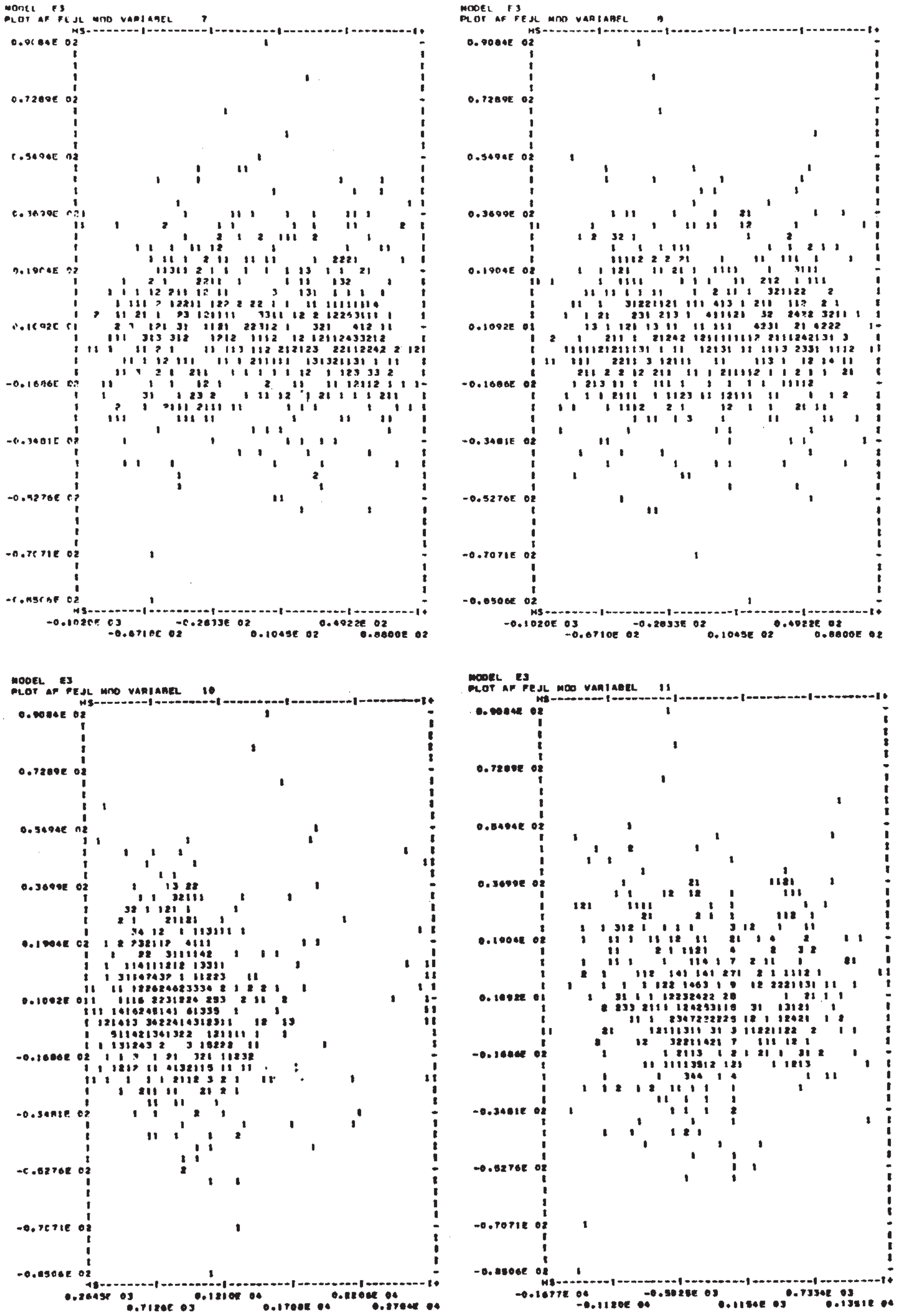

MODEL E3
PLOT AF FEJL MOD VARIABEL 7
0.9084E 02
0.7289E 02
0.5494E 02
0.3699E 02
0.1904E 02
0.1092E 01
-0.1686E 02
-0.3481E 02
-0.5276E 02
-0.7071E 02
-0.8506E 02
-0.1020E 03
-0.2833E 02
0.4922E 02
-0.6710E 02
0.1045E 02
0.8800E 02
MODEL E3
PLOT AF FEJL MOD VARIABEL 9
0.9084E 02
0.7289E 02
0.5494E 02
0.3699E 02
0.1904E 02
0.1092E 01
-0.1686E 02
-0.3481E 02
-0.5276E 02
-0.7071E 02
-0.8506E 02
-0.1020E 03
-0.2833E 02
0.4922E 02
-0.6710E 02
0.1045E 02
0.8800E 02
MODEL E3
PLOT AF FEJL MOD VARIABEL 10
0.9084E 02
0.7289E 02
0.5494E 02
0.3699E 02
0.1904E 02
0.1092E 01
-0.1686E 02
-0.3481E 02
-0.5276E 02
-0.7071E 02
-0.8506E 02
0.2645E 03
0.1210E 04
0.2206E 04
0.7126E 03
0.1708E 04
0.2704E 04
MODEL E3
PLOT AF FEJL MOD VARIABEL 11
0.9084E 02
0.7289E 02
0.5494E 02
0.3699E 02
0.1904E 02
0.1092E 01
-0.1686E 02
-0.3481E 02
-0.5276E 02
-0.7071E 02
-0.8506E 02
-0.1677E 04
-0.5025E 03
0.7334E 03
-0.1120E 04
0.1154E 03
0.1351E 04

Figure 4.3:

We consider the model:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

i.e.

$$\boldsymbol{y} = \mathrm{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

For the i’th row we have:

$$y_i = (x_{i1}, \cdots, x_{ip}) \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \varepsilon_i$$

or

$$y_i = \mathbf{x}_i\boldsymbol{\theta} + \varepsilon_i.$$

We assume that $\boldsymbol{\varepsilon} \in N(\mathbf{0}, \sigma^2\mathbf{I})$ and therefore have the LS estimate

$$\hat{\boldsymbol{\theta}} = (\mathrm{x}'\mathrm{x})^{-1}\mathrm{x}'\boldsymbol{y}$$

The corresponding residual vector is

$$\boldsymbol{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \mathrm{x}\hat{\boldsymbol{\theta}}$$

i.e.

$$\boldsymbol{r} = [\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}']\mathbf{y}$$

The dispersion matrices for $\hat{\boldsymbol{y}}$ and $\boldsymbol{r}$ are

$$D(\hat{\boldsymbol{y}}) = \mathrm{x}D(\hat{\boldsymbol{\theta}})\mathbf{x}' = \sigma^2\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$$

$$\begin{aligned} D(\boldsymbol{r}) \quad &= \sigma^2[\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'][\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'] \\ &= [\mathbf{I} + \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - 2\mathrm{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}']\sigma^2 \\ &= \sigma^2[\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'] \end{aligned}$$

For the i'th row we find

$$\begin{aligned} V(\hat{y}_i) &= \sigma^2 \mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_i' = \sigma^2 h_i \\ V(r_i) &= \sigma^2(1 - \mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_i') = \sigma^2(1 - h_i) \end{aligned}$$

### The deletion formula

Re-calculation of parameter estimates when discarding a single observation can be done using the formula

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

where the involved matrices are assumed to exist. For the case $\mathbf{A} = \mathbf{x}'\mathbf{x}$ and $\boldsymbol{u} = \boldsymbol{v} = \mathbf{x}_i'$ we have

$$(\mathbf{x}'\mathbf{x} - \mathbf{x}_i'\mathbf{x}_i)^{-1} = (\mathbf{x}'\mathbf{x})^{-1} + \frac{(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_i'\mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}}{1 - h_i}$$

If we denote the $\mathbf{x}$ -matrix where the i'th row is removed $\mathbf{x}(i)$ we have that

$$\mathbf{x}(i)'\mathbf{x}(i) = \mathbf{x}'\mathbf{x} - \mathbf{x}_i'\mathbf{x}_i.$$

**PROOF 4.1.** Omitted. ■

We can now state the relevant expressions.

### Cook's D

A confidence region for the parameter $\boldsymbol{\theta}$ is all the vectors $\boldsymbol{\theta}^*$, which satisfy

$$\frac{1}{p\hat{\sigma}^2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\mathbf{x}'\mathbf{x}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \le F(p, n-p)_{1-\alpha}.$$

We use the left hand side as a measure of the distance between the parameter vector and $\hat{\boldsymbol{\theta}}$. We let $\hat{\boldsymbol{\theta}}(i)$ be the estimate, which corresponds to the deletion of the $i$'th observation

$$\mathbf{y}(i) = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n)'$$

and therefore have

$$\hat{\boldsymbol{\theta}}(i) = [\mathbf{x}(i)'\mathbf{x}(i)]^{-1}\mathbf{x}(i)'\mathbf{y}(i).$$

Cook's D then equals

$$\frac{1}{p\hat{\sigma}^2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(i))'\mathbf{x}'\mathbf{x}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(i)).$$

If Cook's D equals e.g. $F_{60\%}$ then this corresponds to the maximum likelihood estimate moving to the 60 % confidence-ellipsoid for $\boldsymbol{\theta}$. This is a relatively large change when just removing a single observation. In the SAS-program REG one can find Cook's D together with other diagnostics statistics. Some are mentioned below.

### RSTUDENT & STUDENT RESIDUAL

RSTUDENT is a so-called "studentised" residual, i.e.

$$\text{RSTUDENT}_i = \frac{r_i}{\hat{\sigma}(i)\sqrt{1-h_i}},$$

where $\hat{\sigma}(i)^2$ is the estimate of variance corresponding to deletion of the $i$'th observation.

SAS also computes a similar statistic, where the $i$'th observation is not excluded

$$\text{STUDENT RESIDUAL} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}.$$

Since both these types of residual are standardised a sensible rule of thumb is that they should lie within $\pm 2$ or $\pm 3$.

### COVRATIO

COVRATIO measures the change in the determinant of the dispersion matrix for the parameter estimate when excluding the $i$'th observation. We find

$$\text{COVRATIO}_i = \frac{\det[\hat{\sigma}(i)^2(\mathbf{x}(i)'\mathbf{x}(i))^{-1}]}{\det[\hat{\sigma}^2(\mathbf{x}'\mathbf{x})^{-1}]}$$

This quantity "should" be close to 1. If it lies far from 1 then the $i$'th observation has a too large influence. As a rule of thumb $\mid \text{COVRATIO}_i - 1 \mid \leq 3p/n$

### DFFITS

DFFITS is - like Cook's distance - a measure of the total change when deleting one single observation. As a rule of thumb they should lie within say $\pm 2$. A similar rule adjusted for number of observations says within $\pm 2\sqrt{p/(n-p)}$.

$$\begin{aligned}\text{DFFITS} &= \frac{\hat{y}_i - \hat{y}(i)_i}{\hat{\sigma}(i)\sqrt{h_i}} \\ &= \frac{\mathbf{x}_i[\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(i)]}{\hat{\sigma}(i)\sqrt{h_i}}.\end{aligned}$$

### DFBETAS

While DFFITS measures changes in the prediction of an observation corresponding to changes in all parameter estimates, then DFBETAS simply measures the change in each individual parameter estimate. As a rule of thumb they should lie within say $\pm 2$. A rule adjusted for number of observations says within $\pm 2/\sqrt{n}$.

We have

$$\text{DFBETAS}_j = \frac{\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}(i)_j}{\hat{\sigma}(i)\sqrt{(\mathbf{x}'\mathbf{x})^{-1}_{jj}}}.$$

### Call in SAS

All the mentioned statistics can be found using simple SAS statements e.g.

```
proc reg data = sundhed;
model ilt = maxpuls loebetid / r influence;
```

Model statements etc. are the same in REG as in GLM. The diagnostic tests come with the options / `r influence`.

## 4.2 Regression using orthogonal polynomials

When performing a regression analysis using polynomials one can often obtain rather large computational savings and numerical stability by introducing the so-called orthogonal polynomials. In the end this will give the same expression for estimates of the mean value as a function of the independent variable but with considerably smaller computational load.

### 4.2.1 Definition and formulation of the model.

We will assume that a polynomial regression model is given i.e. that

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \xi_0(t_1) & \xi_1(t_1) & \cdots & \xi_k(t_1) \\ \vdots & \vdots & & \vdots \\ \xi_0(t_n) & \xi_1(t_n) & \cdots & \xi_k(t_n) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Here $\xi_i$, $i = 0, 1, \ldots, k$ are known polynomials of $i$'th degree in $t$. We assume that

$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

In the usual fashion we can in this model estimate and test hypotheses regarding the parameters $(\alpha, \beta_1, \ldots, \beta_k)$.

As noted before it would be a great advantage to consider the so-called orthogonal polynomials $\xi_i$ since the computational load will be reduced considerably. We introduce these polynomials in

**DEFINITION 4.1.** By a set of orthogonal polynomials corresponding to the values $t_1, \ldots, t_n$ we mean polynomials $\xi_0, \xi_1, \ldots$ where $\xi_i$ is of $i$'th degree which satisfy

$$\sum_{j=1}^{n} \xi_i(t_j) = 0, \qquad i = 1, 2, \ldots, k \tag{4.1}$$

$$\sum_{j=1}^{n} \xi_\mu(t_j)\xi_\gamma(t_j) = 0, \qquad \mu \neq \gamma. \tag{4.2}$$

▲

**REMARK 4.2.** It is seen that $\xi_0$ is a constant, so 4.1 is of course not used for $\xi_0$. For notational reasons we let $\xi_i(t_j) = \xi_{ij}, \forall_{i,j}$. Later we will return to the problem of actually determining orthogonal polynomials. ▼

If we now assume that the polynomials in the model are orthogonal we find using

$$\xi = \begin{bmatrix} \xi_0 & \cdots & \xi_{k1} \\ \vdots & & \vdots \\ \xi_0 & \cdots & \xi_{kn} \end{bmatrix} = \begin{bmatrix} \xi_0(t_1) & \xi_1(t_1) & \cdots & \xi_k(t_1) \\ \vdots & \vdots & & \vdots \\ \xi_0(t_n) & \xi_1(t_n) & \cdots & \xi_k(t_n) \end{bmatrix},$$

that

$$\xi'\xi = \begin{bmatrix} n\xi_0^2 & 0 & \cdots & 0 \\ 0 & \sum \xi_{1j}^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sum \xi_{kj}^2 \end{bmatrix},$$

i.e. $\xi'\xi$ is a diagonal matrix. We therefore find

$$\hat{\boldsymbol{\beta}} = (\xi'\xi)^{-1}\xi'\boldsymbol{Y} = \begin{bmatrix} \bar{Y}/\xi_0 \\ \sum \xi_{1j} Y_j / \sum \xi_{1j}^2 \\ \vdots \\ \sum \xi_{kj} Y_j / \sum \xi_{kj}^2 \end{bmatrix}$$

and

$$\mathrm{D}(\hat{\boldsymbol{\beta}}) = \sigma^2 \begin{bmatrix} 1/n\xi_0^2 & 0 & \cdots & 0 \\ 0 & 1/\sum \xi_{1j}^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & 1/\sum \xi_{kj}^2 \end{bmatrix}.$$

We now have that the estimators for the parameters are uncorrelated and since we are working in a normal model they are therefore also stochastic independent.

We find that the residual sum of squares is

$$\begin{aligned} \mathrm{SS}_{\mathrm{res}} &= \|\boldsymbol{Y} - \xi\,\hat{\boldsymbol{\beta}}\|^2 \\ &= (\boldsymbol{Y} - \xi\,\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \xi\,\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\xi'\xi\,\hat{\boldsymbol{\beta}} \\ &= \sum Y_j^2 - \{\hat{\alpha}^2 n\xi_0^2 + \hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2\} \\ &= \sum (Y_j - \bar{Y})^2 - \{\hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2\}. \end{aligned}$$

From this we immediately have

**THEOREM 4.1.** We have the following partitioning of the total variation

$$\sum(Y_j-\bar{Y})^2 =$$
$$\hat{\beta}_1^2\sum\xi_{1j}^2+\cdots+\hat{\beta}_k^2\sum\xi_{kj}^2+\sum\{Y_j-\bar{Y}-\hat{\beta}_1\xi_1(t_j)-\cdots-\hat{\beta}_k\xi_k(t_j)\}^2,$$

or with an easily understood notation

$$\mathrm{SS}_{\mathrm{tot}} = \mathrm{SS}_{1.\mathrm{grad}} + \cdots + \mathrm{SS}_{k.\mathrm{grad}} + \mathrm{SS}_{\mathrm{res}},$$

i.e. the total sum of squares has been partitioned in terms corresponding to each polynomial plus the residual sum of squares. The degrees of freedom are $n-1$ respectively $1,\ldots,1$ and $n-k-1$. ▲

**PROOF 4.2.** Follows trivially from the above mentioned. ■

Using the partition theorem we furthermore have

**THEOREM 4.2.** The sums of squares which have been stated in the previous theorem are stochastic independent with expected values

$$\begin{aligned}\mathrm{E}(\mathrm{SS}_{i.\mathrm{deg}}) &= \mathrm{E}(\hat{\beta}_i^2\sum_j\xi_i(t_j)^2)\\ &= \sigma^2+\beta_i^2\sum_j\xi_i(t_j)^2, \qquad i=1,\ldots,k.\end{aligned}$$

and

$$\mathrm{E}(\mathrm{SS}_{\mathrm{res}}) = \mathrm{E}[\sum_j(Y_j-\bar{Y}-\cdots-\hat{\beta}_k\xi_k(t_j))^2] = (n-k-1)\sigma^2.$$

Finally

$$\frac{1}{\sigma^2}\mathrm{SS}_{\mathrm{res}} \in \chi^2(n-k-1),$$

and if $\beta_i = 0$ -

$$\frac{1}{\sigma^2}\mathrm{SS}_{i.\mathrm{deg}} \in \chi^2(1).$$

▲

**PROOF 4.3.** Obvious. ■

The theorems contain the necessary results to be able to establish tests for the hypotheses

$$H_{0i} : \beta_i = 0 \quad \text{against} \quad H_{1i} : \beta_i \neq 0.$$

We collect the results in a analysis of variance table

| Variation | SS | $f$ | E(SS/$f$) |
|---|---|---|---|
| Linear | $SS_{1.\text{deg}}$ | 1 | $\sigma^2 + \beta_1^2 \sum_j \xi_1(t_j)^2$ |
| Quadratic | $SS_{2.\text{deg}}$ | 1 | $\sigma^2 + \beta_2^2 \sum_j \xi_2(t_j)^2$ |
| Cubic | $SS_{3.\text{deg}}$ | 1 | $\sigma^2 + \beta_3^2 \sum_j \xi_3(t_j)^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $k$'th order | $SS_{k.\text{deg}}$ | 1 | $\sigma^2 + \beta_k^2 \sum_j \xi_k(t_j)^2$ |
| Residual | $SS_{\text{res}}$ | $n-k-1$ | $\sigma^2$ |
| Total | $SS_{\text{tot}}$ | $n-1$ | |

**REMARK 4.3.** The big advantage of using orthogonal polynomials in the regression analysis is that one without changing any of the previous computations can introduce polynomials of degree $(p+1)$ and degree $(p+2)$ etc. When establishing the order for the describing polynomial we will usually continue (estimation and) testing until 2 successive $\beta_i$ 's = 0 since contributions which are caused by terms of even degree and terms of odd degree are different in nature. This is, however, a rule of thumb which should be used with caution. If we e.g. have an idea which is based on physical considerations that terms of 5th order are important, then we would not stop the analysis just because the 3rd and 4th degree coefficients do not differ significantly from 0. ▼

## 4.2.2 Determination of orthogonal polynomials.

It is readily seen, that multiplication with a constant does not change the orthogonality conditions 4.1 and 4.2. We therefore choose to let

$$\xi_0(t) = \xi_0 = 1.$$

The polynomial of 1st degree is

$$\xi_1(t) = t + a,$$

since we can choose the coefficient for $t$ as 1. From 4.1 we have

$$0 = \sum_{j=1}^{n} \xi_1(t_j) = \sum_{j=1}^{n}(t_j + a) = \sum_{j=1}^{n} t_j + na,$$

or

$$a = -\frac{1}{n}\sum_{j=1}^{n} t_j = -\bar{t},$$

i.e.

$$\xi_1(t) = t - \bar{t}.$$

We can then choose $\xi_2$ as a linear combination of 1, $\xi_1$ $\xi_1^2$, i.e.

$$\xi_2(t) = a_{02} + a_{12}(t - \bar{t}) + a_{22}(t - \bar{t})^2.$$

From 4.1 we have

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \xi_2(t_j) = na_{02} + a_{12}\sum_j (t_j - \bar{t}) + a_{22}\sum_j (t_j - \bar{t})^2 \\ \frac{a_{02}}{a_{22}} &= -\frac{1}{n}\sum_j (t_j - \bar{t})^2. \end{aligned}$$

From 4.2 we have

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \xi_1(t_j)\xi_2(t_j) \\ &= a_{02}\sum_j (t_j - \bar{t}) + a_{12}\sum_j (t_j - \bar{t})^2 + a_{22}\sum_j (t_j - \bar{t})^3 \\ &= a_{12}\sum_j (t_j - \bar{t})^2 + a_{22}\sum_j (t_j - \bar{t})^3. \end{aligned}$$

From this we get

$$\frac{a_{12}}{a_{22}} = -\frac{\sum_j (t_j - \bar{t})^3}{\sum_j (t_j - \bar{t})^2}.$$

$\xi_3, \xi_4$ etc. are found analogously.

The computations are especially simple if the $t_j$ 's are equidistant. Then we let

$$u_j = \frac{t_j - (t_1 - w)}{w},$$

where $w = t_2 - t_1 = t_{i+1} - t_i$. We then have

$$u_i = i, \qquad i = 1, \ldots, n.$$

Corresponding to the values $1, \ldots, n$ we then have the polynomials given by

$$\xi_0(t) = 1 \tag{4.3}$$

$$\xi_1(t) = t - \frac{n+1}{2} \tag{4.4}$$

$$\xi_{i+1}(t) = \xi_1(t)\xi_i(t) - \frac{i^2(n^2 - i^2)}{4(4i^2 - 1)}\xi_{i-1}(t). \tag{4.5}$$

In the table on p. 169 we have given some values of orthogonal polynomials $\xi_1, \ldots, \xi_k$, $k \leq 5$, with $t = 1, \ldots, n$ for $n = 1, \ldots, 8$.

In order to avoid fractional numbers and large values we have chosen to give polynomials where the coefficient to the term of largest degree is a number $\lambda$ which is also seen in the table. Furthermore we have stated the terms

$$D = \sum_{j=1}^{n} \xi_i(j)^2 = \sum_{j=1}^{n} \xi_{ij}^2.$$

We now give an illustrative

**EXAMPLE 4.1.** In the following table corresponding values of reaction temperature and yield of a process (in a fixed time) have been given.

| **Temperature** | **Yield** |
|---|---|
| 200°F | 0.75 oz. |
| 210°F | 1.00 oz. |
| 220°F | 1.35 oz. |
| 230°F | 1.80 oz. |
| 240°F | 2.60 oz. |
| 250°F | 3.60 oz. |
| 260°F | 5.45 oz. |

We will try to describe the yield as a function of temperature using a polynomial. We

| $n$ | 3 | | 4 | | | 5 | | | | 6 | | | | | 7 | | | | | 8 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t$ | $\xi_1$ | $\xi_2$ | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_4$ | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_4$ | $\xi_5$ | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_4$ | $\xi_5$ | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_4$ | $\xi_5$ |
| 1 | $-1$ | 1 | $-3$ | 1 | $-1$ | $-2$ | 2 | $-1$ | 1 | $-5$ | 5 | $-5$ | 1 | $-1$ | $-3$ | 5 | $-1$ | 3 | $-1$ | $-7$ | 7 | $-7$ | 7 | $-7$ |
| 2 | 0 | $-2$ | $-1$ | $-1$ | 3 | $-1$ | $-1$ | 2 | $-4$ | $-3$ | $-1$ | 7 | $-3$ | 5 | $-2$ | 0 | 1 | $-7$ | 4 | $-5$ | 1 | 5 | $-13$ | 23 |
| 3 | 1 | 1 | 1 | $-1$ | $-3$ | 0 | $-2$ | 0 | 6 | $-1$ | $-4$ | 4 | 3 | $-10$ | $-1$ | $-3$ | 1 | 1 | $-5$ | $-3$ | $-3$ | 7 | $-3$ | $-17$ |
| 4 | | | 3 | 1 | 1 | 1 | $-1$ | $-2$ | $-4$ | 1 | $-4$ | $-4$ | 2 | 10 | 0 | $-4$ | 0 | 6 | 0 | $-1$ | $-5$ | 3 | 9 | $-15$ |
| 5 | | | | | | 2 | 2 | 1 | 1 | 3 | $-1$ | $-7$ | $-3$ | $-5$ | 1 | $-3$ | $-1$ | 1 | 5 | 1 | $-5$ | $-3$ | 9 | 15 |
| 6 | | | | | | | | | | 5 | 5 | 5 | 1 | 1 | 2 | 0 | $-1$ | $-7$ | $-4$ | 3 | $-3$ | $-7$ | $-3$ | 17 |
| 7 | | | | | | | | | | | | | | | 3 | 5 | 1 | 3 | 1 | 5 | 1 | $-5$ | $-13$ | $-23$ |
| 8 | | | | | | | | | | | | | | | | | | | | 7 | 7 | 7 | 7 | 7 |
| $D$ | 2 | 6 | 20 | 4 | 20 | 10 | 14 | 10 | 70 | 70 | 84 | 180 | 28 | 252 | 28 | 84 | 6 | 154 | 84 | 168 | 168 | 264 | 616 | 2184 |
| $\lambda$ | 1 | 3 | 2 | 1 | $\frac{10}{3}$ | 1 | 1 | $\frac{5}{6}$ | $\frac{35}{12}$ | 2 | $\frac{3}{2}$ | $\frac{5}{3}$ | $\frac{7}{12}$ | $\frac{21}{10}$ | 1 | 1 | $\frac{1}{6}$ | $\frac{7}{12}$ | $\frac{7}{20}$ | 2 | 1 | $\frac{2}{3}$ | $\frac{7}{12}$ | $\frac{7}{10}$ |

Table 4.1: Values of orthogonal polynomials.

will assume that the assumptions in order to perform a regression analysis are fulfilled. First we transform the temperatures $\tau_i$, $i = 1, \ldots, 7$ by means of the following relation

$$t_i = \frac{\tau_i - (200 - 10)}{10} = \frac{\tau_i - 190}{10}$$

We then get the values $t_1, \ldots, t_7 = 1, \ldots, 7$.

We give the computations in the following table

| $t_\mathrm{j}$ | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_4$ | $\xi_5$ | $y_\mathrm{j}$ |
|---|---|---|---|---|---|---|
| 1 | $-3$ | 5 | $-1$ | 3 | $-1$ | 0.75 |
| 2 | $-2$ | 0 | 1 | $-7$ | 4 | 1.00 |
| 3 | $-1$ | $-3$ | 1 | 1 | $-5$ | 1.35 |
| 4 | 0 | $-4$ | 0 | 6 | 0 | 1.80 |
| 5 | 1 | $-3$ | $-1$ | 1 | 5 | 2.60 |
| 6 | 2 | 0 | $-1$ | $-7$ | $-4$ | 3.60 |
| 7 | 3 | 5 | 1 | 3 | 1 | 5.45 |
| $\sum \xi_\mathrm{ij}^2$ | 28 | 84 | 6 | 154 | 84 | $16.55 = \sum y_\mathrm{j}$ |
| $\sum \xi_\mathrm{ij} y_\mathrm{j}$ | 20.55 | 11.95 | 0.85 | 1.15 | 0.55 | $56.0475 = \sum y_\mathrm{j}^2$ |
| $\lambda$ | 1 | 1 | $\frac{1}{6}$ | $\frac{7}{12}$ | $\frac{7}{20}$ | |

$$\begin{aligned} \sum (y_i - \bar{y})^2 &= 56.0475 - \frac{16.55^2}{7} \\ &= 56.0475 - 39.1289 \\ &= 16.9186 \end{aligned}$$

$$\begin{aligned}
\hat{\alpha} &= \tfrac{16.55}{7} = 2.36 \\
\hat{\beta}_1 &= \tfrac{20.55}{28} = 0.7339 \quad \mathrm{SS}_{1.\mathrm{grad}} = \tfrac{20.55^2}{28} = 15.0822 \\
\hat{\beta}_2 &= \tfrac{11.95}{84} = 0.1423 \quad \mathrm{SS}_{2.\mathrm{grad}} = \tfrac{11.95^2}{84} = 1.7000 \\
\hat{\beta}_3 &= \tfrac{0.85}{6} = 0.1417 \quad \mathrm{SS}_{3.\mathrm{grad}} = \tfrac{0.85^2}{6} = 0.1204 \\
\hat{\beta}_4 &= \tfrac{1.15}{154} = 0.0075 \quad \mathrm{SS}_{4.\mathrm{grad}} = \tfrac{1.15^2}{154} = 0.0086 \\
\hat{\beta}_5 &= \tfrac{0.55}{84} = 0.0065 \quad \mathrm{SS}_{5.\mathrm{grad}} = \tfrac{0.55^2}{84} = 0.0036
\end{aligned}$$

We summarise the result in the following table.

We see that the terms of 1st, 2nd and 3rd degree are significant and the two following are not significant, so we will choose a polynomial of 3rd degree for the description.

| Variation | SS | $f$ | $S^2$ | Test | F-percentile |
|---|---|---|---|---|---|
| Total | 16.9186 | 6 | | | |
| 1. degree | 15.0822 | 1 | 15.0822 | | |
| Residual 1 | 1.8364 | 5 | 0.3673 | 41.06 | 99.8% |
| 2. degree | 1.7000 | 1 | 1.7000 | | |
| Residual 2 | 0.1364 | 4 | 0.0341 | 49.85 | 99.7% |
| 3. degree | 0.1204 | 1 | 0.1204 | | |
| Residual 3 | 0.0160 | 3 | 0.0053 | 22.72 | 98.0% |
| 4. degree | 0.0086 | 1 | 0.0086 | | |
| Residual 4 | 0.0074 | 2 | 0.0037 | 2.32 | 75.0% |
| 5. degree | 0.0036 | 1 | 0.0036 | | |
| Residual 5 | 0.0038 | 1 | 0.0038 | 0.95 | $<$ 50.0% |

From the recursion formulas 4.3, 4.4 and 4.5 we get - since $n = 7$

$$\begin{aligned}
\xi_1(t) &= t - 4 \\
\xi_2(t) &= (t-4)^2 - \frac{48}{12} \\
&= t^2 - 8t + 12 \\
\xi_3(t) &= (t-4)(t^2 - 8t + 12) - \frac{4 \cdot 45}{4 \cdot 15}(t-4) \\
&= t^3 - 12t^2 + 41t - 36.
\end{aligned}$$

Since $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 1/6$ we get the following estimated polynomial

$$\begin{aligned}
\hat{\mu}(t) &= 2.36 + 1 \cdot \hat{\beta}_1 \xi_1(t) + 1 \cdot \hat{\beta}_2 \xi_2(t) + \frac{1}{6}\hat{\beta}_3 \xi_3(t) \\
&= 0.0236t^3 - 0.1409t^2 + 0.5631t + 0.2818.
\end{aligned}$$

Since

$$t_i = \frac{\tau_i - 190}{10},$$

we can get an expression where the original temperatures are given by entering this relationship in the expression for $\hat{\mu}(t)$. We find

$$\mathrm{g}(\tau) = 0.000024\tau^3 - 0.014861\tau^2 + 3.147610\tau - 223.15440.$$

The estimated polynomial is shown together with the original data in the following figure. ◆

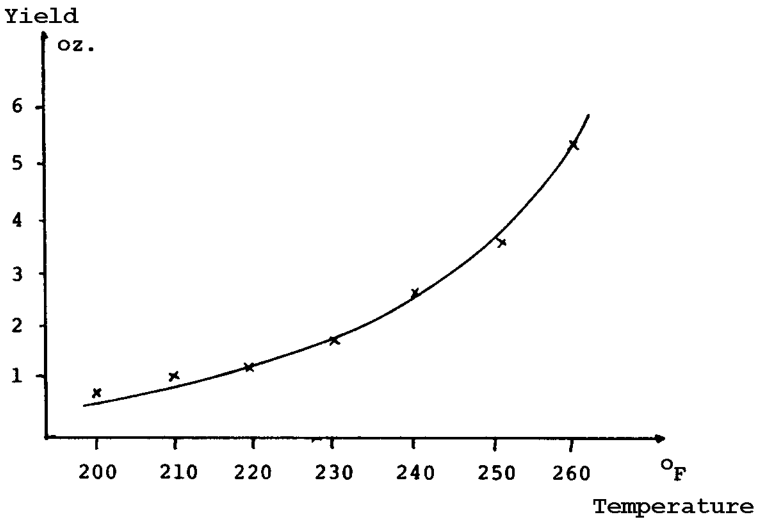


Figure 4.4: The correspondence between temperature and yield by the process given in example 4.1.

## 4.3 Choice of the ”best” regression equation

In this section we will consider the problem of choosing a suitable (small) number of independent variables giving a reasonable description of our data.

### 4.3.1 The Problem.

If we are in the (unpleasant) situation of not being able to formulate a model based upon physical relationships for the phenomena we are studying, we will often simply register all the variables we think could have some effect on our observed values. If we then compute a regression by e.g. polynomials in these independent variables (from a Taylor-approximation point of view) we will very quickly have an enormous number of terms in our regression. If we start off with 10 basic-variables $x_1, \ldots, x_{10}$, then an ordinary second order polynomial in these variables will contain 66 terms. If we include 3rd degree we have on the order of 150 terms. Expressions containing so many terms will (if it is at all possible to estimate all the parameters) be very tedious to work with. If we e.g. wish to determine optimal production conditions for a chemical process we could estimate the response surface and find the maximum for this. This will be extremely difficult if there are many variables involved. We would therefore seek to find a considerably smaller number of terms which will give a reasonably good description of the variation in the material (cf. the section on ridge regression).

It is important, however, to note that an expression found by applying the methods discussed in the following should be used with caution. It will (probably) be an expression which describes the data at hand very well. Whether or not the method is adequate to predict future observations depends upon if the expression also describes the physical conditions well enough. One way of determining this is in the first instance only to base the estimation on half of the data and then compare the other half with the estimated model. If the degree of agreement is reasonable we have the indication that the model is not completely inadequate as a prediction model.

We will use a single illustrative example for all the methods we will describe. In order for it to be possible to overlook (and maybe check) the individual calculations we have only taken a very small part of the original data material. We should therefore not evaluate the suitability of the methods by means of the example, but only use it as an illustration of the principals and the way of going about these. The data are some corresponding measurements of the quality $Y$ of a food additive (measured using viscosity) and some some production parameters $x_1$, $x_2$ $x_3$ (pressure, temperature and degree of neutralisation). In order to simplify the calculations the data are coded, i.e. the variables have had some constants subtracted and been divided by others. We have the following measurements

| $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 4.9 | 0 | 0 | 2 |
| 3.0 | 1 | 0 | 1 |
| 0.2 | 1 | 1 | 0 |
| 2.9 | 1 | 2 | 2 |
| 6.4 | 2 | 1 | 2 |

Experience shows that within a suitably small region of variation of the production parameters it is reasonable to assume that the quality shows a linear dependency on these. We will therefore use the following model

$$\mathrm{E}(Y|\boldsymbol{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

or in matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix},$$

$$\varepsilon \in \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

In the numerical appendix (p. 183) all the $2^3$ regression analysis with $y$ as dependent variable and one of the more of the $x$'s as independent variables are shown. The following models are possible

$$\begin{array}{llllllllll}
M & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & + & \beta_2 x_2 & + & \beta_3 x_3 \\
H_{12} & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & + & \beta_2 x_2 & & \\
H_{13} & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & + & & & \beta_3 x_3 \\
H_{23} & : & \mathrm{E}(Y) & = & \alpha & + & & & \beta_2 x_2 & + & \beta_3 x_3 \\
H_1 & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & & & & \\
H_2 & : & \mathrm{E}(Y) & = & \alpha & + & & & \beta_2 x_2 & & \\
H_3 & : & \mathrm{E}(Y) & = & \alpha & + & & & & & \beta_3 x_3 \\
H_0 & : & \mathrm{E}(Y) & = & \alpha & & & & & &
\end{array}.$$

For each of these 8 models the estimators for $\alpha$ and the $\beta$ 's are shown, we find the projection of the observation vector onto the sub-space corresponding to the model we determine the residual vector, the squared length of the residual vector (the residual sum of squares), the estimate of variance, and the (squared) multiple correlation coefficient. After that we show the analysis of variance tables for the possible sequences of successive testings of hypotheses: that the mean vector is a member of successively smaller (lower dimension) sub-spaces in sequences like

$$M \supseteq H_{12} \supseteq H_2 \supseteq H_0.$$

The above mentioned sequence of sub-spaces corresponds to successive testing of the hypothesis

$$\beta_3 = 0, \quad \beta_1 = 0, \quad \beta_2 = 0.$$

There are 6 (= 3!) possible tables of this type. Finally we show some partial correlation matrices. If we let $y = x_4$ the empirical variance-covariance matrix is (as usual) defined by the $(i, j)$'th element being

$$S_{ij} = \frac{1}{n-1} \sum_{\mu} (x_{i\mu} - \bar{x}_i)(x_{j\mu} - \bar{x}_j).$$

The $(i, j)$'th element in the correlation matrix is then

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}}.$$

Using the formula on p. 84 in section 2 we then compute the partial correlations for given $x_3$ and for given $x_2$, $x_3$.

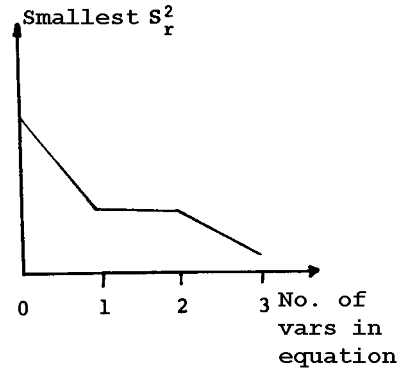
We now have enough background material to mention some of the most popular ways of selecting single independent variables to describe the variation of the dependent variable.

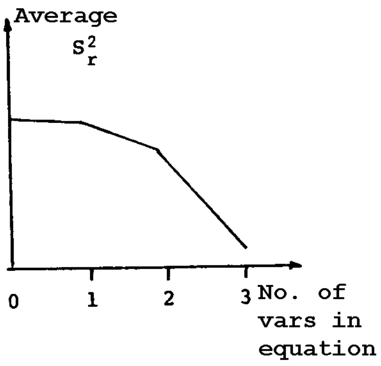
### 4.3.2 Examination of all regressions.

This method can of course only be used if there are reasonably few variables. We summarise the result from the appendix in the following table

| Model | Multiple | Residual variance | Average |
|---|---|---|---|
| | $R^2$ | $S_r^2$ | of $S_r^2$ |
| $H_0 \quad : E(Y) = \alpha$ | 0 | 5.47 | 5.47 |
| $H_1 \quad : E(Y) = \alpha + \beta_1 x_1$ | 5.1% | 6.91 | 5.35 |
| $H_2 \quad : E(Y) = \alpha + \beta_2 x_2$ | 3.8% | 7.01 | |
| $H_3 \quad : E(Y) = \alpha + \beta_3 x_3$ | 70.8% | 2.13 | |
| $H_{12} \quad : E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ | 15.3% | 9.26 | 4.68 |
| $H_{13} \quad : E(Y) = \alpha + \beta_1 x_1 + \beta_3 x_3$ | 76.0% | 2.63 | |
| $H_{23} \quad : E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$ | 80.4% | 2.14 | |
| $M \quad : E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ | 97.1% | 0.634 | 0.634 |

Looking at the multiple correlation coefficient quickly indicates that we do not gain so much by going from one variable $(x_3)$ up to 2 variables. The crucial jump happens when including all 3 variables. Considerations of this type lead us rather to just use $x_3$ i.e. the model $\mathrm{E}(Y|x) = \alpha + \beta_3 x_3$. This decision is strengthened by looking at the residual variance $S_r^2$. We then see that $S_r^2$ for the best equation in one variable is less than for the best equation in two variables which strongly indicates that we should just look at one variable (or use all three). If we besides looking at the smallest $S_r^2$ also look at the average values and depict them by number of included variables we have graphs like





This also indicates that the number of variables in an equation should be either 1 or 3 (there is no significant improvement by going from 1 to 2).

If we only look at the graph with the average values it is not obvious that we should include any independent variable at all. We could therefore test if $\beta_3 = 0$ in the model $H_3 \qquad (\mathrm{E}(y|x) = \alpha + \beta_3 x_3)$

$$\frac{\|p_{H_0}(\boldsymbol{y}) - p_{H_3}(\boldsymbol{y})\|^2/1}{\|\boldsymbol{y} - p_{H_3}\|^2/3} = \frac{21.868 - 6.38}{6.38/3} \simeq 7.28.$$

Therefore we will reject $\beta_3 = 0$ at all levels greater than $8\%$.

As a conclusion of these (rather loose) considerations we will use the model $H_3$:

$$\mathrm{E}(Y|x) = \alpha + \beta_3 x_3 \simeq 0.4 + 2.2 x_3.$$

Here $\simeq$ means estimated at). The estimate of the error (the variance) on the measurements is (estimated with 3 degrees of freedom):

$$s^2 = 2.13.$$

**REMARK 4.4.** It should be added here that the idea of looking at the averages of the residual variances does seem a bit dubious. It has been included merely because the method seems to enjoy widespread use - at least in some parts of the literature. ▼

### 4.3.3 Backwards elimination.

This method is far more economical with respect to computational time than the previous one. Here we start with the full model $M$ and then investigate which of the coefficients which has the smallest F-value for a test of the hypothesis that the coefficient might be 0.

This variable is then excluded and the procedure is repeated with the $k - 1$ remaining variables etc.

We can then stop the procedure when none of the remaining variables have an F-value less than the $1 - \alpha$ quantile in the relevant F-distribution.

We can illustrate the procedure using our example. We collect the data in the following table.

From the table can be seen that this procedure also will end with the model $H_3$: $\mathrm{E}(y) = \alpha + \beta_3 x_3$ when we use an $\alpha$ , greater than $8\%$.

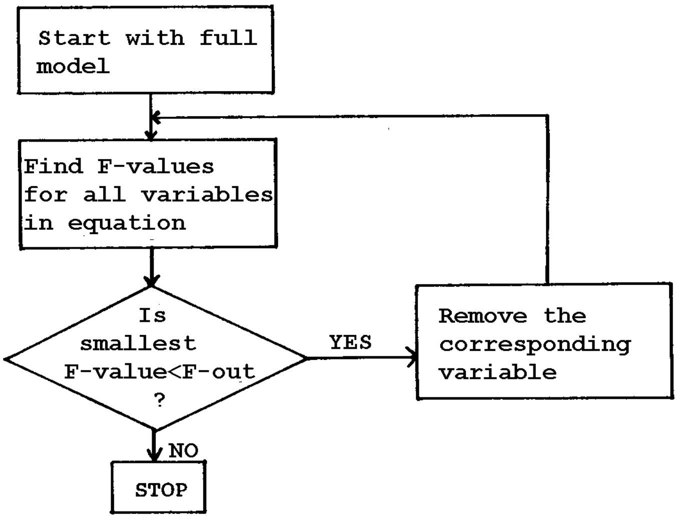

Start with full model
Find F-values for all variables in equation
Is smallest F-value<F-out ?
YES
Remove the corresponding variable
NO
STOP

Figure 4.5: Flow diagram for Backwards-elimination procedure in stepwise regression analysis.

| Step | F-value for test of$\beta_i = 0$ | / Quantile in F-distribution |
|---|---|---|
| Model : $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ | | |
| 1 | $\beta_1 : \frac{3.625/1}{0.634/1} = \quad 5.76$ | $= F(1,1)_{0.71}$ |
| | $\beta_2 : \frac{4.621/1}{0.634/1} = \quad 7.29$ | $= F(1,1)_{0.72}$ |
| | $\beta_3 : \frac{17.879/1}{0.634/1} = \quad 28.20$ | $= F(1,1)_{0.86}$ |
| Remove $x_1$ : Model is now : $E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$ | | |
| 2 | $\beta_2 : \frac{2.095/1}{4.285/2} = \quad 0.98$ | $= F(1,2)_{0.55}$ |
| | $\beta_3 : \frac{16.757/1}{4.285/2} = \quad 7.82$ | $= F(1,2)_{0.88}$ |
| Remove $x_2$ : Model is now : $E(Y) = \alpha + \beta_3 x_3$ | | |
| 3 | $\beta_3 : \frac{15.488/1}{6.38/3} = \quad 7.28$ | $= F(1,3)_{0.92}$ |

The disadvantage with this method is, that we have to solve the full regression model which can be a problem if there many independent variables.

This problem is circumvented by using the following procedure.

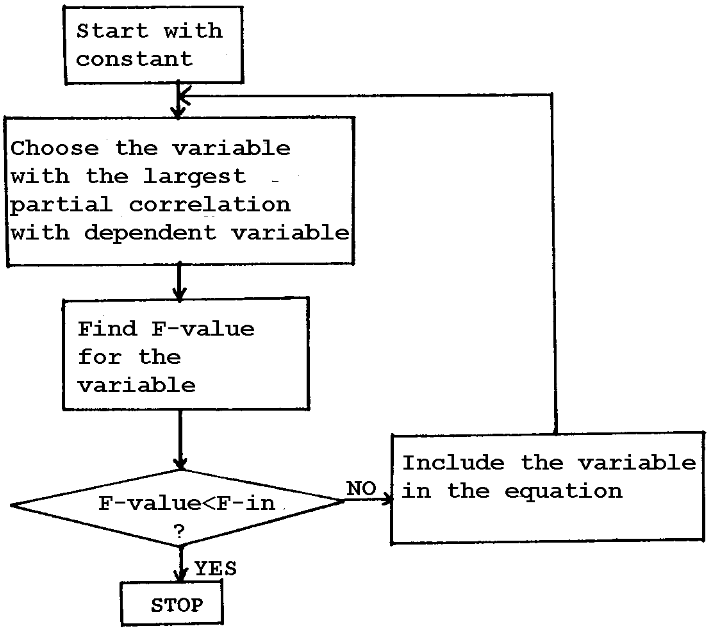

Start with constant
Choose the variable with the largest partial correlation with dependent variable
Find F-value for the variable
F-value<F-in ?
NO
Include the variable in the equation
YES
STOP

Figure 4.6: Flow diagram for Forward-selection procedure in stepwise regression analysis

### 4.3.4 Forward selection

In this procedure we start with the constant term in the equation only. Then we choose the independent variable which shows the greatest correlation with the dependent variable. We then perform an F-test to check if this coefficient is significantly different from 0. If so, then it is included in the model.

Among the independent variables not yet included we now choose the one that has the greatest (absolute) partial correlation coefficient with the dependent variable given the variables already in the equation. We perform an F-test to check if the new variable has contributed to the reduction of the residual variance, i.e. if the coefficient for it is different from 0. If so, continue as before if not stop the analysis.

In our example the steps will be the following

1) From the correlation matrix (p. 188) we see that $x_3$ has the greatest correlation coefficient with y, viz. 0.8416. We test if $\beta_3$ in the model $\mathrm{E}(Y) = \alpha + \beta_3 x_3$ can be assumed to be 0 we have the test statistic (see p. 187).

$$\frac{15.488/1}{6.38/3} = 7.28 \simeq \mathrm{F}(1,3)_{0.92}.$$

If we use $\alpha$ = 10% we continue (since we then reject $\beta_3 = 0$).

2) From the partial correlation matrix given $x_3$ (p. 188) we see that the variable which has the greatest partial correlation coefficient with the $y$'s (given that $x_3$ is in the equation) is $x_2$ ($\rho_{x_2 y|x_3} = -0.5728$). We include $x_2$ and check if $\beta_2$ in the model

$$\mathrm{E}(y) = \alpha + \beta_2 x_2 + \beta_3 x_3$$

can be assumed to be 0. We have the test statistic (see p. 188)

$$\frac{2.095/1}{4.2855/2} = 0.98 \simeq \mathrm{F}(1,2)_{0.55}.$$

Since we were using $\alpha = 10\%$, then this statistic is not significantly different from 0, and we stop the analysis here without including $x_2$ . The resulting model is

$$\mathrm{E}(Y) = \alpha + \beta_3 x_3,$$

where $\alpha$ and $\beta$ are estimated as earlier. We especially note that $x_1$ has not been included in the equation at all.

**REMARK 4.5.** If we had used $\alpha = 50\%$ we would have continued the analysis and considered the partial correlations given $x_2$ and $x_3$. According to the matrix p. 189 the partial correlation coefficient between $y$ and $x_1$ given that $x_2$ and $x_3$ are included in the equation

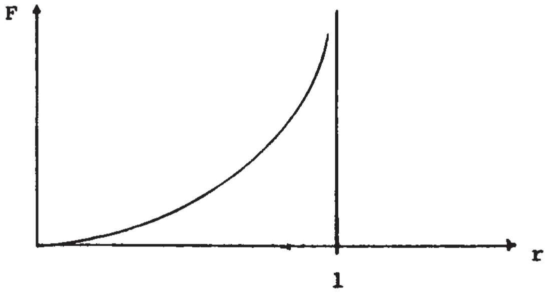
$$\rho_{x_1 y|x_2 x_3} = 0.8956.$$

Now $x_1$ is the only variable not included so it is trivially the one which has the greatest partial correlation with $y$. We now include $x_1$ in the equation and investigate if $\beta_1$ in the model $\mathrm{E}(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ is significantly different from 0. The test statistic is (p. 187)

$$\frac{3.652/1}{0.634/1} = 5.76 \simeq \mathrm{F}(1,1)_{0.71}.$$

In the case we have seen that the equation was extended considerably just by changing $\alpha$. It is important to note that changes in $\alpha$ can have drastic consequences for the resulting model. ▼

**REMARK 4.6.** The procedure of choosing the variable which has the greatest partial correlation with the dependent variable at every step, is equivalent to choosing the variable which has the greatest F-value in the partial F-test. This result comes from



the relation between the partial correlation coefficient and the F-statistic.This is of the form

$$\mathrm{F} = \mathrm{g}(r) = \frac{r^2}{1-r^2} \cdot f,$$

where $f$ is the number of degrees of freedom for the denominator (cf p. 154). This relation is monotonously increasing

If we e.g. in step 2 want to compute the F-test statistic from the correlation matrix we would get

$$\mathrm{F} = \frac{(-0.5728)^2}{1-(-0.5728)^2} \cdot 2 = 0.98.$$

It is further seen that the mentioned criterion is equivalent to at each step always taking the variable which gives the greatest reduction in residual sum of squares. ▼

**REMARK 4.7.** In many of the existing standard regression programmes it is not possible to specify an $\alpha$-value. We must then instead give a fixed number as the limit for the F-test statistics we will accept respectively reject. We must then by looking at a table over F-quantile find a suitable value. If we e.g. wish to have $\alpha = 5\%,$ we see that we should use the value 4 since

$$\mathrm{F}(1, n)_{0.95} \simeq 4,$$

for reasonably large values of $n$. ▼

The 'forward selection' method has its merits compared to the backward elimination method in that we do not have to compute the total equation. The greatest drawback with the method is probably that we do not take into account that some of the variables could be redundant if others enter at a later stage. If we e.g. have that $x_1 = ax_2 + bx_3$ (approximately) and that $x_1$ has been chosen as the most important variable. If we then at a later stage in the analysis also include $x_2$ and $x_3$ then it is obvious that we no longer need $x_1$. It should therefore be removed. This happens in the last method we mention.

### 4.3.5 Stepwise regression.

The name is badly chosen since we could equally well call the last two methods by this name. There are also many authors who use the name stepwise regression as a common name for a number of different procedures. In this text we will specifically have the following method in mind. Choice of the variable to enter the equation is performed like in the forward selection procedure, but at every single step we check each of the variables in the equation as if they were the last included variable. We then compute an F-test statistic for all the variables in the equation. If some of these are smaller than the $1-\alpha$ quantile in the relevant F-distribution then the respective variable is removed. If we look at our standard example we get the following steps $(\alpha_{\text{in}} = 50\%, \alpha_{\text{out}} = 40\%)$.

1) $x_3$ is included as in the forward selection procedure and we test if $\beta_3$ is significantly different from 0. The test statistic and the conclusion are as before.

2) We now include $x_2$ . We compute the partial F-test for $\beta_2$ (in the model $\mathrm{E}(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$):

$$x_2 : \quad \text{F-value} = \frac{2.095/1}{4.285/2} = 0.98 \simeq \mathrm{F}(1,2)_{0.55}.$$

Then we compute a partial F-test for $\beta_3$ (in the model $\mathrm{E}(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$). Using the table p. 187 we find that

$$x_3 : \quad \text{F-value} = \frac{16.757/1}{4.285/2} = 7.82 \simeq \mathrm{F}(1,2)_{0.88}.$$

3) We now again remove $x_2$ from the equation since $0.55 < 0.60$. The difference at this step between the forward selection procedure and the stepwise procedure is that we also compute an F-value for $x_3$ and thereby have a possibility that $x_3$ again will be eliminated from the equation. This was not possible by the ordinary forward selection procedure.

4) The only remaining variable is $x_1$. It has a partial F-value of

$$x_1 : \quad \text{F-value} = \frac{1.125/1}{5.255/2} = 0.43 < \mathrm{F}(1,2)_{0.50},$$

so it does not enter the equation at all.

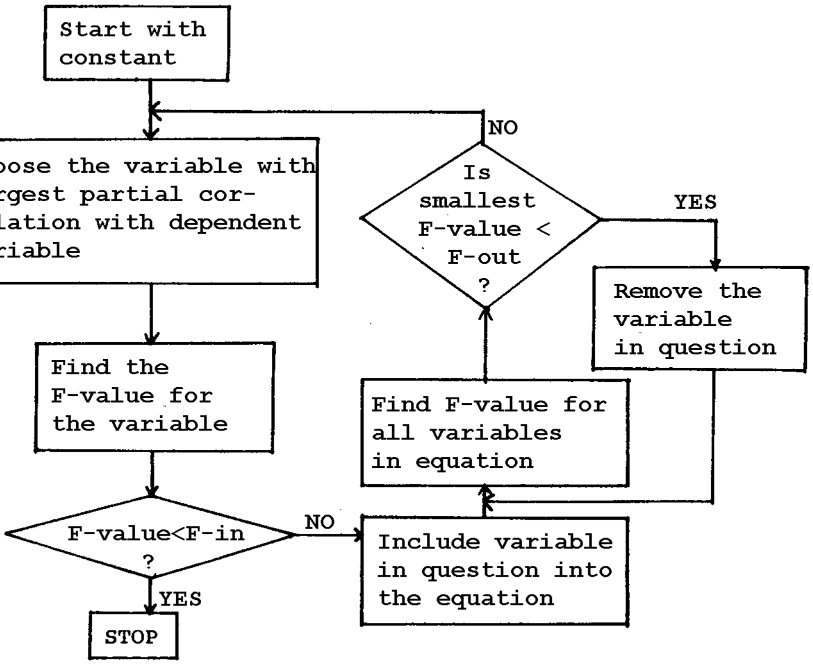

Start with constant
Choose the variable with largest partial correlation with dependent variable
Find the F-value for the variable
F-value<F-in ?
YES
STOP
NO
Include variable in question into the equation
Find F-value for all variables in equation
Is smallest F-value < F-out ?
NO
YES
Remove the variable in question

Figure 4.7: Flow diagram for Stepwise-Regression procedure in stepwise regression analysis.

The analysis stops and we have the model

$$\mathrm{E}(Y) = \alpha + \beta_3 x_3.$$

**REMARK 4.8.** The reason why we investigated the partial F-value under 2, but not under 4 is that $x_1$ does not enter the equation at all since

$$0.43 < \mathrm{F}(1, 2)_{0.50} = \mathrm{F}_{1-\alpha_{ind}}.$$

On the other hand $x_2$ was entered into the equation since

$$0.98 < \mathrm{F}(1, 2)_{0.55} > \mathrm{F}_{1-\alpha_{ind}}.$$

▼

**REMARK 4.9.** Like the section on the forward selection procedure we can note that we are often forced to use fixed F-values instead of $1-\alpha$ quantiles. If we do not use the same level when determining if we want to include more variables as we do when determining if some of the variables should be removed, we will often let the last value be about half as big as the first one i.e.

$$\text{F-out of equation} = \frac{1}{2}\text{F-into equation.}$$

(This is the opposite of what we actually used in the example). ▼

### 4.3.6 Some existing programmes.

Since these programmes are very old we will skip this section and go directly to the

### 4.3.7 Numerical appendix.

In this appendix we will show the calculation of the numbers used in the previous sections. It should not be necessary to go through all these computations but they are shown, so we with the help of these should be able to check our understanding of the different principles.

**A. Data:**

| $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 4.9 | 0 | 0 | 2 |
| 3.0 | 1 | 0 | 1 |
| 0.2 | 1 | 1 | 0 |
| 2.9 | 1 | 2 | 2 |
| 6.4 | 2 | 1 | 2 |

**B. Basic Model:** $\mathrm{E}(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ or

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\boldsymbol{\varepsilon} \in \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

**C. Estimators in sub-models**

i) **Model M:** $\mathrm{E}(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} -0.175 \\ 1.450 \\ -1.400 \\ 2.375 \end{bmatrix} ; p_M(\boldsymbol{y}) = \begin{bmatrix} 4.575 \\ 3.650 \\ -0.125 \\ 3.225 \\ 6.075 \end{bmatrix} ; \boldsymbol{y} - p_M(\boldsymbol{y}) = \begin{bmatrix} 0.325 \\ -0.650 \\ 0.325 \\ -0.325 \\ 0.325 \end{bmatrix}$$

$$\frac{1}{5-4}\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2 = \frac{0.845}{1} = 0.845$$
$$R^2 = \frac{21.868 - 0.633750}{21.868} = 97.1\%$$

**ii) Model** $H_{12}$ : $\mathrm{E}(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3.026 \\ 1.243 \\ -0.987 \end{bmatrix} ; p_{H_{12}}(\boldsymbol{y}) = \begin{bmatrix} 3.026 \\ 4.269 \\ 3.282 \\ 2.295 \\ 4.525 \end{bmatrix} ; \boldsymbol{y} - p_{H_{12}}(\boldsymbol{y}) = \begin{bmatrix} 1.874 \\ -1.269 \\ -3.082 \\ 0.605 \\ -1.875 \end{bmatrix}$$

$$\frac{1}{5-3}\|\boldsymbol{y} - p_{H_{12}}(\boldsymbol{y})\|^2 = \frac{18.512611}{2} = 9.2563$$
$$R^2 = \frac{21.868 - 18.512611}{21.868} = 15.3\%$$

**iii) Model** $H_{13}$**:** $\mathrm{E}(Y) = \alpha + \beta_1 x_1 + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} -0.350 \\ 0.750 \\ 2.200 \end{bmatrix} ; p_{H_{13}}(\boldsymbol{y}) = \begin{bmatrix} 4.05 \\ 2.60 \\ 0.40 \\ 4.80 \\ 5.55 \end{bmatrix} ; \boldsymbol{y} - p_{H_{13}}(\boldsymbol{y}) = \begin{bmatrix} 0.85 \\ 0.40 \\ -1.20 \\ -1.90 \\ 0.85 \end{bmatrix}$$

$$\frac{1}{5-3}\|\boldsymbol{y} - p_{H_{13}}(\boldsymbol{y})\|^2 = \frac{5.2250}{2} = 2.6275$$
$$R^2 = \frac{21.868 - 5.2550}{21.868} = 76.0\%$$

**iv) Model** $H_{23}$**:** $\mathrm{E}(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.945 \\ -0.872 \\ 2.309 \end{bmatrix} ; p_{H_{23}}(\boldsymbol{y}) = \begin{bmatrix} 5.563 \\ 3.254 \\ 0.073 \\ 3.819 \\ 4.691 \end{bmatrix} ; \boldsymbol{y} - p_{H_{23}}(\boldsymbol{y}) = \begin{bmatrix} -0.663 \\ -0.254 \\ 0.127 \\ -0.919 \\ 1.709 \end{bmatrix}$$

$$\frac{1}{5-3}\|\boldsymbol{y} - p_{H_{23}}(\boldsymbol{y})\|^2 = \frac{4.285456}{2} = 2.1427$$
$$R^2 = \frac{21.868 - 4.2855}{21.868} = 80.4\%$$

**v) Model** $H_1$ **:** $\mathrm{E}(Y) = \alpha + \beta_1 x_1$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 2.73 \\ 0.75 \end{bmatrix} ; p_{H_1}(\boldsymbol{y}) = \begin{bmatrix} 2.73 \\ 3.48 \\ 3.48 \\ 3.48 \\ 4.23 \end{bmatrix} ; \boldsymbol{y} - p_{H_1}(\boldsymbol{y}) = \begin{bmatrix} 2.17 \\ -0.48 \\ -3.28 \\ -0.58 \\ 2.17 \end{bmatrix}$$

$$\frac{1}{5-2}\|\boldsymbol{y} - p_{H_1}(\boldsymbol{y})\|^2 = \frac{20.7430}{3} = 6.9143$$
$$R^2 = \frac{21.868 - 20.743}{21.868} = 5.1\%$$

**vi) Model** $H_2$**:** $\mathrm{E}(Y) = \alpha + \beta_2 x_2$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3.914 \\ -0.543 \end{bmatrix}; p_{H_2}(\boldsymbol{y}) = \begin{bmatrix} 3.914 \\ 3.914 \\ 3.371 \\ 2.828 \\ 3.371 \end{bmatrix}; \boldsymbol{y} - p_{H_2}(\boldsymbol{y}) = \begin{bmatrix} 0.986 \\ -0.914 \\ -3.171 \\ 0.072 \\ 3.029 \end{bmatrix}$$

$$\frac{1}{5-2}\|\boldsymbol{y} - p_{H_2}(\boldsymbol{y})\|^2 = \frac{21.042858}{3} = 7.0143$$

$$R^2 = \frac{21.868 - 21.043}{21.868} = 3.8\%$$

**vii) Model** $H_3$**:** $\mathrm{E}(Y) = \alpha + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 2.2 \end{bmatrix}; p_{H_3}(\boldsymbol{y}) = \begin{bmatrix} 4.8 \\ 2.6 \\ 0.4 \\ 4.8 \\ 4.8 \end{bmatrix}; \boldsymbol{y} - p_{H_3}(\boldsymbol{y}) = \begin{bmatrix} 0.1 \\ 0.4 \\ -0.2 \\ -1.9 \\ 1.6 \end{bmatrix}$$

$$\frac{1}{5-2}\|\boldsymbol{y} - p_{H_3}(\boldsymbol{y})\|^2 = \frac{6.38}{3} = 2.1267$$

$$R^2 = \frac{21.868 - 6.38}{21.868} = 70.8\%$$

**viii) Model** $H_0$**:** $\mathrm{E}(Y) = \alpha$

$$\hat{\alpha} = 3.48$$

$$p_{H_0}(\boldsymbol{y}) = \begin{bmatrix} 3.48 \\ 3.48 \\ 3.48 \\ 3.48 \\ 3.48 \end{bmatrix}; \boldsymbol{y} - p_{H_0}(\boldsymbol{y}) = \begin{bmatrix} 1.42 \\ -0.48 \\ -3.28 \\ -0.58 \\ 2.92 \end{bmatrix}$$

$$\frac{1}{5-1}\|\boldsymbol{y} - p_{H_0}(\boldsymbol{y})\|^2 = \frac{21.8680}{4} = 5.4670$$

**D. Successive testings**

1) $H \supseteq H_{12} \supseteq H_1 \supseteq H_0$ i.e. : $\beta_3 = 0, \beta_2 = 0, \beta_1 = 0$

| Variation | SS | d.o.f. |
|---|---|---|
| $H_0 - H_1$ $(\beta_1 = 0)$ | $21.868 - 20.7430 = 1.125$ | 1 |
| $H_1 - H_{12}$ $(\beta_2 = 0)$ | $20.7430 - 18.5126 = 2.230$ | 1 |
| $H - H_{12}$ $(\beta_3 = 0)$ | $18.5126 - 0.6338 = 17.879$ | 1 |
| $M -$ obs | $0.6338 = 0.634$ | 1 |
| $H_0 -$ obs | 21.868 | 4 |

2) $M \supseteq H_{12} \supseteq H_2 \supseteq H_0$ d.v.s. : $\beta_3 = 0, \beta_1 = 0, \beta_2 = 0$

| Variation | SS | d.o.f. |
|---|---|---|
| $H_0 - H_2$ $(\beta_2 = 0)$ | $21.8680 - 21.0429 = 0.825$ | 1 |
| $H_2 - H_{12}$ $(\beta_1 = 0)$ | $21.0429 - 18.5126 = 2.530$ | 1 |
| $H_{12} - M$ $(\beta_3 = 0)$ | $18.5126 - 0.6338 = 17.879$ | 1 |
| $M -$ obs | $0.6338 = 0.634$ | 1 |
| $H_0 -$ obs | 21.868 | 4 |

3) $M \supset H_{13} \supset H_1 \supset H_0$ d.v.s. : $\beta_2 = 0, \beta_3 = 0, \beta_1 = 0$

| Variation | SS | d.o.f. |
|---|---|---|
| $H_0 - H_1$ $(\beta_1 = 0)$ | $21.8680 - 20.7430 = 1.125$ | 1 |
| $H_1 - H_{13}$ $(\beta_3 = 0)$ | $20.7430 - 5.2550 = 15.488$ | 1 |
| $H_{13} - M$ $(\beta_2 = 0)$ | $5.2550 - 0.6338 = 4.621$ | 1 |
| $M -$ obs | $0.6338 = 0.634$ | 1 |
| $H_0 -$ obs | 21.868 | 4 |

4) $M \supseteq H_{13} \supseteq H_3 \supseteq H_0$ d.v.s. : $\beta_2 = 0, \beta_1 = 0, \beta_3 = 0$

| Variation | SS | d.o.f. |
|---|---|---|
| $H_0 - H_3$ $(\beta_3 = 0)$ | $21.8680 - 6.38 = 15.488$ | 1 |
| $H_3 - H_{13}$ $(\beta_1 = 0)$ | $6.38 - 5.2550 = 1.125$ | 1 |
| $H_{13} - M$ $(\beta_2 = 0)$ | $5.2550 - 0.6338 = 4.621$ | 1 |
| $M -$ obs | $0.6338 = 0.634$ | 1 |
| $H_0 -$ obs | 21.868 | 4 |

5) $M \supseteq H_{23} \supseteq H_2 \supseteq H_0$ d.v.s. : $\beta_1 = 0, \beta_3 = 0, \beta_2 = 0$

| Variation | SS | d.o.f. |
|---|---|---|
| $H_0 - H_2$ $(\beta_2 = 0)$ | $21.8680 - 21.0429 = 0.825$ | 1 |
| $H_2 - H_{2\,3}$ $(\beta_3 = 0)$ | $21.0429 - 4.2855 = 16.757$ | 1 |
| $H_{2\,3} - M$ $(\beta_1 = 0)$ | $4.2855 - 0.6338 = 3.652$ | 1 |
| $M -$ obs | $0.6338 = 0.634$ | 1 |
| $H_0 -$ obs | 21.868 | 4 |

6) $M \supset H_{23} \supset H_3 \supset H_0$ d.v.s. : $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0$

| Variation | SS | d.o.f. |
|---|---|---|
| $H_0 - H_3 \quad (\beta_3 = 0)$ | $21.8680 - 6.38 \; = 15.488$ | 1 |
| $H_3 - H_{23} \quad (\beta_2 = 0)$ | $6.38 - 4.2855 \; = \; 2.095$ | 1 |
| $H_{23} - M \quad (\beta_1 = 0)$ | $4.2855 - 0.6338 = \; 3.652$ | 1 |
| $M -$ obs | $0.6338 \; = \; 0.634$ | 1 |
| $H_0 -$ obs | 21.868 | 4 |

**E. Variance-covariance- and correlation- matrix for data.**

$$\text{Variance-covariance matrix} = \frac{1}{5-1}\begin{array}{cc}\left(\begin{array}{cccc}2&1&0&1.50\\1&2.8&0.4&-1.52\\0&0.4&3.2&7.04\\1.50&-1.52&7.04&21.868\end{array}\right) & \begin{array}{c}x_1\\x_2\\x_3\\y\end{array}\\ \begin{array}{cccc}x_1&\quad x_2&\quad x_3&\quad y\end{array} & \end{array}$$

$$\text{correlation matrix} = \begin{array}{cc}\left(\begin{array}{cccc}1&0.4225&0&0.2268\\0.4225&1&0.13393&-0.1942\\0&0.1336&1&0.8416\\0.2268&-0.1942&0.8416&1\end{array}\right) & \begin{array}{c}x_1\\x_2\\x_3\\y\end{array}\\ \begin{array}{cccc}x_1&\quad x_2&\quad x_3&\quad y\end{array} & \end{array}$$

**F. Partial correlations for given $x_3$:**

$$\begin{pmatrix}1&0.4225&0.2268\\0.4225&1&-0.1942\\0.2268&-0.1942&1\end{pmatrix} - \begin{pmatrix}0\\0.1336\\0.8416\end{pmatrix}[1]^{-1}[\;0\quad 0.1336\quad 0.8416\;]$$

$$= \begin{pmatrix}1&0.4225&0.2268\\0.4225&0.9822&-0.3066\\0.2268&-0.3066&0.2917\end{pmatrix},$$

i.e. the correlation matrix is

$$\begin{array}{cc}\left(\begin{array}{ccc}1&0.4263&0.4199\\0.4263&1&-0.5728\\0.4199&-0.5728&1\end{array}\right) & \begin{array}{c}x_1\\x_2\\y\end{array}\\ \begin{array}{ccc}x_1&\quad x_2&\quad y\end{array} & \end{array}$$

First calculated using the above mentioned partial correlation matrix

$$\begin{pmatrix}1&0.4199\\0.4199&1\end{pmatrix} - \begin{pmatrix}0.4263\\0.5728\end{pmatrix}[1]^{-1}[0.4263 - 0.5728] =$$

$$\begin{pmatrix}0.8183&0.6641\\0.6641&0.6718\end{pmatrix},$$

which results in the following correlation matrix

$$\begin{array}{cc} \begin{pmatrix} 1 & 0.8956 \\ 0.8956 & 1 \end{pmatrix} & \begin{matrix} x_1 \\ y \end{matrix} \end{array}$$

As a check we could compute it from the original covariance matrix

$$\begin{aligned} &\begin{pmatrix} 2 & 1.50 \\ 1.50 & 21.868 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ -1.52 & 7.04 \end{pmatrix} \begin{pmatrix} 2.8 & 0.4 \\ 0.4 & 3.2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -1.52 \\ 0 & 7.04 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1.50 \\ 1.50 & 21.868 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ -1.52 & 7.04 \end{pmatrix} \begin{pmatrix} 0.3636 & -0.0455 \\ -0.0455 & 0.3182 \end{pmatrix} \begin{pmatrix} 1 & -1.52 \\ 0 & 7.04 \end{pmatrix} \\ &= \begin{pmatrix} 1.6363 & 2.3727 \\ 2.3727 & 4.2855 \end{pmatrix}, \end{aligned}$$

and the partial correlation matrix is then

$$\begin{array}{cc} \begin{pmatrix} 1 & 0.8960 \\ 0.8960 & 1 \end{pmatrix} & \begin{matrix} x_1 \\ y \end{matrix} \end{array}$$

The deviations in the elements off the diagonal are a result of truncation errors.

## 4.4 Other regression models and solutions

This section is omitted.

# Chapter 5

# CHAPTER OMITTED

# Chapter 6

# Tests in the multidimensional normal distribution

In this chapter we will give a number of generalisations to some of the well known test statistics based on one dimensional normally distributed stochastic variables. In most cases the test statistics will be analogues to the well known ones, except for multiplication being substituted with matrix multiplication, numerical values by the determinant of the matrix etc.

## 6.1 Test for mean value.

### 6.1.1 Hotelling's $T^2$ in the One-Sample Situation

In this section we will consider independent stochastic variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n,$ where

$$\boldsymbol{X}_i \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

i.e. p-dimensionally normally distributed with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. We assume that $\boldsymbol{\Sigma}$ is regular and unknown. We want to test a hypothesis about the mean vector $\boldsymbol{\mu}$ being equal to a given vector $\boldsymbol{\mu}_0$ against all alternatives i.e.

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{against} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

We first repeat some results on the estimators. From theorem 2.27 p. 103 we have the following results on the empirical mean vector $\bar{\boldsymbol{X}}$ and the empirical variance-covariance matrix $\mathbf{S}$

$$\begin{array}{ll} \bar{\boldsymbol{X}} = \frac{1}{n}\sum\limits_{i=1}^{n}\boldsymbol{X}_i & \in \mathrm{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}) \\ \mathbf{S} = \frac{1}{n-1}\sum\limits_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' & \in \mathrm{W}(n-1, \frac{1}{n-1}\boldsymbol{\Sigma}) \\ \bar{\boldsymbol{X}} \text{ and } \mathbf{S} \text{ are stochastic independent.} & \end{array}$$

In the following we will furthermore need the following results on the distribution of certain functions of normally distributed and Wishart distributed stochastic variables.

**LEMMA 6.1.** Let $\boldsymbol{Y}$ be a $p$-dimensional stochastic variable and let $U$ be a $p \times p$ stochastic matrix with

$$\begin{array}{rcl} \boldsymbol{Y} & \in & \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ m\mathbf{U} & \in & \mathrm{W}(m, \boldsymbol{\Sigma}), \end{array}$$

furthermore let $\boldsymbol{Y}$ and $\mathbf{U}$ be stochastically independent. We now let

$$T^2 = \boldsymbol{Y}'\mathbf{U}^{-1}\boldsymbol{Y}.$$

Then the following holds

$$\frac{m-p+1}{mp}T^2 \in \mathrm{F}(p, m-p+1; \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}),$$

i.e. the left hand side is non-centrally F-distributed with non-centrality parameter $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and degrees of freedom equal to $(p, m-p+1)$. If $\boldsymbol{\mu} = \mathbf{0}$, then the non-centrality parameter is 0 i.e. we then have the special case

$$\frac{m-p+1}{mp}T^2 \in \mathrm{F}(p, m-p+1).$$

**PROOF 6.1.** Omitted. See e.g. [2], p. 106. ■

We now have the following main result

**THEOREM 6.1.** We will use the notation

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

where $\bar{\boldsymbol{X}}$, $\boldsymbol{\mu}_0$ and $\mathbf{S}$ are as stated in the introduction to this section. Then the critical area for a ratio test of $H_0$ against $H_1$ at level $\alpha$ is

$$C = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n | \frac{n-p}{(n-1)p} t^2 > \mathrm{F}(p, n-p)_{1-\alpha}\},$$

where $t^2$ is the observed value of $T^2$. ▲

**PROOF 6.2.** From Lemma 6.1 we find that

$$\frac{n-p}{(n-1)p} T^2 \in \mathrm{F}(p, n-p)$$

under $H_0$. From this follows that $C$ is the critical region for a test of $H_0$ versus $H_1$ at level $\alpha$. That this corresponds to a ratio test follows from direct computation by using theorem 1.2 among other things. ■

**REMARK 6.1.** The quantity $T^2$ is often called Hotelling's $T^2$ after Harold Hotelling, who first considered this test statistic. ▼

**REMARK 6.2.** In the one dimensional case we use the test statistic

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}.$$

We now have that $Z^2$ can be written

$$Z^2 = n(\bar{X} - \mu_0)[S^2]^{-1}(\bar{X} - \mu_0),$$

i.e. precisely the same as $T^2$ reduces to in the one-dimensional case. Furthermore note that the square of a student distributed variable $\mathrm{t}(\nu)$ is $\mathrm{F}(1, \nu)$ distributed which means that there (of course) also is a relation between the distribution of the two test statistics. ▼

In order to compute the test statistic it is useful to remember the follow theorem where it is seen that inversion of a matrix can be substituted by the calculation of some determinants.

**THEOREM 6.2.** Let the notation be as above then the following holds true

$$T^2 = \frac{\det[\mathbf{S} + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)']}{\det[\mathbf{S}]} - 1$$

▲

**PROOF 6.3.** Omitted. Purely technical and follows by using theorem 1.2 p. 17 on the matrix

$$\begin{bmatrix} -1 & \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \\ \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) & \mathbf{S} \end{bmatrix}$$

■

We now give an illustrative

**EXAMPLE 6.1.** In the following table values for silicium and aluminium (in %) in 7 samples collected on the moon are given

| | Sample | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Silicium | 19.4 | 21.5 | 19.2 | 18.4 | 20.6 | 19.8 | 18.7 |
| Aluminium | 5.9 | 4.0 | 4.0 | 5.4 | 6.2 | 5.7 | 6.0 |

We are now very interested in testing if these samples can be assumed to come from a population with the same mean values as basalt from our own planet earth. These are

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 22.10 \\ 7.40 \end{pmatrix}.$$

It seems sensible to use Hotelling's $T^2$ to help answer the above question. If we call the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_7,$ we find

$$\begin{aligned} \bar{\boldsymbol{x}} &= \begin{pmatrix} 19.66 \\ 5.31 \end{pmatrix}, \\ \mathbf{s} &= \begin{pmatrix} 1.1795 & -0.3076 \\ -0.3076 & 0.8681 \end{pmatrix}. \end{aligned}$$

Since

$$\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0 = \begin{pmatrix} -2.44 \\ -2.09 \end{pmatrix},$$

then

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' = \begin{pmatrix} 41.68 & 35.70 \\ 35.70 & 30.58 \end{pmatrix},$$

and

$$\mathbf{s} + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' = \begin{pmatrix} 42.86 & 35.39 \\ 35.39 & 31.45 \end{pmatrix}.$$

Then

$$t^2 = \frac{95.49}{0.9293} - 1 = 101.75.$$

The F-test statistic is

$$\frac{7-2}{6 \cdot 2} t^2 = 42.8 > \mathrm{F}(2,5)_{0.999} = 37.1,$$

and the hypothesis is therefore rejected at least at all levels $\alpha$ larger than 0.1%. It therefore does not seem reasonable to assume that the 7 moon samples originate from a population with the same mean value of silicium and aluminium as basalt from our planet earth. ♦

From the result of theorem 6.1 we can easily construct a confidence region for $\boldsymbol{\mu}$. We have with the usual notation

**THEOREM 6.3.** A $(1-\alpha)$ -confidence region for the expectation $\mathrm{E}(\boldsymbol{X})$ is

$$\{\boldsymbol{\mu} | n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \mathbf{s}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} \mathrm{F}(p, n-p)_{1-\alpha}\},$$

i.e. an ellipsoid with centre in $\bar{\boldsymbol{x}}$ and main axes determined by the eigenvectors in the inverse empirical variance-covariance matrix. ▲

**PROOF 6.4.** Trivial from the definition of a confidence area and theorem 6.1. ■

We now continue example 6.1 in the following

**EXAMPLE 6.2.** We will now determine a 95% confidence area for the mean vector. According to theorem 6.3 the confidence area is ordered by the ellipse

$$7(19.66-\mu_1, 5.31-\mu_2)\mathbf{s}^{-1}\begin{pmatrix} 19.66-\mu_1 \\ 5.31-\mu_2 \end{pmatrix} = \frac{12}{5}\mathrm{F}(2,5)_{0.95}$$

or

$$(19.66-\mu_1, 5.31-\mu_2)\mathbf{s}^{-1}\begin{pmatrix} 19.66-\mu_1 \\ 5.31-\mu_2 \end{pmatrix} = 1.9851.$$

We find

$$\mathbf{s}^{-1} = \begin{pmatrix} 0.9341 & 0.3310 \\ 0.3310 & 1.2692 \end{pmatrix}$$

with the eigenvalues 1.4727 and 0.7307 and the corresponding (normed) eigenvectors

$$\begin{pmatrix} 0.5236 \\ 0.8520 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -0.8520 \\ 0.5236 \end{pmatrix}.$$

In the coordinate system with origin in $\bar{\boldsymbol{x}}$ and the above mentioned vectors as unity vectors the ellipse has the equation

$$1.4727y_1^2 + 0.7307y_2^2 = 1.9851$$

or

$$\frac{y_1^2}{1.1610^2} + \frac{y_2^2}{1.6482^2} = 1$$

In figure 6.1 the confidence region and the observations are shown. Furthermore $\boldsymbol{\mu}_0 = (22.10,\ 7.40)'$ is given. It is seen that this observation lies outside the confidence region corresponding to the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ being rejected at all levels greater than 0.01% and therefore especially for $\alpha = 5\%$. ♦

### 6.1.2 Hotelling's $T^2$ in the two-sample situation.

Quite analogous to the t-test in the one dimensional case Hotelling's $T^2$ can be used to investigate if samples from two normal distributions (with the same variance-covariance structure) can be assumed to have the same expected values. We consider independent

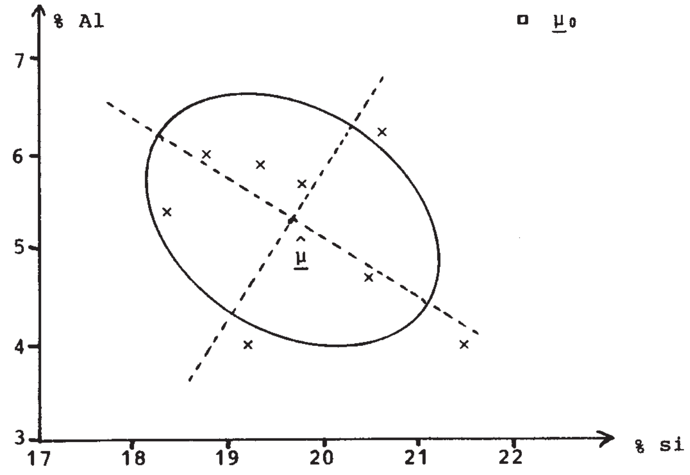


Figure 6.1: Observations and confidence region.

stochastic variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$, where

$$\begin{aligned} \boldsymbol{X}_i &\in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{Y}_i &\in \mathrm{N}_p(\boldsymbol{\nu}, \boldsymbol{\Sigma}), \end{aligned}$$

and we wish to test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\nu} \quad \text{against} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\nu}.$$

We use the notation

$$\begin{aligned} \bar{\boldsymbol{X}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \\ \bar{\boldsymbol{Y}} &= \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{Y}_i \\ \mathbf{S}_1 &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' \\ \mathbf{S}_2 &= \frac{1}{m-1}\sum_{i=1}^{m}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})' \\ \mathbf{S} &= \frac{(n-1)\mathbf{S}_1 + (m-1)\mathbf{S}_2}{n+m-2} \end{aligned}$$

From theorem 2.27 and theorem 2.26 we have

$$\begin{aligned}
\bar{\boldsymbol{X}} &\in \mathrm{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}) \\
\bar{\boldsymbol{Y}} &\in \mathrm{N}_p(\boldsymbol{\nu}, \frac{1}{m}\boldsymbol{\Sigma}) \\
\mathbf{S} &\in \mathrm{W}(n+m-2, \frac{1}{n+m-2}\boldsymbol{\Sigma}).
\end{aligned}$$

We now give the main result on testing $H_0$ against $H_1$ in

**THEOREM 6.4.** We use the same notation as given above. Now, let

$$T^2 = \frac{nm}{n+m}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})'\mathbf{S}^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}).$$

Then the critical region for a test of $H_0$ against $H_1$ at level $\alpha$ is equal to

$$C = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \frac{n+m-p-1}{(n+m-2)p} t^2 > \mathrm{F}(p, n+m-p-1)_{1-\alpha}\}$$

Here $t^2$ is the observed value of $T^2$. ▲

**PROOF 6.5.** From lemma 6.1 and from the above mentioned relationships we find that

$$\frac{n+m-p-1}{(n+m-2)p}T^2 \in \mathrm{F}(p, n+m-p-1;\ (\boldsymbol{\mu}-\boldsymbol{\nu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\nu})),$$

and the result follows readily. ■

Analogous to the one-sample situation we can use the results to determine a confidence region for the difference between mean vectors. We have

**THEOREM 6.5.** We still consider the above mentioned situation and let $\boldsymbol{\mu} - \boldsymbol{\nu} = \boldsymbol{\delta}_o$. Then a $(1-\alpha)$ confidence region for $\boldsymbol{\delta}_o$ is equal to

$$\begin{aligned}
&\{\boldsymbol{\delta} | \frac{nm}{n+m}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}} - \boldsymbol{\delta})'\mathrm{s}^{-1}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}} - \boldsymbol{\delta}) \leq \\
&\quad \frac{(n+m-2)p}{n+m-p-1}\mathrm{F}(p, n+m-p-1)_{1-\alpha}\}.
\end{aligned}$$

▲

**PROOF 6.6.** Follows directly from the definition of a confidence region and from theorem 6.4. ■

**REMARK 6.3.** The confidence region is an ellipsoid with centre in $\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}$ and main axes determined by the eigenvectors in $\mathbf{s}^{-1}$. ▼

**REMARK 6.4.** As mentioned the test results and confidence intervals require that the variance-covariance matrices for the $\boldsymbol{X}$- and for the $\boldsymbol{Y}$ -observations are equal. If this is not the case the above mentioned results are not exact and a different procedure should be used. We will not consider this here but refer to e.g.[2], p. 118. ▼

We will now consider an example on the use of $T^2$ in the two-sample situation.

**EXAMPLE 6.3.** At the Laboratory of Heating- and Climate-technique, DTU, one has measured the following in an experiment

i) the height in cm.

ii) evaporation loss in g/m$^2$ skin during a 3 hour periode

iii) mean temperature in °C. This temperature is found by measuring the skin temperature at 14 different locations every minute for 5 minutes (same locations every time). The mean temperature is then an average of all $14 \times 5 = 70$ measurements,

on 16 men and 16 women. The result of the experiment is given in the table p. 202.

We consider these numbers as realisations of stochastic variables

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{16} \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{16}.$$

We furthermore assume, that the variables are stochastic independent and that

$$\boldsymbol{X}_i \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{Y}_i \in \mathrm{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma}),$$

i.e. the variance-covariance matrices are assumed equal. Later we will discuss whether this hypothesis is reasonable or not.

| Person No. | Height | Evaporation loss | Mean temperature |
|---|---|---|---|
| | in cm | in g/m$^2$skin | in °C |
| 1 | 177 | 18.1 | 33.9 |
| 2 | 189 | 18.8 | 33.2 |
| 3 | 181 | 20.4 | 33.9 |
| 4 | 184 | 19.5 | 33.8 |
| 5 | 183 | 30.5 | 33.3 |
| 6 | 178 | 22.2 | 33.6 |
| 7 | 162 | 19.4 | 39.2 |
| 8 | 176 | 26.7 | 33.2 |
| 9 | 190 | 16.6 | 33.2 |
| 10 | 180 | 45.4 | 33.5 |
| 11 | 179 | 24.0 | 33.9 |
| 12 | 175 | 34.6 | 33.8 |
| 13 | 183 | 21.3 | 33.5 |
| 14 | 177 | 33.3 | 33.9 |
| 15 | 185 | 22.9 | 33.8 |
| 16 | 176 | 18.6 | 33.5 |
| 1 | 160 | 14.6 | 32.9 |
| 2 | 171 | 27.0 | 33.5 |
| 3 | 168 | 27.6 | 32.3 |
| 4 | 171 | 20.2 | 33.1 |
| 5 | 169 | 30.8 | 33.4 |
| 6 | 169 | 17.4 | 33.5 |
| 7 | 167 | 21.1 | 33.0 |
| 8 | 170 | 19.3 | 34.1 |
| 9 | 162 | 21.5 | 33.8 |
| 10 | 160 | 15.2 | 33.0 |
| 11 | 168 | 15.4 | 33.7 |
| 12 | 157 | 25.2 | 33.9 |
| 13 | 161 | 13.9 | 34.8 |
| 14 | 164 | 20.2 | 31.9 |
| 15 | 161 | 25.3 | 39.0 |
| 16 | 180 | 12.6 | 33.5 |

Table 6.1: Data from indoor-climate experiments, laboratory for Heating- and Climate-technique, DTU.

The estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are the empirical mean vectors i.e.

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \begin{pmatrix} 179.7 \\ 24.5 \\ 33.6 \end{pmatrix}$$

and

$$\hat{\boldsymbol{\nu}} = \bar{\boldsymbol{y}} = \begin{pmatrix} 166.1 \\ 20.5 \\ 33.4 \end{pmatrix}.$$

We will now check if the difference between $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\nu}}$ is significant, i.e. whether $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ can be assumed equal.

With the notation chosen in theorem 6.4 we find

$$\mathbf{s} = \begin{pmatrix} 38.5 & -4.3 & -0.8 \\ -4.3 & 45.5 & -0.3 \\ -0.8 & -0.3 & 0.3 \end{pmatrix},$$

and

$$\mathrm{t}^2 = \frac{16 \cdot 16}{16+16}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})'\mathbf{s}^{-1}(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}) = 52.4.$$

The test statistic then becomes

$$\frac{16+16-3-1}{(16+16-2)3} 52.4 = 16.3.$$

Since

$$\mathrm{F}(3,28)_{0.999} = 7.19$$

a hypothesis that $\boldsymbol{\mu} = \boldsymbol{\nu}$ will at least be rejected at all levels greater than 0.1%. We will therefore conclude that there is a fairly large (simultaneous) difference in the three variables for men and for women, a result which probably will not chock anyone when it is remembered that the first variable gives the height.

If we instead only consider the second and third coordinates, i.e. the values for evaporation loss and mean temperature we get the test statistic

$$\frac{16 \cdot 16}{16+16} \frac{16+16-2-1}{(16+16-2)2}(4.0, 0.2) \begin{pmatrix} 45.5 & -0.3 \\ -0.3 & 0.3 \end{pmatrix}^{-1} \begin{pmatrix} 4.0 \\ 0.2 \end{pmatrix} \simeq 0.2.$$

This quantity is to be compared with the quantiles in an F(2,29) -distribution and it is readily seen that a hypothesis that the mean vectors are equal can be accepted at all reasonable levels. ♦

## 6.2 The multidimensional general linear model.

In the previous section we have looked at the one- and two-sample situation for the multidimensional normal distribution. We have seen that the multidimensional results are quite analogous to the one dimensional ones. In this section and in the following we will continue this analogy and derive the results regarding regression and analysis of variance of multidimensional variables.

We consider independently distributed variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$,

$$\boldsymbol{Y}_i \in \mathrm{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}).$$

The variance-covariance matrix $\boldsymbol{\Sigma}$ (and the mean vectors $\boldsymbol{\mu}_i$) are assumed unknown. We arrange the observations in an $n \times p$ data matrix

$$\mathbf{Y} = \begin{bmatrix} \boldsymbol{Y}_1' \\ \vdots \\ \boldsymbol{Y}_n' \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}.$$

Here the single rows represent e.g. repetitions of measurements of a p-dimensional phenomena. In full analogy with the model which we considered in the univariate general linear model we will assume that the mean parameter $\boldsymbol{\mu}_i$ can be written as known linear functions of other (and fewer) unknown parameters $\theta$, i.e.

$$\mathrm{E}(\mathbf{Y}) = \mathbf{x}\,\theta = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{bmatrix}.$$

It is seen that we assume $\mathbf{x}$ known and $\theta$ unknown. This model can be viewed from different angles. If we let the $j$'th column in the $\mathbf{Y}$ matrix equal

$$\mathbf{Y}_{j|} = \begin{bmatrix} Y_{1j} \\ \vdots \\ Y_{nj} \end{bmatrix},$$

then we can write

$$\mathrm{E}(\mathbf{Y}_{j|}) = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_{1j} \\ \vdots \\ \theta_{kj} \end{bmatrix} = \mathbf{x}\,\theta_{j|}.$$

The $n$ measurements on the $j$'th "property" (attribute/variable) will therefore follow an ordinary one dimensional general linear model.

If we instead write the mean value of a single observation $\boldsymbol{Y}_i,$ we find

$$\mathrm{E}(\boldsymbol{Y}_i') = (x_{i1} \cdots x_{ik}) \begin{pmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{pmatrix} = \boldsymbol{x}_i'\theta,$$

where $\boldsymbol{x}_i' = \mathbf{x}_{-i}$ is the $i$'th row in the $\mathbf{x}$ -matrix. This readily gives

$$\mathrm{E}(\boldsymbol{Y}_i) = \theta'\mathbf{x}_i,$$

which is an analogue to the one dimensional regression model.

If the observations are rearranged into a column vector

$$\underset{\sim}{Y} = \mathrm{vc}(\mathbf{Y}) = \begin{bmatrix} \mathbf{Y}_{1|} \\ \vdots \\ \mathbf{Y}_{p|} \end{bmatrix},$$

we find from theorem 2.7, p. 63, that

$$\mathrm{D}(\underset{\sim}{Y}) = \mathbf{\Sigma} \otimes \mathbf{I}_n,$$

where $\mathbf{\Sigma} \otimes \mathbf{I}_n$ is the tensor product of $\mathbf{\Sigma}$ and $\mathbf{I}_n,$ cf. section 1.5.

The first problem is to estimate $\theta$. We have

**THEOREM 6.6.** We consider the above mentioned situation. If the observations $\boldsymbol{Y}_i$ are normally distributed the maximum likelihood estimate of $\theta$ is given by

$$\hat{\theta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}.$$

▲

**PROOF 6.7.** Omitted. See e.g. [2]. ■

**REMARK 6.5.** We see that

$$\hat{\theta}_{j|} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}_{j|},$$

i.e. the estimate for the $j$'th column in $\theta$ is simply equal to the result we get by only considering the one dimensional general linear model for the $j$'th "property". ▼

**REMARK 6.6.** If the observations are not normally distributed one will still be able to use the estimate $\hat{\theta}$, since this of course just like the one dimensional case has a Gauss-Markov property. We will not go into details with this but just mention a couple of results. The least squares properties are that

$$M = (\mathbf{Y} - \mathbf{x}\,\theta)'(\mathbf{Y} - \mathbf{x}\,\theta) - (\mathbf{Y} - \mathbf{x}\,\hat{\theta})'(\mathbf{Y} - \mathbf{x}\,\hat{\theta})$$

is positive semidefinite. From this follows that

$$\text{ch}_i(\mathbf{Y} - \mathbf{x}\,\theta)'(\mathbf{Y} - \mathbf{x}\,\theta) \geq \text{ch}_i(\mathbf{Y} - \mathbf{x}\,\hat{\theta})'(\mathbf{Y} - \mathbf{x}\,\hat{\theta}),$$

where $\text{ch}_i$ corresponds to the $i$'th largest eigenvalue. From this follows again that $\hat{\theta}$ minimises

$$\det(\mathbf{Y} - \mathbf{x}\,\theta)'(\mathbf{Y} - \mathbf{x}\,\theta)$$

and

$$\operatorname{tr}(\mathbf{Y} - \mathbf{x}\,\theta)'(\mathbf{Y} - \mathbf{x}\,\theta).$$

▼

**REMARK 6.7.** Above we have silently assumed that $\mathbf{x}'\mathbf{x}$ has full rank i.e. $\text{rg}(\mathbf{x}) = k < n$. If this is not the case one can by analogy to the one dimensional (univariate) results find solutions by means of pseudo inverse matrices. ▼

After these considerations on the estimation of $\hat{\theta}$ we turn to the estimation of $\mathbf{\Sigma}$.

**THEOREM 6.7.** We consider the situation from theorem 6.6. Then the maximum likelihood estimate for $\mathbf{\Sigma}$ equals

$$\begin{aligned}
\hat{\mathbf{\Sigma}}^* &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \hat{\theta}'\boldsymbol{x}_i)(\boldsymbol{Y}_i - \hat{\theta}'\boldsymbol{x}_i)' \\
&= \frac{1}{n}(\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta}) \\
&= \frac{1}{n}[\mathbf{Y}'\mathbf{Y} - (\mathbf{x}\hat{\theta})'(\mathbf{x}\hat{\theta})].
\end{aligned}$$

The $(i,j)$'th element can also be written

$$\hat{\sigma}^*_{ij} = \frac{1}{n}(\mathbf{Y}_{i|} - \mathbf{x}\hat{\theta}_{i|})'(\mathbf{Y}_{j|} - \mathbf{x}\hat{\theta}_{j|}).$$

▲

**PROOF 6.8.** The many identities between $\hat{\mathbf{\Sigma}}$ 's elements are found by simple matrix manipulations. For the results we refer to [2]. ■

The distribution of the estimators mentioned are given in

**THEOREM 6.8.** We consider the situation from theorems 6.6 and 6.7 and we introduce the usual notations

$$\begin{aligned}
\tilde{\theta} &= \mathrm{vc}(\theta) = \begin{bmatrix} \theta_{|1} \\ \vdots \\ \theta_{|p} \end{bmatrix} \\
\tilde{\hat{\theta}} &= \mathrm{vc}(\hat{\theta}) = \begin{bmatrix} \hat{\theta}_{|1} \\ \vdots \\ \hat{\theta}_{|p} \end{bmatrix}.
\end{aligned}$$

Then we have that $\tilde{\hat{\theta}}$ is normally distributed

$$\tilde{\hat{\theta}} = \mathrm{vc}(\hat{\theta}) \in \mathrm{N}_{pk}(\tilde{\theta}, \mathbf{\Sigma} \otimes (\mathbf{x}'\mathbf{x})^{-1}),$$

and $n\hat{\mathbf{\Sigma}}^*$ is Wishart distributed

$$n\hat{\mathbf{\Sigma}}^* \in \mathrm{W}(n-k, \mathbf{\Sigma}).$$

Finally $\hat{\mathbf{\Sigma}}^*$ and $\tilde{\hat{\theta}}$ and therefore also $\hat{\mathbf{\Sigma}}^*$ and $\hat{\theta}$ are stochastically independent. ▲

**PROOF 6.9.** It is trivial that

$$\mathrm{E}(\hat{\theta}) = \mathrm{E}[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}] = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{x}\,\theta = \theta$$

and from this it follows that $\mathrm{E}(\hat{\tilde{\theta}}) = \tilde{\theta}$. Furthermore $\hat{\tilde{\theta}}$ is of course normally distributed.

Finally we have that

$$\mathrm{D}(\hat{\theta}_{|i}) = \sigma_{ii}(\mathbf{x}'\mathbf{x})^{-1}$$

and

$$C(\hat{\theta}_{|i}, \hat{\theta}_{|j}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'C(\mathbf{Y}_{|i}, \mathbf{Y}_{|j})\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} = \sigma_{ij}(\mathbf{x}'\mathbf{x})^{-1}.$$

From this the result concerning the variance covariance matrix for $\hat{\tilde{\theta}}$ is readily seen.

The result concerning the distribution of $\hat{\mathbf{\Sigma}}^*$ and concerning the independence of $\hat{\theta}$ and $\hat{\mathbf{\Sigma}}^*$ are quite analogous to the corresponding one dimensional results but we will not look further into these matters here. The reader is referred to e.g. [2]. ■

From the theorem we readily find

**COROLLORY 6.1.** The unbiased estimate for $\mathbf{\Sigma}$ is equal to

$$\hat{\mathbf{\Sigma}} = \frac{n}{n-k}\hat{\mathbf{\Sigma}}^* = \frac{1}{n-k}(\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta}).$$

**PROOF 6.10.** Trivial when you remember that

$$\mathrm{E}(\mathrm{W}(k, \mathbf{\Delta})) = k\mathbf{\Delta}.$$

■

Q.E.D.

We now turn to testing the parameters in the model.

We have

**THEOREM 6.9.** We consider the above mentioned situation including the assumption of the normality of the observations. Furthermore we consider the hypothesis

$$H_0 : \mathbf{A}\,\theta\,\mathbf{B}' = \mathbf{C} \quad \text{against} \quad H_1 : \mathbf{A}\,\theta\,\mathbf{B}' \neq \mathbf{C},$$

where $\mathbf{A}(r \times k)$, $\mathbf{B}(s \times p)$ and $\mathbf{C}(r \times s)$ are given (known) matrices. We introduce

$$\begin{aligned} \mathbf{\Delta} &= \mathbf{A}\,\hat{\theta}\,\mathbf{B}' - \mathbf{C} \\ \mathbf{R} &= n\hat{\mathbf{\Sigma}}^* = (\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta}) \end{aligned}$$

and

$$\begin{aligned} \mathbf{S} &= \mathbf{B}\,\mathbf{R}\,\mathbf{B}' \\ \mathbf{H} &= \mathbf{\Delta}'[\mathbf{A}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{A}']^{-1}\mathbf{\Delta}. \end{aligned}$$

Since the ratio test for testing $H_0$ against $H_1$ is equivalent to the test given by the critical region

$$\{\mathbf{y}|\frac{\det(\mathbf{s})}{\det(\mathbf{s}+\mathbf{h})} \leq \mathrm{U}(s,r,n-k)_\alpha\},$$

where $\mathrm{U}(s,r,n-k)_\alpha$ is the $\alpha$ quantile in the null-hypothesis distribution of the test statistic (see below). ▲

**PROOF 6.11.** Omitted. The essential part of the proof is that it can be shown that $\mathbf{S}$ and $\mathbf{H}$ are independent Wishart distributed variables if $H_0$ is true. For more detail we refer to the literature. As it is seen indirectly from the formulation of the theorem the null-hypothesis distribution of

$$u = \frac{\det(\mathbf{S})}{\det(\mathbf{S}+\mathbf{H})}$$

only depends on $s$, $r$ and $n-k$. The quantity is termed in the literature as **Wilk's $\mathbf{\Lambda}$** or **Anderson's $U$**. Since the distribution contains three parameters it is somewhat difficult to use in practise and we therefore give an approximation to an F-distribution in the following ■

**THEOREM 6.10.** Let $U$ be U(p,q,r)-distributed and let

$$\begin{aligned} t &= \begin{cases} 1 & p^2+q^2=5 \\ \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} & p^2+q^2 \neq 5 \end{cases} \\ v &= \frac{1}{2}(2r+q-p-1). \end{aligned}$$

Then

$$\mathrm{F} = \frac{1 - U^{\frac{1}{t}}}{U^{\frac{1}{t}}} \cdot \frac{vt + 1 - \frac{1}{2}pq}{pq}$$

is approximately distributed as

$$\mathrm{F}(pq, vt + 1 - \frac{1}{2}pq).$$

If either $p$ or $q$ are equal to 1 or 2, then the approximation is exact. ▲

**PROOF 6.12.** Omitted. ■

We shall now illustrate the introduced concept in the following example.

**EXAMPLE 6.4.** In the period 1968-69 the Royal Veterinary and Agricultural University's Experimental Farm for crop growing, Højbakkegård, conducted an experiment concerning the growth of lucerne. They investigated the offsprings from 176 crossings. In order to establish the “quality” of the single crossings 9 properties were measured on each one. The 9 variables are given in the following table.

As mentioned, the 5 first variables are graded on a numerical scale. This method is chosen since it is very difficult to measure the respective variables directly, and experience shows that it gives satisfactory results.

| Variable No. & name | Unit of measure | Explanation |
|---|---|---|
| 1: Type of growth | Grade $1 - 9$ | $1 =$ growth is lying down,<br>$9 =$ growth is upright |
| 2: Regrowth after winter | ” | $1 =$ worst, $9 =$ best |
| 3: Ability to creep | ” | $1 =$ no runners,<br>$9 =$ most runners |
| 4: Activity | ” | $1 =$ weakest, $9 =$ strongest |
| 5: Time of blooming | ” | $1 =$ latest blooming,<br>$9 =$ earliest blooming |
| 6: Plant height | cm | |
| 7: Seed weight | g per plant | |
| 8: Plant weight | g per plant after drying | |
| 9: Percent seed | % | Calculated per plant by means of (7) and (8) |

The following analyses are based on the average values for the 9 variables based on numbers from between 15 and 20 plants (most of the results are based on 20 plants). In the following table a section of these numbers is shown.

| Obs.No. | Variable No. and name | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| No. of crossing | Type of growth | Re-growth | Ability to creep | Activity | Blooming | Plant-height | Seed weight | Plant weight | Percent seed |
| 1 | 4.11 | 5.00 | 3.05 | 6.17 | 3.67 | 50.00 | 3.47 | 120.10 | 2.75 |
| 2 | 3.08 | 4.75 | 4.17 | 7.50 | 5.17 | 61.50 | 0.82 | 111.33 | 0.75 |
| 3 | 3.12 | 4.00 | 3.35 | 6.53 | 3.99 | 55.29 | 0.86 | 97.47 | 0.81 |
| ⋮ | | | | | | | | | |
| 176 | 4.00 | 4.40 | 4.60 | 7.40 | 2.90 | 50.00 | 0.66 | 153.50 | 0.44 |

The main goal with the experiment was to examine the variation among the 9 variables. More specifically one was e.g. interested in how variable 3 (ability to creep) and variable 4 (activity) varies together with the others. The two variables mentioned are usually of great importance for the development of a plant and it is therefore of importance what the relation is to the other variables.

As a first orientation we will compute the empirical correlation matrix. It is found to be

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.000 | −0.033 | 0.116 | 0.018 | 0.131 | −0.207 | 0.035 | −0.087 | 0.041 |
| 2 | −0.033 | 1.000 | 0.711 | 0.515 | 0.125 | 0.199 | −0.025 | 0.348 | −0.066 |
| 3 | 0.116 | 0.711 | 1.000 | 0.440 | 0.022 | 0.039 | −0.133 | 0.218 | −0.157 |
| 4 | 0.018 | 0.515 | 0.440 | 1.000 | 0.201 | 0.517 | 0.071 | 0.689 | −0.081 |
| 5 | 0.131 | 0.125 | 0.022 | 0.201 | 1.000 | 0.496 | 0.987 | 0.168 | 0.486 |
| 6 | −0.207 | 0.199 | 0.039 | 0.517 | 0.496 | 1.000 | 0.453 | 0.559 | 0.367 |
| 7 | 0.035 | −0.025 | −0.133 | 0.071 | 0.487 | 0.453 | 1.000 | 0.360 | 0.947 |
| 8 | −0.087 | 0.348 | 0.218 | 0.689 | 0.168 | 0.559 | 0.360 | 1.000 | 0.128 |
| 9 | 0.041 | −0.066 | −0.157 | −0.081 | 0.486 | 0.367 | 0.947 | 0.128 | 1.000 |

We note that variable 1 (type of growth) is only vaguely correlated with the other variables. On the other hand e.g. variables 2 and 3 (re-growth and ability to creep) and (of course) 7 and 9 (weight of seed and percentage of seed) are very strongly correlated.

As mentioned we are especially interested in variable 3's and variable 4's variation with the other variables. We note that there are a number of fairly large correlations but it is difficult to get an impression solely based on these. We will therefore try if it is possible to express these two variables as linear functions of the others i.e.

$$\begin{aligned}
\mathrm{E}(Y_1) &= \sum_{i=1}^{k} \theta_{i1} x_i \\
\mathrm{E}(Y_2) &= \sum_{i=1}^{k} \theta_{i2} x_i
\end{aligned}$$

where we now have used the variable notations

$$\begin{aligned}
Y_1 &= \text{Ability to “creep”} \\
Y_2 &= \text{Activity} \\
x_1 &= \text{Type of growth} \\
x_2 &= \text{Re growth after winter} \\
x_3 &= \text{Time of blooming} \\
x_4 &= \text{Height of plant} \\
x_5 &= \text{Weight of seed} \\
x_6 &= \text{Weight of plant} \\
x_7 &= \text{Percentage of seed}
\end{aligned}$$

We are obviously talking about a multidimensional general linear model. If we let $\theta = (\theta_{ij})$, we get

$$\hat{\theta} = \begin{bmatrix} 0.28400 & 0.42731 \\ 0.79508 & 0.22230 \\ -0.02573 & 0.02607 \\ -0.01151 & 0.06290 \\ -0.14467 & -0.16756 \\ 0.00307 & 0.01103 \\ 0.10614 & 0.03463 \end{bmatrix}.$$

If we assume

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \in \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),$$

then the unbiased estimate of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.85897 & 0.07870 \\ 0.07870 & 0.29444 \end{bmatrix}.$$

The matrix $(\mathbf{x}'\mathbf{x})^{-1}$ is found to be

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 1.55920 | −0.16549 | −0.47258 | −0.05010 | 0.41826 | −0.00235 | −0.42289 |
| −0.16549 | 0.85139 | −0.17981 | −0.01327 | 0.63774 | −0.01759 | −0.69467 |
| −0.47258 | −0.17981 | 1.77862 | −0.10728 | −0.29340 | 0.01164 | −0.02184 |
| −0.05010 | −0.01327 | −0.10728 | 0.02253 | 0.12325 | −0.00441 | −0.17012 |
| 0.41826 | 0.63774 | −0.29340 | 0.12325 | 5.25546 | −0.08437 | −7.04885 |
| −0.00235 | −0.01759 | 0.01164 | −0.00441 | −0.08437 | 0.00243 | 0.11182 |
| −0.42289 | −0.69467 | −0.02184 | −0.17012 | −7.04885 | 0.11182 | 10.11541 |

From this we can easily compute the variance and covariance on the single $\theta$ -values. Because we have

$$\mathrm{D}(\hat{\hat{\theta}}) = \mathbf{\Sigma} \otimes (\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} \sigma_{11}(\mathbf{x}'\mathbf{x})^{-1} & \sigma_{12}(\mathbf{x}'\mathbf{x})^{-1} \\ \sigma_{21}(\mathbf{x}'\mathbf{x})^{-1} & \sigma_{22}(\mathbf{x}'\mathbf{x})^{-1} \end{pmatrix},$$

and therefore e.g.

$$\hat{\mathrm{V}}(\hat{\theta}_{42}) = 0.2944 \cdot 0.02253 = 0.0066.$$

These results can be used in the construction of ordinary t-tests for the single coefficients. We will, however, not consider this here. Instead we will give a couple of examples of how to construct simultaneous tests. Let us e.g. consider the hypothesis

$$H_0 : \theta_{41} = \theta_{42} = 0$$

against all alternatives. This hypotheses must be brought into the form given in theorem 6.9. This can be done by choosing

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \\ \mathbf{B} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

and

$$\mathbf{C} = \begin{pmatrix} 0 & 0 \end{pmatrix}.$$

Then we will have

$$\mathbf{A}\,\theta\,\mathbf{B}' = \begin{pmatrix} \theta_{41} & \theta_{42} \end{pmatrix}.$$

By the use of a standard programme (BDX63) we get the F-test statistic

$$\mathrm{F} = 53.66$$

with degrees of freedom

$$(f_1, f_2) = (2, 168).$$

The test statistic is in this case exact F-distributed, since $s = 2$ and $r = 1$. It is seen that the observed F-value is significant at all reasonable levels.

As another example consider the hypothesis

$$\theta_1 = \begin{bmatrix} \theta_{51} & \theta_{52} \\ \theta_{61} & \theta_{62} \\ \theta_{71} & \theta_{72} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

against all alternatives. This hypothesis can be transformed into the form of theorem 6.9 by choosing

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\ \mathbf{B} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

and

$$\mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix};$$

since then we obtain

$$\mathbf{A}\,\theta\,\mathbf{B}' = \theta_1.$$

With the previously mentioned standard programme we find

$$\mathrm{F} = 10.63\,; \quad (f_1, f_2) = (6, 336).$$

Once again we have a clear significance.

As a last example consider the hypothesis

$$\theta_{62} = \theta_{72} = 0$$

against all alternatives. This is brought into the standard form by choosing

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \\ \mathbf{B} &= \begin{pmatrix} 0 & 1 \end{pmatrix}\end{aligned}$$

and

$$\mathbf{C} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The F-test statistic has $(2, 169)$ degrees of freedom and is found to be 27.4. The values shown are therefore significant. ♦

We will now specialise the results from the previous section to generalisations of the univariate one- and two-sided analysis of variance. First

### 6.2.1 One-sided multi-dimensional analysis of variance

We consider observations

$$\begin{matrix} \boldsymbol{Y}_{11}, & \cdots, & \boldsymbol{Y}_{1n_1} \\ \vdots & & \vdots \\ \boldsymbol{Y}_{k1}, & \cdots, & \boldsymbol{Y}_{kn_k} \end{matrix}.$$

These are assumed to be stochastically independent with

$$\boldsymbol{Y}_{ij} \in \mathrm{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, k\,; \quad j = 1, \ldots, n_i,$$

i.e. $p$-dimensional normal distributed with the same variance-covariance matrix. We wish to test hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k$$

against

$$H_1 : \exists i, j(\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j).$$

Analogously to the univariate one-sided analysis of variance we define sums of squares deviation matrices

$$\begin{aligned}
\mathbf{T} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}})(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}})' \\
\mathbf{W} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_i)(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_i)' \\
\mathbf{B} &= \sum_{i=1}^{k} n_i(\bar{\boldsymbol{Y}}_i-\bar{\boldsymbol{Y}})(\bar{\boldsymbol{Y}}_i-\bar{\boldsymbol{Y}})'
\end{aligned}$$

Here we have with $n = \sum_i n_i$

$$\begin{aligned}
\bar{\boldsymbol{Y}}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{Y}_{ij} \\
\bar{\boldsymbol{Y}} &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\boldsymbol{Y}_{ij}.
\end{aligned}$$

After a bit of algebra we see that "total" matrix $\mathbf{T}$ is the sum of the "between groups" matrix $\mathbf{B}$ and the "within groups" matrix $\mathbf{W}$ i.e.

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

i.e. as in the one-dimensional case we have a partitioning of the total variation in the variation between groups and the variation within groups.

It is trivial that we as an unbiased estimate of the variance-covariance matrix $\boldsymbol{\Sigma}$ can use

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-k}\mathbf{W}.$$

If the hypothesis is true then $\mathbf{T}$ will also be proportional with such an estimate. If the hypothesis is not true then $\mathbf{T}$ will be "larger". Therefore the following theorem seems intuitively reasonable.

**THEOREM 6.11.** The ratio test for the test of the hypothesis $H_0$ against $H_1$ is given by the critical region

$$\{\boldsymbol{y}_{11},\ldots,\boldsymbol{y}_{kn_k} \,\Big|\, \frac{\det(\mathbf{w})}{\det(\mathbf{t})} \le U(p,k-1,n-k)_\alpha\}.$$

▲

**PROOF 6.13.** Omitted. Is found by special choices of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ matrices in theorem 6.9. ■

Just as the case for the one-dimensional analysis of variance the results are displayed using an analysis of variance table.

| Source of variation | SS – matrix | Degrees of freedom |
|---|---|---|
| Deviation from hypothesis = variation between groups | $\mathbf{B} = \sum_i n_i(\bar{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}})(\bar{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}})'$ | $k-1$ |
| Error = variation within groups | $\mathbf{W} = \sum_i \sum_j (\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_i)(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_i)'$ | $n-k$ |
| Total | $\mathbf{T} = \sum_i \sum_i (\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}})'$ | $n-1$ |

As it is done in univariate ANOVA it is of course possible to determine expected values of the $\mathbf{B}$ and $\mathbf{T}$ matrices even without $H_0$ being true. We will, however, not pursue this further here.

### 6.2.2 Two-sided multidimensional analysis of variance

In this case we will only look at a two-sided analysis of variance with 1 observation per cell. We will therefore assume that we have observations

$$\begin{matrix} \boldsymbol{Y}_{11}, & \ldots, & \boldsymbol{Y}_{1m} \\ \vdots & & \vdots \\ \boldsymbol{Y}_{k1}, & \ldots, & \boldsymbol{Y}_{km} \end{matrix},$$

which are assumed to be $p$-dimensional normal distributed with the same variance-covariance matrix $\mathbf{\Sigma}$ and with mean values

$$\mathrm{E}(\boldsymbol{Y}_{ij}) = \boldsymbol{\mu}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j,$$

where the parameters $\boldsymbol{\alpha}_i$ $\boldsymbol{\beta}_j$ satisfy

$$\sum_i \boldsymbol{\alpha}_i = \sum_j \boldsymbol{\beta}_j = \mathbf{0}.$$

We now want to test the hypothesis

$$H_0 : \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_k = \mathbf{0}$$

against

$$H_1 : \exists i(\boldsymbol{\alpha}_i \neq \mathbf{0})$$

and

$$K_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_m = \mathbf{0}$$

against

$$K_1 : \exists j(\boldsymbol{\beta}_j \neq \mathbf{0}).$$

Analogous to the sums of squares of the one-dimensional (univariate) analysis of variance we define the matrices

$$\begin{aligned}
\mathbf{T} &= \sum_{i=1}^{k}\sum_{j=1}^{m}(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{..})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{..})' \\
\mathbf{Q}_1 &= \sum_{i=1}^{k}\sum_{j=1}^{m}(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{.j} + \bar{\boldsymbol{Y}}_{..})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{.j} + \bar{\boldsymbol{Y}}_{..})' \\
\mathbf{Q}_2 &= m\sum_{i=1}^{k}(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})' \\
\mathbf{Q}_3 &= k\sum_{j=1}^{m}(\bar{\boldsymbol{Y}}_{.j} - \bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{.j} - \bar{\boldsymbol{Y}}_{..})'.
\end{aligned}$$

Here we have used the usual notation

$$\begin{aligned}
\bar{\boldsymbol{Y}}_{..} &= \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m}\boldsymbol{Y}_{ij} \\
\bar{\boldsymbol{Y}}_{i.} &= \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{Y}_{ij}, \qquad i = 1, \ldots, k \\
\bar{\boldsymbol{Y}}_{.j} &= \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{Y}_{ij}, \qquad j = 1, \ldots, m.
\end{aligned}$$

We see in this case that we also have the usual partitioning of the total variation

$$\mathbf{T} = \mathbf{Q}_1 + \mathbf{Q}_2 + \mathbf{Q}_3,$$

i.e. the total variation $(\mathbf{T})$ is partitioned in the variation between rows $(\mathbf{Q}_2)$, and the variation between columns $(\mathbf{Q}_3)$ and the residual variation (interaction variation) $(\mathbf{Q}_1)$.

We now have

**THEOREM 6.12.** The ratio test at level $\alpha$ for test of $H_0$ against $H_1$ is given by the critical region

$$\{\boldsymbol{y}_{11},\ldots,\boldsymbol{y}_{km} \left| \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1+\mathbf{q}_2)} \right. \leq U(p,k-1,(k-1)(m-1))_\alpha\}.$$

The ratio test at level $\alpha$ for test of $K_0$ against $K_1$ is given by the critical region

$$\{\boldsymbol{y}_{11},\ldots,\boldsymbol{y}_{km} \left| \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1+\mathbf{q}_3)} \right. \leq U(p,m-1,(k-1)(m-1))_\alpha\}.$$

▲

**PROOF 6.14.** Omitted. Follows readily from theorem 6.9. See e.g. [2]. ■

We collect the results in a usual analysis of variance table

| Source of variation | SS-matrix | Degrees of freedom | Test statistic |
|---|---|---|---|
| Differences between columns | $\mathbf{Q}_3 = k\sum_j(\bar{\boldsymbol{Y}}_{.j}-\bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{.j}-\bar{\boldsymbol{Y}}_{..})'$ | $m-1$ | $\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1+\mathbf{Q}_3)}$ |
| Differences between rows | $\mathbf{Q}_2 = m\sum_i(\bar{\boldsymbol{Y}}_{i.}-\bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{i.}-\bar{\boldsymbol{Y}}_{..})'$ | $k-1$ | $\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1+\mathbf{Q}_2)}$ |
| Residual | $\mathbf{Q}_1 = \sum_i\sum_j(\bar{\boldsymbol{Y}}_{ij}-\bar{\boldsymbol{Y}}_{i.}-\bar{\boldsymbol{Y}}_{.j}+\bar{\boldsymbol{Y}}_{..})\times$<br>$(\bar{\boldsymbol{Y}}_{ij}-\bar{\boldsymbol{Y}}_{i.}-\bar{\boldsymbol{Y}}_{.j}+\bar{\boldsymbol{Y}}_{..})'$ | $(k-1)(m-1)$ | |
| Total | $\mathbf{T} = \sum_i\sum_j(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{..})(\boldsymbol{Y}_{ij}-\bar{\boldsymbol{Y}}_{..})'$ | $km-1$ | |

The matrix $\frac{1}{(k-1)(m-1)}\mathbf{Q}_1$ can be used as a unbiased estimate of $\boldsymbol{\Sigma}$.

We now give an illustrative example.

**EXAMPLE 6.5.** At the Royal Veterinary and Agricultural University's experimental farm, Højbakkegård, an experiment concerning the yield of crops was conducteded in the period 1956-58 as part of an international study. Experiments on 10 plant types were performed. The kinds of yield which were of interest were the amounts of

dry matter
green matter
nitrogen.

Each type of plant was grown in 6 blocks (i.e. plots of soil with different quality). In order to reduce the amount of data we will limit ourselves to three plants and to the year 1957. The results of the experiment considered are given below.

| Type of plant | Type of yield | Block No. 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Marchi-giana | Dry matter | 9.170 | 10.683 | 10.063 | 8.104 | 10.018 | 9.570 |
| | nitrogen | 0.286 | 0.335 | 0.315 | 0.259 | 0.319 | 0.304 |
| | green matter | 40.959 | 47.677 | 44.950 | 36.919 | 45.859 | 43.838 |
| Kayseri | Dry matter | 9.403 | 10.914 | 11.018 | 11.385 | 13.387 | 12.848 |
| | nitrogen | 0.285 | 0.330 | 0.333 | 0.339 | 0.400 | 0.383 |
| | green matter | 42.475 | 49.546 | 50.152 | 51.718 | 60.758 | 58.334 |
| Atlan-tic | Dry matter | 11.349 | 10.971 | 9.794 | 8.944 | 11.715 | 11.903 |
| | nitrogen | 0.369 | 0.357 | 0.319 | 0.291 | 0.379 | 0.386 |
| | green matter | 52.475 | 50.757 | 45.151 | 42.221 | 55.505 | 56.364 |

Yield in 1000 kg/ha

We wish to analyse how the yield varies with the blocks, the type of plants and the type of yield.

We will first analyse each type of yield by itself. For this we base the analysis on a two-sided analysis of variance. The model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (i = 1, 2, 3, \quad j = 1, \dots, 6),$$

and we are therefore assuming that each observation $y_{ij}$ can be written as a sum of $\mu$ (level), $\alpha_i$ (effect of plant), $\beta_j$ (effect of block) and $\varepsilon_{ij}$ (residual, being a small randomly varying quantity).

If we first consider dry matter we get

$$y_{11} = 9.170, \quad y_{12} = 10.683, \dots, \quad y_{36} = 11.903.$$

The analysis of variance table was (found by means of SSP-ANOVA)

| Source of variation | Sums of squares | Degrees of freedom | Mean squares | F-values |
|---|---|---|---|---|
| $A$ | 11.218244 | 5 | 2.243648 | 2.25 |
| $B$ | 10.945597 | 2 | 5.472798 | 5.49 |
| $AB$ | 9.970109 | 10 | 0.997010 | |
| Total | 32.133936 | 17 | | |

The test statistic for the hypothesis $\beta_1 = \cdots = \beta_6 = 0$ is

$$\mathrm{F} = \frac{s_3^2}{s_1^2} = 2.25 < 3.33 = \mathrm{F}_{95\%}(5, 10)$$

i.e. we cannot reject that the $\beta$ s equal 0.
Correspondingly the test statistic for the hypothesis $\alpha_1 = \alpha_2 = \alpha_3 = 0$ equals

$$\mathrm{F} = \frac{s_2^2}{s_1^2} = 5.49 > 4.10 = \mathrm{F}_{95\%}(2, 10).$$

At a 5% level we therefore reject that the $\alpha$ s all equal 0. However, we note that

$$\mathrm{F}_{97.5\%}(2, 10) = 5.46,$$

so there is no significance at the 2.5% level.

If we perform the corresponding computations on the nitrogen yield we get, using as observations: $y'_{ij} = y_{ij} \cdot 1000$:

| Source of variation | Sums of squares | Degrees of freedom | Mean squares | F-values |
|---|---|---|---|---|
| $A$ | 10802.27734 | 5 | 2160.45532 | 2.60 |
| $B$ | 8030.77734 | 2 | 4015.38867 | 4.83 |
| $AB$ | 8310.55469 | 10 | 831.05542 | |
| Total | 27143.60938 | 17 | | |

Here we again find that there is no difference between blocks but there is possibly a difference between plants. This difference is, however, not significant at the 2.5% level.

The corresponding computations on yield of green matter was (again using coded observations: $y'_{ij} = 1000 y_{ij}$):

| Source of variation | Sums of squares | Degrees of freedom | Mean squares | F-values. |
|---|---|---|---|---|
| $A$ | 261702416 | 5 | 52340480 | 2.75 |
| $B$ | 260173824 | 2 | 130086912 | 6.83 |
| $AB$ | 190600448 | 10 | 19060032 | |
| Total | 712476672 | 17 | | |

Here we again have that there is no difference between blocks. We also find a difference between plants at the 5% level but not at the 1% level since

$$\mathrm{F}_{99\%}(2,10) = 7.56.$$

We therefore see that the three types of yield show more or less the same sort of variation: There is no difference between blocks but there is difference between plants. These are, however, not significant at a small levels of $\alpha$.

Now the three forms of yield are known to be strongly interdependent. Therefore we will expect that the analysis of variance would give more or less similar results and it would therefore be interesting to examine the variation and the yield when we take this dependency into consideration. Such a type of analysis can be performed by a three dimensional two-sided analysis of variance i.e. we use the model

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_{ij}, \qquad i = 1,2,3, \quad j = 1,\dots,6,$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \quad \boldsymbol{\alpha}_i = \begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \alpha_{3i} \end{pmatrix}, \quad \boldsymbol{\beta}_j = \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix},$$

and the observations are

$$\boldsymbol{Y}_{ij} = \begin{pmatrix} \text{content of green matter} & \text{in plant } i \text{ in blok } j \\ \text{content of nitrogen} & -''- \\ \text{content of dry matter} & -''- \end{pmatrix}.$$

The observed values are

$$\boldsymbol{y}_{11} = \begin{pmatrix} 40.959 \\ 0.286 \\ 9.170 \end{pmatrix}, \dots, \quad \boldsymbol{y}_{36} = \begin{pmatrix} 56.364 \\ 0.386 \\ 11.903 \end{pmatrix}.$$

In this way we can aggregate the three analysis of variances shown above into one.

With the notation from p. 218 the matrices $\mathbf{Q}_1$, $\mathbf{Q}_2$ and $\mathbf{Q}_3$ are found to be

$$\mathbf{q}_2 = \begin{bmatrix} 260.18359 & & \\ 1.38547 & 0.00803 & \\ 52.37032 & 0.26262 & 10.94564 \end{bmatrix}$$

$$\mathbf{q}_3 = \begin{bmatrix} 261.70239 & & \\ 1.67129 & 0.01080 & \\ 53.97473 & 0.34801 & 11.21827 \end{bmatrix}$$

$$\mathbf{q}_1 = \begin{bmatrix} 190.59937 & & \\ 1.25512 & 0.00831 & \\ 43.45444 & 0.28667 & 9.97013 \end{bmatrix}$$

The matrices have been found by means of the BMD-programme BMDX69. Still by means of the programme mentioned we find

| Source | ln(Generalized variance) | U-statistic | Degrees of freedom | Approximate F-statistic | Degrees of freedom |
|---|---|---|---|---|---|
| $I$ | $-1.89908$ | 0.003315 | 3 2 10 | 43.6455 | 6 16.00 |
| $J$ | $-4.84194$ | 0.062894 | 3 5 10 | 2.5843 | 15 22.49 |
| Full model | $-7.60824$ | | | | |

Here $I$ corresponds to the variation between plants and $J$ to the variation between blocks.

The (in this case exact) F-test statistic for a test of the hypothesis $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = 0$, (i.e.. the hypothesis that all plants are equal) is 43.6. The number of degrees of freedom is (6,16). Since

$$\mathrm{F}(6,16)_{0.9995} = 7.74,$$

we therefore have a very strong rejection of the hypothesis.

Since

$$\mathrm{F}(15,22)_{0.975} = 2.50,$$

we see that now also the hypothesis of the blocks being equal is rejected at the level $\alpha = 2.5\%$.

The conclusion on the multi-dimensional analysis of variance is therefore that there is a clear difference in the yield for the three types of plants. It is on the other hand more uncertain if there are differences between the blocks.

We note a difference from three one-dimensional analyses. In these cases we only have moderate or no significance for the hypothesis of the plant yields being equal. We therefore have different results by considering the simultaneous analysis instead of the three marginal ones. ♦

## 6.3 Tests regarding variance-covariance matrices

In this section we will briefly give some of the tests for hypothesis on variance covariance matrices. On one hand corresponding to a hypothesis about the variance covariance matrix having a given structure or is equal to a given matrix, or on the other hand corresponding to a hypothesis that several variance covariance matrices are equal.

### 6.3.1 Tests regarding a single variance covariance matrix

First we will give a test that k-groups of normally distributed variables are independent. We are considering a $\boldsymbol{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and we divide $\boldsymbol{X}$ in k components we the dimensions $p_1, \ldots, p_k$, i.e.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_k \end{bmatrix}.$$

The corresponding partitioning of the parameters is

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_k \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \vdots & & \vdots \\ \boldsymbol{\Sigma}_{k1} & \cdots & \boldsymbol{\Sigma}_{kk} \end{bmatrix}.$$

Our hypothesis is now that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are independent i.e. that variance covariance has the form

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{kk} \end{bmatrix}.$$

If we define $\hat{\mathbf{\Sigma}}$ computed on the basis of n realisations of $\boldsymbol{X}$ in the usual way and if we partition $\hat{\mathbf{\Sigma}}$ analogously to the partitioning of $\mathbf{\Sigma}$, we have

**THEOREM 6.13.** We consider the above mentioned situation and let

$$V = \frac{\det(\hat{\mathbf{\Sigma}})}{\prod_{i=1}^{k} \det(\hat{\mathbf{\Sigma}}_{ii})}.$$

Then the coefficient test for test of the hypothesis $\mathbf{\Sigma} = \mathbf{\Sigma}_0$ is given by the critical region

$$\{V \leq v_\alpha\}.$$

When finding the boundary of the critical region we can use that

$$\begin{aligned} &P\{-m \ln V \leq v\} \\ &\quad \simeq P\{\chi^2(f) \leq v\} + \frac{\gamma_2}{m^2}[P\{\chi^2(f+4) \leq v\} - P\{\chi^2(f) \leq v\}], \end{aligned}$$

where

$$\begin{aligned} m &= n - \frac{3}{2} - \frac{p^3 - \sum p_i^3}{3(p^2 - \sum p_i^2)} \\ \gamma_2 &= \frac{p^4 - \sum p_i^4}{48} - \frac{5(p^2 - \sum p_i^2)}{96} - \frac{(p^3 - \sum p_i^3)^2}{72(p^2 - \sum p_i^2)}. \end{aligned}$$

$$f = \frac{1}{2}[p^2 - \sum p_i^2], \qquad p = \sum p_i$$

If k = 2, the V is distributed as $U(p_1, p_2, n-1-p_2)$. ▲

**PROOF 6.15.** Omitted. See e.g. [2]. ■

In the above mention situation we looked at a test for a variance covariance matrix having a certain structure. We will now turn around and look at a test for the hypothesis that a variance covariance matrix is proportional with a given matrix. We briefly give the result in

**THEOREM 6.14.** We consider independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with $\boldsymbol{X}_i \in \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, and we let

$$\mathbf{A} = \sum (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'.$$

The quotient test statistic for a test of $H_0 : \mathbf{\Sigma} = \sigma^2 \mathbf{\Sigma}_0$, where $\mathbf{\Sigma}_0$ is known and $\sigma^2$ unknown against all alternatives is

$$\mathrm{W} = \frac{[\det(\mathbf{A}\,\mathbf{\Sigma}_0^{-1})]^{\frac{n}{2}}}{[\operatorname{tr}\mathbf{A}\,\mathbf{\Sigma}_0^{-1}/p]^{\frac{pn}{2}}}.$$

When determining the critical region we can use that

$$\begin{aligned} &P\{-(n-1)\rho \ln \mathrm{W} \le z\} \\ &\quad \simeq P\{\chi^2(f) \le z\} + \omega_2 [P\{\chi^2[f+4] \le z\} - P\{\chi^2(f) \le z\}], \end{aligned}$$

where

$$\begin{aligned} \rho &= 1 - \frac{2p^2 + p + 2}{6p(n-1)} \\ f &= \frac{1}{2} p(p+1) - 1 \\ \omega_2 &= \frac{(p+2)(p-1)(p-2)(2p^3 + 6p^2 + 3p + 2)}{288 p^2 n^2 \rho^2}. \end{aligned}$$

▲

**PROOF 6.16.** Omitted. See e.g. [2]. ■

Finally we will consider the situation where we wish to test that a variance covariance matrix is equal to a given matrix. Then the following holds true

**THEOREM 6.15.** We consider independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with $\boldsymbol{X}_i \in \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, and we let

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

The quotient test statistic for a test of $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0$, where $\mathbf{\Sigma}_0$ is known against all alternatives is

$$\lambda_1 = (\frac{e}{n})^{pn/2} [\det(\mathbf{A}\,\mathbf{\Sigma}_0^{-1})]^{\frac{n}{2}} \exp(-\frac{1}{2} \operatorname{tr}(\mathbf{A}\,\mathbf{\Sigma}_0^{-1})).$$

When determining the critical region we can use that

$$P\{-2\ln\lambda_1 \le v\} \simeq P\{\chi^2(\frac{1}{2}p(p+1)) \le v\}.$$

▲

**PROOF 6.17.** Omitted. See e.g. [2]. ■

## 6.4 Test for equality of several variance-covariance matrices

We will in this section consider the problem of testing the assumption of equal variance covariance matrices in Hotelling's two sample situation and in the multidimensional analysis of variance.

We will assume that there are independent observations

$$\begin{array}{llll} \boldsymbol{X}_{11}, & \ldots, & \boldsymbol{X}_{1n_1}, & \boldsymbol{X}_{1j} \in \mathrm{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ \vdots & & \vdots & \\ \boldsymbol{X}_{k1}, & \ldots, & \boldsymbol{X}_{kn_k}, & \boldsymbol{X}_{kj} \in \mathrm{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{array},$$

and we wish to test the hypothesis

$$H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k \quad \text{against} \quad H_1 : \exists i, j : \boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j.$$

We let

$$\begin{aligned} n &= \sum n_i, \\ \mathbf{A}_i &= \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)', \end{aligned}$$

and

$$\mathbf{A} = \sum_{i=1}^{k} \mathbf{A}_i,$$

cf. section 6.2.1, where the notation $\mathbf{W}$ is used instead of $\mathbf{A}$.

We then have

**THEOREM 6.16.** As a test statistic for the test of $H_0$ against $H_1$ we can use

$$\mathrm{W}_1 = \frac{\prod_{i=1}^{k}[\det(\mathbf{A}_i)]^{\frac{(n_i-1)}{2}}}{[\det \mathbf{A}]^{\frac{(n-k)}{2}}} \cdot \frac{(n-k)^{\frac{p(n-k)}{2}}}{\prod_{i=1}^{k}(n_i-1)^{\frac{p(n_i-1)}{2}}}.$$

The critical region is of the form

$$\{\mathrm{W}_1 \le w_\alpha\}$$

and in the determination of this we can use that

$$\begin{aligned} &P\{-2\rho \ln \mathrm{W}_1 \le z\} \approx \\ &\quad P\{\chi^2(f) \le z\} + \omega_2[P\{\chi^2(f+4) \le z\} - P\{\chi^2(f) \le z\}], \end{aligned}$$

where

$$\begin{aligned} f &= \frac{1}{2}(k-1)p(p+1), \\ \rho &= 1 - (\sum_i \frac{1}{n_i} - \frac{1}{n})\frac{2p^2+3p-1}{6(p+1)(k-1)}, \\ \omega_2 &= \frac{1}{48\rho^2}p(p+1)[(p-1)(p+2)(\sum_i \frac{1}{n_i^2} - \frac{1}{n^2}) - 6(k-1)(1-\rho)^2]. \end{aligned}$$

▲

**PROOF 6.18.** Omitted. See e.g. [2]. ■

# Chapter 7
# Discriminant analysis

In this section we will address the problem of classifying an individual in one of two (or more) known populations based on measurements of some characteristics of the individual.

We first consider the problem of discriminating between two groups (classes).

## 7.1 Discrimination between two populations

### 7.1.1 Bayes and minimax solutions

We consider the **populations** $\pi_1$ and $\pi_2$ and wish to conclude whether a given individual is a member of group one or group two. We perform measurements of $p$ different characteristics of the individual and hereby get the result

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

If the individual comes from $\pi_1$ the frequency function of $\boldsymbol{X}$ is $\mathrm{f}_1(\boldsymbol{x})$ and if it comes from $\pi_2$ it is $\mathrm{f}_2(\boldsymbol{x})$.

Let us furthermore assume that we have given a **loss function** L:

| | | Choise: | |
|---|---|---|---|
| | | $\pi_1$ | $\pi_2$ |
| Truth | $\pi_1$ | 0 | $\mathrm{L}(1,2)$ |
| | $\pi_2$ | $\mathrm{L}(2,1)$ | 0 |

We will assume that there is no loss if we take the correct decision.

In certain situations one also knows approximately what the **prior probability** is to have an individual from each of the groups i.e. we haven given a prior distribution g:

$$\mathrm{g}(\pi_1) = p_1, \quad \mathrm{g}(\pi_2) = p_2.$$

We now seek a **decision function** d: $R^p \to \{\pi_1, \pi_2\}$. d is defined by

$$\mathrm{d}(\boldsymbol{x}) = \mathrm{d}_{R_1}(\boldsymbol{x}) = \begin{cases} \pi_1 & \text{if } \boldsymbol{x} \in R_1 \\ \pi_2 & \text{if } \boldsymbol{x} \in R_2 = \complement R_1. \end{cases}$$

We divide $R^p$ in two regions $R_1$ and $R_2$. If our observation lies in $R_1$ we will choose $\pi_1$ and if our observation lies in $R_2$ we will choose $\pi_2$.

If we have a **prior distribution** we define the posterior distribution k by

$$\mathrm{k}(\pi_i|\boldsymbol{x}) = \frac{\mathrm{f}_i(\boldsymbol{x})\mathrm{g}(\pi_i)}{p_1\mathrm{f}_1(\boldsymbol{x}) + p_2\mathrm{f}_2(\boldsymbol{x})} = \frac{p_i\mathrm{f}_i(\boldsymbol{x})}{p_1\mathrm{f}_1(\boldsymbol{x}) + p_2\mathrm{f}_2(\boldsymbol{x})},$$

cf. p. 6.6 in Vol. 1.

The expected loss in this distribution is

$$\begin{aligned} \mathrm{E}\boldsymbol{x}(\mathrm{L}(\pi_i, \mathrm{d}_{R_1}(\boldsymbol{x}))) &= \mathrm{L}(\pi_1, \mathrm{d}_{R_1}(\boldsymbol{x}))\mathrm{k}(\pi_1|\boldsymbol{x}) + \mathrm{L}(\pi_2, \mathrm{d}_{R_1}(\boldsymbol{x}))\mathrm{k}(\pi_2|\boldsymbol{x}) \\ &= \begin{cases} \mathrm{L}(\pi_2, \pi_1)\mathrm{k}(\pi_2|\boldsymbol{x}), & \boldsymbol{x} \in R_1 \\ \mathrm{L}(\pi_1, \pi_2)\mathrm{k}(\pi_1|\boldsymbol{x}), & \boldsymbol{x} \in R_2 \end{cases} . \end{aligned}$$

The Bayes solution is defined by minimising this quantity for any $\boldsymbol{x}$ (p. 6.9 in Vol. 1), i.e. we define $R_1$ by

$$\begin{aligned} \boldsymbol{x} \in R_1 \quad &\Leftrightarrow \quad \mathrm{L}(2,1)\mathrm{k}(\pi_2|\boldsymbol{x}) \le \mathrm{L}(1,2)\mathrm{k}(\pi_1|\boldsymbol{x}) \\ &\Leftrightarrow \quad \frac{\mathrm{L}(1,2)\mathrm{f}_1(\boldsymbol{x})p_1}{\mathrm{L}(2,1)\mathrm{f}_2(\boldsymbol{x})p_2} \ge 1 \\ &\Leftrightarrow \quad \frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \ge \frac{\mathrm{L}(2,1)}{\mathrm{L}(1,2)}\frac{p_2}{p_1}. \end{aligned}$$

These considerations are collected in

**THEOREM 7.1.** The Bayes solution to the classification problem is given by the region

$$R_1 = \{\boldsymbol{x} | \frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq \frac{\mathrm{L}(2,1)}{\mathrm{L}(1,2)} \frac{p_2}{p_1}\}.$$

▲

**REMARK 7.1.** This result is exactly the same as the one given in theorem 5, chapter 6 in Vol. 1. ▼

If we do not have a prior distribution we can instead determine a minimax strategy i.e. determine $R_1$ so that the maximal risk is minimised. The risk is (cf. p. 6.3, Vol 1)

$$\begin{aligned} R(\pi_1, \mathrm{d}_{R_1}) &= \mathrm{E}_{\pi_1} \mathrm{L}(\pi_1, \mathrm{d}_{R_1}(\boldsymbol{X})) = \mathrm{L}(1,2) P\{\boldsymbol{X} \in R_2 | \pi_1\}. \\ R(\pi_2, \mathrm{d}_{R_1}) &= \mathrm{E}_{\pi_2} \mathrm{L}(\pi_2, \mathrm{d}_{R_1}(\boldsymbol{X})) = \mathrm{L}(2,1) P\{\boldsymbol{X} \in R_1 | \pi_2\}. \end{aligned}$$

One can now show (see e.g. the proof for theorem 4, chapter 6 in Vol. 1)

**THEOREM 7.2.** The minimax solution for the classification problem is given by the region

$$R_1 = \{\boldsymbol{x} | \frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq c\},$$

where $c$ is determined by

$$\mathrm{L}(1,2) P\{\frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} < c | \pi_1\} = \mathrm{L}(2,1) P\{\frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq c | \pi_2\}.$$

▲

**REMARK 7.2.** The relation for determinating $c$ can be written

$$\begin{aligned} & \mathrm{L}(1,2) \cdot \text{ (the probability of misclassification if } \pi_1 \text{ is true)} \\ = & \mathrm{L}(2,1) \cdot \text{ (the probability of misclassification if } \pi_2 \text{ is true)} \end{aligned}$$

Since the first is an increasing and the second is a decreasing function of $c$ it is obvious that we will minimise the maximal risk when we have equality. If we do not have any idea about the size of the losses we can let them both equal one. The minimax solution gives us the region which minimises the maximal probability of misclassification. ▼

We will now consider the important special case where $\mathrm{f}_1$ and $\mathrm{f}_2$ are normal distributions.

### 7.1.2 Discrimination between two normal populations

If $\mathrm{f}_1$ and $\mathrm{f}_2$ are normal with the same variance-covariance matrix we have

**THEOREM 7.3.** Let $\pi_1 \simeq \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\pi_2 \simeq \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Then we have

$$\frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq c \Leftrightarrow \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \geq \ln c.$$

▲

**PROOF 7.1.** We introduce the inner product $(\cdot|\cdot)$ and the norm $\|\quad\|$ by

$$(\boldsymbol{x}|\boldsymbol{y}) = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

and

$$\|x\|^2 = (\boldsymbol{x}|\boldsymbol{x}).$$

We then have

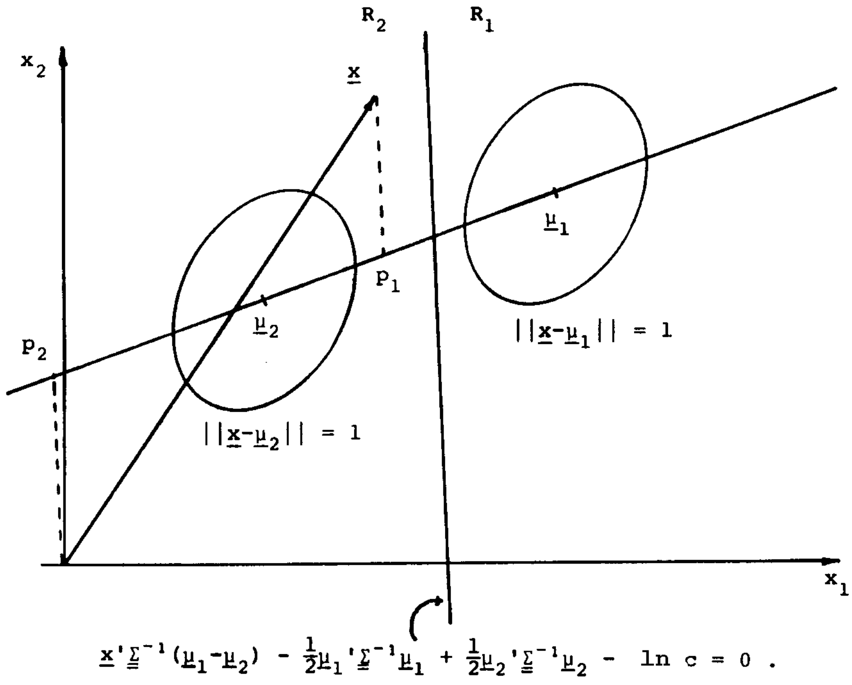
$$\mathrm{f}_i(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^p\sqrt{\det\boldsymbol{\Sigma}}}\exp(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2).$$

From this we readily get

$$\begin{aligned}
& \frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq c \Leftrightarrow \ln\frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq \ln c \\
\Leftrightarrow \quad & -\|\boldsymbol{x} - \boldsymbol{\mu}_1\|^2 + \|\boldsymbol{x} - \boldsymbol{\mu}_2\|^2 \geq 2\ln c \\
\Leftrightarrow \quad & -(\boldsymbol{x} - \boldsymbol{\mu}_1|\boldsymbol{x} - \boldsymbol{\mu}_1) + (\boldsymbol{x} - \boldsymbol{\mu}_2|\boldsymbol{x} - \boldsymbol{\mu}_2) \geq 2\ln c \\
\Leftrightarrow \quad & 2(\boldsymbol{x}|\boldsymbol{\mu}_1) - 2(\boldsymbol{x}|\boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1|\boldsymbol{\mu}_1) + (\boldsymbol{\mu}_2|\boldsymbol{\mu}_2) \geq 2\ln c \\
\Leftrightarrow \quad & 2(\boldsymbol{x}|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1|\boldsymbol{\mu}_1) + (\boldsymbol{\mu}_2|\boldsymbol{\mu}_2) \geq 2\ln c.
\end{aligned}$$

By using the connexion between $(|)$ and $\boldsymbol{\Sigma}^{-1}$ we find that the theorem readily follows.

■



**REMARK 7.3.** The expression $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq c$ is seen to define a subset of $R^p$ which is delimited by a hyper-plane (for $p = 2$ a straight line and for $p = 3$ a plane).

The vector $p_1\vec{p}_2$ is the orthogonal projection (NB! The orthogonal projection with respect to $\boldsymbol{\Sigma}^{-1}$) of $\boldsymbol{x}$ onto the line which connects $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. (It can be shown that the slope of the projection lines etc. are equal to the slope of the ellipse- (ellipsoid-) tangents in the at the points where they intersect the line $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$). Since the length of a projection of a vector is equal to the inner product between the vector and a unit vector on the line we see that we have classified the observation as coming from $\pi_1$ iff the projection of $\boldsymbol{x}$ is large enough (computed with sign). Otherwise we will classify the observation as coming from $\pi_2$.

The function

$$\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \ln c$$

is called the discriminator or the discriminant function.

We then have that the discriminator is the linear projection which - after the addition of suitable constants - minimises the expected loss (the Bayes situation) or the probability of misclassification (the minimax situation). ▼

In order to make the reader more confident with the content - we will now give a slightly different interpretation of a discriminator. If we let

$$\boldsymbol{\delta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

we have the following

**THEOREM 7.4.** The vector $\boldsymbol{\delta}$ has the property that it maximises the function

$$\varphi(\boldsymbol{d}) = \frac{[\mathrm{E}_1(\boldsymbol{X}'\boldsymbol{d}) - \mathrm{E}_2(\boldsymbol{X}'\boldsymbol{d})]^2}{\mathrm{V}(\boldsymbol{X}'\boldsymbol{d})} = \frac{[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{d}]^2}{\boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{d}}.$$

▲

**PROOF 7.2.** The proof is not very interesting but fairly simple. Since we readily have that $\varphi(k \cdot \boldsymbol{d}) = \varphi(\boldsymbol{d})$ we can determine extremes for $\varphi$ by determining extremes for the numerator under the following constraint

$$\boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{d} = 1.$$

We introduce a Lagrange multiplier $\lambda$ and seek the maximum of

$$\psi(\boldsymbol{d}) = [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{d}]^2 - \lambda(\boldsymbol{d}'\boldsymbol{\Sigma}\boldsymbol{d} - 1).$$

Now we have that

$$\frac{\partial\psi}{\partial\boldsymbol{d}} = 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{d} - 2\lambda\boldsymbol{\Sigma}\boldsymbol{d}.$$

If we let this equal 0, we have

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{d} = \lambda\boldsymbol{\Sigma}\boldsymbol{d},$$

i.e.

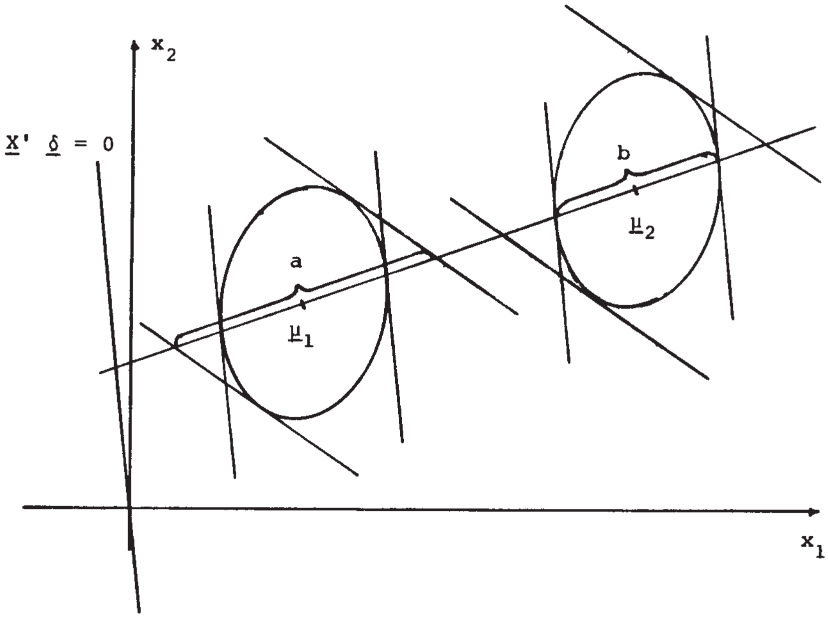
$$\boldsymbol{d} = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{d}}{\lambda}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = k \cdot \boldsymbol{\delta},$$

where $k$ is a scalar. ■

**REMARK 7.4.** The content of the theorem is that the linear function determined by $\boldsymbol{\delta}$

$$\boldsymbol{X}'\boldsymbol{\delta} = \delta_1 X_1 + \cdots + \delta_p X_p,$$

is the projection that "moves" $\pi_1$ furthest possible away from $\pi_2$ or - in analysis of variance terms - the projection which maximises the variance between populations divided by the total variance.



The geometrical content of the theorem is indicated in the above figure where

b: is the projection of the ellipse onto the line $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ in the direction determined by $\boldsymbol{x}'\boldsymbol{\delta} = 0$

a: is the projection of the ellipse onto the line $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ in a different direction.

It is seen that the projection determined by $\boldsymbol{\delta}$ onto the line which connects $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is the one which "moves" the projection of the contour ellipsoids of the two populations distribution furthest possible away from each other. ▼

We now give a theorem which is very useful in the determination of misclassification probabilities.

**THEOREM 7.5.** We consider the criterion in theorem 7.3

$$Z = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2.$$

It can be proved that

$$Z \in \begin{cases} \mathrm{N}(+\frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2, \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2), & \text{if } \pi_1 \text{ is true} \\ \mathrm{N}(-\frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2, \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2), & \text{if } \pi_2 \text{ is true} \end{cases} .$$

▲

**PROOF 7.3.** The proof is straight forward. Let us e.g. consider the case $\pi_1$ true. We then have that $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}_1$ and then

$$\begin{aligned} \mathrm{E}(Z) &= \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \\ &= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2. \end{aligned}$$

$$\begin{aligned} \mathrm{V}(Z) &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2. \end{aligned}$$

The result regarding $\pi_2$ is shown analogously. ■

We will now consider some examples.

**EXAMPLE 7.1.** We consider the case where

$$\begin{aligned} \pi_1 &\leftrightarrow \mathrm{N}(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}) \\ \pi_2 &\leftrightarrow \mathrm{N}(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}), \end{aligned}$$

and we want to determine a “best” discriminator function. Since we know nothing about the prior probabilities and so on, we will use the function which corresponds to the constant $c$ in theorem 7.3 being 1. Since

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix},$$

we get the following function

$$(x_1x_2)\begin{pmatrix}2 & -1\\ -1 & 1\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix}-\frac{1}{2}(2\cdot 16+1\cdot 4-2\cdot 8)+\frac{1}{2}(2\cdot 1+1\cdot 1-2\cdot 1)=0$$

or

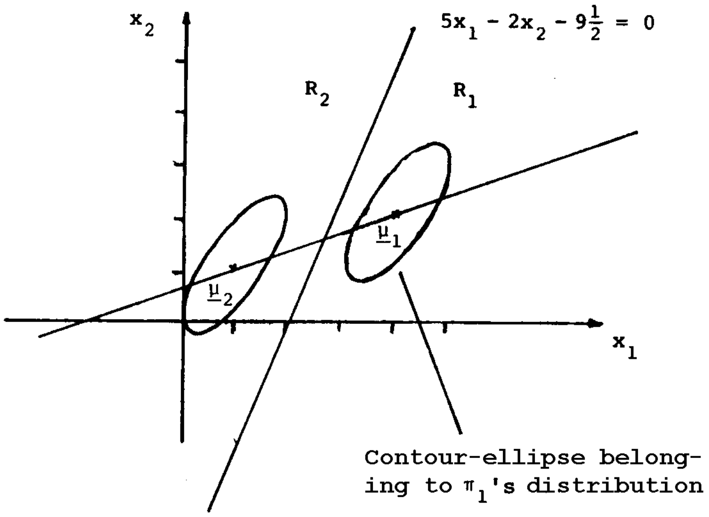
$$5x_1-2x_2-9\frac{1}{2}=0.$$

If we enter an arbitrary point, e.g. $\begin{pmatrix}5\\6\end{pmatrix}$ we get

$$5\cdot 5-2\cdot 6-9\frac{1}{2}=3\frac{1}{2}>0.$$

This point is therefore classified as coming from $\pi_1$.

We have indicated the situation in the following figure



♦

If we have a loss function, the procedure is a bit different which is seen from

**EXAMPLE 7.2.** Let us assume that we have losses assigned to the different decisions:

|  |  | Choise: |  |
|---|---|---|---|
|  |  | $\pi_1$ | $\pi_2$ |
| Truth: | $\pi_1$ | 0 | 2 |
|  | $\pi_2$ | 1 | 0 |

Since we have no prior probabilities we will determine the minimax solution. We will need

$$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = 2 \cdot 9 + 1 \cdot 1 - 2 \cdot 3 \cdot 1 = 13.$$

From theorem 7.2 follows that we must determine $c$ so

$$\begin{aligned}
& 2 \cdot P\left\{\frac{\mathrm{f}_1(\boldsymbol{X})}{\mathrm{f}_2(\boldsymbol{X})} < c|\pi_1\right\} = P\left\{\frac{\mathrm{f}_1(\boldsymbol{X})}{\mathrm{f}_2(\boldsymbol{X})} \geq c|\pi_2\right\} \\
\Leftrightarrow \quad & 2 \cdot P\{Z < \ln c|\pi_1\} = P\{Z \geq \ln c|\pi_2\} \\
\Leftrightarrow \quad & 2 \cdot P\{\mathrm{N}(\frac{1}{2}13, 13) < \ln c\} = P\{\mathrm{N}(-\frac{1}{2}13, 13) \geq \ln c\} \\
\Leftrightarrow \quad & 2 \cdot P\left\{\mathrm{N}(0,1) < \frac{\ln c - 6.5}{\sqrt{13}}\right\} = P\left\{\mathrm{N}(0,1) \geq \frac{\ln c + 6.5}{\sqrt{13}}\right\}.
\end{aligned}$$

By trying with different values of $c$ we see that

$$c \simeq 0.5617.$$

Using this value the misclassification probabilities are

$$\text{If } \pi_1 \text{ is true:} \quad P\{\mathrm{N}(0,1) < \tfrac{\ln 0.5617 - 6.5}{\sqrt{13}}\} \simeq 0.025.$$

$$\text{If } \pi_2 \text{ is true:} \quad P\{\mathrm{N}(0,1) < \tfrac{\ln 0.5617 + 6.5}{\sqrt{13}}\} \simeq 0.050.$$

The discriminating line is now determined by

$$5x_1 - 2x_2 - 9\frac{1}{2} = \ln 0.5617,$$

or

$$5x_1 - 2x_2 - 8.92 = 0.$$

This line intersects the line connecting $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ in $(2.36, 1.46)$ i.e. it is moved towards $\boldsymbol{\mu}_2$ compared to the mid-point $(2.5, 1.5)$. It is also obvious that the line is moved parallelly in this direction since we see from the loss matrix that it is more serious to be wrong if $\pi_1$ is true than if $\pi_2$ is true. We must therefore expand $R_1$ i.e. move the limiting line towards $\boldsymbol{\mu}_2$. ♦

We must stress that it is of importance that the variance-covariance matrices for the two populations are equal. If this is not the case we will get a completely different result which will be seen from the following example.

**EXAMPLE 7.3.** Let us assume that the variance-covariance matrix for population 2 is changed to an identity matrix i.e.

$$\begin{aligned}\pi_1 &\leftrightarrow \mathrm{N}\left(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right) \\ \pi_2 &\leftrightarrow \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)\end{aligned}$$

Again we want to classify an observation $\boldsymbol{X}$ which comes from one of the above mentioned distributions. Since the variance covariance matrices are not equal we cannot use the result in theorem 7.3 but have to start from the beginning with theorem 7.2.

For $c > 0$ we have

$$\frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq c \quad \Leftrightarrow$$

$$-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + (\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \geq 2\ln c.$$

Since

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) &= 2(x_1-4)^2 + (x_2-2)^2 - 2(x_1-4)(x_2-2) \\ &= 2x_1^2 + x_2^2 - 2x_1x_2 - 12x_1 + 4x_2 + 20,\end{aligned}$$

and

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) &= (x_1-1)^2 + (x_2-1)^2 \\ &= x_1^2 + x_2^2 - 2x_1 - 2x_2 + 2,\end{aligned}$$

then

$$\frac{\mathrm{f}_1(\boldsymbol{x})}{\mathrm{f}_2(\boldsymbol{x})} \geq c \Leftrightarrow -x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 \geq 2\ln c.$$

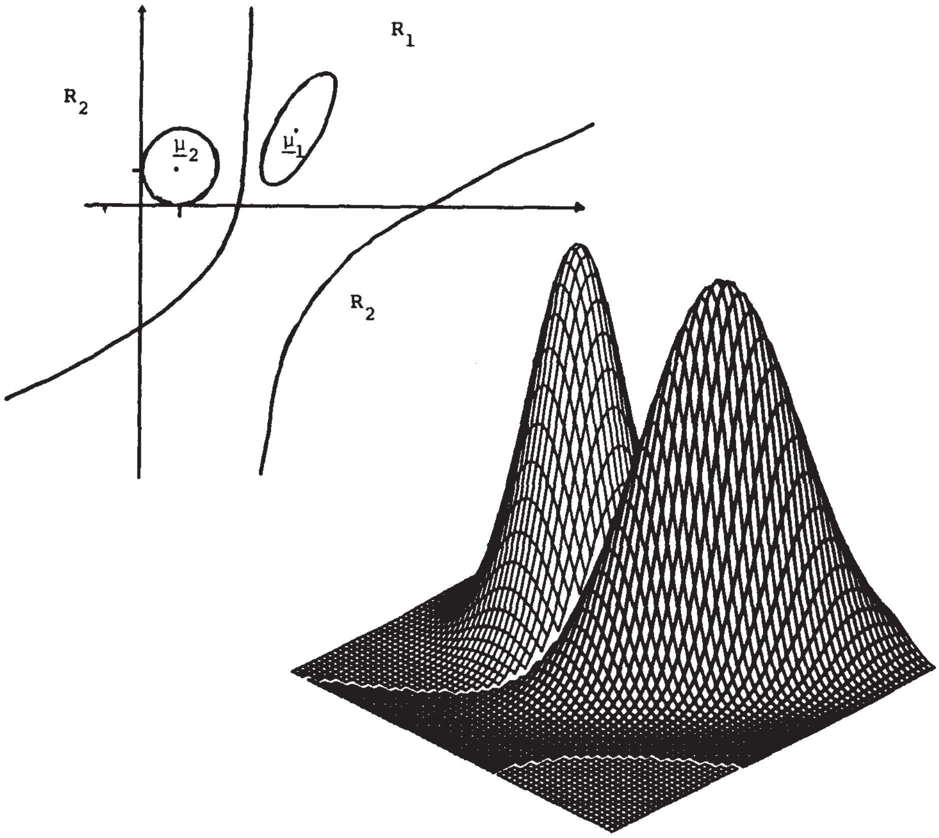
If we choose $c = 1$, we note that the curve which separates $R_1$ and $R_2$ is the hyperbola

$$\{\boldsymbol{x}| - x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 = 0\}.$$

It has centre in $(3, -2)$ and asymptotes

$$x_1 - 3 = 0,$$

$$x_1 - 2x_2 - 7 = 0.$$



These curves are shown in the above figure together with the contour ellipses for the two normal distributions. Note e.g. that a point such as $(9, 0)$ is in $R_2$ and therefore will be classified as coming from the distribution with centre in $(1, 1)$. Furthermore the frequency functions are shown.

♦

We will not consider the problem of misclassification probabilities in cases as the above mentioned where we have quadratic discriminators.

### 7.1.3 Discrimination with unknown parameters

If one does not know the two distributions $\mathrm{f}_1$ and $\mathrm{f}_2$ one must estimate them based on some observations and then construct discriminators from the estimated distributions the same way we did for the exact distributions.

Let us consider the normal case

$$\begin{aligned} \pi_1 &\leftrightarrow \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ \pi_2 &\leftrightarrow \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \end{aligned}$$

where the parameters are unknown. If we have observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ which we know come from $\pi_1$ and observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ which we know come from $\pi_2$ we can estimate the parameters as follows

$$\begin{aligned} \hat{\boldsymbol{\mu}}_1 &= \frac{1}{n_1}\sum_i \boldsymbol{X}_i = \bar{\boldsymbol{X}} \\ \hat{\boldsymbol{\mu}}_2 &= \frac{1}{n_2}\sum_i \boldsymbol{Y}_i = \bar{\boldsymbol{Y}} \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{n_1+n_2-2}(\sum_i(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' + \sum_i(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})') \end{aligned}$$

In complete analogy to the theorem on p. 232 we have the discriminator

$$\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) - \frac{1}{2}\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1 + \frac{1}{2}\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2 \,.$$

The exact distribution of this quantity if we substitute $\boldsymbol{x}$ with a stochastic variable $\boldsymbol{X} \in \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ is fairly complicated but for large sample sizes it is asymptotically equal to the distribution of $Z$ in theorem 7.5 so for reasonable sample sizes we can use the theory we have derived.

The estimated norm between the expected values is

$$\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2\|^2 \simeq \mathrm{D}^2 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) = \|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2\|^2_{\hat{\boldsymbol{\Sigma}}^{-1}}.$$

This is called **Mahalanobis**’ distance. It should here be noted that a number of authors use the expression Mahalanobis’ distance also about the quantity $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. This is after the Indian statistician P.C. Mahalanobis who developed discriminant analysis at the same time as the English statistician R.A. Fisher in the 1930’s.

By means of $D^2$ we can test if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ since

$$Z = \frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \cdot \frac{n_1 n_2}{n_1+n_2}\mathrm{D}^2$$

is $\mathrm{F}(p, n_1 + n_2 - p - 1)$-distributed if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. If $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ then $Z$ has a larger mean value so the critical region corresponds to large values of $Z$. This test is of course equivalent to Hotelling's $T^2$-test in section 6.1.2.

We give an example (data come from K.R. Nair: A biometric study of the desert locust, Bull. Int. Stat. Inst. 1951).

**EXAMPLE 7.4.** In a study of desert locusts one measured the following biometric characteristics they were

| | |
|---|---|
| $x_1$: | length of hind femur |
| $x_2$: | maximum width of the head in the genal region |
| $x_3$: | length of pronotum at the scull |

The two species which were examined are gregaria and an intermediate phase between gregaria and solotaria.

The following mean values were found.

| | Mean values | |
|---|---|---|
| | Gregaria<br>$n_1 = 20$ | Intermediate phase<br>$n_2 = 72$ |
| $x_1$ | 25.80 | 28.35 |
| $x_2$ | 7.81 | 7.41 |
| $x_3$ | 10.77 | 10.75 |

The estimated variance-covariance matrix is

| | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| $x_1$ | 4.7350 | 0.5622 | 1.4685 |
| $x_2$ | 0.5622 | 0.1413 | 0.2174 |
| $x_3$ | 1.4685 | 0.2174 | 0.5702 |

We are now interested in determining a discrimination function for classification of future locusts by means of measurements of $x_1$, $x_2$, $x_3$.

However, first it would be reasonable to check if the three measurements from the two populations are different at all i.e. we must investigate if it can be assumed that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. We have

$$\mathrm{D}^2 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) = 9.7421.$$

This value is inserted in the test statistic p. 241 and we get

$$Z = \frac{20 + 72 - 3 - 1}{3(20 + 72 - 2)} \cdot \frac{20 \cdot 72}{20 + 72} \cdot 9.7421 = 49.70.$$

Since

$$\mathrm{F}(3,88)_{0.999} \simeq 6,$$

we will reject the hypothesis of the two mean values being equal. It is therefore reasonable to try constructing a discriminator.

We have

$$\boldsymbol{x}'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) = -2.7458x_1 + 6.6217x_2 + 4.5820x_3$$

and

$$\frac{1}{2}(\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2) = 25.3506.$$

Since there is no information on prior probabilities we will use $c = 1$, i.e. : $\ln c = 0$, and we will therefore use the function

$$\mathrm{d}(\boldsymbol{x}) = -2.7458x_1 + 6.6217x_2 + 4.582x_3 - 25.3506$$

in classifying the two possible locust species.

If we for instance have caught a specimen and measured the characteristics

$$\boldsymbol{x} = \begin{pmatrix} 27.06 \\ 8.03 \\ 11.36 \end{pmatrix}$$

we get $\mathrm{d}(\boldsymbol{x}) = 5.5715 > 0$ meaning we will classify the individual as being a gregaria.

♦

### 7.1.4 Test for best discrimination function

We remind ourselves that the best discriminator

$$\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2),$$

can be found by maximising the function

$$\hat{\varphi}(\boldsymbol{d}) = \frac{[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\boldsymbol{d}]^2}{\boldsymbol{d}'\hat{\boldsymbol{\Sigma}}\boldsymbol{d}}.$$

The maximum value is

$$\hat{\varphi}(\hat{\boldsymbol{\delta}}) = \frac{[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)]^2}{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)} = \mathrm{D}^2,$$

i.e. Mahalanobis' $\mathrm{D}^2$ is the maximum value of $\hat{\varphi}(\boldsymbol{d})$. For an arbitrary (fixed) $\boldsymbol{d}$ we now let

$$\mathrm{D}_1^2 = \hat{\varphi}(\boldsymbol{d}) = \frac{[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\boldsymbol{d}]^2}{\boldsymbol{d}'\hat{\boldsymbol{\Sigma}}\boldsymbol{d}}.$$

We can then test the hypothesis that the linear projection determined by $\boldsymbol{d}$ is the best discriminator by means of the test statistic

$$Z = \frac{n_1 + n_2 - p - 1}{p - 1} \cdot \frac{n_1 n_2 (\mathrm{D}^2 - \mathrm{D}_1^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 \, \mathrm{D}_1^2},$$

which is $\mathrm{F}(p - 1, n_1 + n_2 - p - 1)$-distributed under the hypothesis. Large values of $Z$ are critical.

We will not consider the reason why the distribution for the null-hypothesis looks the way it does but just note that $Z$ gives a measure of how much the "distance" between the two populations is reduced by using $\boldsymbol{d}$ instead of $\hat{\boldsymbol{\delta}}$. If this reduction is too big i.e. if Z is large we will not be able to assume that $\boldsymbol{d}$ gives essentially as good a discrimination between the two populations as $\hat{\boldsymbol{\delta}}$.

**EXAMPLE 7.5.** In the following table we give averages of 50 measurements of different characteristics of two different types of Iris, Iris versicolor and Iris setosa. (The data come from Fisher's investigations in 1936.)

| | Versicolor | Setosa | Difference |
|---|---|---|---|
| Sepal length | 5.936 | 5.006 | 0.930 |
| Sepal width | 2.770 | 3.428 | −0.658 |
| Petal length | 4.260 | 1.462 | 2.789 |
| Petal width | 1.326 | 0.246 | 1.080 |

The estimated variance-covariance matrix (based on 98 degrees of freedom) is

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.19534 & 0.09220 & 0.099626 & 0.03306 \\ & 0.12108 & 0.04718 & 0.02525 \\ & & 0.12549 & 0.039586 \\ & & & 0.02511 \end{bmatrix}$$

From this it readily follows that

$$\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) = \begin{bmatrix} -3.0692 \\ -18.0006 \\ 21.7641 \\ 30.7549 \end{bmatrix}.$$

Mahalanobis' distance between the mean values is

$$\mathrm{D}^2 = [0.930, -0.658, 2.789, 1.080] \begin{bmatrix} -3.0692 \\ -18.0006 \\ 21.7641 \\ 30.7549 \end{bmatrix} = 103.2119.$$

We first test if we can assume that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. The test statistic is

$$\frac{50+50-4-1}{4(50+50-2)} \frac{50 \cdot 50}{50+50} \cdot 103.2119 = 625.3256$$

$$> \mathrm{F}(4,95)_{0.9995} \simeq 5.5.$$

It is therefore not reasonable to assume $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

By looking at the differences between the components in $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ we note that the number for versicolor is largest except for $x_2$ (the sepal width). Since we are looking for a linear projection which takes a large value for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ we could try with the projection

$$\boldsymbol{x}'\boldsymbol{d}_0 = x_1 - x_2 + x_3 + x_4,$$

where $\boldsymbol{d}_0$ here is the vector $\begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$.

We will now test if it can be assumed that the best discriminator has the form

$$\boldsymbol{\delta} = \text{constant} \cdot \begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix} = \text{constant} \cdot \boldsymbol{d}_0.$$

We determine the value of $\varphi$ corresponding to $\boldsymbol{d}_0$:

$$\frac{[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\boldsymbol{d}_0]^2}{\boldsymbol{d}_0'\hat{\boldsymbol{\Sigma}}\boldsymbol{d}_0} = 61.9479.$$

The test statistic becomes

$$\frac{50+50-4-1}{4-1}\cdot\frac{50\cdot 50(103.2119-61.9479)}{(50+50)(50+50-2)+50\cdot 50\cdot 61.9479}$$

$$= 1984 > \mathrm{F}(3,95)_{0.9995} \simeq 6.5.$$

We must therefore reject the hypothesis and note that we cannot assume that the best discriminator is of the form $x_1 - x_2 + x_3 + x_4$. ♦

### 7.1.5 Test for further information

Given one has obtained measurements of a number of variables for some individuals with the objective of determining a discriminant function. Often the question arises if it is really necessary with all the measurements, or if one can do with fewer variables in order to separate the populations from each other. One could e.g. think it might be sufficient just to measure the length of sepal and petal in order to discriminate between Iris versicolor and Iris setosa.

We will formulate these thoughts a bit more precisely. In order to perform a discrimination we measure the variables $X_1, \ldots, X_p$. We now will formulate a test in order to investigate if it might be possible that the last $q$ variables are unnecessary for the discrimination.

We still assume that there are $n_1$ observations from $\pi_1$ and $n_2$ observations from population $\pi_2$. We let

$$\begin{bmatrix} X_1 \\ \vdots \\ X_{p-q} \end{bmatrix} = \boldsymbol{X}_1 \quad \text{and} \quad \begin{bmatrix} X_{p-q+1} \\ \vdots \\ X_p \end{bmatrix} = \boldsymbol{X}_2,$$

and we perform the same partitioning of mean vectors and variance-covariance matrix

$$\begin{aligned} \boldsymbol{\mu}_i &= \begin{bmatrix} \boldsymbol{\mu}_i^{(1)} \\ \boldsymbol{\mu}_i^{(2)} \end{bmatrix} \\ \boldsymbol{\Sigma} &= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \end{aligned}$$

We now compute Mahalanobis' distance between the populations, first using full information i.e. all $p$ variables and then using the reduced information i.e. only the first

$p-q$ variables. We then have

$$D_p^2 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$$

and

$$D_{p-q}^2 = (\hat{\boldsymbol{\mu}}_1^{(1)} - \hat{\boldsymbol{\mu}}_2^{(1)})'\hat{\boldsymbol{\Sigma}}_{11}^{-1}(\hat{\boldsymbol{\mu}}_1^{(1)} - \hat{\boldsymbol{\mu}}_2^{(1)}).$$

A test for the hypothesis that the last $q$ variables do not contribute to a better discrimination is based on

$$Z = \frac{n_1+n_2-p-1}{q}\frac{n_1n_2(D_p^2 - D_{p-q}^2)}{(n_1+n_2)(n_1+n_2-2)+n_1n_2D_{p-q}^2}.$$

It can be shown that $Z \in \mathrm{F}(q, n_1+n_2-p-1)$ if $H_0$ is true. We omit the proof, but just state that $Z$ "measures" relatively the larger distance there is between populations when going from $p-q$ variables to $p$ variables. It is therefore also intuitively reasonable that we reject the hypothesis that it is sufficient with $p-q$ variables if $Z$ is large.

We now give an illustrative

**EXAMPLE 7.6.** We will investigate if it is sufficient only to measure the length of sepal and petal in order to discriminate the types of Iris given in example 7.5.

We now perform an ordinary discriminant analysis on the data given that we disregard the width measurements. The resulting Mahalanobis' distance is

$$D_2^2 = 76.7082,$$

so the test statistic for the hypothesis is

$$\frac{50+50-4-1}{2}\frac{50\cdot 50(103.2119-76.7082)}{(50+50)(50+50-2)+50\cdot 50\cdot 76.7082}$$
$$= 15.6132 > \mathrm{F}(2,95)_{0.9995} \simeq 8.25.$$

We must therefore conclude that there is actually extra information in the width measurements which can help us in discriminating setosa from versicolor. ♦

# 7.2 Discrimination between several populations

## 7.2.1 The Bayes solution

The main idea of the generalisation in this section is that one compares the populations pairwise as in the previous section to finally choose the most probable population.

We consider the populations

$$\pi_1, \dots, \pi_k$$

Based on measurements of $p$ characteristics (or variables) of a given individual we wish to classify it as coming from one of the populations $\pi_1, \dots, \pi_k$.

The observed measurement is

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

If the individual comes from $\pi_i$ then the frequency function for $\boldsymbol{X}$ is $\mathrm{f}_i(\boldsymbol{x})$.

We assume that a loss function L is given as shown in the following table.

| | | Choise | | | |
|---|---|---|---|---|---|
| | | $\pi_1$ | $\pi_2$ | $\cdots$ | $\pi_k$ |
| | $\pi_1$ | $0$ | $\mathrm{L}(1,2)$ | $\cdots$ | $\mathrm{L}(1,k)$ |
| | $\pi_2$ | $\mathrm{L}(2,1)$ | $0$ | $\cdots$ | $\mathrm{L}(2,k)$ |
| Truth | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| | $\pi_k$ | $\mathrm{L}(k,1)$ | $\mathrm{L}(k,2)$ | $\cdots$ | $0$ |

Finally we assume we have a prior distribution

$$\mathrm{g}(\pi_i) = p_i, \qquad i = 1, \dots, k.$$

For an individual with the observation $\boldsymbol{x}$ we define the discriminant value or discriminant score for the $i$'th population as

$$S_i^*(\boldsymbol{x}) = S_i^* = -[p_1\mathrm{f}_1(\boldsymbol{x})\mathrm{L}(1,i) + \cdots + p_k\mathrm{f}_k(\boldsymbol{x})\mathrm{L}(k,i)]$$

(note that $\mathrm{L}(i,i) = 0$ so the sum has no term $p_i\mathrm{f}_i(\boldsymbol{x})$). Since the posterior probability for $\pi_\nu$ is

$$\begin{aligned}\mathrm{k}(\pi_\nu|\boldsymbol{x}) &= \frac{p_\nu \mathrm{f}_\nu(\boldsymbol{x})}{p_1\mathrm{f}_1(\boldsymbol{x})+\cdots+p_k\mathrm{f}_k(\boldsymbol{x})}\\ &= \frac{p_\nu \mathrm{f}_\nu(\boldsymbol{x})}{\mathrm{h}(\boldsymbol{x})},\end{aligned}$$

we note that by choosing the $i$'th population then $S_i^*$ is a constant $(-\mathrm{h}(\boldsymbol{x}))$ times the expected loss with respect to the posterior distribution of $\pi$. Since the proportionality factor $-\mathrm{h}(\boldsymbol{x})$ is negative we note that the Bayes' solution to the decision problem is to choose the population which has the largest discriminant value (discriminant score) i.e. choose $\pi_\nu$ if

$$S_\nu^* \geq S_i^*, \qquad \forall i.$$

If all losses $\mathrm{L}(i,j)$ $(i \neq j)$ are equal we can simplify the expression for the discriminant score. We prefer $\pi_i$ compared to $\pi_j$ if

$$S_i^* > S_j^*,$$

i.e. if

$$\begin{aligned}&-(\sum_\nu p_\nu \mathrm{f}_\nu(\boldsymbol{x}) - p_i\mathrm{f}_i(\boldsymbol{x})) > -(\sum_\nu p_\nu \mathrm{f}_\nu(\boldsymbol{x}) - p_j\mathrm{f}_j(\boldsymbol{x}))\\ \Leftrightarrow\quad & p_i\mathrm{f}_i(\boldsymbol{x}) > p_j\mathrm{f}_j(\boldsymbol{x}).\end{aligned}$$

In this case we can therefore choose the discriminant score

$$S_i' = p_i\mathrm{f}_i(\boldsymbol{x}).$$

In this case the **Bayes' rule** is that we choose the population which has the largest posterior distribution i.e. choose group $i$, if $S_i' > S_j', \forall j \neq i$. This rule is not only used where the losses are equal but also where it has not been possible to determine such losses. If the $p_i$'s are unknown and it is not possible to estimate them one usually uses the discriminant score

$$S_i'' = \mathrm{f}_i(\boldsymbol{x}),$$

i.e. choose the population where the observed probability is the largest.

The minimax solutions are determined by choosing the strategy which makes all the misclassification probabilities equally large. (Still assuming that all losses are equal.) However, we will not go into more detail about this here.

### 7.2.2 The Bayes' solution in the case with several normal distributions

We will now consider the case where

$$\pi_i \quad \leftrightarrow \quad \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

i.e.

$$\mathrm{f}_i(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}_i}} \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)),$$

for $i = 1, \ldots, k$.

Since we get the same decision rule by choosing monotone transformations of our discriminant scores we will take the logarithm of the $\mathrm{f}_i$'s and disregard the common factor $(2\pi)^{-\frac{p}{2}}$. This gives (assuming that the losses are equal)

$$S_i' = -\frac{1}{2}\ln(\det \boldsymbol{\Sigma}_i) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln p_i.$$

This function is quadratic in $\boldsymbol{x}$ and is called a quadratic discriminant function. If all the $\boldsymbol{\Sigma}_i$ are equal then the terms

$$-\frac{1}{2}\ln\det\boldsymbol{\Sigma} - \frac{1}{2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

are common for all $S_i$'s and can therefore be omitted. We then get

$$S_i = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i.$$

This is seen to be a linear (affine) function in $\boldsymbol{x}$ and is called a linear discriminant function. If there are only two groups we note that we choose group 1 if

$$\begin{aligned} & S_1' > S_2' \Leftrightarrow S_1 - S_2 > 0 \\ \Leftrightarrow \quad & \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \geq \ln\frac{p_2}{p_1}, \end{aligned}$$

i.e. the same result as p. 232.

The posterior probability for the $\nu$'th group becomes

$$\mathrm{k}(\pi_\nu|\boldsymbol{x}) = \frac{\exp(S_\nu)}{\sum_{i=1}^k \exp(S_i)}$$

It is of course possible to describe the decision rules by dividing $R^p$ into sets $R_1, \ldots, R_k$ so that we choose $\pi_i$ exactly when $\boldsymbol{x} \in R_i$. Among other things this can be seen from the following

**EXAMPLE 7.7.** We consider populations $\pi_1$, $\pi_2$ and $\pi_3$ given by normal distributions with expected values

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\mu}_3 = \begin{pmatrix} 2 \\ 6 \end{pmatrix},$$

and common variance-covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

cf. the example p. 236. Assuming that all $p_i$ are equal so that we may disregard them in the discriminant scores - we then have

$$\begin{aligned} S_1' &= (x_1 x_2) \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \end{pmatrix} - \frac{1}{2}(4,2) \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \end{pmatrix} \\ &= 6x_1 - 2x_2 - 10 \end{aligned}$$

$$\begin{aligned} S_2' &= (x_1 x_2) \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2}(1,1) \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= x_1 - \frac{1}{2} \end{aligned}$$

$$\begin{aligned} S_3' &= (x_1 x_2) \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 6 \end{pmatrix} - \frac{1}{2}(2,6) \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 6 \end{pmatrix} \\ &= -2x_1 + 4x_2 - 10. \end{aligned}$$

We now choose to prefer $\pi_1$ to $\pi_2$ if

$$\begin{aligned} u_{12}(\boldsymbol{x}) &= (6x_1 - 2x_2 - 10) - (x_1 - \frac{1}{2}) \\ &= 5x_1 - 2x_2 - 9\frac{1}{2} \\ &> 0. \end{aligned}$$

We choose to prefer $\pi_1$ to $\pi_3$ if

$$\begin{aligned} u_{13}(\boldsymbol{x}) &= (6x_1 - 2x_2 - 10) - (-2x_1 + 4x_2 - 10) \\ &= 8x_1 - 6x_2 \\ &> 0, \end{aligned}$$

and finally we will choose to prefer $\pi_2$ to $\pi_3$ if

$$\begin{aligned} u_{23}(\boldsymbol{x}) &= (x_1 - \frac{1}{2}) - (-2x_1 + 4x_2 - 10) \\ &= 3x_1 - 4x_2 + 9\frac{1}{2} \\ &> 0. \end{aligned}$$

It is now evident that we will choose $\pi_1$ if both $u_{12}(\boldsymbol{x}) > 0$ and $u_{13}(\boldsymbol{x}) > 0$ and analogously with the others.

We can therefore define the regions

$$\begin{aligned} R_1 &= \{\boldsymbol{x} | u_{12}(\boldsymbol{x}) > 0 \quad \wedge \quad u_{13}(\boldsymbol{x}) > 0\} \\ R_2 &= \{\boldsymbol{x} | u_{12}(\boldsymbol{x}) < 0 \quad \wedge \quad u_{23}(\boldsymbol{x}) > 0\} \\ R_3 &= \{\boldsymbol{x} | u_{13}(\boldsymbol{x}) < 0 \quad \wedge \quad u_{23}(\boldsymbol{x}) < 0\}, \end{aligned}$$

and we have that we will choose $\pi_i$ exactly when $\boldsymbol{x} \in R_i$.

We have sketched the situation in the following figure.

One can easily prove that the lines will intersect in a point. It is, however, also possible to make a simple reasoning for this. Let us assume that the situation is as in figure 7.1.

We now note that

$$u_{ij} > 0 \Leftrightarrow S_i' > S_j' \Leftrightarrow \mathrm{f}_i > \mathrm{f}_j.$$

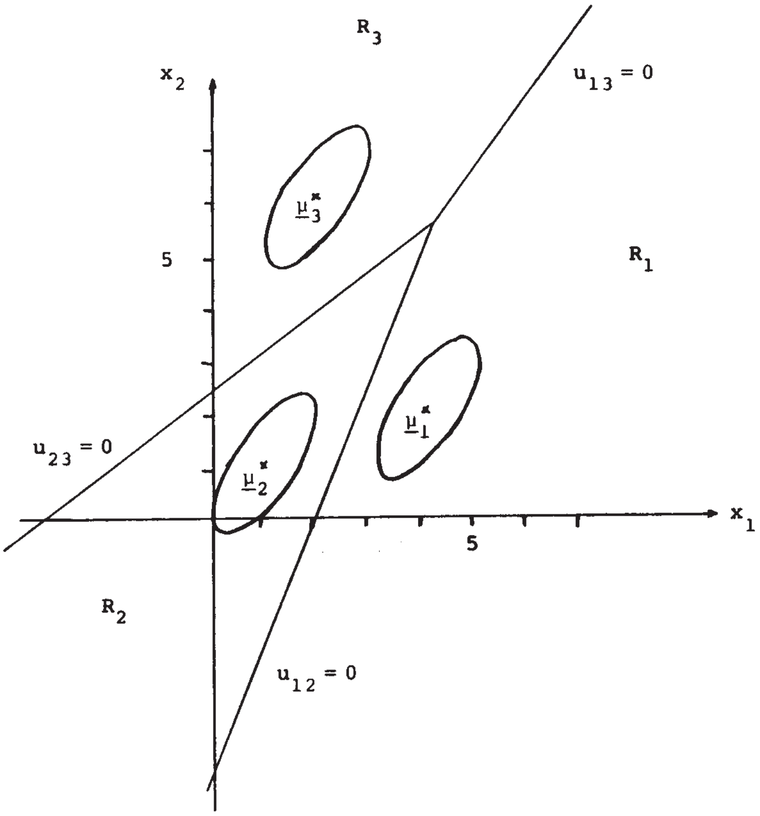
For the point $\boldsymbol{x}$ we have

$$\left.\begin{array}{lll} u_{23}(\boldsymbol{x}) < 0 & \text{i.e.} & \mathrm{f}_2(\boldsymbol{x}) < \mathrm{f}_3(\boldsymbol{x}) \\ u_{13}(\boldsymbol{x}) > 0 & \text{i.e.} & \mathrm{f}_1(\boldsymbol{x}) > \mathrm{f}_3(\boldsymbol{x}) \end{array}\right\} \Rightarrow \mathrm{f}_1(\boldsymbol{x}) > \mathrm{f}_2(\boldsymbol{x})$$

$$u_{12}(\boldsymbol{x}) < 0 \quad \text{i.e.} \quad \mathrm{f}_1(\boldsymbol{x}) < \mathrm{f}_2(\boldsymbol{x})$$

We have now established a contradiction i.e. the three lines determined by $u_{12}$, $u_{13}$ and $u_{23}$ must intersect each other in one single point. ♦

If the parameters are unknown and instead are estimated they are normally substituted in the estimating expressions in the above mentioned relations cf. the procedure in section 7.1.3.



### 7.2.3 Alternative discrimination procedure for the case of several populations.

In the previous section we have given one form of the generalisation of discriminant analysis from 2 to several populations. We will now describe another procedure which instead generalises theorem 7.4.

We still consider $k$ groups with $n_1, \ldots, n_k$ observations in each. The group averages are called $\bar{\boldsymbol{X}}_1, \ldots, \bar{\boldsymbol{X}}_k$. We define an "among groups" (or between groups) matrix

$$\mathbf{A} = \sum_{i=1}^{k} n_i(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})',$$

a ”within groups” matrix

$$\mathbf{W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)'$$

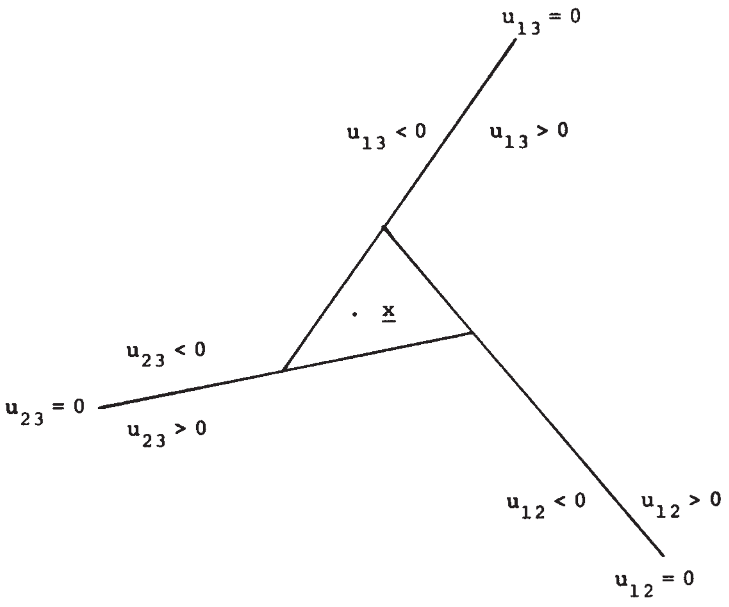


Figure 7.1:

and a "total" matrix

$$\mathbf{T} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})'.$$

A fundamental equation is that

$$\mathbf{T} = \mathbf{A} + \mathbf{W}.$$

We can now go ahead with the discrimination. We seek a best discriminator function where best means that the function should maximise the ratio between variation among groups and variation within groups. I.e. we seek a function $\boldsymbol{y} = \boldsymbol{d}'\boldsymbol{x}$ so

$$\varphi(\boldsymbol{d}) = \frac{\boldsymbol{d}'\mathbf{A}\,\boldsymbol{d}}{\boldsymbol{d}'\mathbf{W}\,\boldsymbol{d}} \qquad (\boldsymbol{d} \text{ is chosen so } \boldsymbol{d}'\boldsymbol{d} = 1)$$

is maximised. We note from theorem 1.23 that the maximum value is the largest eigenvalue $\lambda_1$ and the corresponding eigenvector $\boldsymbol{d}_1$ to

$$\det(\mathbf{A} - \lambda\mathbf{W}) = 0$$

or

$$\det(\mathbf{W}^{-1}\mathbf{A} - \lambda\mathbf{I}) = 0.$$

We then seek a new discriminant function $\boldsymbol{d}_2$ so

$$\varphi(\boldsymbol{d}_2) = \frac{\boldsymbol{d}_2\mathbf{A}\,\boldsymbol{d}_2}{\boldsymbol{d}_2\mathbf{W}\,\boldsymbol{d}_2}$$

is maximised under the constraint that

$$\boldsymbol{d}_2'\boldsymbol{d}_1 = 0 \quad \text{or} \quad \boldsymbol{d}_1 \perp \boldsymbol{d}_2 \quad \text{and} \quad \boldsymbol{d}_2'\boldsymbol{d}_2 = 1.$$

This corresponds to the second largest eigenvalue for $\mathbf{W}^{-1}\mathbf{A}$ and the corresponding eigenvector.

In this way one can continue until one gets an eigenvalue for $\mathbf{W}^{-1}\mathbf{A}$ which is 0 (or until $\mathbf{W}^{-1}\mathbf{A}$ is exhausted).

A plot of the projections of the single observations (centered by the total mean) onto the $\boldsymbol{d}_1, \boldsymbol{d}_2$ plane is very useful as a means of visualisation. This plan separates the points best in the sense described above.

The coordinates of the projections are

$$[\boldsymbol{d}_1'(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}), \quad \boldsymbol{d}_2'(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})].$$

Another useful plot consists of the vectors

$$\begin{pmatrix} d_{11} \\ d_{21} \end{pmatrix}, \ldots, \begin{pmatrix} d_{1p} \\ d_{2p} \end{pmatrix}.$$

These show with which weight the value of each single variable contributes to the plot on the $(\boldsymbol{d}_1, \boldsymbol{d}_2)$-plane.

E.g. in the programme BMD07M - STEPWISE DISCRIMINANT ANALYSIS - the plane $(\boldsymbol{d}_1, \boldsymbol{d}_2)$ is denoted the first two canonical variables.

As the name indicates variables can in this programme be included or removed from the analysis in a way which is completely analogous to a stepwise regression analysis (The version which is called STEPWISE REGRESSION). Apart from controlling the inclusion and removal of variables by means of F-tests there are a number of intuitive criteria which are very well described in the BMD manual p. 243.

It should also be mentioned here that Wilk's $\Lambda$ for the test of the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k \quad \text{against} \quad H_1 : \exists i, j : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j,$$

is

$$\Lambda = \frac{\det \mathbf{W}}{\det \mathbf{T}} = \prod_{j=1}^{p} \frac{1}{1+\lambda_j}.$$

The distribution of this quantity can be approximated by a $\chi^2$- or F-distribution. The latter is probably the numerically best approximation. These are given in the BMD manual p. 242. Cf. with section 6.2.1.

**EXAMPLE 7.8.** In the following table we give mean values and standard deviations for the content of different elements of 208 washed soil samples collected in Jameson Land. The variable Sum gives the sum of the content of Y and La.

| Variable | Mean Value | Standard deviation |
|---|---|---|
| B | 73 | 141 |
| Ti | 40563 | 22279 |
| V | 678 | 491 |
| Cr | 1135 | 1216 |
| Mn | 2562 | 2081 |
| Fe | 225817 | 122302 |
| Co | 62 | 26 |
| Ni | 116 | 54 |
| Cu | 69 | 56 |
| Ga | 21 | 10 |
| Zr | 14752 | 14771 |
| Mo | 29 | 20 |
| Sn | 56 | 99 |
| Pb | 351 | 786 |
| Sum | – | – |

A distributional analysis showed that the data were best approximated by LN-distributions. Therefore all numbers were logarithmically transformed and were furthermore standardised in order to obtain a mean of 0 and a variance of 1. The problem is to how great an extent the content of the elements characterises the different geologic periods involved in the area. The number of measurements from the different periods are given below.

| Period | Number |
|---|---|
| Jura | 17 |
| Trias | 80 |
| Perm | 30 |
| Carbon | 9 |
| Devon | 31 |
| Tertiære intrusives | 35 |
| Caledonsk crystallic | 4 |
| Eleonora Bay Formation | 2 |

In order to examine this some discriminant analyses were performed. We will not pursue this further here. We will simply illustrate the use of the previously mentioned plot, see figure 7.2.

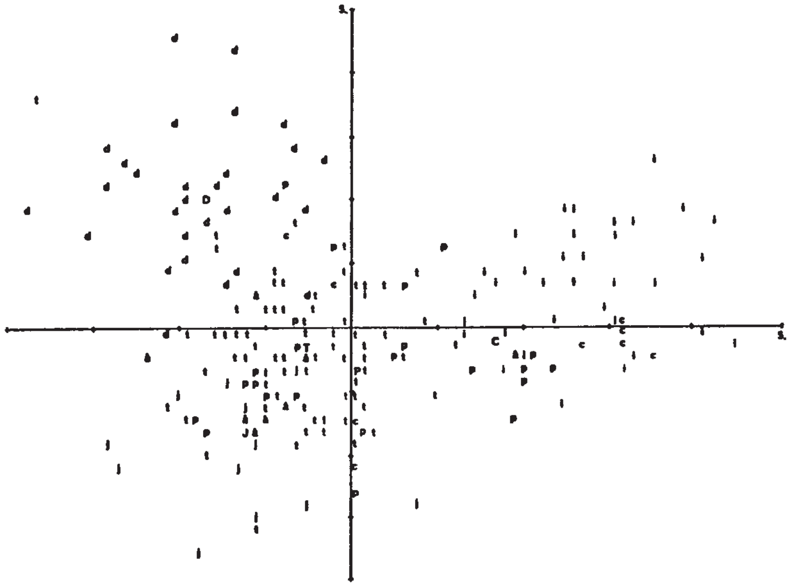


Figure 7.2:

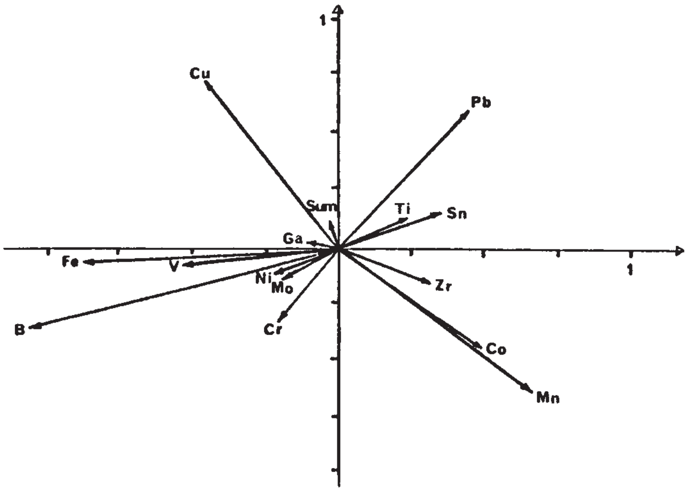


Figure 7.3:

In figure 7.3 the coefficients for the ordinary variables on the two "canonical" variables are given.

By comparing the two figures one can e.g. see that Cu is fairly specific for Devon, and overall the figures give quite a good impression of what the distribution of elements is for the different periods. ♦

# Chapter 8

# Principal components canonical variables and correlations and factor analysis

In this chapter we will give a first overview of some of the methods which can be used to show the underlying structure in a multidimensional data material.

Principal components simply correspond to the results of an eigenvalue analysis of the variance covariance matrix for a multi-dimensional stochastic variable. The method has its origin from around the turn of the century (Karl Pearson), but it was not until the thirties it got its precise formulation by Harold Hotelling.

Factor analysis was originally developed by psychologists - Spearman (1904) and Thurstone at the beginning of the previous century. Because of this the terminology has unfortunately largely been determined by the terminology of the psychologists. Around 1940 Lawley developed the maximum likelihood solutions to the problems in factor analysis - developments which later have been refined by Jöreskog and who in this period introduced factor analysis as a “statistical method”.

The canonical variables and correlations also date back to Harold Hotelling. The concept resembles principal components a lot, however, we are now considering at the correlation between two variables instead of just transforming a single one.

## 8.1 Principal components

### 8.1.1 Definition and simple characteristics

We consider a multi-dimensional, stochastic variable

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix},$$

which has the variance-covariance (dispersion-) matrix

$$\mathrm{D}(X) = \boldsymbol{\Sigma},$$

and without loss of generality we can assume it has the mean value $\mathbf{0}$.

We will sort the eigenvalues in $\boldsymbol{\Sigma}$ descending order and will denote them

$$\lambda_1 \geq \cdots \geq \lambda_k.$$

The corresponding orthonormal eigenvectors are denoted

$$\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k,$$

and we define the orthogonal matrix $\mathbf{P}$ by

$$\mathbf{P} = (\boldsymbol{p}_1 \cdots \boldsymbol{p}_k).$$

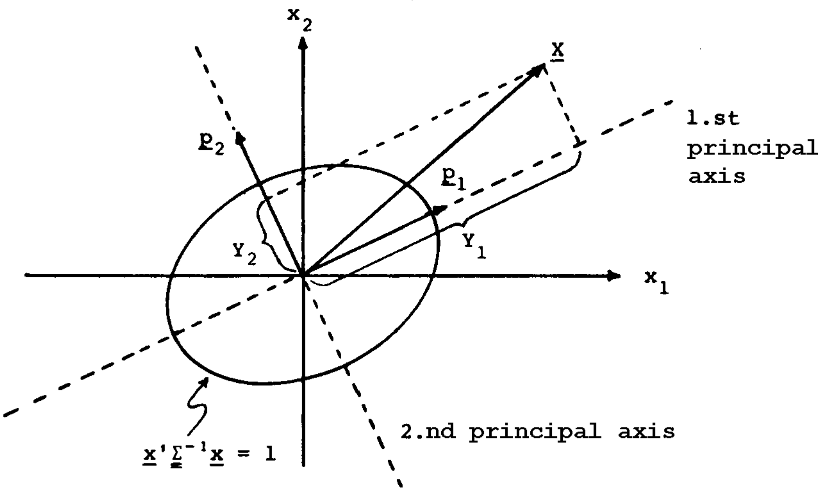
We then have the following

**DEFINITION 8.1.** By the $\boldsymbol{i}$'th principal axis of $\boldsymbol{X}$ we mean the direction of the eigenvector $\boldsymbol{p}_i$ corresponding to the $i$'th largest eigenvalue. ▲

**DEFINITION 8.2.** By the $i$'th principal component of $\boldsymbol{X}$ we will understand $\boldsymbol{X}$'s projection $Y_i = \boldsymbol{p}_i'\boldsymbol{X}$ on the $i$'th principal axis.

The vector

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} = \mathbf{P}'\boldsymbol{X}$$



is called the **vector of principal components**.

The situation has been sketched geometrically in the figure above where we have drawn the unit ellipsoid corresponding to the variance-covariance structure i.e. the ellipsoid with the equation

$$\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = 1.$$

It is seen that the principal axes are the main axes in this ellipsoid. ▲

A number of theorems hold about the characteristics of the principal components. Most of these theorems are statistical reformulations of a number of the results corresponding to symmetrical positive semidefinite matrices which are given in chapter 1.

**Theorem 8.1.** The principal components are uncorrelated and the variance of the $i$'th component is $\lambda_i$ i.e. the $i$'th largest eigenvalue.

▲

**Proof 8.1.** From the theorems 2.5 and 1.10 we have

$$\mathrm{D}(\boldsymbol{Y}) = \mathrm{D}(\mathbf{P}'\boldsymbol{X}) = \mathbf{P}'\boldsymbol{\Sigma}\,\mathbf{P} = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix},$$

and the result follows readily. ■

Further we have

**THEOREM 8.2.** The generalised variance of the principal components is equal to the generalised variance of the original observations. ▲

**PROOF 8.2.** From the definition p. 105 we have

$$\mathrm{GV}(\boldsymbol{X}) = \det \boldsymbol{\Sigma}$$

and

$$\mathrm{GV}(\boldsymbol{Y}) = \det \boldsymbol{\Lambda} = \lambda_1 \cdots \lambda_k,$$

■

A similar result is the following

**THEOREM 8.3.** The total variance i.e. the sum of variance of the original variables is equal to the sum of the variance of the principal components i.e.

$$\sum_i \mathrm{V}(X_i) = \sum_i \mathrm{V}(Y_i)$$

▲

**PROOF 8.3.** Since

$$\sum \mathrm{V}(X_i) = \mathrm{tr}\, \boldsymbol{\Sigma}$$

and

$$\sum \mathrm{V}(Y_i) = \mathrm{tr}\, \boldsymbol{\Lambda}$$

the result follows from the note above. ■

Finally we have

**THEOREM 8.4.** The first principal component is the linear combination (with normed coefficients) of the original variables which has the largest variance. The $m$'th principal components is the linear combination (with normed coefficients) of the original variables which is uncorrelated with the first $m-1$ principal components and then has the largest variance. Formally expressed:

$$\sup_{\|b\|=1} \mathrm{V}(\boldsymbol{b}'\boldsymbol{X}) = \lambda_1,$$

and the supremum is given when $\boldsymbol{b} = \boldsymbol{p}_1$. Further we have

$$\sup_{\substack{\boldsymbol{b} \perp \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{m-1} \\ \|\boldsymbol{b}\| = 1}} \mathrm{V}(\boldsymbol{b}'\boldsymbol{X}) = \lambda_m,$$

and the supremum is given by $\boldsymbol{b} = \boldsymbol{p}_m$ ▲

**PROOF 8.4.** Since

$$\mathrm{V}(\boldsymbol{b}'\boldsymbol{X}) = \boldsymbol{b}'\boldsymbol{\Sigma}\,\boldsymbol{b},$$

and

$$\begin{aligned}\mathrm{Cov}(Y_i, \boldsymbol{b}'\boldsymbol{X}) &= \mathrm{Cov}(\boldsymbol{p}_i'\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{X}) = \boldsymbol{p}_i'\boldsymbol{\Sigma}\,\boldsymbol{b} \\ &= \lambda_i \boldsymbol{p}_i'\boldsymbol{b},\end{aligned}$$

so that

$$\mathrm{Cov}(Y_i, \boldsymbol{b}'\boldsymbol{X}) = 0 \Leftrightarrow \boldsymbol{p}_i \perp \boldsymbol{b},$$

the theorem is just a reformulation of theorem 1.15 p. 36. ■

**REMARK 8.1.** From the theorem we have that if we seek the linear combination of the original variables which explains most of the variation in these, then the first principal component is the solution. If we seek the $m$ variables which explain most of the original variation, then the solution is the $m$ first principal components. A measure of how well these describe the original variation is found by means of theorems 8.1 and 8.3 which show that the $m$ first principal components describe the fraction

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_m + \cdots + \lambda_k}$$

of the original variation.

A better and more qualified measure of how good the "recreation ability": is found by trying to reconstruct the original $\boldsymbol{X}$ from the vector

$$\boldsymbol{Y}^* = (Y_1, \ldots, Y_m, 0, \ldots, 0)'.$$

Since

$$\boldsymbol{Y} = \mathbf{P}'\boldsymbol{X} \Leftrightarrow \boldsymbol{X} = \mathbf{P}\,\boldsymbol{Y},$$

It is tempting to try with

$$\boldsymbol{X}^* = \mathbf{P}\,\boldsymbol{Y}^*\ .$$

We find

$$\begin{aligned}
\mathrm{D}(\boldsymbol{X}^*) &= \mathbf{P}\,\mathrm{D}(\boldsymbol{Y}^*)\mathbf{P}' \\
&= (\boldsymbol{p}_1 \cdots \boldsymbol{p}_k) \begin{pmatrix} \lambda_1 & & \cdots & & 0 \\ & \ddots & & & \\ \vdots & & \lambda_m & & \vdots \\ & & & \ddots & \\ 0 & & \cdots & & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{p}_1' \\ \vdots \\ \boldsymbol{p}_k' \end{pmatrix} \\
&= \lambda_1 \boldsymbol{p}_1 \boldsymbol{p}_1' + \cdots + \lambda_m \boldsymbol{p}_m \boldsymbol{p}_m'.
\end{aligned}$$

The spectral decomposition of $\boldsymbol{\Sigma}$ is (p. 31)

$$\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{p}_1 \boldsymbol{p}_1' + \cdots + \lambda_m \boldsymbol{p}_m \boldsymbol{p}_m' + \lambda_{m+1} \boldsymbol{p}_{m+1} \boldsymbol{p}_{m+1}' + \cdots + \lambda_k \boldsymbol{p}_k \boldsymbol{p}_k',$$

which means that

$$\boldsymbol{\Sigma} - \mathrm{D}(\boldsymbol{X}^*) = \lambda_{m+1} \boldsymbol{p}_{m+1} \boldsymbol{p}_{m+1}' + \cdots + \lambda_k \boldsymbol{p}_k \boldsymbol{p}_k'.$$

If there is a large difference between the eigenvalues then the smallest ones will be negligible and the difference between the original variance-covariance matrix and the one "reconstructed" from the first $m$ principal components is therefore small. ▼

### 8.1.2 Estimation and Testing

If the variance covariance matrix is unknown but is estimated on the basis of $n$ observations, then one estimates the principal components and their variances simply by using the estimated variance covariance matrix as if it were known. If all the eigenvalues in $\mathbf{\Sigma}$ are different it can be shown that the eigenvalue and eigenvectors we get in this way are maximum likelihood estimates of the true parameters (see e.g. [2]).

There is, however, a very common problem here since it can be shown that the principal components are dependent of the scales of measurements our original variables have been measured in. Therefore one often chooses only to consider the normed (standardised) variables i.e.

$$Y_{\ell\, i} = \frac{X_{\ell i} - \bar{X}_\ell}{\sqrt{\sum_i (\bar{X}_{\ell\, i} - \bar{X}_\ell)^2/(n-1)}},$$

where

$$\boldsymbol{X}_i = \begin{pmatrix} X_{1\,i} \\ \vdots \\ X_{k\,i} \end{pmatrix}, \qquad i = 1, \ldots, n.$$

This transformation corresponds to analysing the empirical correlation matrix instead of analysing the empirical variance covariance matrix.

If one decides to use only some of the principal components in the further analysis one could e.g. choose a strategy such as to retain as many of the components needed to account for at least e.g. $90\%$ of the total variation.

Another criterion would be to test a hypothesis like

$$H_0 : \lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_k$$

against the alternative that we have a distinct "greater than" ($>$) among the $k-m$ last eigenvalues.

If we are using the estimated variance covariance matrix $\hat{\mathbf{\Sigma}}$, the test statistic becomes

$$Z_1 = -n' \ln \frac{\det \hat{\mathbf{\Sigma}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n' \ln \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}^{k-m}},$$

where

$$n' = n - m - \frac{1}{6}(2(k-m) + 1 + \frac{2}{k-m}),$$

and

$$\hat{\lambda} = (\operatorname{tr} \hat{\mathbf{\Sigma}} - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k)/(k-m).$$

The critical region using a test at level $\alpha$ is approximately

$$\{(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) | z_1 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

If we instead are using the estimated **correlation matrix $\hat{\mathbf{R}}$** we get the criterion

$$Z_2 = -n \ln \frac{\det \hat{\mathbf{R}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n \ln \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}^{k-m}},$$

where

$$\hat{\lambda} = (k - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k)/(k-m).$$

The critical region for a test at level $\alpha$ becomes approximately equal to

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | z_2 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

However, it should be noted that this approximation is far worse than the corresponding approximation for the variance covariance matrix.

A discussion of the above mentioned tests can be found in [15].

We now give an example.

**EXAMPLE 8.1.** The example is based on an example from [6] p. 486. The background material is measurements of seven variables on 25 boxes with randomly generated sides. The seven variables are

| | |
|---|---|
| $X_1$ : | longest side |
| $X_2$ : | second longest side |
| $X_3$ : | smallest side |
| $X_4$ : | longest diagonal |
| $X_5$ : | radius in the circumscribed sphere divided by radius in the inscribed sphere |
| $X_6$ : | longest side + second longest side)/shortest side |
| $X_7$ : | surface area/volume. |

In the following table we have shown some of the observations of the seven variables.

| Box | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|
| 1 | 3.760 | 3.660 | 0.540 | 5.275 | 9.768 | 13.741 | 4.782 |
| 2 | 8.590 | 4.990 | 1.340 | 10.022 | 7.500 | 10.162 | 2.130 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 24 | 8.210 | 3.080 | 2.420 | 9.097 | 3.753 | 4.657 | 1.719 |
| 25 | 9.410 | 6.440 | 5.110 | 12.495 | 2.446 | 3.103 | 0.914 |

We will now consider the question: Which things about a box determine how we perceive its size?

In order to answer this question we will perform a principal component analysis of the above mentioned data. By such an analysis we hope to find out if the above mentioned 7 variables, which all in one way or another are related to "size" or "form" vary freely in the 7 dimensional space or if they are more or less concentrated in some subspaces.

We first give the empirical-variance covariance matrix for the variables. It is

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} 5.400 & 3.260 & 0.779 & 6.391 & 2.155 & 3.035 & -1.996 \\ 3.260 & 5.846 & 1.465 & 6.083 & 1.312 & 2.877 & -2.370 \\ 0.779 & 1.465 & 2.774 & 2.204 & -3.839 & -5.167 & -1.740 \\ 6.391 & 6.083 & 2.204 & 9.107 & 1.610 & 2.782 & -3.283 \\ 2.155 & 1.312 & -3.839 & 1.610 & 10.710 & 14.770 & 2.252 \\ 3.035 & 2.877 & -5.167 & 2.782 & 14.770 & 20.780 & 2.622 \\ -1.996 & -2.370 & -1.740 & -3.283 & 2.252 & 2.622 & 2.594 \end{bmatrix}$$

Then we determine the eigenvectors and eigenvalues for $\hat{\mathbf{\Sigma}}$. The eigenvectors are given in descending order together with the fraction and the cumulated fraction of the total variance that the eigenvalues contribute:

| Eigenvalue $\hat{\lambda}_i, i = 1, \cdots, 7$ | Percentage of total variance | Cumulated percentage of total variance |
|---|---|---|
| 34.490 | 60.290 | 60.290 |
| 19.000 | 33.210 | 93.500 |
| 2.540 | 4.440 | 97.940 |
| 0.810 | 1.410 | 99.350 |
| 0.340 | 0.600 | 99.950 |
| 0.033 | 0.060 | 100.010 |
| 0.003 | 0.004 | 100.014 |

Computational errors in the determination of the eigenvalues lead to deviations like the cumulated sum being more than $100\%$.

The corresponding coordinates of the eigenvectors are shown in the following table.

| Variable | $\hat{\boldsymbol{p}}_1$ | $\hat{\boldsymbol{p}}_2$ | $\hat{\boldsymbol{p}}_3$ | $\hat{\boldsymbol{p}}_4$ | $\hat{\boldsymbol{p}}_5$ | $\hat{\boldsymbol{p}}_6$ | $\hat{\boldsymbol{p}}_7$ |
|---|---|---|---|---|---|---|---|
| $X_1$ | 0.164 | 0.422 | 0.645 | −0.090 | 0.225 | 0.415 | −0.385 |
| $X_2$ | 0.142 | 0.447 | −0.713 | −0.050 | 0.395 | 0.066 | −0.329 |
| $X_3$ | −0.173 | 0.257 | −0.130 | 0.629 | −0.607 | 0.280 | −0.211 |
| $X_4$ | 0.170 | 0.650 | 0.146 | 0.212 | 0.033 | −0.403 | 0.565 |
| $X_5$ | 0.546 | −0.135 | 0.105 | 0.165 | −0.161 | −0.596 | −0.513 |
| $X_6$ | 0.768 | −0.133 | −0.149 | −0.062 | −0.207 | 0.465 | 0.327 |
| $X_7$ | 0.073 | −0.313 | 0.065 | 0.719 | 0.596 | 0.107 | 0.092 |

It is seen that the first eigenvector is the direction which corresponds to more than $60\%$ of the total variation, has especially numerically large 5th and 6th coordinates. This means that the first principal component

$$Y_1 = 0.164X_1 + \ldots + 0.546X_5 + 0.768X_6 + 0.073X_7$$

is especially sensitive to variations in $X_5$ and $X_6$. These two variables: The ratio between the radius in the circumscribed sphere and the radius in the inscribed sphere and the ratio between the sum of the two longest sides and the shortest side both have something to do with how "flat" a box is. The larger these two variables, the flatter the box. Therefore, the first principal component measures the difference in "flatness" of the boxes. The second eigenvector has large positive coordinates for the first 4 variables and a fairly large negative coordinate for the last variable. If the second principle component

$$Y_2 = 0.422X_1 + 0.447X_2 + 0.257X_3 + 0.650X_4 + \cdots - 0.313X_7,$$

is large then one or more of the variables $X_1, \ldots, X_4$ must be large while $X_7$ is small. Now we know that a cube is the box which for a given volume has the smallest surface. Therefore we also know that if a box deviates a lot from a cube then it will have a large $X_7$- value, and this corresponds to a very strong reduction of $Y_2$. A large $Y_2$- value therefore indicates that most of the sides are large - and furthermore - more or less equal. We therefore conclude that $Y_2$ measures a more general perception of size.

In the following figure we have depicted the boxes in a coordinate system where the axes are the first two principal axes. The coordinates for a single box then become the values of the first and the second principal component for that specific box.

For the first box we e.g. find

$$\begin{aligned} Y_1 &= 0.164 \cdot 3.760 + \cdots + 0.073 \cdot 4.782 = 18.18 \\ Y_2 &= 0.422 \cdot 3.760 + \cdots - 0.313 \cdot 4.782 = 2.15. \end{aligned}$$

At the coordinate $(18.18, 2.15)$ we have then drawn a picture of box No. 1, etc..

From this graph we also very clearly see the interpretation we have given the principal components. To the left in the graph corresponding to small values of component No.1 we have shown the "fattest" boxes and to the right the "flattest". At the top of the graph corresponding to big values of component No. 2 we have the big boxes and at the bottom we have the small ones.

On the other hand we do not seem to have any precise discrimination between the oblong boxes and the more flat boxes. This discrimination is first seen when we also consider the third principal component. It is

$$Y_3 = 0.645X_1 - 0.713X_2 + \cdots + 0.065X_7.$$

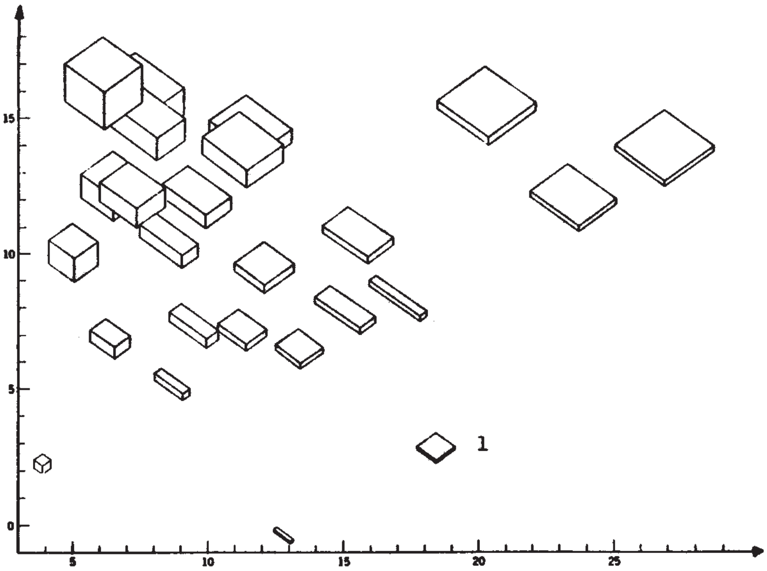


Figure 8.1:

This component puts a large positive weight on variable No. 1 the length of the largest side and a large negative weight on the length of the second largest side. An oblong box will have $X_1 >> X_2$ and therefore $Y_3$ will be relatively large for such a box. If the base of the box corresponding to the two largest sizes is close to a square then $Y_3$ will be close to 0 for the respective box.

The three first principal components then take care of about $98\%$ of the total variation and by means of these we can partition a box's "size characteristics" in three uncorrelated components: one corresponding to the flatness of the box $(Y_1)$, one which corresponds to a more general concept of size $(Y_2)$, and one which corresponds to "the degree of oblong-ness" $(Y_3)$. Now the initial question of: What is "the size of a box" should at least be partly illustrated. ♦

The next example is based on some investigations by Agterberg et al. (see [1] p. 128).

**EXAMPLE 8.2.** The Mount Albert peridotit intrusion is part of the Appalachtic ultramafic belt in the Quebec province. A number of mineral samples were collected and the values of the 4 following variables were determined:

$X_1$ : mol$\%$ forsterit (= Mg-olivin)
$X_2$ : mol$\%$ enstatit (= Mg-ortopyroxen)
$X_3$ : dimension of unit-cell of chrome-spinal
$X_4$ : specific density of mineral sample.

Using between 99 and 156 observations the following correlation matrix between the variables was estimated:

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.00 & 0.32 & 0.41 & -0.31 \\ 0.32 & 1.00 & 0.68 & -0.38 \\ 0.41 & 0.68 & 1.00 & -0.36 \\ -0.31 & -0.38 & -0.36 & 1.00 \end{bmatrix}.$$

It is quite obvious that we should analyse the correlation matrix rather than the variance-covariance matrix. Because we are analysing variables which are measured in non-comparable units we must standardise the numbers.

The eigenvalues and the corresponding eigenvectors are

$$\begin{aligned} \hat{\lambda}_1 &= 2.25; \quad \hat{\boldsymbol{p}}_1 = \begin{bmatrix} 0.43 \\ 0.55 \\ 0.57 \\ -0.44 \end{bmatrix} \\ \hat{\lambda}_2 &= 0.74; \quad \hat{\boldsymbol{p}}_2 = \begin{bmatrix} -0.66 \\ 0.49 \\ 0.37 \\ 0.44 \end{bmatrix} \\ \hat{\lambda}_3 &= 0.70; \quad \hat{\boldsymbol{p}}_3 = \begin{bmatrix} 0.60 \\ -0.02 \\ 0.16 \\ 0.78 \end{bmatrix} \\ \hat{\lambda}_4 &= 0.31; \quad \hat{\boldsymbol{p}}_4 = \begin{bmatrix} -0.14 \\ -0.68 \\ 0.72 \\ -0.06 \end{bmatrix} \end{aligned}$$

All the eigenvectors have fairly large coordinates in most places so there does not seem to be any obvious possibility of giving an intuitive interpretation of the principal components.

The first principal component corresponds to 2.25/4 = 56.25% of the total variation.

It would be interesting to know if the three smallest eigenvectors of the correlation matrix can be considered as being of the same magnitude.

The test statistic we will use is

$$Z = -n \ln \frac{0.74 \cdot 0.70 \cdot 0.31}{[(0.74 + 0.70 + 0.31)/3]^3} = 0.2120n,$$

where $n$ is the number of observations on which we have based the correlation matrix on. Since this number is not the same for all the different correlation coefficients the theoretical background for the test disappears so to speak. However, if we disregard that problem, then the number of degrees of freedom in the $\chi^2$-distribution with which to compare the test statistic becomes

$$f = \frac{1}{2}(4-1+2)(4-1-1) = 5.$$

Since

$$\chi^2(5)_{0.995} = 16.7,$$

and since $0.21n$ for $n$ approximately equal to 100 is quite a lot larger than this value it would be reasonable to conclude that the three smallest eigenvectors in the (true) correlation matrix are not of the same order of magnitude. ♦

## 8.2 Canonical variables and canonical correlations

In the following we will discuss dependency between groups of variables where we in the last section only looked at dependency (correlation structure) between single variables.

We consider a stochastic variable $\boldsymbol{X}$

$$\boldsymbol{X} \in \mathrm{N}_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $p \leq q$ and $\boldsymbol{X}$ and the parameters have been partitioned as follows:

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

If we on the basis of $n$ observations of $\boldsymbol{X}$ wish to investigate if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent this could be done as shown in chapter 6 by investigating

$$\frac{\det(\mathbf{S})}{\det(\mathbf{S}_{11})\det(\mathbf{S}_{22})},$$

which is $U_{p,q,n-l-q}$ distributed for $H_0$. We will now try to consider the problem for another point of view. We will consider two one-dimensional variables $U$ and $V$ given by

$$U = \boldsymbol{a}'\boldsymbol{X}_1 \quad \text{and} \quad V = \boldsymbol{b}'\boldsymbol{X}_2.$$

Then we have

$$D\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}' \\ \boldsymbol{b}' \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{a},\boldsymbol{b}) = \begin{bmatrix} \boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a} & \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b} \\ \boldsymbol{b}'\boldsymbol{\Sigma}_{21}\boldsymbol{a} & \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b} \end{bmatrix},$$

and the correlation between $U$ and $V$ is

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

Now we have

$$\boldsymbol{\Sigma}_{12} = 0 \Leftrightarrow \forall \boldsymbol{a},\boldsymbol{b}: \quad \rho(\boldsymbol{a},\boldsymbol{b}) = 0.$$

The accept region for the hypothesis $\rho(\boldsymbol{a},\boldsymbol{b}) = 0$ is of the form (cf. chapter 2)

$$r^2(\boldsymbol{a},\boldsymbol{b}) \le r_\beta^2,$$

where $r(\boldsymbol{a},\boldsymbol{b})$ is the empirical correlation coefficient and $r_\beta^2$ is a suitable quantile in the distribution of the 0 hypothesis. We therefore have an accept of $\boldsymbol{\Sigma}_{12} = 0$ if

$$\forall \boldsymbol{a},\boldsymbol{b}: \quad r^2(\boldsymbol{a},\boldsymbol{b}) \le r_\beta^2,$$

which is obviously equivalent to

$$\max_{\boldsymbol{a},\boldsymbol{b}} r^2(\boldsymbol{a},\boldsymbol{b}) \le r_\beta^2.$$

We now have that the 2 groups are independent if the maximal (empirical) correlation coefficient between a linear combination of the first group and a linear combination from the second group is suitable small. This maximum correlation coefficient is called the first (empirical) canonical correlation coefficient and the corresponding variables the first (empirical) canonical variables.

It is now obvious as in the case of the principal components can continue the definition. We can define the second canonical correlation coefficient as the maximum correlation between the linear combination of $\boldsymbol{X}_1$'s and $\boldsymbol{X}_2$'s so that these combinations are independent of the previous ones etc.. Formerly we have

**DEFINITION 8.3.** Let $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$ be a stochastic variable where $\boldsymbol{X}_1$ has $p$ components and $\boldsymbol{X}_2$ $q$ components $(p \le q)$. **The $\boldsymbol{r}$'th pair of canonical variables** is the pair of linear combinations linearkombinationer $U_r = \boldsymbol{\alpha}_r'\boldsymbol{X}_1$ and $V_r = \boldsymbol{\beta}_r'\boldsymbol{X}_2$ which each has the variance 1 and which are uncorrelated with the previous $r-1$ pairs of canonical variables and which have maximum correlation. The correlation is the $\boldsymbol{r}$'th canonical correlation. ▲

Now we have the problem of determining the canonical variables and correlations. We have the following theorem:

**THEOREM 8.5.** Let the situation be given in the above mentioned definition and let lad $\mathrm{D}(X) = \mathbf{\Sigma}$ be partitioned analogously

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

Then the $r$'th canonical correlation is equal to the $r$'th largest root $\lambda_r$ of

$$\det \begin{pmatrix} -\lambda\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & -\lambda\mathbf{\Sigma}_{22} \end{pmatrix} = 0,$$

and the coefficients in the $r$'th pair of canonical variables satisfies

(i) $\begin{pmatrix} -\lambda\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & -\lambda\mathbf{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_r \\ \boldsymbol{\beta}_r \end{pmatrix} = \mathbf{0}$

(ii) $\boldsymbol{\alpha}_r' \mathbf{\Sigma}_{11} \boldsymbol{\alpha}_r = 1$

(iii) $\boldsymbol{\beta}_r' \mathbf{\Sigma}_{22} \boldsymbol{\beta}_r = 1.$

▲

**PROOF 8.5.** We are talking of a maximisation problem with restrictions and one can solve the problem by using a Lagrange multiplier technique see e.g. [2]p. 289. ■

One can also determine the correlations and the coefficients by solving an eigenvalue problem since we have

**THEOREM 8.6.** Let the situation be as in the previous theorem then we have

$$\begin{aligned} (\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} - \lambda_r^2\mathbf{\Sigma}_{11})\boldsymbol{\alpha}_r &= \mathbf{0} \\ \det(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} - \lambda_r^2\mathbf{\Sigma}_{11}) &= 0 \end{aligned}$$

respectively

$$\begin{aligned} (\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} - \lambda_r^2\mathbf{\Sigma}_{22})\boldsymbol{\beta}_r &= \mathbf{0} \\ \det(\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} - \lambda_r^2\mathbf{\Sigma}_{22}) &= 0 \end{aligned}$$

▲

**PROOF 8.6.** Omitted see e.g. [2]. ■

Corresponding to the estimation we have nothing special to add to the previous. If we insert the maximum likelihood estimates for $\mathbf{\Sigma}$ in the previous theorems we get the maximum likelihood estimates of the parameters. Most often one will probably insert the unbiased estimate $\mathbf{S}$ and one then gets what one can call the empirical values (English: Sample values) for the parameters involved.

In most kinds of canned software there exists programmes for the evaluation of canonical correlations and variables. E.g. we can mention BMDP6M: Canonical Correlation Analysis from the BMDP-package.

## 8.3 Factor analysis

Once again we will consider the analysis of the correlation structure for a single multidimensional variable but contrary to the case in the section on principal components we here assume an underlying model of the structure.

### 8.3.1 Model and assumptions

It is assumed that we have an observation

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix},$$

which - considering the situation historically - can be thought of as a single person's scores in e.g. $k$ different types of intelligence tests or the reactions of a person to $k$ different stimuli.

One then has a model for how one thinks that these reactions (scores) depend on some underlying factors or more specifically that

$$\boldsymbol{X} = \mathbf{A}\,\boldsymbol{F} + \boldsymbol{G},$$

or in more detail

$$\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} G_1 \\ \vdots \\ G_k \end{bmatrix}.$$

Here we call $\boldsymbol{F}$ the vector of **common** factors, they are also called **factor scores**. These are not observable. Examples of these are characteristics like three dimensional intelligence, verbal intelligence etc.

The elements of the $\mathbf{A}$ matrix are called factor loadings and they give the weights of how the single factors enter the description of the different variables. If one e.g. assumes that $\mathbf{A}$ describes 3-dimensional intelligence and verbal intelligence and that $F_1$ is the result of a test of a 3-dimensional kind and $F_m$ the result of a reading test, well then one will obviously have that $X_1$ is large and $X_k$ is small and vice-versa that $a_{k1}$ is small and $a_{km}$ is large corresponding to the 3-dimensional intelligence being deterministic of a person's scores in the solving of 3-dimensional problems and analogously for the verbal intelligence.

The vector $\boldsymbol{G}$ is called the vector of unique factors and can be thought of as composed of some specific factors i.e. factors which are special for these specific tests and of errors i.e. non-describable deviations. Obviously these factors are not observable either.

Here we must stipulate that both $\boldsymbol{X}$ and $\boldsymbol{F}$ and $\boldsymbol{G}$ are assumed to be stochastic. Therefore we are not considering a general linear model with the parameters $F_1, \ldots, F_m$.

In order to make this difference quite clear we therefore give the model in the case where we have several observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. We then have the $n$ models

$$\begin{bmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \begin{bmatrix} F_{1i} \\ \vdots \\ F_{mi} \end{bmatrix} + \begin{bmatrix} G_{1i} \\ \vdots \\ G_{ki} \end{bmatrix},$$

Here we note that $\boldsymbol{F}_i$ and $\boldsymbol{G}_i$ change value when the observations $\boldsymbol{X}_i$ change value. We can aggregate the above models into

$$\begin{bmatrix} X_{11} \cdots X_{1n} \\ \vdots \quad\quad \vdots \\ X_{k1} \cdots X_{kn} \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \quad\quad \vdots \\ a_{k1} \cdots a_{km} \end{bmatrix} \begin{bmatrix} F_{11} \cdots F_{1n} \\ \vdots \quad\quad \vdots \\ F_{m1} \cdots F_{mn} \end{bmatrix} + \begin{bmatrix} G_{11} \cdots G_{1n} \\ \vdots \quad\quad \vdots \\ G_{k1} \cdots G_{kn} \end{bmatrix}.$$

It is assumed that $\boldsymbol{F}$ and $\boldsymbol{G}$ are uncorrelated and that

$$\mathrm{D}(\boldsymbol{F}) = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = \mathbf{I} = \mathbf{I}_m,$$

and

$$\mathrm{D}(\boldsymbol{G}) = \begin{pmatrix} \delta_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \delta_k \end{pmatrix} = \boldsymbol{\Delta}.$$

Furthermore, we assume that the observations are standardised in such a way that $\mathrm{V}(X_i) = 1$, $\forall i$ i.e. that the variance-covariance matrix for $\boldsymbol{X}$ is equal to its correlation matrix which is denoted

$$\mathrm{D}(\boldsymbol{X}) = \mathbf{R} = \begin{pmatrix} 1 & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & 1 \end{pmatrix}.$$

From the original factor equation we find by means of theorem 2.5 p. 60, that

$$\mathbf{R} = \mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta}.$$

From this we especially find that for $j = 1, \ldots, k$ we have

$$\mathrm{V}(X_j) = a_{j1}^2 + \cdots + a_{jm}^2 + \delta_j = 1.$$

Here we introduce the notation

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2, \qquad j = 1, \ldots, k.$$

These quantities are called **communalities** and $h_j^2$ describes how large a proportion of $X_j$'s variance is due to the $m$ common factors. Correspondingly $\delta_j$ gives the uniqueness in $X_j$'s variance. I.e. the proportion of $X_j$'s variance which is not due to the $m$ common factors.

Finally the $(i, j)$'th factor weight gives the correlation between the $i$'th variable and the $j$'th factor i.e.

$$\mathrm{Cov}(X_i, F_j) = \mathrm{Cov}(\sum_{\nu} a_{i\nu} F_\nu + G_i, F_j) = a_{ij}.$$

It can be shown [7] that

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2 \geq r_{j|1\ldots k}^2,$$

i.e. that the $j$'th communality is always larger than or equal to the square of the multiple correlation coefficient between $X_j$ and the rest of the variables. This is not strange when remembering that this quantity exactly equals the proportion of $X_j$'s variance which is described by the variance in the other $X_i$'s.

We now turn to the more basic problem of estimating the factors. What we are interested in determining is $\mathbf{A}$. We find

$$\mathbf{A}\,\mathbf{A}' = \mathbf{R} - \boldsymbol{\Delta}.$$

The diagonal elements in this matrix are

$$1 - \delta_j = h_j^2, \qquad j = 1, \ldots, k.$$

We do not know these but we could estimate them e.g. by inserting the squares of the multiple correlation coefficient. If we insert these we get a matrix

$$\mathbf{V} = \begin{bmatrix} r^2_{1|2\cdots k} & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & r^2_{k|1\cdots k-1} \end{bmatrix},$$

in which the elements outside the diagonal are equal to the original correlation matrix $\mathbf{R}$'s elements. This matrix is still symmetric but not necessarily positive semidefinite. However, since it is still an estimate of one, we will (silently) assume that it still is positive semidefinite.

Independently of how the communalities have been estimated the resulting "correlation matrix" is called $\mathbf{V}$. $\mathbf{V}$ could e.g. be the above mentioned.

We will call the eigenvalues of $\mathbf{V}$ and the corresponding normed orthogonal eigenvectors respectively

$$\lambda_1 \geq \cdots \geq \lambda_k,$$

and

$$\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k.$$

If we let

$$\mathbf{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k),$$

we then have from theorem 1.10 p. 30, that

$$\mathbf{P}'\mathbf{V}\,\mathbf{P} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}.$$

Since $\mathbf{P}$ is orthogonal (as a consequence of being orthonormal) we get

$$\mathbf{V} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}' = (\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})',$$

where

$$\mathbf{\Lambda}^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda}_k \end{pmatrix}.$$

We now define

$$\mathbf{\Lambda}_*^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ & & \vdots \\ \vdots & & \sqrt{\lambda}_m \\ & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

I.e. $\mathbf{\Lambda}_*^{\frac{1}{2}}$ consists of the first $m$ columns in i $\mathbf{\Lambda}^{\frac{1}{2}}$ corresponding to the $m$ largest eigenvalues. We then see that

$$\begin{aligned} (\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})' &= \mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}}\mathbf{\Lambda}_*^{\frac{1}{2}\prime}\mathbf{P}' \\ &= \mathbf{P}\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \lambda_m & \vdots \\ 0 & \cdots & 0 \end{pmatrix}\mathbf{P}' \\ &\simeq \mathbf{V}, \end{aligned}$$

cf. the analogous considerations p. 264.

Since $\mathbf{V}$ is an estimate of $\mathbf{A}\,\mathbf{A}'$, we then have

$$\mathbf{A}\,\mathbf{A}' \simeq (\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})',$$

so it would be natural to choose $\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}}$ as an estimate of $\mathbf{A}$. This solution is called the principle factor solution for our estimation problem.

We will gather our considerations in the following

**THEOREM 8.7.** We consider the factor model $\boldsymbol{X} = \mathbf{A}\,\boldsymbol{F}+\boldsymbol{G}$ where $\boldsymbol{X}$ is $k$-dimensional and $\boldsymbol{F}$ $m$-dimensional. The correlation matrix of $\boldsymbol{X}$ is denoted $\mathbf{R}$, and $\mathbf{V}$ is the matrix which we find by substituting the ones in the diagonal of $\mathbf{R}$ with estimates of the communalities. These should be chosen in the interval $[r^2, 1]$ where $r^2$ is the multiple correlation coefficient between the relevant variable and the rest of the variables. Usually one chooses either $r^2$ or 1. The principle factor solution to the estimation problem is then

$$\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}} = (\sqrt{\lambda_1}\boldsymbol{p}_1, \ldots, \sqrt{\lambda_m}\boldsymbol{p}_m),$$

where $\lambda_i$, $i = 1, \ldots, m$ are the $m$ largest eigenvalues of $\mathbf{V}$ and where $\boldsymbol{p}_i$, $i = 1, \ldots, m$ are the corresponding normed eigenvectors. ▲

**REMARK 8.2.** In the theorem we assume that the number of factors $m$ is known. If this is not the case it is common to retain those which correspond to eigenvalues larger than 1. Other authors recommend that one retains one, two or three because that will usually be the upper limit to how many factors one can give a reasonable interpretation.

▼

### 8.3.2 Factor rotation

Once again we consider the expression

$$\mathbf{A}\,\mathbf{A}' \simeq (\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}})'$$

If $\mathbf{Q}$ is an arbitrary $m \times m$ orthonormal matrix i.e. $\mathbf{Q}\,\mathbf{Q}' = \mathbf{I}$ then we have

$$\begin{aligned}(\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}}\mathbf{Q})(\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}}\mathbf{Q})' &= (\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}})\mathbf{Q}\,\mathbf{Q}'(\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}})' \\ &= (\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\boldsymbol{\Lambda}_*^{\frac{1}{2}})' \\ &= \mathbf{A}\,\mathbf{A}'.\end{aligned}$$

This means that we can have as many estimates of the $\mathbf{A}$-matrix as we want by multiplying the principle factor solution by an orthonormal matrix.

The problem is then how to choose the $\mathbf{Q}$-matrix in a reasonable way. The main principle is that one wants the $\mathbf{A}$-matrix to become "simple" (without explaining what this means).

One of the most often used criterions is the one introduced by Kaiser, the Varimax criterion. It says that we must choose $\mathbf{Q}$ in such a way that the quantity

$$\sum_j m \left\{ \sum_i \left( \frac{a_{ij}^2}{h_i^2} \right)^2 - \frac{1}{m} \left[ \sum_i \left( \frac{a_{ij}^2}{h_i^2} \right) \right]^2 \right\}$$

is maximised. It is seen that the expression is the empirical variance of the terms $a_{ij}^2/h_i^2$. The maximisation will therefore mean that many of the $a_{ij}$'s become 0 (approximately) and many become large (close to $\pm 1$). This corresponds to a simple structure which will be easy to interpret.

Another rotation principle is the so-called quartimax principle. Here we try to make the rows in the factor matrix simple so that the single variables have a simple relation with the factors.

Contrary to this the Varimax criterion tries to make the columns simple corresponding to easily interpretable factors.

Before we continue with the theory we give an example.

**EXAMPLE 8.3.** We will now try to perform a factor analysis on the data given in example 8.1.

First we determine the correlation matrix. From the estimate of the variance-covariance matrix p. 267 we find

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.000 & 0.580 & 0.201 & 0.911 & 0.283 & 0.287 & -0.533 \\ 0.580 & 1.000 & 0.364 & 0.834 & 0.166 & 0.261 & -0.609 \\ 0.201 & 0.364 & 1.000 & 0.439 & -0.704 & -0.681 & -0.649 \\ 0.911 & 0.834 & 0.439 & 1.000 & 0.163 & 0.202 & -0.676 \\ 0.283 & 0.166 & -0.704 & 0.163 & 1.000 & 0.990 & 0.427 \\ 0.287 & 0.261 & -0.681 & 0.202 & 0.990 & 1.000 & 0.357 \\ -0.533 & -0.609 & -0.649 & -0.676 & 0.427 & 0.357 & 1.000 \end{bmatrix}$$

Completely analogously with the procedure in example 8.1 we then determine the eigenvalues and vectors for $\hat{\mathbf{R}}$ (note that in this case our choice of $\mathbf{V}$ is simply $\hat{\mathbf{R}}$). We find

| Eigenvalue | Percentage of | Cumulated percent- |
|---|---|---|
| $\hat{\lambda}_i, 1, \ldots, 7$ | total variance | age of total variance |
| 3.3946 | 48.495 | 48.495 |
| 2.8055 | 40.078 | 88.573 |
| 0.4373 | 6.247 | 94.820 |
| 0.2779 | 3.971 | 98.791 |
| 0.0810 | 1.157 | 99.948 |
| 0.0034 | 0.049 | 99.996 |
| 0.0003 | 0.004 | 100.000 |

The coordinates of the corresponding eigenvectors are shown in the following table.

| Variable | $\hat{p}_1$ | $\hat{p}_2$ | $\hat{p}_3$ | $\hat{p}_4$ | $\hat{p}_5$ | $\hat{p}_6$ | $\hat{p}_7$ |
|---|---|---|---|---|---|---|---|
| $X_1$ | $0.405$ | $0.293$ | $-0.667$ | $0.089$ | $-0.227$ | $0.410$ | $-0.278$ |
| $X_2$ | $0.432$ | $0.222$ | $0.698$ | $-0.034$ | $-0.437$ | $0.144$ | $-0.254$ |
| $X_3$ | $0.385$ | $-0.356$ | $0.148$ | $0.628$ | $0.512$ | $0.188$ | $-0.108$ |
| $X_4$ | $0.494$ | $0.232$ | $-0.119$ | $0.210$ | $-0.105$ | $-0.588$ | $5.536$ |
| $X_5$ | $-0.128$ | $0.575$ | $0.209$ | $0.111$ | $0.389$ | $-0.423$ | $-0.556$ |
| $X_6$ | $-0.097$ | $0.580$ | $0.174$ | $-0.006$ | $0.355$ | $0.500$ | $0.498$ |
| $X_7$ | $-0.481$ | $0.130$ | $0.018$ | $0.735$ | $-0.455$ | $0.033$ | $0.049$ |

We now assume that the number of factors is 2 (the assumption is not based on any deep consideration of the structure of the problem. The number 2 is chosen because there are only two eigenvalues larger than 1).

From theorem 8.7 the estimated principal factor solution to the problem is $(\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{p}}_1, \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{p}}_2)$, where

$$\begin{pmatrix} \sqrt{\hat{\lambda}_1}\hat{\boldsymbol{p}}'_1 \\ \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{p}}'_2 \end{pmatrix} = \begin{pmatrix} 0.747 & 0.795 & 0.710 & 0.910 & -0.235 & -0.178 & -0.886 \\ 0.491 & 0.373 & -0.596 & 0.389 & 0.963 & 0.971 & 0.218 \end{pmatrix}.$$

E.g. we find

$$\hat{h}_7^2 = (-0.886)^2 + 0.218^2 = 0.833$$

The vector of estimated communalities is

$$\hat{\boldsymbol{h}}^{2\prime} = [\ 0.798 \quad 0.771 \quad 0.860 \quad 0.979 \quad 0.983 \quad 0.976 \quad 0.833\ ],$$

and we see that e.g. the variation in variable 4 (the length of the longest diagonal) is described by the variation of the two factors by a proportion of 97.9%.

On the other hand the quantities $\hat{\delta}_j = 1 - \hat{h}_j^2$ give the uniqueness value i.e. the fraction of the variance of $X_j$'s which is not explained by the two common factors but which is assigned to the $j$'th unique factor (cf. p. 275). We find

$$\boldsymbol{\delta}' = [\ 0.202 \quad 0.229 \quad 0.140 \quad 0.021 \quad 0.017 \quad 0.024 \quad 0.167\ ].$$

A more qualified measure of the ability to describe the variation in the data material of the two factors is found by recomputing the correlation matrix only from the factors.

We therefore compute the so-called residual correlation matrix

$$\hat{\mathbf{Z}} = \hat{\mathbf{R}} - \hat{\mathbf{A}}\hat{\mathbf{A}}',$$

as a more detailed measure of the factors ability to describe the original variability in the material. We get

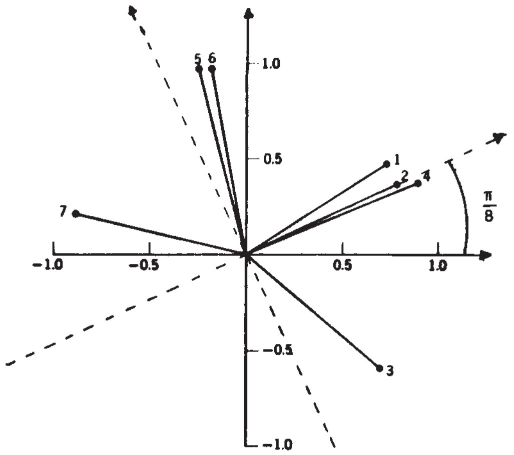
$$\hat{\mathbf{Z}} = \begin{bmatrix} 0.202 & -0.196 & -0.037 & 0.041 & -0.914 & -0.057 & 0.021 \\ -0.196 & 0.229 & 0.071 & -0.035 & -0.006 & 0.041 & 0.015 \\ -0.037 & 0.021 & 0.140 & 0.024 & 0.037 & 0.025 & 0.111 \\ 0.041 & -0.035 & 0.024 & 0.021 & 0.002 & -0.013 & 0.046 \\ -0.014 & -0.006 & 0.037 & 0.002 & 0.017 & 0.012 & 0.009 \\ -0.057 & 0.041 & 0.025 & -0.013 & 0.012 & 0.024 & -0.013 \\ 0.021 & 0.015 & 0.111 & 0.046 & 0.009 & -0.013 & 0.167 \end{bmatrix}.$$

The more $\hat{\mathbf{Z}}$ deviates from the **0**-matrix the poorer the factors describe the original material.

Apart from using the variance-covariance matrix in example 8.1 while we use the correlation matrix here, then the biggest difference in the analysis is that we have multiplied the factors by the square root of the eigenvalues corresponding to each factor. In this way the length of each factor becomes proportional to the proportion of the total variance which it explains.

We will now see if we can obtain factors which are easier to interpret by rotating the factors.

First we depict the factor weights (given on p. 281) $\hat{a}_{ij}$ in a two-dimensional coordinate system. We find



It is noted that most of the variables have large first and second coordinates.

It seems to be possible to obtain a simple structure by rotating the coordinate system about $\frac{\pi}{8}(=22\frac{1}{2}^{\circ})$ anti-clockwise (dashed coordinate system).

This corresponds to multiplication by the matrix

$$\begin{pmatrix} \cos\frac{\pi}{8} & -\sin\frac{\pi}{8} \\ \sin\frac{\pi}{8} & \cos\frac{\pi}{8} \end{pmatrix} = \begin{pmatrix} 0.9239 & -0.3827 \\ 0.3827 & 0.9239 \end{pmatrix},$$

cf. section 1.4.1.

The new factors or rather factor weights then become

$$\begin{bmatrix} 0.747 & 0.491 \\ 0.795 & 0.373 \\ 0.710 & -0.596 \\ 0.910 & 0.389 \\ -0.235 & 0.963 \\ -0.178 & 0.971 \\ -0.886 & 0.218 \end{bmatrix} \begin{bmatrix} 0.9239 & -0.3827 \\ 0.3827 & 0.9239 \end{bmatrix} = \begin{bmatrix} 0.878 & 0.168 \\ 0.877 & 0.040 \\ 0.428 & -0.822 \\ 0.990 & 0.011 \\ 0.151 & 0.980 \\ 0.207 & 0.965 \\ -0.735 & 0.540 \end{bmatrix}.$$

These new factor weights are simpler than the original ones in the sense that we have more values close to $\pm 1$ and close to $0$. Later we will see that this solution found visually is quite close to the Varimax-solution. ♦

Apart from the Varimax-principle there are as mentioned a large number of other methods for orthogonal rotation of factors which are not within the scope of this description. The interested reader is referred to the literature (e.g. [8] and [4]).

There also exists a number of rotation methods which allow relaxation of the assumption of orthogonality. These rotation methods are called "oblique rotations". The philosophy behind these is that the factors are not necessarily independent but may be correlated. Use of these methods demands thorough knowledge of the subject. We again refer to [8] and [4].

### 8.3.3 Computation of the factor scores

If we in the above mentioned example 8.3 wish to make a diagram analogous to the one mentioned on p. 269 then we must compute the factor scores for the single boxes. This is a bit more complicated than it was when we did the principal component analysis. Then we just had to compute the values of the principal components on the different axes. The reason that we cannot just perform the analogue operation is the existence of the specific factors.

We have the model (cf p. 274)

$$\boldsymbol{X} = \mathbf{A}\,\boldsymbol{F} + \boldsymbol{G},$$

where

$$\begin{aligned} \mathrm{D}(\boldsymbol{F}) &= \mathbf{I} \\ \mathrm{D}(\boldsymbol{G}) &= \boldsymbol{\Delta}, \end{aligned}$$

and where $\boldsymbol{F}$ and $\boldsymbol{G}$ are uncorrelated.

Therefore we have

$$\mathrm{D}\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{F} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta} & \mathbf{A} \\ \mathbf{A}' & \mathbf{I} \end{pmatrix}.$$

As previously mentioned, since we have that

$$\mathrm{Cov}(X_i, F_j) = a_{ij},$$

we now have that the matrices outside the diagonal are the $\mathbf{A}$-matrix and its transposed respectively.

The estimate of this variance-covariance matrix is

$$\begin{bmatrix} \hat{\mathbf{A}}\,\hat{\mathbf{A}}' + \hat{\boldsymbol{\Delta}} & \hat{\mathbf{A}} \\ \hat{\mathbf{A}}' & \mathbf{I} \end{bmatrix}.$$

Assuming that the underlying distributions are normal, the conditional distribution of $\mathbf{F}$ given $\boldsymbol{X}$ has the mean value

$$\boldsymbol{\mu}_F + \mathbf{A}'(\mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x)$$

(cf. section 2.2.3).

Since our computations are performed on the standardised x-values it is reasonable to assume that $\boldsymbol{\mu}_x = \mathbf{0}$. The level for the factor scale is arbitrary but it is usually set equal to 0 so that we have the expression

$$\mathbf{A}'(\mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta})^{-1}\boldsymbol{x}$$

for the conditional mean value of $\boldsymbol{F}$.

As an estimate of the $i$'th observation of the factor score of $\boldsymbol{X}_i$ we then have

$$\hat{\boldsymbol{F}}_i = \hat{\mathbf{A}}'(\hat{\mathbf{A}}\,\hat{\mathbf{A}}' + \hat{\boldsymbol{\Delta}})^{-1}\boldsymbol{X}_i. \tag{8.1}$$

Now the $\mathbf{A}$-matrix will often have a large number of rows which means we have to invert a fairly large matrix. This can be circumvented by the following identity

$$(\mathbf{A}\mathbf{A}' + \boldsymbol{\Delta})^{-1}\mathbf{A} = \boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1},$$

which gives

$$\hat{\boldsymbol{F}}_i = (\mathbf{I} + \hat{\mathbf{A}}'\hat{\boldsymbol{\Delta}}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\boldsymbol{\Delta}}^{-1}\boldsymbol{X}_i. \tag{8.2}$$

The validity of the identity is found by the following relationships

$$\begin{array}{lrcl} & (\mathbf{A}\mathbf{A}' + \boldsymbol{\Delta})^{-1}\mathbf{A} & = & \boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1} \\ \Leftrightarrow & \mathbf{A} & = & (\mathbf{A}\mathbf{A}' + \boldsymbol{\Delta})\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1} \\ & & = & \mathbf{A}(\mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A} + \mathbf{I})(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1}, \end{array}$$

and the last relationship is trivially fulfilled.

Now $\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A}$ is an $m \times m$ matrix where $m$ is the number of factors i.e. often not more than 2-3-4 so the inversion problem is not overwhelming. On the other hand as mentioned $(\mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta})$ is a $k \times k$ matrix where $k$ is the number of variables i.e. often far larger than $m$.

If $k$ is only of moderate size we can use the first expression for

$F_i$ directly. Here one should utilise that

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \boldsymbol{\Delta}$$

(cf. p. 276). This gives the expression which is equivalent to (8.1)

$$\hat{\boldsymbol{F}}_i = \hat{\mathbf{A}}'\hat{\mathbf{R}}^{-1}\boldsymbol{X}_i \tag{8.3}$$

Finally we must stipulate that there are a number of other methods of determining the factor scores see e.g. [8] or [16]. It must also be noted that the problem is treated rather weakly in the main part of the literature. The main reason is probably that this problem does not have great interest for psychologists and sociologists who for many years have been the main users of factor analysis. Howeverm in a number of technical/natural science (and sociological) uses one is often interested in classifying single measurements by the size of the factor scores. We will see a use of this in section 8.3.4.

We will now illustrate the computation of factor scores on our box example.

**EXAMPLE 8.4.** In example 8.3, p. 280 we found a rotated factor solution with two factors. The rotated factor weights were

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.878 & 0.168 \\ 0.877 & 0.040 \\ 0.428 & -0.828 \\ 0.990 & 0.011 \\ 0.151 & 0.980 \\ 0.207 & 0.965 \\ -0.735 & 0.540 \end{bmatrix}.$$

In order to determine the factor scores for the single boxes we must first find the communalities and the uniqueness values. We find

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\hat{h}_j^2$ | 0.7991 | 0.7707 | 0.8589 | 0.9802 | 0.9832 | 0.9741 | 0.8318 |
| $\hat{\delta}_j$ | 0.2009 | 0.2293 | 0.1411 | 0.0198 | 0.0168 | 0.0259 | 0.1682 |
| $1/\hat{\delta}_j$ | 4.9776 | 4.3611 | 7.0872 | 50.5051 | 59.5238 | 38.6100 | 5.9453 |

Here we have (cf. p. 276)

$$\hat{h}_j^2 = \hat{a}_{j1}^2 + \hat{a}_{j2}^2 = 1 - \hat{\delta}_j.$$

We note that the given communalities are equal to those we found on p. 281 for the unrotated factors. This always holds and can be used as a check in the computation of the rotated factors.

Since we have

$$\hat{\mathbf{\Delta}} = \text{diag}(\hat{\delta}_j),$$

i.e.

$$\hat{\mathbf{\Delta}}^{-1} = \text{diag}(\frac{1}{\hat{\delta}_j}),$$

we then have

$$(\mathbf{I} + \hat{\mathbf{A}}'\hat{\mathbf{\Delta}}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\mathbf{\Delta}}^{-1} =$$
$$\begin{bmatrix} 0.0669 & 0.0597 & 0.0593 & 0.7839 & 0.0244 & 0.0510 & -0.0750 \\ -0.0002 & -0.0059 & 0.0655 & -0.0943 & 0.5770 & 0.3641 & 0.0415 \end{bmatrix}$$

Equation (8.2) assumes that the variables $X$ are standardised. We must therefore first determine the mean value and the standard deviation for each of the 7 variables. These are

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\bar{X}_j$ | 7.1000 | 4.7730 | 2.3488 | 9.1338 | 5.4582 | 7.1674 | 2.3462 |
| $s_j$ | 2.3238 | 2.4178 | 1.6656 | 3.0178 | 3.2733 | 4.5581 | 1.6105 |

The standardised values for e.g. the first box becomes

$$\boldsymbol{z} = (-1.4373 \quad -0.4603 \quad -1.0860 \quad -1.2787 \quad 1.3167 \quad 1.4422 \quad 1.5124)',$$

where e.g. the second value is found as

$$z_2 = \frac{3.660 - 4.773}{2.4178} = -0.4603.$$

We now easily find the factor scores corresponding to the first box as

$$\hat{\boldsymbol{F}}_1 = (\mathbf{I} + \hat{\mathbf{A}}'\hat{\boldsymbol{\Delta}}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\boldsymbol{\Delta}}^{-1}\boldsymbol{z} = \begin{pmatrix} -1.20 \\ 1.40 \end{pmatrix}.$$

The others are found analogously.

In the following figure we have shown the 25 boxes in a 2-dimensional coordinate system so that each box is placed at the coordinates corresponding to its factor scores (cf. p. 269).

We note (cf. example 8.3) that the two factors describe "thickness" and "size". However, we also note that the "importance" of the two concepts has been switched compared to example 8.1.

♦

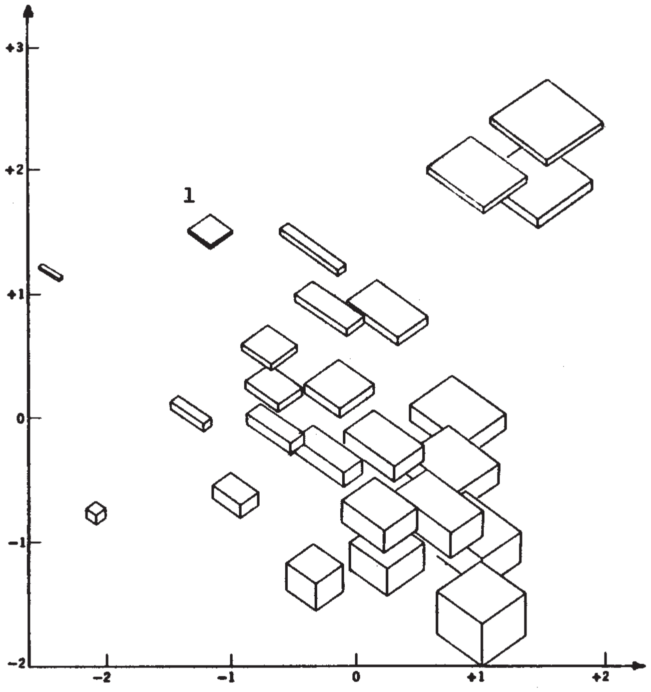
### 8.3.4 A case study

This section is omitted

### 8.3.5 Briefly on maximum likelihood factor analysis

After the appearance of efficient maximisation methods (e.g. Davison-Fletcher-Powell's method) it has become possible to perform maximum likelihood estimation of the factor scores. This is from a statistical point of view somewhat more satisfactory than e.g. the principal factor method. Furthermore, the maximum likelihood solution has a scale-invariance property which is very satisfactory.

We will not concern ourselves with the important numerical and technical problems in determining the maximum likelihood solution but more consider the scale-invariance.



We denote the empirical variance-covariance matrix $\mathbf{S}$ and if we assume normality of the observations we have that $\mathbf{S}$ is Wishart distributed with the parameters $(n-1, \frac{1}{n-1}\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ equals $\mathrm{D}(\boldsymbol{X}_i)$ i.e. the density is

$$c_1(\det \mathbf{S})^{\frac{1}{2}(n-k-2)}(\det \boldsymbol{\Sigma})^{-\frac{1}{2}(n-1)} \exp(-\frac{1}{2}(n-1)\operatorname{tr}(\mathbf{S}\,\boldsymbol{\Sigma}^{-1})),$$

where $c_1$ is an integration constant which only depends on $n$ and $k$. The logarithm of the likelihood function is therefore (disregarding the terms which do not depend on $\boldsymbol{\Sigma}$):

$$\ln \mathrm{L}(\boldsymbol{\Sigma}) = -\frac{1}{2}(n-1)\ln(\det \boldsymbol{\Sigma}) - \frac{1}{2}(n-1)\operatorname{tr}(\mathbf{S}\,\boldsymbol{\Sigma}^{-1}).$$

Here we now introduce the usual $m$ factor model

$$\mathrm{D}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}' + \boldsymbol{\Delta},$$

where $\mathbf{A}$ and $\boldsymbol{\Delta}$ are as in section 8.3.3. Note that we are not assuming that $\boldsymbol{\Sigma}$ has ones on the diagonal. This gives

$$\begin{aligned}\ln \mathrm{L}(\mathbf{A}, \mathbf{\Delta}) \quad &= \quad -\frac{1}{2}(n-1)\ln(\det(\mathbf{A}\mathbf{A}' + \mathbf{\Delta})) \\ &\quad\;\; -\frac{1}{2}(n-1)\,\mathrm{tr}(\mathbf{S}(\mathbf{A}\mathbf{A}' + \mathbf{\Delta})^{-1}).\end{aligned}$$

Maximisation of this function with respect to $\mathbf{A}$ and $\mathbf{\Delta}$ gives the ML-solution to our factor analysis. Concerning the technical problems which remain, we refer to [10].

By partial differentiation of the logarithm of the likelihood function, and after long and tedious algebraic manipulations, one obtains the equation:

$$\hat{\mathbf{A}} = (\hat{\Delta} + \hat{\mathbf{A}}\hat{\mathbf{A}}')\mathbf{S}^{-1}\hat{\mathbf{A}}, \tag{8.4}$$

see e.g. [16].

If we perform a scale-transformation of the $\boldsymbol{X}$'s i.e. we introduce

$$\boldsymbol{Z}_i = \mathbf{C}\,\boldsymbol{X}_i,$$

we then have

$$\mathbf{S}_z = \mathbf{C}\,\mathbf{S}_x\mathbf{C}'$$

where $z$ and $x$ as subscripts shows whether the different quantities have been computed on the base of the $\boldsymbol{Z}_i$'s or the $\boldsymbol{X}_i$'s. With the same convention of notation we then have

$$\hat{\mathbf{A}}_z = (\hat{\mathbf{\Delta}}_z + \hat{\mathbf{A}}_z\hat{\mathbf{A}}_z')\mathbf{C}'^{-1}\mathbf{S}_x^{-1}\mathbf{C}^{-1}\hat{\mathbf{A}}_z.$$

If we pre-multiply by $\mathbf{C}^{-1}$ we get

$$\mathbf{C}^{-1}\hat{\mathbf{A}}_z = [\mathbf{C}^{-1}\hat{\mathbf{\Delta}}_z\mathbf{C}'^{-1} + \mathbf{C}^{-1}\hat{\mathbf{A}}_z(\mathbf{C}^{-1}\hat{\mathbf{A}}_z)']\mathbf{S}_x^{-1}\mathbf{C}^{-1}\hat{\mathbf{A}}_z. \tag{8.5}$$

By comparing (8.4) and (8.5) we find that if $\mathbf{A}$ is a solution to (8.4) then

$$\mathbf{A}_z = \mathbf{C}^{-1}\mathbf{A}$$

will be a solution to (8.5). This means that a scaling of the $\boldsymbol{X}$s (the observations) with the matrix $\mathbf{C}$ implies that the factor weights are scaled by $\mathbf{C^{-1}}$.

If we retain the assumption of normality we can test if the factor model is valid i.e. test

$$H_0 : \mathbf{\Sigma} = \mathbf{\Delta} + \mathbf{A}\mathbf{A}' \quad \text{against} \quad H_1 : \mathbf{\Sigma} \ \text{arbitrary.}$$

The ratio test will then be equivalent to the test given by the test statistic

$$Z = (n - 1 - \frac{1}{6}(2k + 5) - \frac{2}{3}m) \ln \frac{|\hat{\mathbf{\Delta}} + \hat{\mathbf{A}}\hat{\mathbf{A}}'|}{|\mathbf{S}|}$$

and we will reject for

$$Z > \chi^2(\frac{1}{2}\{(k - m)^2 - k - m\}).$$

Finally we will draw the attention to certain standard programmes e.g. in the BMDP package which give possibilities of performing maximum likelihood factor analysis.

**EXAMPLE 8.5.** In the following table we have shown the result of a principle factor solution (PCA), and a maximum likelihood solution (ML) and finally a little Jiffy solution (see [11]).

The data material consists of 198 samples of Portland cement where each sample is analysed for 15 variables (contents of different cement minerals, fine grainedness etc.). The 15 variables have only been given by their respective numbers because we do not consider the interpretation here but only the comparison of the three methods. In the table, weights, which are numerically less than 0.25, have been set equal to 0 to ease the interpretation.

We note that the three methods give remarkably similar results. For factor three we note that the PCA solution differs somewhat from the ML and the LJIF solutions.

| Variable | $Factor1$ | | | $Factor2$ | | | $Factor3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | PCA | ML | LJIF | PCA | ML | LJIF | PCA | ML | LJIF |
| 1 | $-0.26$ | $0$ | $0$ | $0.95$ | $0.91$ | $0.95$ | $0$ | $0.36$ | $0$ |
| 2 | $0$ | $0$ | $0$ | $-0.98$ | $-1.00$ | $-0.99$ | $0$ | $0$ | $0$ |
| 3 | $-0.50$ | $0.93$ | $1.08$ | $0$ | $0$ | $0$ | $-0.40$ | $-0.34$ | $-0.72$ |
| 4 | $0.94$ | $-0.78$ | $-0.80$ | $0$ | $0$ | $0$ | $0$ | $-0.62$ | $-0.32$ |
| 5 | $0$ | $0.29$ | $0.34$ | $0$ | $0$ | $0$ | $-0.48$ | $0$ | $0$ |
| 6 | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $-0.25$ |
| 7 | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ |
| 8 | $0.53$ | $-0.32$ | $-0.32$ | $0$ | $0$ | $0$ | $0.27$ | $-0.31$ | $0$ |
| 9 | $0.90$ | $-0.72$ | $-0.76$ | $0$ | $0$ | $0$ | $0$ | $-0.45$ | $0$ |
| 10 | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0.72$ | $0$ | $0$ |
| 11 | $0$ | $-0.28$ | $-0.31$ | $0$ | $0$ | $0$ | $0.82$ | $0$ | $0$ |
| 12 | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $-0.78$ | $0$ | $0$ |
| 13 | $-0.73$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0$ | $0.98$ | $0.95$ |
| 14 | $-0.86$ | $0.97$ | $1.05$ | $0$ | $0$ | $0$ | $-0.31$ | $0$ | $0$ |
| 15 | $0$ | $0.25$ | $0$ | $0.93$ | $0.93$ | $0.92$ | $0$ | $0$ | $-0.35$ |

♦

### 8.3.6 Q-mode analysis

In the form of factor analysis we have regarded up till now - the so-called R-modus analysis - one investigates the correlations between the different variables. The samples of the individuals etc. are used as repetitions and these are used to estimate the different correlations. If we call the observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ and let

$$\mathbf{X}' = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kn} \end{bmatrix},$$

where the rows corresponds to the single variables and the columns to the individuals. If we assume that the observations have been normalised so they have mean value 0 and variance 1 we get the correlation matrix as

$$\mathbf{R} = \mathbf{X}'\mathbf{X},$$

cf. theorem 2.19. In a **dual** way we could of course define

$$\mathbf{Q} = \mathbf{X}\mathbf{X}',$$

and then interpret it as an expression for the correlation between individuals and then perform a factor analysis on these. The results of such a procedure will be a classification of individuals into groups which are close to each other.

We give a small example which comes from [14].

**EXAMPLE 8.6.** We consider 12 stream sediment samples collected in Jameson Land in East Greenland. They are analysed for 7 elements which are Cu, Ni, V, Pb, Zr, Ca and Ba. An ordinary R-modus analysis showed that the two first factors described $42\% + 37\% = 79\%$ of the variation. In the following figure we have shown the rotated factor weights.

Then a $Q$-modus analysis was performed as mentioned above. This gave a first factor which described $38\%$ of the total variation and a second factor which described $26\%$ of the total variation.

From the figure with the factor weights we now get a direct comparison of the different samples. This could also be obtained through $R$-modus analysis but we would then have to go via the factor scores.

Analysis of this kind is used in mineral prospecting in the attempt to determine which samples are to be declared non-normal and thereby interesting. ♦

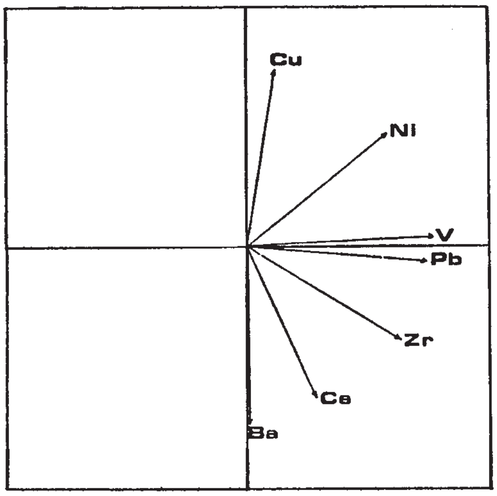


Figure 8.2: Factor weights in R-modus analysis.

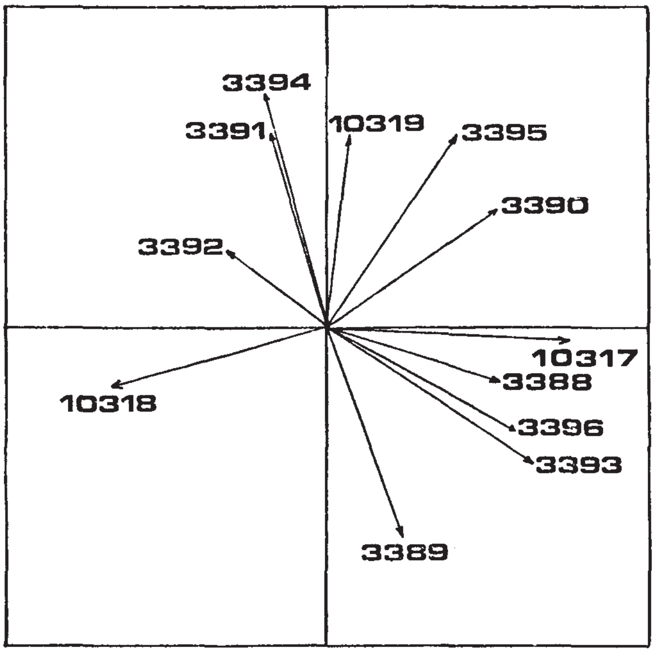


Figure 8.3: Factor weights in Q-modus analysis.

When performing a $Q$-modus analysis one will often end up with a large amount of computations since the $Q$-matrix is of the order $n \times n$, where $n$ is the number of individuals. One can then draw advantage of the theorems in section 1.4.2. From these we see that the eigenvalues which are different from 0 in R and Q are equal and there is a simple relationship between the eigenvectors. Since $R$ is only of the order $k \times k$ and the number of variables usually is considerably less than the number of individuals it is possible to save a lot of numerical work.

Finally we remark that $Q$-modus analysis often is not performed on $\mathbf{X}\mathbf{X}'$ but on another matrix containing some more or less arbitrarily chosen similarity measures. The technique is, however, unchanged and one can still obtain computational savings by using the above mentioned relation between $R$-modus and $Q$-modus analysis. For special choices of similarity measures one often calls this a principal coordinate analysis.

An attempt to do both analyses at one time is found in the so-called correspondence analysis which is due to the Frenchman Benzécri (1973).

### 8.3.7 Some standard programmes

Principal component analysis is merely an eigenvalue analysis of the variance-covariance matrix or of an estimated variance covariance matrix. Such an analysis is therefore performed by means of a standard programme for the solution of the eigenvalue problem for an symmetric positive semidefinite matrix.

There are, however, also a number of standard programmes for the computation of principal components. Here we can e.g. mention the programmes BMD01M nd BMD02M from the BMD system.

BMD01M, PRINCIPAL COMPONENT ANALYSIS, computes a principal component solution on the standardised data i.e. we are analysing the empirical correlation matrix. Output from this programme includes correlation coefficients and eigenvalues including the cumulated fractions of the total variance and the eigenvectors i.e. the principal axes. Finally the rank of each observation (standardised) is given by size of the single principal components.

BMD02M, REGRESSION ON PRINCIPAL COMPONENTS, computes the same quantities as BMD01M and furthermore computes regressions of each of the dependent variables on the first, the first two, the first three and all principal components.

Most standard programmes for the computation of factor solutions use the principal factor solution mentioned in this book followed by rotation.

One of the largest systems is the programme complex which is given in the SPSS manual (Statistical Package for the Social Sciences). In this system we find a number of factorisation routines. The most often used are probably the principal factor methods. These are found in two versions. One where one just uses the ordinary principal factor solution and one where one iteratively estimates the communalities by means of the squared multiple correlation coefficient estimate the number of necessary factors maybe exclude certain of these reestimate the communalities etc. until the difference between two sets of estimated communalities is less than a certain limit.

Among a number of other methods there is also a method by Rao which was developed in a more classical statistical sense (see Rao (1955)[18]). Here the more usual estimates of and test of the number of necessary factors etc. are performed.

Of the orthogonal rotation principles there are three, they are quartimax, Varimax (see p. 279) and equimax. Furthermore there is a procedure which performs the so-called oblique rotation (by the oblimin principle).

Computation of factor scores is performed by a principle which is in relationship to the one mentioned in section 8.3.6.

The BMD-programme BMD08M, FACTOR ANALYSIS, is also very large. The factorisation routines are, however, all of the principal factor type. They operate on both the correlation and the variance covariance matrices. Possibilities exist for different types of communality estimates and the above mentioned iterative estimation procedure can be utilised.

There are a number of rotational principles including the orthogonal (among other quartimax and Varimax) and as "oblique" (oblimin-types).

Computation of factor scores is performed by the same principles as mentioned in section 8.3.6.

In the BMDP packages factor analysis programme one can also perform a maximum likelihood estimation.

The SSP sample programme FACTO performs a principle factor solution and rotates the factors by the Varimax-method. The programme is more or less identical to the old factor analysis programme from the BMD system i.e. BMD03M. The output includes the usual quantities, however, not the factor scores. Some of the users will be shown below. The rest of this chapter is neglected because the programme etc. are obsolete.

# Bibliography

[1] AGTERBERG, F. P. *Geomathematics. Mathematical background and geo-science applications*. Elsevier, Amsterdam 1973.

[2] ANDERSON, T. W. *An Introduction to Multivariable Statistical Analysis*. John Wiley & sons, New York 1958.

[3] BOURBAKI, N. *Algèbre*. Hermann, Paris 1967, ch. 2 Algèbre Lineaire.

[4] CATTELL, R. Factor analysis: An introduction to essentials. i.the purpose and underlying models. ii.the role of factor analysis in research. *Biometrics 21* (1965), 190–215, 405–435.

[5] DAVIES, O. L., Ed. *Design and Analysis of Industrial Experiments*, second ed. Oliver and Boyd, London 1967.

[6] DAVIS, J. C. *Statistics and data analysis in geology*. John Wiley, New York 1973.

[7] DWYER, P. S. *The contribution of an orthogonal multiple facto solution to multiple correlation*, vol. 4. John Wiley & Sons, 1939.

[8] HARMAN, H. H. *Modern Factor Analysis*, second ed. The University of Chicago Press, Chicago 1967.

[9] JOHNSON, R. M. On a theorem stated by eckart and young. In *Psychometrika*, vol. 28. 1963, pp. 259–263.

[10] JÖRESKOG, K. G. Some contributions to maximum likelihood factor analysis. In *Psycometrika*, vol. 32. 1967.

[11] KAISER, H. F. *The varimax criterion for analytic rotation in factor analysis*. 1958.

[12] KENDALL, M. G., AND STUART, A. *The Advanced Theory of Statistics*, vol. 2. Charles Griffin & Co., London 1967.

[13] KNUDSEN, J. G. En statistisk analyse af cementstyrke. Eksamensprojekt, DTU,IMSOR, 1975.

[14] LARSEN, P. M. Geokemisk oversigtsprospektering. multivarable statistiske metoders anvendelighed ved interpretation af regionale geokemiske data. Eksamensprojekt, DTU,IMSOR, 1976.

[15] LAWLEY, D. N. *The estimation of factor loadings by the method of maximum likelihood*. 1940.

[16] MORRISON, D. F. *Multivariate Statistical Methods*. McGraw-Hill, New York 1967.

[17] PEDERSEN, S. T., AND SKJØTH, P. Statistisk analyse af data fra cementfabrikation. Eksamensprojekt, DTU,IMSOR, 1976.

[18] RAO, C. R. *Estimation and tests of significance in factor analysis*. 1955.

[19] RAO, C. R., AND MITRA, S. K. *Generalized Inverse of Matrices and Its Applications*. John Wiley, New York 1971.

[20] SPLIID, H. *User Guide for Stepwise Regression Program REGRGO*. IMSOR, Lyngby 1974.

[21] WILKINSON, J. H. Error analysis of direct methods of matrix inversion. *Journal of the Association of Computing Machinery 8* (1961), 281–330.

# Index

# Appendix A

# The Greek alphabet

| | *Letter name* | *Pronounciation* | *Equivalent to* |
|---|---|---|---|
| Α $\alpha$ | alfa | [a][a:] | a |
| Β $\beta$ | bēta | [b] | b |
| $\Gamma$ $\gamma$ | gamma | [g] | g |
| $\Delta$ $\delta$ | delta | [d] | d |
| Ε $\epsilon$ | epsilon | [e] | e |
| Ζ $\zeta$ | zēta | [ts,s] | z |
| Η $\eta$ | ēta | [æ:] | ē |
| $\Theta$ $\vartheta$ $\theta$ | thēta | [$\theta$,th,t] | th,t |
| Ι $\iota$ | iōta | [i][i:] | i |
| Κ $\kappa$ | kappa | [k] | k |
| $\Lambda$ $\lambda$ | lambda | [l] | l |
| Μ $\mu$ | my | [m] | m |
| Ν $\nu$ | ny | [n] | n |
| $\Xi$ $\xi$ | ksi | [ks] | ks(x) |
| Ο ο | omikron | [o] | o |
| $\Pi$ $\pi$ | pi | [p] | p |
| Ρ $\rho$ | ro | [r] | r |
| $\Sigma$ $\sigma$ $\varsigma$ | sigma | [s] | s |
| Τ $\tau$ | tau | [t] | t |
| $\Upsilon$ $\upsilon$ | ypsilon | [y][y:] | y |
| $\Phi$ $\phi$ | fi | [f] | f(ph) |
| Χ $\chi$ | khi | [x,ç,kh,k] | ch(kh) |
| $\Psi$ $\psi$ | psi | [ps] | ps |
| $\Omega$ $\omega$ | ōmega | [å:] | ō |