Adjustment theory

an introduction

Adjustment theory

an introduction

P.J.G. Teunissen

Delft University of Technology

Department of Mathematical Geodesy and Positioning

VSSD

First edition 2000-2006

Published by:
VSSD
Leeghwaterstraat 42, 2628 CA Delft, The Netherlands
tel. +31 15 278 2124, telefax +31 15 278 7585, e-mail: hlf@vssd.nl
internet: http://www.vssd.nl/hlf
URL about this book: **http://www.vssd.nl/hlf/a030.htm**

A collection of digital pictures and an elctronic version can be made available for lecturers who adopt this book. Please send a request by e-mail to hlf@vssd.nl

ISBN 978-90-407-1974-5 Electronic version ISBN 978-90-6562-215-0
NUR 930

Keywords: adjustment theory, geodesy

Preface

As in other physical sciences, empirical data are used in Geodesy and Earth Observation to make inferences so as to describe the physical reality. Many such problems involve the determination of a number of unknown parameters which bear a linear (or linearized) relationship to the set of data. In order to be able to check for errors or to reduce for the effect these errors have on the final result, the collected data often exceed the minimum necessary for a unique determination (redundant data). As a consequence of (measurement) uncertainty, redundant data are usually inconsistent in the sense that each sufficient subset yields different results from another subset. To obtain a unique solution, consistency needs to be restored by applying corrections to the data. This computational process of making the data consistent such that the unknown parameters can be determined uniquely, is referred to as *adjustment*. This book presents the material for an introductory course on Adjustment theory. The main goal of this course is to convey the knowledge necessary to perform an optimal adjustment of redundant data. Knowledge of linear algebra, statistics and calculus at the undergraduate level is required. Some prerequisites, repeatedly needed in the book, are summarized in the Appendices A-F.
Following the *Introduction*, the elementary adjustment principles of 'least-squares', unbiased minimum variance' and 'maximum likelihood' are discussed and compared in *Chapter 1*. These results are generalized to the multivariate case in chapters 2 and 3. In *Chapter 2* the model of observation equations is presented, while the model of condition equations is discussed in *Chapter 3*. Observation equations and condition equations are dual to each other in the sense that the first gives a parametric representation of the model, while the second gives an implicit representation. Observables that are either stochastically or functionally related to another set of observables are referred to as y^R-variates. Their adjustment is discussed in *Chapter 4*. In *Chapter 5* we consider mixed model representations and in *Chapter 6* the partitioned models. Models may be partitioned in different ways: partitioning of the observation equations, partitioning of the unknown parameters, partitioning of the condition equations, or all of the above. These different partitionings are discussed and their relevance is shown for various applications. Up tot this point all the material is presented on the basis of the assumption that the data are linearly related. In geodetic applications however, there are only a few cases where this assumption holds true. In the last chapter, *Chapter 7*, we therefore start extending the theory from linear to nonlinear situations. In particular an introduction to the linearization and iteration of nonlinear models is given.

Many colleagues of the Department of Mathematical Geodesy and Positioning whose assistance made the completion of this book possible are greatly acknowledged. C.C.J.M. Tiberius took care of the editing, while the typing was done by Mrs. J. van der Bijl and Mrs. M.P.M. Scholtes, and the drawings by Mr. A. Smit. Various lecturers have taught the book's material over the past years. In particular the feedback and valuable recommendations of the lecturers C.C.J.M. Tiberius, F. Kenselaar, G.J. Husti and H.M. de Heus are acknowledged.

P.J.G. Teunissen
April, 2003

Contents

Introduction

The *calculus of observations* (in Dutch: *waarnemingsrekening*) can be regarded as the part of mathematical statistics that deals with the results of measurement. Mathematical statistics is often characterized by saying that it deals with the making of decisions in uncertain situations. In first instance this does not seem to relate to measurements. After all, we perform measurements in order to obtain sharp, objective and unambiguous quantitative information, information that is both correct and precise, in other words accurate information. Ideally we would like to have 'mathematical certainty'. But mathematical certainty pertains to statements like:

if a=b and b=c and c=d, then a=d.

If a, b, c, and d are measured terrain distances however, a surveyor to be sure will first compare d with a before stating that a=d. Mathematics is a product of our mind, whereas with measurements one is dealing with humans, instruments, and other materialistic things and circumstances, which not always follow the laws of logic. From experience we know that measurements produce results that are only sharp to a certain level. There is a certain vagueness attached to them and sometimes even, the results are simply wrong and not usable.

In order to eliminate uncertainty, it seems obvious that one seeks to improve the instruments, to educate people better and to try to control the circumstances better. This has been successfully pursued throughout the ages, which becomes clear when one considers the very successful development of geodetic instruments since 1600. However, there are still various reasons why measurements will always remain uncertain to some level:

1. 'Mathematical certainty' relates to abstractions and not to physical, technical or social matters. A mathematical model can often give a useful description of things and relations known from our world of experience, but still there will always remain an essential difference between a model and the real world.
2. The circumstances of measurement can be controlled to some level in the laboratory, but in nature they are uncontrollable and only partly descriptive.
3. When using improved methods and instrumentation, one still, because of the things mentioned under 1 and 2, will have uncertainty in the results of measurement, be it at another level.
4. Measurements are done with a purpose. Dependent on that purpose, some uncertainty is quite acceptable. Since better instruments and methods usually cost more, one will have to ask oneself the question whether the reduction in uncertainty is worth the extra costs.

Summarizing, it is fundamentally and practically impossible to obtain absolute certainty from measurements, while at the same time this is also often not needed on practical grounds. Measurements are executed to obtain quantitative information, information which is accurate enough for a specific purpose and at the same time cheap enough for that purpose. It is because of this intrinsic uncertainty in measurements, that we have to consider the calculus of observations. The calculus of observations deals with:

1. The description and analysis of measurement processes.
2. Computational methods that take care, in some sense, of the uncertainty in measurements.
3. The description of the quality of measurement and of the results derived therefrom.
4. Guidelines for the design of measurement set-ups, so as to reach optimal working procedures.

The calculus of observations is essential for a geodesist, its meaning comparable to what mechanics means for the civil-engineer or mechanical-engineer. In order to shed some more light on the type of problems the calculus of observations deals with, we take some surveying examples as illustration. Some experience in surveying will for instance already lead to questions such as:

1. Repeating a measurement does usually not give the same answer. How to describe this phenomenon?
2. Results of measurement can be corrupted by systematic or gross errors. Can these errors be traced and if so, how?
3. Geometric figures (e.g. a triangle) are, because of 2, usually measured with *redundant* (*overtallig*) observations (e.g. instead of two, all three angles of a triangle are measured). These redundant observations will usually not obey the mathematical rules that hold true for these geometric figures (e.g. the three angular measurements will usually not sum up to π). An *adjustment* (*vereffening*) is therefore needed to obtain consistent results again. What is the best way to perform such an adjustment?
4. How to describe the effect of 1 on the final result? Is it possible that the final result is still corrupted with non-detected gross errors?

Once we are able to answer questions like these, we will be able to determine the required measurement set-up from the given desired quality of the end product. In this book we mainly will deal with points 1 and 3, the elementary aspects of adjustment theory.

Functional and stochastic model

In order to analyze scientific or practical problems and/or make them accessible for a quantitative treatment, one usually tries to rid the problem from its unnecessary frills and describes its essential aspects by means of notions from an applicable mathematical theory. This is referred to as the making of a *mathematical model*. On the one hand the model should give a proper description of the situation, while, on the other hand, the model should not be needlessly complicated and difficult to apply. The determination of a proper model can be considered the ’art’ of the discipline. Well-tried and well-tested experience, new experience from experiments, intuition and creativity, they all play a role in formulating the model.

Here we will restrict ourselves to an example from surveying. Assume that we want to determine the relative position of terrain-’points’, whereby their differences in height are of no interest to us. If the distances between the ’points’ are not too large, we are allowed, as experience tells us, to describe the situation by projecting the ’points’ onto a level surface (niveauvlak) and by treating this surface as if it was a plane. For the description of the relative position of the points, we are then allowed to make use of the two-dimensional Euclidean geometry (vlakke meetkunde), the theory of which goes back some two thousand years. This theory came into being as an abstraction from how our world was seen and experienced. But for us the theory is readily available, thus making it possible to anticipate on its use, for instance by using sharp and clear markings for the terrain-'points' such that they can be treated as if they were mathematically defined points. In this example, the two-dimensional Euclidean geometry acts as our *functional model*. That is, notions and rules from this theory (such as the sine- and cosine rule) are used to describe the relative position of the terrain-'points'.

Although the two-dimensional Euclidean geometry serves its purpose well for many surveying tasks, it is pointed out that the gratuitously use of a 'standard' model can be full of risks. Just like the laws of nature (natuurwetten), a model can only be declared 'valid' on the basis of experiments. We will come back to this important issue in some of our other courses.

When one measures some of the geometric elements of the above mentioned set of points, it means that one assigns a number (the observation) according to certain rules (the measurement procedure) to for instance an angle α (a mathematical notion). From experience we know that a measurement of a certain quantity does not necessarily give the same answer when repeated. In order to describe this phenomenon, one could repeat the measurements a great many times (e.g. 100). And on its turn the experiment itself could also be repeated. If it then turns out that the histograms of each experiment are sufficiently alike, that is, their shapes are sufficiently similar and their locations do not differ too much, one can describe the measurement variability by means of a stochastic variable. In this way notions from mathematical statistics are used in order to describe the inherent variability of measurement processes. This is referred to as the *stochastic model*.

In practice not every measurement will be repeated of course. In that case one will have to rely on the extrapolation of past experience. When measuring, not only the measurement result as a number is of interest, but also additional information such as the type of instrument used, the measurement procedure followed, the observer, the weather, etc. Taken together this information is referred to as the *registration* (*registratie*). The purpose of the registration is to establish a link with past experience or experiments such that a justifiable choice for the stochastic model can be made.

In surveying and geodesy one often may assume as stochastic model that the results of measurement (the observations) of a certain measurement method can be described as being an independent sample drawn (aselecte trekking) from a normal (Gaussian) distribution with a certain standard deviation. The mathematical *expectation* (*verwachting*) of the normally distributed stochastic variable (the observable, in Dutch: de waarnemingsgrootheid), is then assumed to equal the unknown angle etc. From this it follows, if the three angles α, β and γ of a triangle are measured, that the expectations of the three angular stochastic variables obey the following relation of the functional model:

$$\alpha + \beta + \gamma = \pi$$

For the individual observations themselves this is usually not the case. A sketch of the relation between terrain situation, functional model, measurement, experiment and stochastic model is given in figure 0.1. The functional and stochastic model together from the mathematical model.

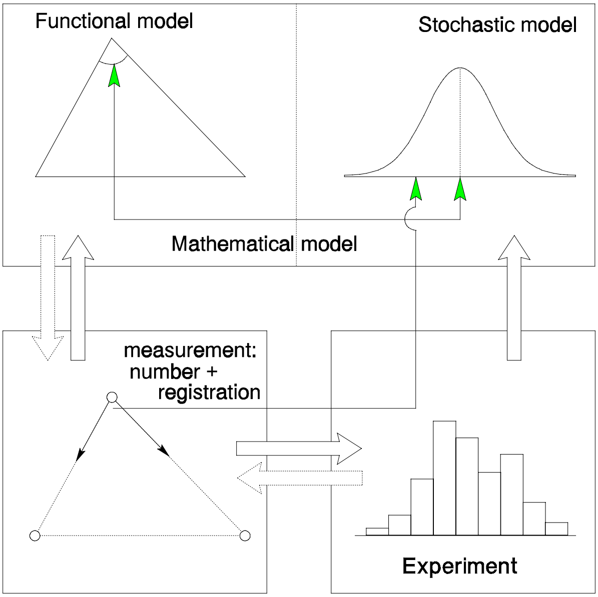

Figure 0.1: Diagram of the fundamental relations in adjustment theory

1 Linear estimation theory: an introduction

1.1 Introduction

Before stating the general problem, let us define a simple example that contains most of the elements that will require our consideration.

Suppose that an unknown distance, which we denote as x, has been measured twice. The two measurements are denoted as y_1 and y_2, respectively. Due to all sorts of influences (e.g. instrumental effects, human effects, etc.) the two measurements will differ in general. That is $y_1 \neq y_2$ or $y_1 - y_2 \neq 0$. One of the two measurements, but probably both, will therefore also differ from the true but unknown distance x. Thus, $y_1 \neq x$ and $y_2 \neq x$. Let us denote the two differences $y_1 - x$ and $y_2 - x$, the so-called measurement errors, as e_1 and e_2 respectively. We can then write:

$$y_i = x + e_i \quad , \quad i = 1,2$$

or in vector notation:

$$(1) \qquad \boxed{\underset{2\times1}{y} = \underset{2\times1}{a} \; \underset{1\times1}{x} + \underset{2\times1}{e}}$$

with:

$$y = (y_1,y_2)^*$$
$$a = (1,1)^*$$
$$e = (e_1,e_2)^*$$

(* denotes the operation of transposition; thus y^* is the transpose of y).

Our problem is now, how to estimate the unknown x from the given measurement or observation vector y. It is to this and related questions that we address our attention. In this introductory chapter we will restrict ourselves to the case of two observations and one unknown. In the next section we will introduce the least-squares principle. Least-squares can be regarded as a deterministic approach to the estimation problem. In the sections following we will introduce additional assumptions regarding the probabilistic nature of the observations and measurement errors. This leads then to the so-called best linear unbiased estimators and the maximum likelihood estimators.

1.2 Least-squares estimation (orthogonal projection)

The equation $\underset{2x1}{y} = \underset{2x1}{a} \; \underset{1x1}{x} + \underset{2x1}{e}$, in which y and a are given, and x and e are unknown, can be visualized as in figure 1.1a.

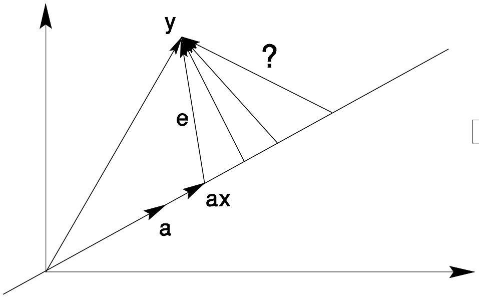

Figure 1.1a: x and e unknown

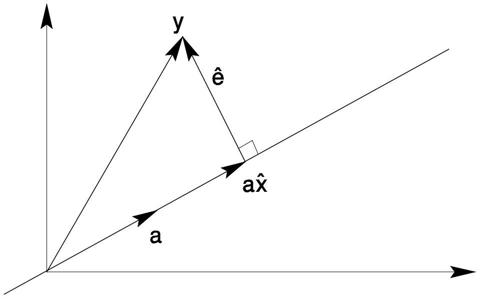

Figure 1.1b: $\hat{x}$ as estimate of x
$\hat{e}$ as estimate of e

From the geometry of the figure it seems intuitively appealing to estimate x as $\hat{x}$, such that$a\hat{x}$ is as close as possible to the given observation vector y. From figure 1.1b follows that $a\hat{x}$ has minimum distance to y if $\hat{e} = y - a\hat{x}$ is orthogonal to a. That is if:

$$
\begin{array}{lll}
 & a^*(y - a\hat{x}) = 0 & \text{"orthogonality"} \\
 & \Downarrow & \\
 & (a^*a)\hat{x} = a^*y & \text{"normal equation"} \\
 & \Downarrow & \\
 & \hat{x} = (a^*a)^{-1}a^*y & \text{"LSQ-estimate of } x\text{"} \\
 & \Downarrow & \\
(2) & \hat{y} = a\hat{x} = [a(a^*a)^{-1}a^*]y & \text{"LSQ-estimate of observations"} \\
 & \Downarrow & \\
 & \hat{e} = y - \hat{y} = [I - a(a^*a)^{-1}a^*]y & \text{"LSQ-estimate of measurement errors"} \\
 & \Downarrow & \\
 & \hat{e}^*\hat{e} = y^*[I - a(a^*a)^{-1}a^*]y & \text{"Sum of squares"}
\end{array}
$$

The name *"Least-Squares" (LSQ)* is given to the estimate $\hat{x} = (a^*a)^{-1}a^*y$ of x, since $\hat{x}$ minimizes the sum of squares $(y-ax)^*(y-ax)$. We will give two different proofs for this:

***First proof*:** (algebra)

$$
\begin{aligned}
(y-ax)^*(y-ax) &= [y - a\hat{x} - a(x-\hat{x})]^*[y - a\hat{x} - a(x-\hat{x})] \\
&= [\hat{e} - a(x-\hat{x})]^*[\hat{e} - a(x-\hat{x})] \\
&= \hat{e}^*\hat{e} \;-\; \hat{e}^*a(x-\hat{x}) \;-\; (x-\hat{x})^*a^*\hat{e} \;+\; (x-\hat{x})^*a^*a(x-\hat{x}) \\
&= \hat{e}^*\hat{e} \;+\; (x-\hat{x})^*a^*a\;(x-\hat{x}) \\
&= \text{minimal for } x = \hat{x}.
\end{aligned}
$$

End of proof.

***Second proof*:** (calculus)

Call $E(x) \doteq (y-ax)^*(y-ax) = y^*y - 2y^*ax + a^*ax^2.$

From calculus we know that $\hat{x}$ is a solution of the minimization problem $\min\limits_{x} E(x)$ if:

$$\frac{dE}{dx}(\hat{x}) = 0 \text{ and } \frac{d^2E}{dx^2}(\hat{x}) > 0 \ .$$

With:

$$\frac{dE}{dx}(\hat{x}) = -2a^*y + 2a^*a\hat{x} = 0 \quad \Rightarrow \quad (a^*a)\hat{x} = a^*y$$

and:

$$\frac{d^2E}{dx^2}(\hat{x}) = 2a^*a > 0$$

the proof follows. End of proof.

***Example 1*:** Let us consider the problem defined in section 1.1:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}.$$

With $a = (1,1)^*$ the normal equation $(a^*a)\hat{x} = a^*y$ reads:

$$2\hat{x} = y_1 + y_2 \ .$$

Hence the least-squares estimate $\hat{x} = (a^*a)^{-1}a^*y$ is:

$$\hat{x} = \tfrac{1}{2}(y_1 + y_2).$$

Thus, to estimate x, one adds the measurements and divides by the number of measurements. Hence the least-squares estimate equals in this case the arithmetic average. The least-squares estimates of the observations and observation errors follow from $\hat{y} = a\hat{x}$ and $\hat{e} = y - \hat{y}$ as:

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} y_1 + y_2 \\ y_1 + y_2 \end{pmatrix}, \text{ and } \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} y_1 - y_2 \\ y_2 - y_1 \end{pmatrix}.$$

Finally the square of the length of $\hat{e}$ follows as:

$$\hat{e}^*\hat{e} = \tfrac{1}{2}(y_1 - y_2)^2.$$

End of example.

The matrices $a(a^*a)^{-1}a^*$ and I - $a(a^*a)^{-1}a^*$ that occur in (2) are known as *orthogonal projectors*. They play a central role in estimation theory. We will denote them as:

(3) $$\boxed{P_a = a(a^*a)^{-1}a^* \quad , \quad P_a^{\perp} = I - a(a^*a)^{-1}a^*}$$

The matrix P_a projects *onto* a line spanned by a and *along* a line orthogonal to a. To see this, note that we can write $P_a y$ with the help of the cosine rule as (see figure 1.2):

$$P_a y = a(a^*a)^{-1}a^*y = a(a^*a)^{-1}\|a\|\,\|y\|\cos\alpha = \|y\|\cos\alpha\,\frac{a}{\|a\|} \ .$$

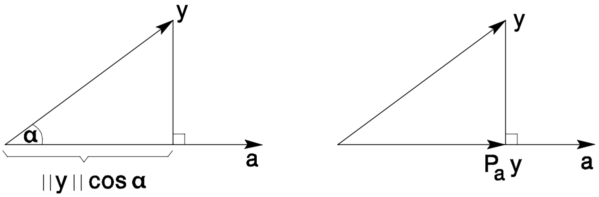

Figure 1.2

In a similar way we can show that $P_a^{\perp}$ projects *onto* a line orthogonal to a and *along* a line spanned by a (see figure 1.3).

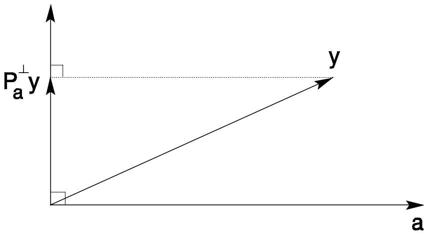

Figure 1.3

If we denote the vector that is orthogonal to a as b, thus b^*a=0, the projectors P_a and $P_a^{\perp}$ can alternatively be represented as:

(4) $$\boxed{P_a = P_b^{\perp} = I - b(b^*b)^{-1}b^* \quad , \quad P_a^{\perp} = P_b = b(b^*b)^{-1}b^*}$$

To see this, consult figure 1.4 and note that with the help of the cosine rule we can write:

$$P_a^{\perp}y = \|y\|\cos\beta\,\frac{b}{\|b\|} = b\|b\|^{-2}b^*y = b(b^*b)^{-1}b^*y = P_b y .$$

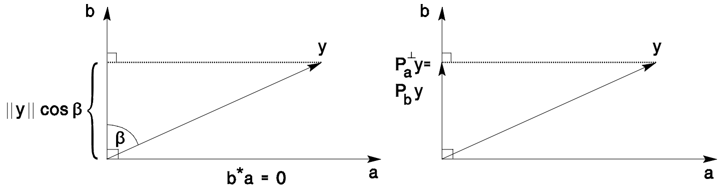

Figure 1.4

With (3) and (4) we have now two different representations of the orthogonal projectors $P_a = P_b^{\perp}$ and $P_a^{\perp} = P_b$. The representations of (3) are used in (2), which followed from solving equation (1) in a least-squares sense. We will now show that the representations of (4) correspond to solving the equation:

(5)
$$\boxed{\underset{1\times2}{b^*}\underset{2\times1}{y} = \underset{1\times2}{b^*}\underset{2\times1}{e}}$$

in a least-squares sense. Equation (5) follows from pre-multiplying equation (1) with b^* and noting that $b^*a=0$. We know that the least-squares solution of (1) follows from solving the minimization problem:

$$\min_x \|y - ax\|^2 .$$

We have used the notation $\|.\|^2$ here to denote the square of the length of a vector. The minimization problem $\|y\text{-}ax\|^2$ can also be written as:

$$\min_z \|y - z\|^2 \quad \text{subject to} \quad z = ax .$$

The equation $z=ax$ is the *parametric representation* of a line through the origin with direction vector a. From linear algebra we know that this line can also be represented in *implicit form* with the help of the normal vector b. Thus we may replace $z=ax$ by $b^*z=0$. This gives:

$$\min_z \|y - z\|^2 \quad \text{subject to} \quad b^*z = 0 .$$

With $e=y$ - z, this may also be written as:

$$\min_e \|e\|^2 \quad \text{subject to} \quad b^*y = b^*e .$$

The least-squares estimate $\hat{e}$ of the unknown e in (5) thus follows as the vector $\hat{e}$ which satisfies (5) and has minimal length. Figure 1.5 shows that $\hat{e}$ is obtained as the orthogonal projection of y onto the vector b. The least-squares estimates of (5) read therefore:

$$
\begin{aligned}
&\hat{e} = P_b y = [b(b^*b)^{-1}b^*]y \\
&\hat{y} = y - \hat{e} = P_b^{\perp} y = [I - b(b^*b)^{-1}b^*]y
\end{aligned}
\tag{6}
$$

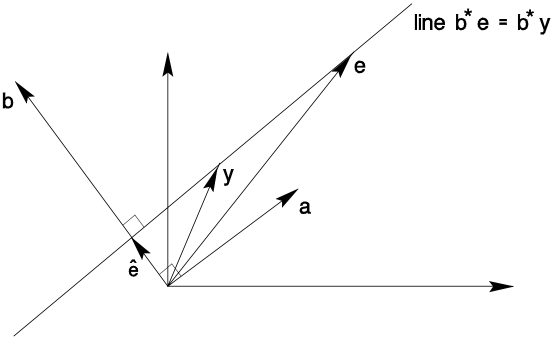

Figure 1.5

Example 2**:** Let us consider the problem defined in section 1.1:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} .$$

The form corresponding to (5) is:

$$(1 \;\; -1)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (1 \;\; -1)\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} .$$

This equation expresses the fact that the difference between the two observed distances y_1 and y_2 is unequal to zero in general because of the presence of measurement errors e_1 and e_2. Since the vector b is given in this case as:

$$\boldsymbol{b} = (1-1)^*,$$

we have:

$$\boldsymbol{P}_b = \tfrac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \text{ and } \boldsymbol{P}_b^{\perp} = \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The least-squares estimates read therefore:

$$\begin{pmatrix}\hat{e}_1\\ \hat{e}_2\end{pmatrix} = \boldsymbol{P}_b\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \frac{1}{2}\begin{pmatrix}y_1-y_2\\ y_2-y_1\end{pmatrix} \quad\text{and}\quad \begin{pmatrix}\hat{y}_1\\ \hat{y}_2\end{pmatrix} = \boldsymbol{P}_b^{\perp}\begin{pmatrix}y_1\\ y_2\end{pmatrix} = \frac{1}{2}\begin{pmatrix}y_1+y_2\\ y_1+y_2\end{pmatrix},$$

which is identical to the results obtained earlier.

End of example.

1.3 Weighted least-squares estimation

Up to now, we have regarded the two observations y_1 and y_2 as equally accurate, and solved:

$$\underset{2\times1}{y} = \underset{2\times1}{a}\,\underset{1\times1}{x} + \underset{2\times1}{e}$$

by looking for the value $\hat{x}$ that minimized:

$$E(x) = (y_1-a_1x)^2 + (y_2-a_2x)^2 = (y-ax)^*\ I\ (y-ax)\ .$$

The least-squares principle can however be generalized by introducing a *symmetric, positive-definite weight matrix W* so that:

$$E(x) = w_{11}(y_1-a_1x)^2 + 2w_{12}(y_1-a_1x)(y_2-a_2x) + w_{22}(y_2-a_2x)^2 = (y-ax)^*\ W\ (y-ax)\ .$$

The elements of the 2×2 matrix W can be chosen to emphasize (or de-emphasize) the influence of specific observations upon the estimate $\hat{x}$. For instance, if $w_{11}>w_{22}$, then more importance is attached to the first observation, and the process of minimizing $E(x)$ tries harder to make $(y_1\text{-}a_1x)^2$ small. This seems reasonable if one believes that the first observation is more trustworthy than the second; for instance if observation y_1 is obtained from a more accurate instrument than observation y_2.

The solution $\hat{x}$ of:

$$\min_x\ E(x) = \min_x\ (y-ax)^*W(y-ax),$$

is called the *weighted* least-squares (WLSQ) estimate of x. As we know the solution can be obtained by solving:

$$\frac{dE}{dx}(\hat{x}) = 0,$$

and checking whether:

$$\frac{d^2E}{dx^2}(\hat{x}) > 0\ .$$

From:

$$E(x) = y^*Wy - 2y^*Wax + a^*Wax^2$$

follows:

$$\frac{dE}{dx}(\hat{x}) = -2a^*Wy + 2a^*Wa\hat{x} = 0$$

and:

$$\frac{d^2E}{dx^2}(\hat{x}) = 2a^*Wa \ .$$

Since the weight matrix W is assumed to be positive-definite we have $a^*Wa>0$ (recall from linear algebra that a matrix M is said to be *positive-definite* if it is symmetric and $z^*Mz>0 \forall\ z\neq 0$). From the first equation we therefore get:

$$a^*W(y-a\hat{x}) = 0$$
$$\Downarrow$$
$$(a^*Wa)\hat{x} = a^*Wy \qquad \text{"normal equation"}$$
$$\Downarrow$$
$$\hat{x} = (a^*Wa)^{-1}a^*Wy \qquad \text{"WLSQ-estimate of } x\text{"}$$
$$\Downarrow$$
(7) $$\hat{y} = a\hat{x} = [a(a^*Wa)^{-1}a^*W]y \qquad \text{"WLSQ-estimate of } y\text{"}$$
$$\Downarrow$$
$$\hat{e} = y-\hat{y} = [I-a(a^*Wa)^{-1}a^*W]y \qquad \text{"WLSQ-estimate of } e\text{"}$$
$$\Downarrow$$
$$\hat{e}^*W\hat{e} = y^*[W-Wa(a^*Wa)^{-1}a^*W]y \qquad \text{"weighted sum of squares"}$$

Compare this result with (2).

***Example 3*:** Consider again the problem defined in section 1.1:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}.$$

As weight matrix we take:

$$W = \begin{pmatrix} w_{11} & 0 \\ 0 & w_{22} \end{pmatrix} .$$

With $a=(1\ 1)^*$ the normal equation $(a^*Wa)\hat{x}=a^*Wy$ reads:

$$(w_{11} + w_{22})\hat{x} = w_{11}y_1 + w_{22}y_2 \ .$$

Hence, the weighted least-squares estimate $\hat{x}=(a^*Wa)^{-1}a^*Wy$ is:

$$\hat{x} = \frac{w_{11}y_1 + w_{22}y_2}{w_{11} + w_{22}} \ .$$

Thus instead of the arithmetic average of y_1 and y_2, as we had with $W=I$, the above estimate is a *weighted* average of the data. This average is closer to y_1 than to y_2 if $w_{11}>w_{22}$. The WLSQ-estimate of the observations, $\hat{y}=a\hat{x}$, is:

$$\begin{pmatrix}\hat{y}_1\\ \hat{y}_2\end{pmatrix} = \frac{1}{w_{11} + w_{22}}\begin{pmatrix}w_{11}y_1 + w_{22}y_2\\ w_{11}y_1 + w_{22}y_2\end{pmatrix},$$

and the WLSQ-estimate of e, $\hat{e}=y - \hat{y}$, is:

$$\begin{pmatrix}\hat{e}_1\\ \hat{e}_2\end{pmatrix} = \frac{1}{w_{11}+w_{22}}\begin{pmatrix}w_{22}(y_1 - y_2)\\ w_{11}(y_2 - y_1)\end{pmatrix}$$

Note that $|\hat{e}_2|>|\hat{e}_1|$ if $w_{11}>w_{22}$.

Finally the weighted sum of squares $\hat{e}^*W\hat{e}$ is:

$$\hat{e}^*W\hat{e} = \frac{w_{11}w_{22}}{w_{11} + w_{22}}(y_1 - y_2)^2.$$

End of example.

In section 1.2 we introduced the least-squares principle by using the intuitively appealing geometric principle of orthogonal projection. In this section we introduced the weighted least-squares principle as the problem of minimizing the weighted sum of squares $(y\text{-}ax)^*W(y\text{-}ax)$. An interesting question is now, whether it is also possible to give a geometric interpretation to the *weighted* least-squares principle. This turns out to be the case. First note that the equation:

$$z^*Wz = \text{constant} = c$$

or:

$$w_{11}z_1^2 + 2w_{12}z_1z_2 + w_{22}z_2^2 = c$$

describes an *ellipse*. If $w_{11}=w_{22}$ and $w_{12}=0$, the ellipse is a circle. If $w_{11}\neq w_{22}$ and $w_{12}=0$, the ellipse has its major and minor axis parallel to the z_1- and z_2- axes, respectively. And if $w_{12}\neq 0$, the ellipse may have an arbitrary orientation (see figure 1.6).

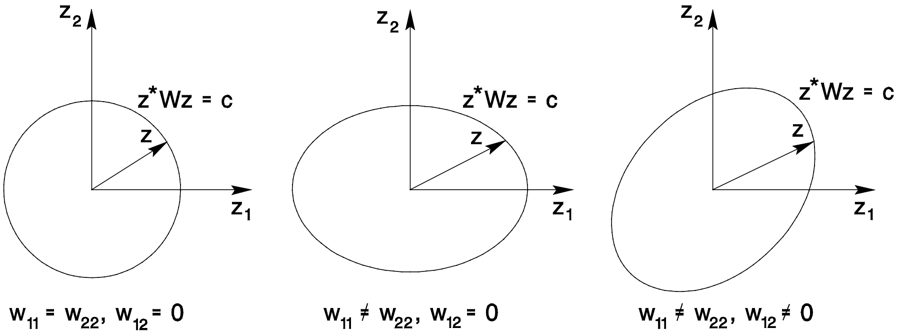
z_2
$z^*Wz = c$
z
z_1
$w_{11} = w_{22}, w_{12} = 0$
$w_{11} \neq w_{22}, w_{12} = 0$
$w_{11} \neq w_{22}, w_{12} \neq 0$

Figure 1.6

If we define a function *F*(*z*) as *F*(*z*)=z^*Wz, then we know from calculus that the *gradient* of *F*(*z*) evaluated at z_0 is a vector which is *normal* to the ellipse *F*(z)=*c* at z_0. The gradient of *F*(*z*) evaluated at z_0 is:

$$\begin{pmatrix} \dfrac{\partial F}{\partial z_1} \\ \dfrac{\partial F}{\partial z_2} \end{pmatrix}_{z_o} = \begin{pmatrix} 2w_{11}z_1 + 2w_{12}z_2 \\ 2w_{12}z_1 + 2w_{22}z_2 \end{pmatrix}_{z_o} = 2\,Wz_o \ .$$

This vector is thus normal or orthogonal to the ellipse at z_0 (see figure 1.7).

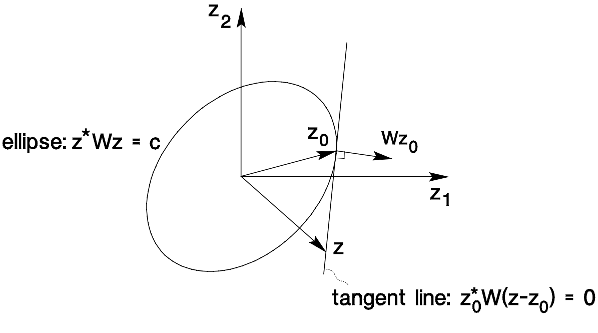

Figure 1.7

From this follows that the tangent line of the ellipse z^*Wz=*c* at z_0 is given by:

$$z_o^*W(z-z_o) = 0 \ .$$

Comparing this evaluation with the first equation of (7) shows that the WLSQ-estimate of x is given by that $\hat{x}$ for which y-$a\hat{x}$ is orthogonal to the normal Wa at a of the ellipse $z^*Wz=a^*Wa$. Hence, y-$a\hat{x}$ is parallel to the tangent line of the ellipse $z^*Wz=a^*Wa$ at a (see figure 1.8). Note that if $W=I$, the ellipse becomes a circle and the unweighted least-squares estimate is obtained through orthogonal projection (see figure 1.9).

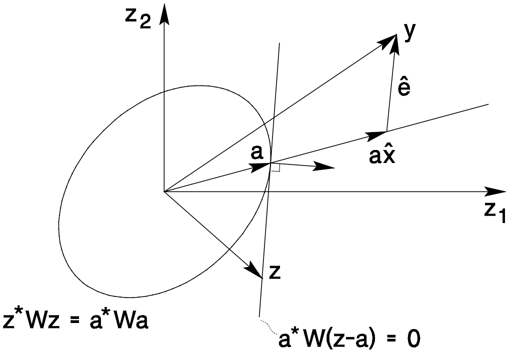

Figure 1.8

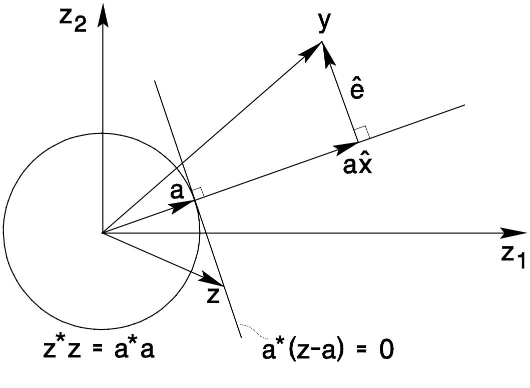

Figure 1.9

The matrices $a(a^*Wa)^{-1}a^*W$ and I-$a(a^*Wa)^{-1}a^*W$ that occur in (7) are known as *oblique projectors*.

We will denote them as:

$$\boxed{\begin{aligned} P_{a,(Wa)^\perp} &= a(a^*Wa)^{-1}a^*W\ , \\ P^\perp_{a,(Wa)^\perp} &= P_{(Wa)^\perp,a} = I\text{-}a(a^*Wa)^{-1}a^*W \end{aligned}} \quad (8)$$

The matrix $\boldsymbol{P}_{a,(Wa)^{\perp}}$ projects *onto* a line spanned by *a* and *along* a line orthogonal to *Wa*.

To see this, note that we can write $P_{a,(Wa)^{\perp}}y$ with the help of the cosine rule as (see figure 1.10):

$$\boldsymbol{P}_{a,(Wa)^{\perp}}y \;=\; a(a^{*}Wa)^{-1}a^{*}Wy \;=\; a\frac{\|Wa\|\|y\|\cos\alpha}{\|Wa\|\|a\|\cos\beta} \;=\; \|y\|\frac{\cos\alpha}{\cos\beta}\frac{a}{\|a\|}$$

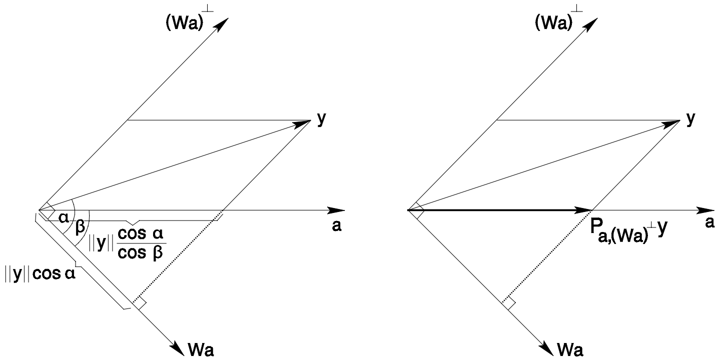

Figure 1.10

In a similar way we can show that $\boldsymbol{P}^{\perp}_{a,(Wa)^{\perp}}$ projects *onto* a line orthogonal to *Wa* and *along* a line spanned by *a* (see figure 1.11).

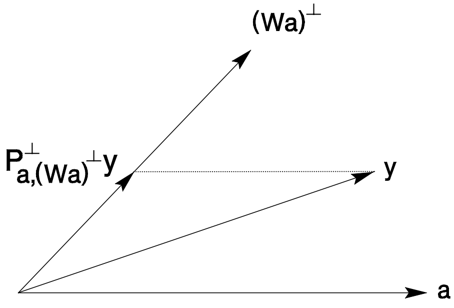

Figure 1.11

Above we have denoted the vector orthogonal to *Wa* as $(Wa)^{\perp}$; thus $((Wa)^{\perp})^{*}(Wa)=0$. With the help of the vector *b*, which is defined in section 1.2 as being orthogonal to *a*, thus $b^{*}a=0$, we can write instead of $(Wa)^{\perp}$ also $W^{-1}b$, because $(W^{-1}b)^{*}(Wa)=0$ (the inverse W^{-1} of *W* exists, since we assumed *W* to be positive-definite). With the help of $W^{-1}b$ we can now represent the two projectors of (8) alternatively as:

(9)
$$\boxed{\begin{aligned} P_{a,(Wa)^{\perp}} &= P_{a,W^{-1}b} = I - W^{-1}b(b^{*}W^{-1}b)^{-1}b^{*} \ , \\ P_{(Wa)^{\perp},a} &= P_{W^{-1}b,a} = W^{-1}b(b^{*}W^{-1}b)^{-1}b^{*} \end{aligned}}$$

To see this, consult figure 1.12 and note that with the help of the cosine rule we can write:

$$P_{W^{-1}b,a}y = \|y\|\frac{\cos\alpha}{\cos\beta}\ \frac{W^{-1}b}{\|W^{-1}b\|} = W^{-1}b\frac{\|y\|\cos\alpha}{\|W^{-1}b\|\cos\beta} = W^{-1}b\frac{b^{*}y}{(W^{-1}b)^{*}b} \ ,$$

or:

$$P_{W^{-1}b,a}y = W^{-1}b(b^{*}W^{-1}b)^{-1}b^{*}y \ .$$

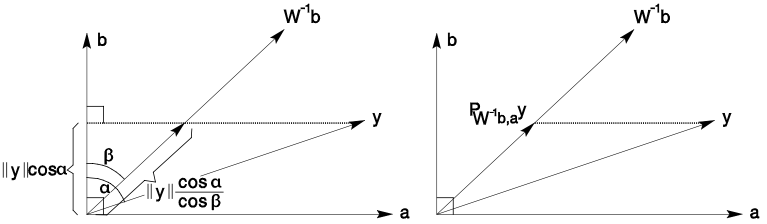

Figure 1.12

Note that if $W{=}I$, then $W^{-1}{=}I$ and equations (8) and (9) reduce to equations (3) and (4) respectively. In section 1.2 we showed that for $W{=}I$ the representations of (9) correspond to solving equation (5) in an unweighted least-squares sense. We will now show that the representations of (9) correspond to solving (5) in a weighted least-squares sense. From the equivalence of the following minimization problems:

$$\min_{x}\ (y-ax)^{*}W(y-ax),$$

$$\min_{z}\ (y-z)^{*}W(y-z) \qquad \text{subject to} \qquad z = ax,$$

$$\min_{z}\ (y-z)^{*}W(y-z) \qquad \text{subject to} \qquad b^{*}z = 0,$$

$$\min_{e}\ (e)^{*}W(e) \qquad \text{subject to} \qquad b^{*}y = b^{*}e,$$

follows that the weighted least-squares estimate of e is given by that $\hat{e}$ on the line $b^{*}y{=}b^{*}e$ for which $e^{*}We$ is minimal. We know that $e^{*}We$ = constant = c is the equation of an ellipse. For different values of c we get different ellipses and the larger the value of c is, the larger the ellipse (see figure 1.13).

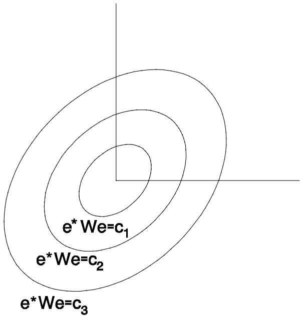

Figure 1.13 $c_3>c_2>c_1$

The solution $\hat{e}$ is therefore given by the *point of tangency* of the line $b^*y=b^*e$ with one of the ellipses $e^*We=c$ (see figure 1.14).

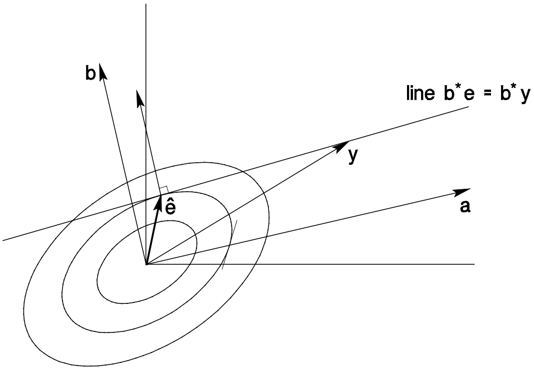

Figure 1.14

At this point of tangency the normal of the ellipse is parallel to the normal of the line $b^*e=b^*y$. But since the normal of the line $b^*e=b^*y$ is given by b, it follows that the normal of the ellipse at $\hat{e}$ is also given by b. From this we can conclude that $\hat{e}$ has to be parallel to $W^{-1}b$. Hence we may write:

$$\hat{e} = W^{-1}b\alpha \ .$$

The unknown scalar α may now be determined from the fact that $\hat{e}$ has to lie on the line $b^*e=b^*y$. Pre-multiplying $\hat{e}=W^{-1}b\alpha$ with b^* gives for α

$$\alpha = (b^*W^{-1}b)^{-1}b^*\hat{e} = (b^*W^{-1}b)^{-1}b^*y \ .$$

Substituting this in the equation $\hat{e}=W^{-1}b\alpha$ shows that the weighted least-squares estimates of equation (5) read:

$$\begin{aligned} \hat{e} &= \boldsymbol{P}_{W^{-1}b,a}y = [W^{-1}b(b^*W^{-1}b)^{-1}b^*]y \\ \hat{y} &= y-\hat{e} = \boldsymbol{P}^{\perp}_{W^{-1}b,a}y = [I-W^{-1}b(b^*W^{-1}b)^{-1}b^*]y \end{aligned} \tag{10}$$

Compare this result with equation (6).

Example 4**:** Consider again the problem defined in section 1.1:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

The form corresponding to (5) is:

$$(1\ -1)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (1\ -1)\begin{pmatrix} e_1 \\ e_2 \end{pmatrix}.$$

thus the vector b is given as:

$$b = (1\ -1)^* \ .$$

As weight matrix we take:

$$W = \begin{pmatrix} w_{11} & 0 \\ 0 & w_{22} \end{pmatrix} .$$

Its inverse reads therefore:

$$W^{-1} = \begin{pmatrix} w_{11}^{-1} & 0 \\ 0 & w_{22}^{-1} \end{pmatrix} .$$

The two projectors $\boldsymbol{P}_{W^{-1}b,a}$ and $\boldsymbol{P}^{\perp}_{W^{-1}b,a}$ can now be computed as:

$$\boldsymbol{P}_{W^{-1}b,a} = W^{-1}b(b^*W^{-1}b)^{-1}b^* = \frac{w_{11}w_{22}}{w_{11}+w_{22}}\begin{pmatrix} w_{11}^{-1} & -w_{11}^{-1} \\ -w_{22}^{-1} & w_{22}^{-1} \end{pmatrix},$$

and

$$P^{\perp}_{W^{-1}b,a} = I - W^{-1}b(b^{*}W^{-1}b)^{-1}b^{*} = \frac{w_{11}w_{22}}{w_{11} + w_{22}} \begin{pmatrix} w_{22}^{-1} & w_{11}^{-1} \\ w_{22}^{-1} & w_{11}^{-1} \end{pmatrix} .$$

The weighted least-squares estimates $\hat{e}$ and $\hat{y}$ read therefore:

$$\begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix} = P_{W^{-1}b,a}y = \frac{1}{w_{11}+w_{22}} \begin{pmatrix} w_{22}(y_1 - y_2) \\ w_{11}(y_2 - y_1) \end{pmatrix} ,$$

and

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = P^{\perp}_{W^{-1}b,a}y = \frac{1}{w_{11}+w_{22}} \begin{pmatrix} w_{11}y_1 + w_{22}y_2 \\ w_{11}y_1 + w_{22}y_2 \end{pmatrix} .$$

These results are identical to the results obtained in example 3. End of example.

1.4 Weighted least-squares is also orthogonal projection

In the previous two sections we have seen that unweighted and weighted least-squares estimation could be interpreted as *orthogonal* and *oblique* projection respectively (see figure 1.15).

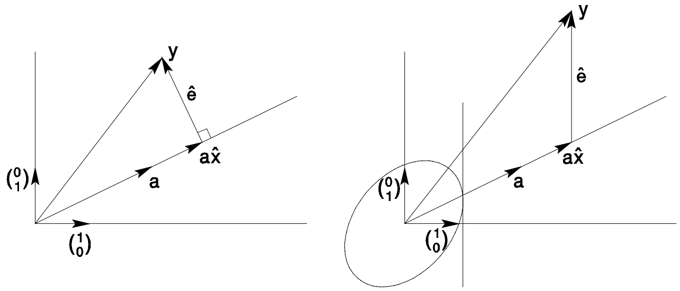

Figure 1.15

It is important to realize however that these geometric interpretations are very much dependent on the choice of base vectors with the help of which the vectors y, a and e are constructed. In all the cases considered so far we have implicitly assumed to be dealing with the orthogonal base vectors $(1,0)^{*}$ and $(0,1)^{*}$. That is, the vectors y, a and e of the equation:

$$y = ax + e,$$

were constructed from the scalars y_1, y_2, a_1, a_2, e_1 and e_2 as:

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

$$(11) \qquad a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = a_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

$$e = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = e_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + e_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Let us now investigate what happens if instead of the orthonormal set $(1,0)^*$ and $(0,1)^*$, a different set, i.e. a non-orthonormal set, of base vectors is chosen. Let us denote these new base vectors as d_1 and d_2. With the help of the base vectors d_1 and d_2, the vectors y, a and e can then be constructed from the scalars y_1, y_2, a_1, a_2, e_1 and e_2 as:

$$y = y_1 d_1 + y_2 d_2 \ ,$$

$$(12) \qquad a = a_1 d_1 + a_2 d_2 \ ,$$

$$e = e_1 d_1 + e_2 d_2 \ .$$

Again the equation $y{=}ax{+}e$ holds. This is shown in figure 1.16a. Note that the vectors of (12) are not the same as those of (11). They do however have the same components.

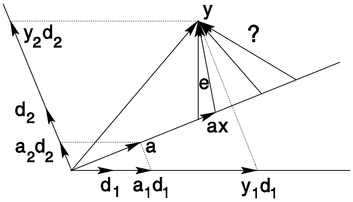

Figure 1.16a

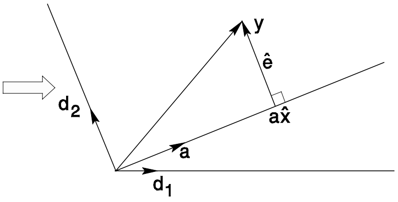

Figure 1.16b

Looking at figure 1.16 it seems again, like in section 1.2, intuitively appealing to estimate x as $\hat{x}$ such that $a\hat{x}$ is as close as possible to given vector y. This is the case if the vector $y\text{-}a\hat{x}$ is orthogonal to the vector a; that is, if:

$$(13) \qquad a^*(y - a\hat{x}) = 0. \qquad \text{"orthogonality"}$$

With (12) this gives:

$$(a_1 d_1 + a_2 d_2)^* \left[y_1 d_1 + y_2 d_2 - (a_1 d_1 + a_2 d_2)\hat{x} \right] = 0,$$

or

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^* \begin{pmatrix} d_1^* d_1 & d_1^* d_2 \\ d_2^* d_1 & d_2^* d_2 \end{pmatrix} \left[\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \hat{x} \right] = 0,$$

or

(14)
$$\boxed{\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^* D^*D \left[\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \hat{x} \right] = 0,}$$

with

$$D = (d_1 \; d_2).$$

The matrix D^*D of (14) is symmetric, since $(D^*D)^*=D^*D$, and it is positive-definite, since $z^*D^*Dz = (Dz)^*(Dz)>0$ for all $z\neq 0$. Hence, the matrix D^*D may be interpreted as a weight matrix. By interpreting the matrix D^*D of (14) as a weight matrix we see that the solution of (13) can in fact be interpreted as a *weighted least-squares* solution, i.e. a solution of the minimization problem:

$$\min_x \; [\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} x]^* \; D^*D \; [\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} x].$$

We have thus shown that if in a weighted least-squares problem the weight matrix W can be written as the product of a matrix D^* and its transpose D, then the weighted least-squares estimate $\hat{y}=a\hat{x}$ can be interpreted as being obtained through *orthogonal* projection of $y=y_1d_1+y_2d_2$ onto the line with direction vector $a=a_1d_1+a_2d_2$, where d_1 and d_2 are the column vectors of matrix D (see figure 1.17).

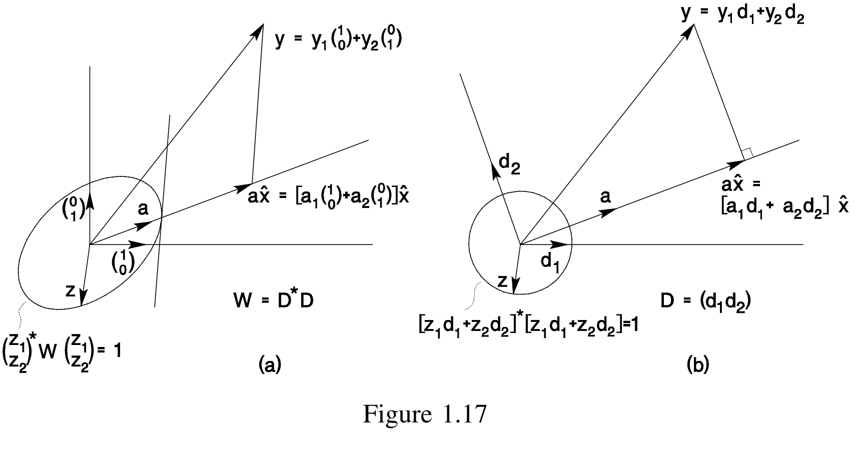

Figure 1.17

The crucial question is now whether *every* weight matrix W, i.e. every symmetric and positive-definite matrix, can be written as a product of a matrix and its transpose, i.e. as $W=D^*D$. If this is the case, then every weighted least-squares problem can be interpreted as a problem of orthogonal projection! The answer is: *yes*, every weight matrix W can be written as $W=D^*D$. For a proof, see the *intermezzo* at the end of this section.

The weight matrix $W{=}D^*D$ can be interpreted as defining an *inner product* (Strang, 1988). In the unweighted case, $W{=}I$, the inner product of two vectors u and v with components u_1,u_2 and v_1,v_2 respectively, is given as:

$$(u,v)_I = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}^* I \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \left[u_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right]^* \left[v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right] .$$

This is the so-called *standard inner product*, where the inner product is represented by the unit matrix I. The square of the length of the vector u reads then:

$$\|u\|_I^2 = u_1^2 + u_2^2 .$$

In the weighted case, $W{=}D^*D$, the inner product of u and v is given as:

(15)
$$(u,v)_W = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}^* D^*D \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \left(u_1 d_1 + u_2 d_2\right)^* \left(v_1 d_1 + v_2 d_2\right) .$$

The square of the length of the vector u reads then:

$$\|u\|_W^2 = u_1^2 d_1^* d_1 + 2 u_1 u_2 d_1^* d_2 + u_2^2 d_2^* d_2 .$$

Note that the two ways in which $(u,v)_W$ is written in (15) correspond with the two figures 1.17a and 1.17b respectively. In figure 1.17a orthonormal base vectors are used and the metric is *visualized* by the ellipse:

$$(z,z)_W = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}^* W \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = 1 .$$

In figure 1.17b however, the ellipse is replaced by a circle:

$$(z,z)_W = \left(z_1 d_1 + z_2 d_2\right)^* I \left(z_1 d_1 + z_2 d_2\right) = 1 ,$$

and the metric is *visualized* by the non-orthonormal base vectors d_1 and d_2.

Now that we have established the fact that weighted least-squares estimation can be interpreted as orthogonal projection, it must also be possible to interpret the oblique projectors of equation (8) as orthogonal projectors. We will show that the matrix $P_{a,(Wa)^\perp}$ of (8) can be interpreted as the matrix which projects orthogonally onto a line spanned by the vector a, where orthogonality is measured with respect to the inner product defining weight matrix W. From figure 1.18 follows that:

$$\hat{y} = \|y\|_W \cos\alpha \frac{a}{\|a\|_W} = a \|a\|_W^{-2} (a,y)_W ,$$

or

$$\hat{y}_1 d_1 + \hat{y}_2 d_2 = (a_1 d_1 + a_2 d_2)\left[\begin{pmatrix}a_1\\a_2\end{pmatrix}^* W \begin{pmatrix}a_1\\a_2\end{pmatrix}\right]^{-1} \begin{pmatrix}a_1\\a_2\end{pmatrix}^* W \begin{pmatrix}y_1\\y_2\end{pmatrix},$$

or

$$\begin{pmatrix}\hat{y}_1\\\hat{y}_2\end{pmatrix} = \begin{pmatrix}a_1\\a_2\end{pmatrix}\left[\begin{pmatrix}a_1\\a_2\end{pmatrix}^* W \begin{pmatrix}a_1\\a_2\end{pmatrix}\right]^{-1} \begin{pmatrix}a_1\\a_2\end{pmatrix}^* W \begin{pmatrix}y_1\\y_2\end{pmatrix},$$

and in this expression we recognize indeed the matrix $P^{\perp}_{a,(Wa)}$ of equation (8). From now on we will assume that it is clear from the context which inner-product is taken and therefore also denote the matrix $P^{\perp}_{a,(Wa)}$ as P_a.

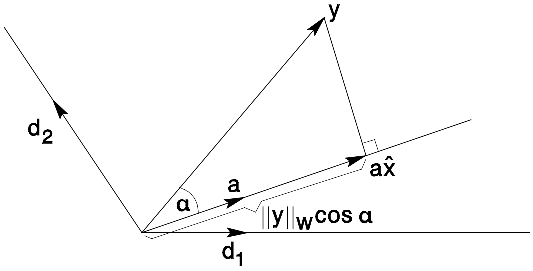

Figure 1.18

***Intermezzo: Cholesky decomposition*:**

If W is an $n \times n$ symmetric positive definite matrix, there is a nonsingular lower triangular matrix L with positive diagonal elements such that $W=LL^*$.

***Proof*:** The proof is by induction. Clearly, the result is true for n=1 and the induction hypothesis is that it is true for any (n-1)×(n-1) symmetric positive definite matrix.

Let

$$W = \begin{pmatrix} A & a \\ a^* & b_{nn} \end{pmatrix},$$

where A is a (n-1)×(n-1) matrix. Since W is symmetric positive definite, it will be clear that A is also symmetric positive definite. Thus, by the induction hypothesis, there is a nonsingular lower triangular matrix $L_{n\text{-}1}$ with positive diagonal elements such that $A=L_{n\text{-}1}L^*_{n\text{-}1}$.

Let

$$L = \begin{pmatrix} L_{n-1} & 0 \\ l^* & \beta \end{pmatrix},$$

where l and β are to be obtained by equating LL^* to W:

$$LL^* = \begin{pmatrix} L_{n-1} & 0 \\ l^* & \beta \end{pmatrix}\begin{pmatrix} L^*_{n-1} & l \\ 0 & \beta \end{pmatrix} = \begin{pmatrix} A & a \\ a^* & b_{nn} \end{pmatrix} = W.$$

This implies that l and β must satisfy:

$$L_{n-1}l = a \quad , \quad l^*l + \beta^2 = b_{nn}.$$

Since $L_{n\text{-}1}$ is nonsingular, the first equation determines l as $l{=}L^{-1}{}_{n\text{-}1}\, a$. The second equation allows a positive β provided that b_{nn} - $l^*l{>}0$. In order to proof that b_{nn} - $l^*l{>}0$ we make use of the fact that W is positive definite, i.e.:

$$x^*Wx = (x_1^*\, x_2)\begin{pmatrix} A & a \\ a^* & b_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1^*\, x_2)\begin{pmatrix} L_{n-1}L^*_{n-1} & a \\ a^* & b_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} > 0 \quad \forall x \neq 0$$

Substitution of $x_1 = -(L^*_{n-1})^{-1}l$ and $x_2{=}1$ shows that b_{nn} - l^*l >0.

End of proof.

A simple example of a Cholesky decomposition is:

$$\begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{3} \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 0 & \sqrt{3} \end{pmatrix}.$$

1.5 The mean and covariance matrix of least-squares estimators

From now on we shall make no distinction between unweighted and weighted least-squares estimation, and just speak of least-squares estimation. From the context will be clear whether $W{=}I$ or not.

In our discussions so far no assumptions regarding the probabilistic nature of the observations or measurement errors were made. We started by writing the two equations:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} x + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

in vector notation as:

$$y = ax + e,$$

and showed that the least-squares estimates of x, y and e follow from solving the minimization problem:

$$\min_x \; (y-ax)^* W (y-ax)$$

as:

(16)
$$\begin{aligned} \hat{x} &= (a^*Wa)^{-1}a^*Wy, \\ \hat{y} &= a\hat{x}, \\ \hat{e} &= y - a\hat{x}. \end{aligned}$$

Vector $\hat{e}$ contains the least-squares residuals. Since functions of random variables are again random variables, it follows that if the observation vector is assumed to be a random vector $\underline{y}$ (note: the underscore indicates that we are dealing with random variables) and substituted for y in (16), the left-hand sides of (16) also become random variables:

(17)
$$\boxed{\begin{aligned} \underline{\hat{x}} &= (a^*Wa)^{-1}a^*W\underline{y}, \\ \underline{\hat{y}} &= a\underline{\hat{x}}, \\ \underline{\hat{e}} &= \underline{y} - a\underline{\hat{x}} \end{aligned}}$$

These random variables are called least-squares *estimators*, whereas if $\underline{y}$ is replaced by its sample value y one speaks of least-squares *estimates*.

From your lecture notes in statistics you know how to derive the probability density functions of the estimators of (17) from the probability density function $p_{\underline{y}}(y)$ of $\underline{y}$. You also know that if the probability density function $p_{\underline{y}}(y)$ of $\underline{y}$ is *"normal"* or *"Gaussian"*, then also $\underline{\hat{x}}$, $\underline{\hat{y}}$ and $\underline{\hat{e}}$ are normally distributed. This follows, since $\underline{\hat{x}}$, $\underline{\hat{y}}$ and $\underline{\hat{e}}$ are *linear functions of* $\underline{y}$. You know furthermore how to compute the means and covariance matrices of $\underline{y}$. This is particularly straightforward in case of linear functions (see also appendix E and F).

Application of the *"propagation law of means"* to (17) gives:

(18)
$$\begin{aligned} E\{\underline{\hat{x}}\} &= (a^*Wa)^{-1}a^*WE\{\underline{y}\} \\ E\{\underline{\hat{y}}\} &= aE\{\underline{\hat{x}}\} = P_aE\{\underline{y}\} \\ E\{\underline{\hat{e}}\} &= E\{\underline{y}\} - aE\{\underline{\hat{x}}\} = (I-P_a)E\{\underline{y}\} = P_a^{\perp}E\{\underline{y}\} \end{aligned}$$

where $E\{.\}$ denotes *mathematical expectation*. The result (18) holds true irrespective the type of probability distribution one assumes for $\underline{y}$. Let us now, without assuming anything about the type of the probability distribution of $\underline{y}$, be a bit more specific about the mean $E\{\underline{y}\}$ of $\underline{y}$. We will assume that the randomness of the observation vector $\underline{y}$ is a consequence of the random nature of the measurement errors. Thus we assume that $\underline{y}$ can be written as the sum of a deterministic, i.e. non-random, component ax and a random component $\underline{e}$:

(19)
$$\boxed{\underline{y} = ax + \underline{e}}$$

Furthermore we assume that the mean $E\{\underline{e}\}$ of $\underline{e}$ is zero. This assumption seems acceptable if the measurement errors are zero "on the average".

With (19) the assumption $E\{\underline{e}\}=0$ implies that:

(20)
$$\boxed{E\{\underline{y}\} = ax}$$

Substitution of (20) into (18) gives then:

(21)
$$\boxed{\begin{aligned} E\{\underline{\hat{x}}\} &= x \\ E\{\underline{\hat{y}}\} &= ax = E\{\underline{y}\} \\ E\{\underline{\hat{e}}\} &= E\{\underline{y}\}-ax = 0 \end{aligned}}$$

This important result shows that if (19) with (20) holds, the least-squares estimators $\underline{\hat{x}}$, $\underline{\hat{y}}$ and $\underline{\hat{e}}$ are *unbiased* estimators. Note that this unbiasedness property of least-squares estimators is independent of the choice for the weight matrix W ! Let us now derive the covariance matrices of the estimators $\underline{\hat{x}}$, $\underline{\hat{y}}$ and $\underline{\hat{e}}$. We will denote the covariance matrix of $\underline{y}$ as Q_y, the variance of $\underline{\hat{x}}$ as $\sigma^2_{\hat{x}}$ and the covariance matrices of $\underline{\hat{y}}$ and $\underline{\hat{e}}$ as $Q_{\hat{y}}$ and $Q_{\hat{e}}$ respectively. Application of the *"propagation law of covariances"* to (17) gives:

(22)
$$\boxed{\begin{aligned} \sigma^2_{\hat{x}} &= (a^*Wa)^{-1}a^*WQ_yWa(a^*Wa)^{-1} \\ Q_{\hat{y}} &= a\sigma^2_{\hat{x}}a^* = P_aQ_yP_a^* \\ Q_{\hat{e}} &= Q_y - aQ_{\hat{x}y} - Q_{y\hat{x}}a^* + a\sigma^2_{\hat{x}}a^* \\ & \quad Q_y - P_aQ_y - Q_yP_a^* + P_aQ_yP_a^* \end{aligned}}$$

As you know from your lecture notes in Statistics, the variance of a random variable is a measure for the spread of the probability density function around the mean value of the random variable. It is a measure of the variability one can expect to see when independent samples are drawn from the probability distribution. It is therefore in general desirable to have estimators with small variances. This is particularly true for our least-squares estimators. Together with the property of unbiasedness, a small variance would namely imply that there is a high probability that a sample from the distribution of $\underline{\hat{x}}$ will be close to the true, but unknown, value x. Since we left the choice for the weight matrix W open up till now, let us investigate what particular weight matrix makes $\sigma^2_{\hat{x}}$ minimal. Using the decomposition $Q_y=D^*D$, it follows from (22) that $\sigma^2_{\hat{x}}$ can be written as:

$$\sigma^2_{\hat{x}} = \frac{a^*WD^*DWa}{(a^*D^{-1}DWa)^2} = \frac{\|DWa\|^2}{\|DWa\|^2\|(D^*)^{-1}a\|^2\cos^2\alpha} = \frac{1}{\|(D^*)^{-1}a\|^2\cos^2\alpha}$$

where α is the angle between the two vectors DWa and $(D^*)^{-1}a$. From this result follows that the variance $\sigma^2_{\hat{x}}$ is minimal when the angle α is zero; that is, when the two vectors DWa and $(D^*)^{-1}a$ are parallel. The variance $\sigma^2_{\hat{x}}$ is therefore minimal when $DWa=(D^*)^{-1}a$ or $D^*DWa=a$ or $Q_yWa=a$. Hence, the variance is minimal when we choose the weight matrix W as:

(23)
$$\boxed{W = Q_y^{-1}}$$

With this choice for the weight matrix, the variance $\sigma^2_{\hat{x}}$ becomes:

(24)
$$\sigma^2_{\hat{x}} = (a^* Q_y^{-1} a)^{-1} .$$

1.6 Best Linear Unbiased Estimators (BLUE's)

Up till now we have been dealing with least-squares estimation. The least-squares estimates were derived from the *deterministic* principles of "orthogonality" and "minimum distance". That is, no probabilistic considerations were taken into account in the derivation of least-squares estimates.
Nevertheless, it was established in the previous section that least-squares estimators do have some "optimal properties" in a probabilistic sense. They are *unbiased estimators* independent of the choice for the weight matrix W, and their *variance can be minimized* by choosing the weight matrix as the inverse of the covariance matrix of the observations. Moreover, they are *linear* estimators, i.e. linear functions of the observations.

Estimators which are linear, unbiased and have minimum variance are called best linear unbiased estimators or simply BLUE's. "Best" in this case means minimum variance. Thus least-squares estimators are BLUE's in case the weight matrix equals the inverse of the covariance matrix of the observations. In order to find out how large the class of best linear unbiased estimators is (is there only one? or are there more than one?), we will now derive this class. Thus instead of starting from the least-squares principle of "orthogonality" we start from the conditions of linearity, unbiasedness and best. Again we assume that the mean of $\underline{y}$ satisfies:

(25)
$$E\{\underline{y}\} = ax .$$

Our problem is to find an estimator $\underline{\hat{x}}$ of the unknown parameter x, where $\underline{\hat{x}}$ is a function of $\underline{y}$. Since we want the estimator $\underline{\hat{x}}$ to be a *linear* function of $\underline{y}$, $\underline{\hat{x}}$ must be of the form:

(26) $\underline{\hat{x}} = l^* \underline{y}$ "L-property"

Furthermore, we want $\underline{\hat{x}}$ to be an *unbiased* estimator of x for all the values x may take. This means that:

(27) $E\{\underline{\hat{x}}\} = x \ , \ \forall x$ "U-property"

And finally, we want to find those estimators $\underline{\hat{x}}$, which in the class of estimators that satisfy (26) and (27), have *minimum variance:*

(28) $\sigma^2_{\hat{x}}$ minimal in the class of LU-estimators "B-property"

It will be clear that the class of linear estimators is infinite. Let us therefore look at the class of linear and unbiased estimators.

From taking the expectation of (26) follows, with (25) and (27) that:

$$E\{\underline{\hat{x}}\} = l^{*}E\{\underline{y}\} = l^{*}ax = x \quad \forall x$$

or

(29) $$\boxed{l^{*}a = 1}$$

Thus for LU-estimators the vector l has to satisfy (29) (see figure 1.19).

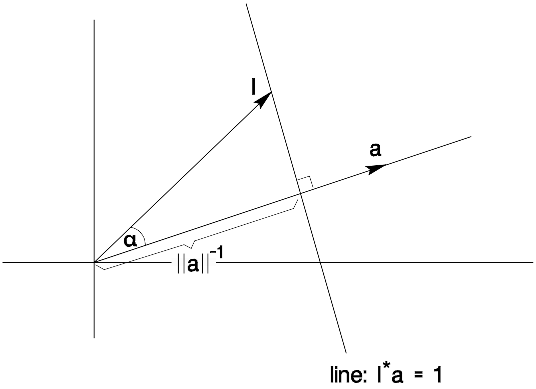

Figure 1.19: $l^{*}a = \|l\|\|a\|\cos\alpha = 1$

It will be clear that, although the class of permissable vectors l has been narrowed down to the line $l^{*}a$=1, the class of LU-estimators is still infinite. For instance, every vector l of the form:

$$l = c(a^{*}c)^{-1} \ ,$$

where c is arbitrary, is permitted.

Let us now include the "B-property". The variance of the estimator $\underline{\hat{x}}$ follows from applying the "propagation law of variances" to (26). This gives:

(30) $$\sigma_{\hat{x}}^{2} = l^{*}Q_{y}l$$

The "B-property" implies now that we have to find the vectors l, let us denote them by $\hat{l}$, that minimize $\sigma_{\hat{x}}^{2}$ of (30) and at the same time satisfy (29). We thus have the minimization problem:

(31) $$\boxed{\min_{l}\ l^{*}Q_{y}l \quad \text{subject to} \quad l^{*}a = 1}$$

With the decomposition Q_y=$D^{*}D$, this can also be written as:

$$\min_{l}\ (Dl)^{*}(Dl) \quad \text{subject to} \quad (Dl)^{*}(D^{*})^{-1}a = 1$$

or as:

(32) $$\min_{\bar{l}} \ \bar{l}^*\bar{l} \quad \text{subject to} \quad \bar{l}^*(D^*)^{-1}a = 1 \ ,$$

where:

(33) $$\bar{l} = Dl.$$

Geometrically, the minimization problem (32) boils down to finding those vectors $\hat{\bar{l}}$ whose endpoints lie along the line $\bar{l}^*(D^*)^{-1}a=1$ and who have minimum length (see figure 1.20).

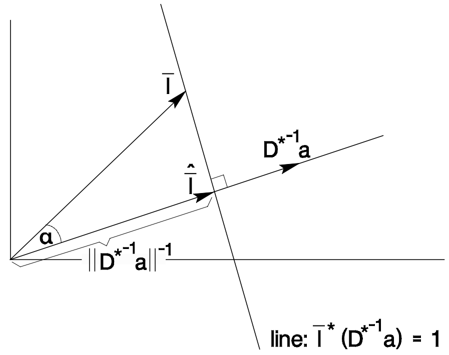

Figure 1.20: $\bar{l}^*(D^*)^{-1}a = \|\bar{l}\|\ \|(D^*)^{-1}a\| \cos\alpha=1$

From figure 1.20 it is clear that there is only one such vector $\hat{\bar{l}}$. This vector lies along vector $D^{*-1}a$ and has length $\|(D^*)^{-1}a\|^{-1}$. Thus:

(34) $$\hat{\bar{l}} = (D^*)^{-1}a\beta \ ,$$

where the positive scalar β follows from:

$$\hat{\bar{l}}^*\ \hat{\bar{l}} = \|(D^*)^{-1}a\|^2\ \beta^2 = \|(D^*)^{-1}a\|^{-2}$$

as:

(35) $$\beta = \|(D^*)^{-1}a\|^{-2}$$

Substitution of (35) into (34) gives with (33):

$$\hat{l} = D^{-1}(D^*)^{-1}a\ \left[a^*D^{-1}(D^*)^{-1}a\right]^{-1}$$

or:

(36) $$\boxed{\hat{l} = Q_y^{-1}a\ (a^*Q_y^{-1}a)^{-1}}$$

The best linear unbiased estimator $\hat{\underline{x}}$ of x follows therefore with (26) as:

(37) $$\hat{\underline{x}} = (a^*Q_y^{-1}a)^{-1}\ a^*Q_y^{-1}\underline{y}$$

From this result we can draw the important conclusion that the best linear unbiased estimator is *unique* and equal to the least-squares estimator if $W=Q_y^{-1}$!

We will now give an *alternative* derivation of the solution (37). Equation (29) is the equation of a line with a direction vector orthogonal to a (see figure 1.20a). Since the vector b was defined as being orthogonal to a, $b^*a=0$, the line $a^*l=1$ can be represented parametrically as:

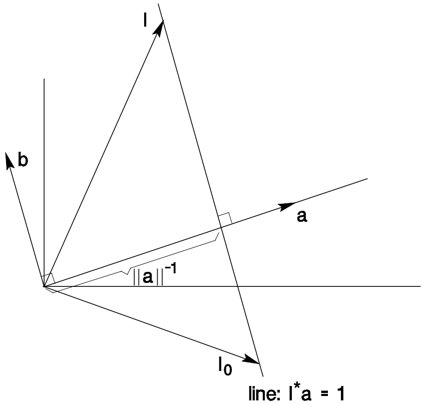

Figure 1.20a: $l^*a=1$ is equivalent to $l=l_o+b\alpha$

(37a)
$$\underset{2\times1}{l} = \underset{2\times1}{l_o} + \underset{2\times1}{b}\ \underset{1\times1}{\alpha} \ ,$$

where l_o is a particular solution of (29). Hence, the condition $a^*l=1$ in (31) may be replaced by $l=l_o+b\alpha$:

$$\underset{l}{\text{min.}}\ l^*Q_y l \quad \text{subject to} \quad l = l_o + b\alpha$$

This may also be written as:

$$\underset{\alpha}{\text{min.}}\ (l_o + b\alpha)^*Q_y(l_o + b\alpha).$$

The solution of this minimization problem is:

(37b)
$$\hat{\alpha} = -(b^*Q_y b)^{-1}b^*Q_y l_o .$$

This is easily verified by using an algebraic proof like the one given in section 1.2. Substitution of (37b) into (37a) gives:

$$\hat{l} = \left[I - b(b^*Q_y b)^{-1}b^*Q_y\right]l_o$$

The best linear unbiased estimator $\underline{\hat{x}}$ of x follows therefore with (26) as:

(37c)
$$\underline{\hat{x}} = l_o^* \left[I - Q_y b (b^* Q_y b)^{-1} b^* \right] \underline{y} .$$

From equations (8) and (9) in section 1.3 we know that:

$$I - Q_y b (b^* Q_y b)^{-1} b^* \; = \; a (a^* Q_y^{-1} a)^{-1} a^* Q_y^{-1} .$$

Substitution in (37c) gives therefore:

$$\underline{\hat{x}} = l_o^* a (a^* Q_y^{-1} a)^{-1} a^* Q_y^{-1} \underline{y}$$

or:

$$\underline{\hat{x}} = (a^* Q_y^{-1} a)^{-1} a^* Q_y^{-1} \underline{y} ,$$

since $l_o^* a=1$. This concludes our alternative derivation of (37).

1.7 The Maximum Likelihood (ML) method

In the previous sections we have seen the two principles of *least-squares* estimation and *best linear unbiased* estimation at work. The least-squares principle is a deterministic principle where no probability assumptions concerning the vector of observations is made. In case $y=ax+e$, the least-squares principle is based on minimizing the quadratic form $(y\text{-}ax)^* \, W(y\text{-}ax)$, with the weight matrix W.

In contrast with the least-squares principle, the principle of best linear unbiased estimation does make use of some probability assumptions. In particular it is based on assumptions concerning the mean and covariance matrix of the random vector of observables $\underline{y}$. However, only the mean and covariance matrix of $\underline{y}$ need to be specified, not its complete probability distribution. We will now present a method of estimation that in contrast with the BLUE's method needs the complete probability distribution of $\underline{y}$.

Let $p(y;x)$ denote the probability density function of $\underline{y}$. For a fixed value of x the function $p(y;x)$ is a function of y and for different values of x the function $p(y;x)$ may take different forms (see figure 1.21).

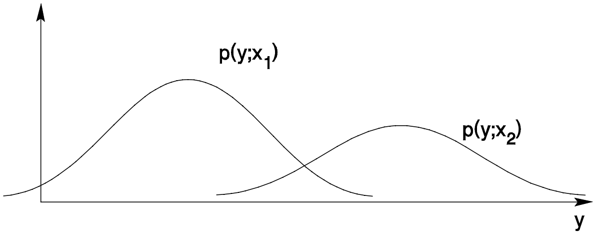

Figure 1.21: Two one-dimensional probability density functions $p(y;x_1)$, $p(y;x_2)$

Our objective is now to determine or estimate the parameter x on the basis of the observation vector y. That is, we wish to determine the correct value of x that produced the observed y. This suggests considering for each possible x how probable the observed y would be if x were the true value. The higher this probability, the more one is attracted to the explanation that the x in question produced y, and the more likely the value of x appears. Therefore, the expression $p(y;x)$ considered for fixed y as a function of x has been called the *likelihood* of x. Note that the likelihood of x is a probability density or a differential probability. The estimation principle is now to maximize the likelihood of x, given the observed value of y:

(38)
$$\boxed{\max_{x}.\ p(y;x)}$$

Thus in the *method of maximum likelihood* we choose as an estimate of x the value which maximizes $p(y;x)$ for the given observed y. Note that in the ML-method one needs to know the complete mathematical expression for the probability density function $p(y;x)$ of $\underline{y}$. Thus it does not suffice, as is the case with the BLUE's method, to specify only the mean and covariance matrix of $\underline{y}$.

As an important example we consider the case that $p(y;x)$ is the probability density function of the normal distribution. Let us therefore assume that $\underline{y}$ is normally distributed with mean $E\{\underline{y}\}=ax$ and covariance matrix Q_y. The probability density function of $\underline{y}$ reads then in case it is a two-dimensional vector:

(39)
$$p(y;x) = (2\pi)^{-1}|Q_y|^{-1/2}\ \exp[-\tfrac{1}{2}(y-ax)^*Q_y^{-1}(y-ax)].$$

As it is often more convenient to work with ln $p(y;x)$ then with $p(y;x)$ itself, the required value for x is found by solving:

(40)
$$\max_{x}.\ \ln p(y;x).$$

this is permitted since ln $p(y;x)$ and $p(y;x)$ have their maxima at the same values of x. Taking the logarithm of (39) gives:

$$\ln p(y;x) = -\ln(2\pi) - \tfrac{1}{2}\ln|Q_y| - \tfrac{1}{2}(y-ax)^* Q_y^{-1}(y-ax). \tag{41}$$

Since the first two terms on the right-hand side of (41) are constant, it follows that maximization of ln $p(y;x)$ amounts to maximization of $-1/2(y\text{-}ax)^* Q_y^{-1}(y\text{-}ax)$. But this is equivalent to minimization of $(y\text{-}ax)^* Q_y^{-1}(y\text{-}ax)$. Hence we have obtained the important conclusion that maximum likelihood estimators, in case $p(y;x)$ is the normal distribution, are identical to least-squares estimators for which the weight matrix is $W{=}Q_y^{-1}$!

Let us now exemplify the above discussion geometrically. From (39) follows that the contours of constant probability density are ellipses:

$$p(y;x) = \text{constant} \;\Leftrightarrow\; (y-ax)^* Q_y^{-1}(y-ax) = \text{constant} = c.$$

For different values of c we get different ellipses. The *smaller* the probability density, the *larger* the value of c and the larger the ellipses are (see figure 1.22).

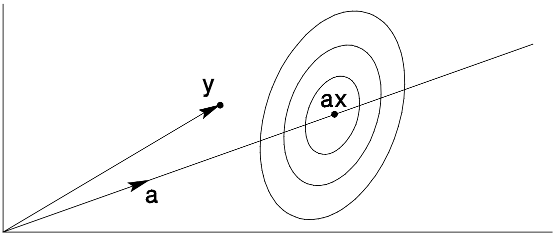

Figure 1.22: Contours of $p(y;x)$ = constant

By varying the values of x the family of contour lines centred at ax translate along the line with direction vector a (see figure 1.23).

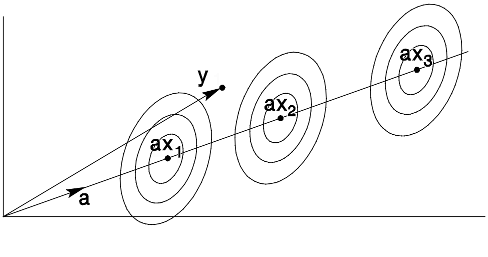

Figure 1.23

From the figure it is clear that the likelihood of x obtains its maximum there where the observation point y is a point of tangency of one of the ellipses (see figure 1.24).

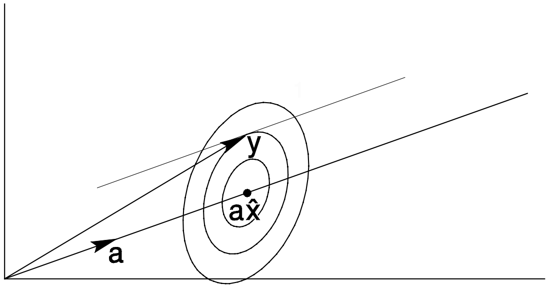

Figure 1.24: The ML-estimate $\hat{x}$

Hence, after translating the family of ellipses along the line with direction vector a such that y becomes a point of tangency, the corresponding centre $a\hat{x}$ of the ellipses gives the maximum likelihood estimate $\hat{x}$. At the point of tangency y, the normal to the ellipse is $Q_y^{-1}(y\text{-}a\hat{x})$ (see figure 1.25).

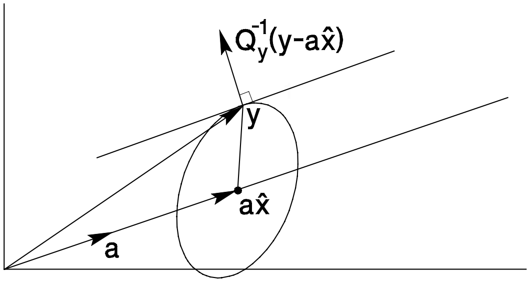

Figure 1.25: The normal $Q_y^{-1}(y\text{-}a\hat{x})$

The normal $Q_y^{-1}(y\text{-}a\hat{x})$ is orthogonal to the tangent line at y. But this tangent line is parallel to the line with direction vector a. Thus we have $a^*Q_y^{-1}(y\text{-}a\hat{x})=0$, which shows once again that in case of the normal distribution, ML-estimators and LSQ-estimators with $W=Q_y^{-1}$ are identical!

1.8 Summary

In this chapter we have discussed three different methods of estimation: the *LSQ*-method, the *BLUE*-method and the *ML*-method. For an overview of the results see table 1. The LSQ-method, being a deterministic method, was applied to the model:

(42)
$$\underset{2\times1}{y} = \underset{2\times1}{a}\ \underset{1\times1}{x} + \underset{2\times1}{e}\ .$$

In case the measurement errors are considered to be random, $\underline{e}$, the observables also become random, $\underline{y}$, and (42) has to be written as:

(43) $$\underline{y} = ax + \underline{e} .$$

With the assumptions that $E\{\underline{e}\}=0$ and $E\{\underline{e}\underline{e}^*\}=Q_y$, we get with (43) the model:

(44) $$E\{\underline{y}\} = ax \;\; ; \;\; E\{(\underline{y} - ax)(\underline{y} - ax)^*\} \; = \; Q_y \; .$$

These two assumption were the starting point of the BLUE-method. If in addition to (44), it is also assumed that $\underline{y}$ is normally distributed then the model becomes:

(45) $$\underline{y} \sim N(ax,Q_y)$$

It was shown that for this model the ML-estimator is identical to the BLU-estimator and the LSQ-estimator if $W=Q_y^{-1}$.

We have restricted ourselves in this introductory chapter to the case of two observations y_1, y_2 and one unknown x. In the next chapter we will start to generalize the theory to higher dimensions. That is, in the next chapter we will start developing the theory for m observations y_1, y_2,...,y_m and n unknowns x_1, x_2,...,x_n. In that case model (45) generalizes to:

(46) $$\underset{m\times 1}{\underline{y}} \sim N(\underset{m\times n}{A} \; \underset{n\times 1}{x} \; , \underset{m\times m}{Q_y}) ,$$

where A is an $m \times n$ matrix, i.e. it has m rows and n columns.

Least Squares Estimation (LSQ)	Best Linear Unbiased Estimation (BLUE)	Maximum Likelihood Estimation (ML)
model: $\boxed{y = ax + e}$ *estimation principle*: $\min_{x} (y-ax)^{*}W(y-ax)$ *solution:* $\hat{x} = (a^{*}Wa)^{-1}a^{*}Wy$ $\hat{y} = a\hat{x} = \boldsymbol{P}_a y$ $\hat{e} = y - a\hat{x} = \boldsymbol{P}_a^{\perp} y$ $\hat{e}^{*}W\hat{e} = y^{*}W\boldsymbol{P}_a^{\perp} y$ $\boldsymbol{P}_a = a(a^{*}Wa)^{-1}a^{*}W$ $\boldsymbol{P}_a^{\perp} = I - \boldsymbol{P}_a$	*model*: $\boxed{\begin{array}{c} \underline{y} = ax + \underline{e} \\ E\{\underline{e}\} = 0 \; ; \; E\{\underline{e}\,\underline{e}^{*}\} = Q_y \end{array}}$ or $\boxed{E\{\underline{y}\} = ax \; ; \; E\{(\underline{y}-ax)(\underline{y}-ax)^{*}\} = Q_y}$ *estimation principle*: L: $\underline{\hat{x}} = l^{*}\underline{y}$; U: $E\{\underline{\hat{x}}\} = x$; $\boldsymbol{B}$: $\sigma_{\hat{x}}^{2}$ min. *solution*: $\underline{\hat{x}} = (a^{*}Q_y^{-1}a)^{-1}a^{*}Q_y^{-1}\underline{y}$; $\sigma_{\hat{x}}^{2} = (a^{*}Q_y^{-1}a)^{-1}$ $\underline{\hat{y}} = a\underline{\hat{x}} = \boldsymbol{P}_a\underline{y}$; $Q_{\hat{y}} = \boldsymbol{P}_a Q_y$ $\underline{\hat{e}} = y - a\underline{\hat{x}} = \boldsymbol{P}_a^{\perp}\underline{y}$; $Q_{\hat{e}} = \boldsymbol{P}_a^{\perp} Q_y$ $\underline{\hat{e}}^{*}Q_y^{-1}\underline{\hat{e}} = \underline{y}^{*}Q_y^{-1}\boldsymbol{P}_a^{\perp}\underline{y}$; $\boldsymbol{P}_a = a(a^{*}Q_y^{-1}a)^{-1}a^{*}Q_y^{-1}$ $\boldsymbol{P}_a^{\perp} = I - \boldsymbol{P}_a$	*model*: $\boxed{\underline{y} \sim p(y;x)}$ *estimation principle*: $\max_{x} p(y;x)$ In case $\boxed{\underline{y} \sim N(ax,Q_y)}$ the solution is identical to BLUE and to LSQ if $W = Q_y^{-1}$

Table 1.1