Delft University of Technology

$$f(x) = \frac{1}{2\pi|Q_{xx}|^{1/2}} e^{-\frac{1}{2}x^T Q_{xx}^{-1} x}$$

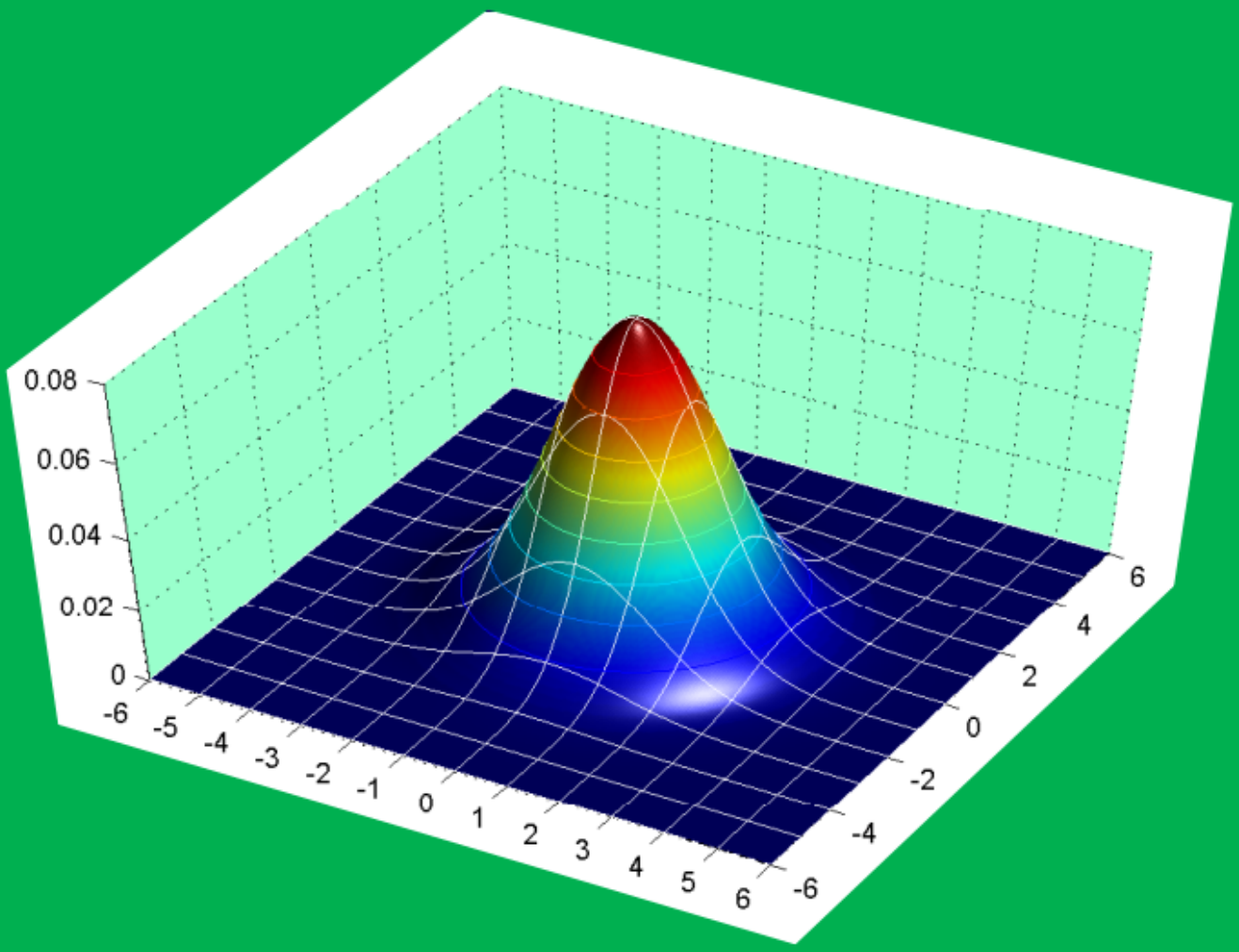
$$\hat{\underline{x}} = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} \underline{y}$$

Primer on Mathematical Geodesy

CTB3310 / CTB3425

Christian Tiberius

$$\hat{\sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{x})^2$$



TU Delft Delft University of Technology

Challenge the future

The front cover shows a two dimensional normal Probability Density Function.

C.C.J.M. Tiberius

Lecture notes Primer on Mathematical Geodesy
CTB3310 Surveying and Mapping
CTB3425 Monitoring and Stability of Dikes and Embankments

July 2015 / March 2018 / September 2020

Publisher:
Faculty of Civil Engineering and Geosciences
Delft University of Technology
P.O. Box 5048
Stevinweg 1
2628 CN Delft
The Netherlands

Copyright ©2015 by C.C.J.M. Tiberius
All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, scanning, recording or by any information storage and retrieval system, without written permission from the copyright owner.

The text has been type set using the MikTex 2.9 implementation of LaTeX. Graphs were produced with Matlab (R2011b), and diagrams with Adobe Illustrator, the one-and-only Mayura Draw, and Inkscape.

Contents

1 Introduction **1**

2 Random variable **3**
2.1 Introduction . . . 3
2.2 Random variable . . . 4
2.3 Histogram . . . 4
2.4 Probability Density Function . . . 6
2.4.1 Normal distribution . . . 6
2.5 Moments: mean and variance . . . 7
2.5.1 Formal moments . . . 7
2.5.2 Empirical moments . . . 8
2.5.3 Empirical moments: precision [*] . . . 9
2.6 Mean square error: accuracy . . . 9
2.6.1 Empirical MSE . . . 11
2.6.2 Example on bias, precision and accuracy . . . 11
2.7 Probabilities and intervals . . . 13
2.8 Exercises and worked examples . . . 13

3 Multi-variate: random vector **17**
3.1 Probability density function and moments . . . 17
3.1.1 Multi-variate normal distribution . . . 19
3.2 Mean and variance propagation laws . . . 19
3.3 Example . . . 19
3.4 Non-linear mean and variance propagation laws . . . 21
3.5 Exercises and worked examples . . . 21

4 Observation modelling and parameter estimation **23**
4.1 Introduction . . . 23
4.2 Observation modelling . . . 23
4.2.1 Example . . . 24
4.2.2 Redundancy . . . 24
4.3 Parameter estimation . . . 24
4.3.1 Example . . . 25
4.3.2 Least-squares estimate . . . 25
4.3.3 Example . . . 25
4.3.4 Minimum variance estimator . . . 26
4.3.5 Example . . . 27
4.4 Non-linear observation equations . . . 27
4.4.1 Linearization . . . 28
4.4.2 Estimation . . . 28
4.4.3 Example . . . 29

5 Land-surveying **31**
5.1 Levelled height-difference . . . 31
5.2 Azimuth and angle measurements . . . 31
5.3 Distance measurements . . . 32
5.4 Idealisation . . . 33
5.5 Analysis of measurement set-up: confidence ellipse . . . 33
5.6 Example: resection with distances . . . 38
5.7 Elementary measurement set-up . . . 39
5.7.1 Levelling . . . 39
5.7.2 Intersection with azimuths . . . 42
5.7.3 Polar coordinates . . . 42
5.7.4 Intersection with distances . . . 44

6 Validation **45**
6.1 Least-squares residuals . . . 45
6.1.1 Overall consistency test . . . 46
6.1.2 Simplification . . . 46
6.1.3 Discussion . . . 46
6.1.4 Example: repeated measurements . . . 47
6.2 Example . . . 47
6.3 Observation testing and outlook [*] . . . 50
6.4 Example — continued . . . 51
6.5 Exercises and worked examples . . . 52

7 Interpolation [*] **55**
7.1 Introduction . . . 55
7.2 Deterministic interpolation . . . 56
7.2.1 Inverse distance interpolation . . . 57
7.2.2 Triangular interpolation . . . 57
7.3 Stochastic interpolation . . . 58
7.3.1 Kriging . . . 60
7.3.2 Ordinary Kriging . . . 61
7.3.3 Simple Kriging . . . 62
7.4 Exercises and worked examples . . . 64

8 Conclusion **69**

A Several proofs [*] **71**
A.1 Mean and variance of normal distribution . . . 71
A.2 Mean and variance propagation laws . . . 72
A.3 Non-linear mean and variance propagation laws . . . 72
A.4 Least-squares . . . 74
A.5 Concerns on non-linear estimation . . . 74
A.6 Line-fitting with observed independent variable . . . 74
A.7 Ordinary Kriging . . . 75

B Normal distribution: table **79**

C Chi-square distribution: table **81**

Concluding remarks **83**

Bibliography **85**

Preface

This hand-out on the calculus of observations and mathematical geodesy has originally been compiled for the course *Surveying and Mapping* (CTB3310) in the 3rd year of the BSc-program for Civil Engineering.

Mathematical Geodesy provides us with the 'data-processing engine' for land-surveying in particular, and geo-science in general. Basically it addresses the question of how to go from measurements to information, and interpretation. In practice, surveying equipment is used for instance to measure distances. But, typically one is not interested in knowing these distances. Instead, one is interested in the position of a certain object, which can be determined using these measured distances. And, being aware that in reality measurements are never perfect, one would like to know the quality of the resulting position, for instance, what is the uncertainty in the outcome one can expect under ordinary circumstances, or, to phrase it differently, how close can one expect the obtained outcome to be to the actual true position?

This hand-out keeps a quite informal and introductive style. The subjects are not treated with full mathematical rigor. This hand-out has been written with an *educational* goal in mind. The idea is to give students in Civil Engineering a first *insight* in the area of calculus of observations (in Dutch: waarnemingsrekening). The approach to 'handling measurement data' in this hand-out is widely applicable in engineering.

It is assumed that students have completed a course in Linear Algebra as well as an introductory course in Statistics (Kansrekening en Statistiek), for instance by using the textbook [1].

Chapters and sections marked with a [*] contain material which is optional for the course CTB3310.

Colleague Roderik Lindenbergh provided the original outline for the chapter on Interpolation, and colleague Marc Schleiss carefully reviewed this chapter. Former colleague Siavash Shakeri is acknowledged for a first proof-reading of this primer and for providing some of the graphs and exercises. Former colleague Davide Imparato designed the front cover. The author does welcome notifications of corrections and suggestions for improvement, as well as feedback in general.

Delft, November 2014 / March 2018

1

Introduction

The development of mathematical theory, together with its implementation and use, needed to process, analyze, integrate and validate geodetic data, such as measurements for land-surveying, is referred to as the discipline of *Mathematical Geodesy*. It is about the calculus of observation, the validation of measurements and mathematical models, and the analysis of spatial and temporal phenomena in geosciences, i.e. about parameter estimation, testing and reliability, and about interpolation and prediction.

An important step in this discipline was made with the discovery of the method of least-squares in the 18th century. It occurred in the fields of astronomy and geodesy, as scientists and mathematicians sought to provide solutions to the challenges of navigation the Earth's oceans.

The German Carl Friederich Gauss (1777-1855) proposed the method of least-squares, which allows to make an optimal combination of redundant measurements for the purpose of determining values for parameters of interest. It provided an optimal solution to an inconsistent system of equations. The French mathematicien Adrien-Marie Legendre (1752-1833) published in 1805 on the method of least-squares (a translation of the French term 'méthode des moindres carrés') in the context of the determination of the orbits of comets (Nouvelles Méthodes pour la Détermination des Orbites des Comètes). Later, in 1809, in a volume on celestial mechanics, Gauss published the least-squares method, and claimed he had been using this method already back in 1795, hence earlier than Legendre, but who published it first. Many thoughts, studies and investigations have been spent in the meantime on this dispute between Legendre and Gauss, but generally it is believed that Gauss should indeed be regarded as the first inventor of the method of least-squares. Legendre came to the same method, independently, and delivered it in a clear publication. Later, Gauss provided a probabilistic justification of the method of least-squares, and proposed the normal, or Gaussian distribution. Today, more than 200 years later, the least-squares method is frequently applied in a wide variety of scientific and engineering applications.

Further developments in mathematical geodesy followed on advances — in the early 20th century — in the field of statistical inference, which is about drawing conclusions from data which are subject to random variation, for example, observational errors or sampling variation. Statisticians R.A. Fisher (1890-1962), and J. Neyman (1894-1981) and E. Pearson (1895-1980) introduced concepts as statistical confidence and hypothesis testing.

Delft University of Technology professors J.M. Tienstra (1895-1951) and W. Baarda (1917-2005) founded the 'Delft School' of Mathematical Geodesy. World-reknown contributions have been made on the subject of statistical hypothesis testing, and theory was developed, for instance leading to the concept of reliability, and applied to land surveying networks, enabling

design and analysis of these networks. Later — by professor P.J.G. Teunissen — extensions have been made to the (automated) data quality control of dynamic measurement systems in navigation and satellite navigation (GPS) applications. In recent years substantial developments have taken place on the theory for data processing and analysis of interferometric measurement techniques, such as high-precision GPS and radar-interferometry, while in parallel a clear framework on prediction theory was set up, covering the subject of interpolation, and, the reliability theory is being extended to risk evaluation for safety-critical applications, such as the navigation of unmanned vehicles, and the monitoring of land-slide and subsidence, and other (natural) hazards.

2

Random variable

2.1. Introduction

Suppose we would like to measure the inner-width of a pre-fab tunnel-element, just delivered to the construction site, in order to check whether it has been built according to requirements and specifications. To measure the distance in-between the two concrete side-walls, we use a decent laser-disto-meter, see Figure 2.1. We take a measurement and the result reads 7.451 m. We ask a colleague to do the same, and when he/she does so, the outcome is 7.453 m. When a second colleague does, we get 7.454 m, and a third colleague gets 7.452 m, and so on. The point is that such a measurement is not perfect. If a series of measurements is carried out, even under unchanged circumstances, we will see a certain *spread* in the results; the outcomes do not give all the same, exactly true, answer. This is for a number of reasons. First of all, the instrument is an electronic device, with (free) electrons moving around in its circuits (as we are not operating at zero Kelvin temperature); they will cause (small) errors in the reading of the distance.

Second, the concrete wall on one side to which we hold the instrument will not be perfectly smooth, and so will be the other wall, which has to reflect the laser-pulse, used for measuring the distance. This may cause small differences between two measurements when we hold the instrument, take a measurement, remove the instrument, and put the instrument again etc. In addition, you may not hold the instrument completely still during the measurement. Finally, there are external circumstances which can cause errors in the measurement process, such as reduced visibility in the tunnel, due to dust and smoke, and motions and vibrations of the tunnel element itself. Generally, conditions at a construction-site are sub-optimal … for

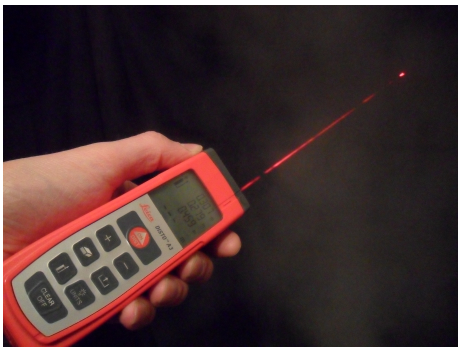


Figure 2.1: Laser distometer for measuring distances.

carrying out accurate measurements. A measurement does not give exactly the true answer; but, hopefully, its value is close though.

2.2. Random variable

The whole exercise of using a certain instrument, carrying out the measurement in a certain way (e.g. holding the instrument here, or there) and obtaining the result, is captured - in a mathematical sense - in a random variable. The bottom-line is that the outcome of our measurement (generally) will be close to the desired, true value (which we do not know), but it will not be perfect - it will contain some amount of error. In the sequel, we will use the following notation: y for the measurement value, x for the unknown (true) distance, and e for the (random) measurement error, so that we have

$$y = x + e \tag{2.1}$$

A random variable is a mathematical concept, it is denoted by a symbol with an underscore, such as $\underline{y}$. We can obtain a realization of this random variable, by actually taking a measurement, a sample y, or with an index y_1, where the index denotes that this is the first measurement. We can repeat the measurement a number of times to obtain y_1, y_2, ..., y_N. Later, we refer to $\underline{y}$ as the *observable*, the 'thing which *can* be observed', and y_i is one of the *observations* (y_i has a single, fixed value, namely the numerical outcome of that specific measurement, for example y_3=7.450 m).

Then (2.1) can be written as

$$\underline{y} = x + \underline{e} \tag{2.2}$$

The observable $\underline{y}$ equals the true, but unknown x (distance, for instance), plus a **random measurement error** $\underline{e}$, and e is referred to as an un-observable statistical error (we will never know this error in practice). We typically assume that the random measurement error $\underline{e}$ is — on average — equal to zero. Individual realizations $e_1, e_2, \ldots, e_N$ are not equal to zero, but the average over a large number of realizations of $\underline{e}$ will be close to zero.

The parameters we are trying to measure in this course are of the continuous type. The unknown inner-width of the tunnel-element x is a distance, which can have any real-value, hence $x \in R$. To allow for automated processing of the measured distance, the observation available to the user is presented and stored using a finite number of decimals. The observation has been digitized and actually become discrete, though we will not address the subject of quantization here. By approximation, it still is a continuous quantity.

2.3. Histogram

When the measurement has been repeated a number of times, all measurements together can be presented in terms of a histogram. The range (or a part of it) of variable y is divided into k intervals (bins or classes) of equal length h, the bin width. With a chosen origin y_o, we have the following intervals around y_o

$$\begin{array}{ll}
j = 1 & [y_o - \frac{k}{2}h, y_o - (\frac{k}{2} - 1)h\rangle \\
\vdots & \vdots \\
\vdots & [y_o - 2h, y_o - h\rangle \\
j = \frac{k}{2} & [y_o - h, y_o\rangle \\
j = \frac{k}{2} + 1 & [y_o, y_o + h\rangle \\
\vdots & [y_o + h, y_o + 2h\rangle \\
\vdots & \vdots \\
j = k & [y_o + (\frac{k}{2} - 1)h, y_o + \frac{k}{2}h\rangle
\end{array} \tag{2.3}$$

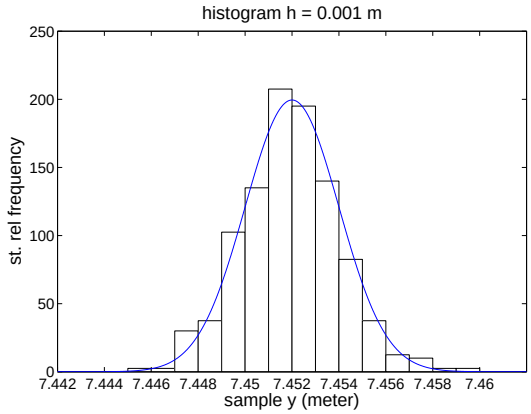


Figure 2.2: Histogram with binsize $h = 0.001$ m, $k = 20$ bins, and center $y_o = 7.452$ m. The vertical-axis gives the standardized relative frequency. The data, with sample size $N = 400$, were generated here from a normal distribution with $x = 7.452$ m and $\sigma = 0.002$ m. The curve of this theoretical probability density function is imposed.

where we assumed k to be even.

The N samples, assumed to be all in $[y_o - \frac{k}{2}h, y_o + \frac{k}{2}h\rangle$, are divided over the bins. The observed (absolute) frequencies (or cell counts) in the k bins are denoted by f_j, with $\sum_{j=1}^{k} f_j = N$.

The histogram, see also chapter 17 in [1], is now created by plotting

$$\hat{f}(y) = \frac{f_j}{Nh} \quad \text{with} \quad y \in \text{ interval } j \tag{2.4}$$

as a function of y; $\hat{f}(y) = 0$ outside the k intervals.

The function $\hat{f}(y)$ is an interval-wise (width h) constant function. Figure 2.2 gives an example. The $\frac{f_j}{N}$ are the relative frequencies, and h in the denominator assures that $\int_{-\infty}^{\infty} \hat{f}(y)dy = 1$, that is, the area under the histogram equals 1. This enables direct comparison of histograms of different data sets (with different bin widths h, and sample sizes N).

For the set of distance measurements, most of the measurements are close to the value of 7.452 m. The further we go away from 7.452 m, the fewer observed values we see. This is typical behaviour in practice. This behaviour is formalized in a *probability density function* (PDF) of the random variable. The PDF is a mathematical formula, describing the uncertain outcome of our distance measurement. It gives the probability density as a function of the value of the observed parameter y. The PDF describes the distribution of the random variable, and it is a mathematical model for the histogram.

The theoretical probability density function $f(y)$ can—for comparison—be directly imposed on the histogram $\hat{f}(y)$, as is done in Figure 2.2.

The bin width h controls the amount of 'smoothing'. In practice one has to match the interval $[y_o - \frac{k}{2}h, y_o + \frac{k}{2}h\rangle$ with $y_{\min}$ and $y_{\max}$, and as a rule of thumb one often sets the number of bins to $k = \sqrt{N}$. Also in practice, in order to focus on the core of the distribution in case there are outlying measurements, one may want to match the interval with e.g. $q_{.01}$ and $q_{.99}$, the 1-st and 99-th (empirical) percentiles respectively, see (2.20), rather than $y_{\min}$ and $y_{\max}$.

The histogram is useful for presentation of the data. It gives a first impression of the probability density which lies at the basis of the samples. One might for instance visually judge — preliminary — whether normality is not unlikely. One must however, be very careful.

The picture can be manipulated, certain features can be masked or over-emphasized, by the choice of the origin y_o and the bin width h.

2.4. Probability Density Function

A histogram is useful to visualize a (big) set of repeated measurements. In practice, where 'time is money', one does not want to repeat measurements. Though one still would like to have some measure about the uncertainty that can be expected in the outcome, in particular, how big the chance is that the outcome is off from the truth, by more than a certain amount. Therefore, the histogram is captured by a Probability Density Function (PDF), denoted by $f(y)$, for the random variable $\underline{y}$. It gives a mathematical expression for the probability density, as a function of the value y of the random variable.

A probability density function $f(y)$ has to satisfy two general requirements:

$$f(y) \geq 0 \; \forall \; y \text{ and } \int_{-\infty}^{\infty} f(y)dy = 1$$

The (cumulative) probability distribution function is denoted by $F(y)$ and can be found by

$$F(y) = \int_{-\infty}^{y} f(y)dy \tag{2.5}$$

or the other way around

$$f(y) = \frac{\partial F(y)}{\partial y} \tag{2.6}$$

$F(y)$ is a monotonic non-decreasing function, and provides a mapping from the R into $[0, 1]$. It holds that the probability $P[\underline{y} \leq k] = F(k)$.

2.4.1. Normal distribution

When $\underline{y}$ has a normal or Gaussian distribution, the probability density function reads

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}} \tag{2.7}$$

where x is the mean, and σ the standard deviation (σ^2 the variance); a proof can be found in Appendix A.1; the function is completely specified by these two parameters x and σ. The density is a (unimodal) bell-shaped curve, see Figure 2.3 on the left. Often this distribution is denoted as $\underline{y} \sim N(x, \sigma^2)$, where x represents the mean, and σ^2 the variance.

In practice, the manufacturer of the equipment (e.g. of the laser-disto-meter), has analyzed histograms of numerous repeated measurements, in the field, and in the lab, and provides users with information, stating that the error in the measurements will follow a normal distribution, with zero mean (hence on average, the measurement is correct), and a certain standard deviation, e.g. σ=0.002 m. With the random variables in (2.2) we have $\underline{y} = x + \underline{e}$ and hence, $\underline{e} \sim N(0, \sigma^2)$, and $\underline{y} \sim N(x, \sigma^2)$.

When random variable $\underline{z}$ has the following distribution $\underline{z} \sim N(0, 1)$, it is said to be *standard* normally distributed, see Figure 2.3. The cumulative probability distribution function (CDF) F of $\underline{z}$ is also denoted as Φ. Appendix B provides a table of the standard normal distribution. Given are the probabilities α, as the right tail probabilities $\alpha = 1 - \Phi(r_\alpha)$, where $P[\underline{y} \leq r_\alpha] = \Phi(r_\alpha)$.

In practice not all observables are normally distributed. There are actually many different probability density functions — the normal one is certainly not the only one. Later, in section 5.5, the Chi-squared distribution will be introduced.

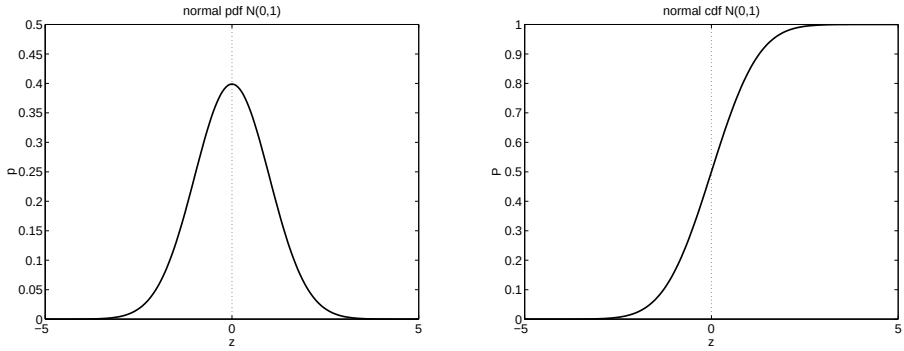


Figure 2.3: Standard normal, or Gaussian distribution, $\underline{z} \sim N(0,1)$ with $x = 0$ and $\sigma = 1$. On the left the probability density function $f(z)$. On the right the cumulative probability distribution function $F(z)$; as variable z runs from $-\infty$ to ∞, the probability runs from 0 to 1.

2.5. Moments: mean and variance

In this section we present several characteristic measures of probability density functions. Most common are the mean and the variance. We will consider them from a theoretical (formal) point of view (section 2.5.1), as well as from a practical (empirical) point of view (section 2.5.2).

2.5.1. Formal moments

The expectation of $\underline{y}$ about some constant ϑ reads

$$E'(\underline{y}) = \int_{-\infty}^{+\infty} (y - \vartheta) f(y)\, dy \tag{2.8}$$

When we take, as usual, the expectation about zero ($\vartheta = 0$), we obtain the well known first moment, or mean,

$$E(\underline{y}) = \int_{-\infty}^{+\infty} y f(y)\, dy \tag{2.9}$$

and it holds that $E'(\underline{y}) = E(\underline{y}) - \vartheta$.

When $\underline{y}$ is distributed as $\underline{y} \sim N(x, \sigma^2)$, as in (2.7), it can be shown that $E(\underline{y}) = x$, see Appendix A.1. It gives the location (the center) of the normal curve.

The second central moment, or variance is

$$D(\underline{y}) = E((\underline{y} - E(\underline{y}))^2) = \int_{-\infty}^{+\infty} (y - E(\underline{y}))^2 f(y)\, dy \tag{2.10}$$

The word '*central*' refers to the fact that the moment is taken about the mean $E(\underline{y})$. We denote the variance by $D(.)$ (dispersion), rather than $Var(.)$ as done for instance in [1], and often the symbol σ^2 is used for the variance of a random variable (and σ for standard deviation). Later, we will use the index y to denote the variance of $\underline{y}$, hence σ_y^2.

When $\underline{y}$ is distributed as $\underline{y} \sim N(x, \sigma^2)$, as in (2.7), it can be shown that indeed the variance equals $D(\underline{y}) = \sigma^2$, see Appendix A.1. The standard deviation σ describes the width of the normal curve. It presents the spread in the result; the standard deviation is a numerical measure of the uncertainty.

2.5.2. Empirical moments

We will now consider the empirical equivalents of the moments discussed above. Therefore we assume to have N realizations of the random variable y. The measurement has been repeated (under unchanged conditions), and as a result we have N outcomes $y_1, y_2, \ldots, y_N$. These N measurements are used to come up with estimates for the mean and variance, i.e to come up with the sample mean and sample variance.

The average deviation from some known constant ϑ reads

$$\hat{x}' = \frac{1}{N}\sum_{i=1}^{N}(y_i - \vartheta) \tag{2.11}$$

and all outcomes $y_1, y_2, \ldots, y_N$ are first 'corrected for' ϑ , and then the average is taken. The well known first **sample** moment is the (arithmetic) **mean** about zero ($\vartheta = 0$)

$$\hat{x} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2.12}$$

which is denoted by x with a 'hat'-symbol, meaning that it is an estimate for the unknown true mean x of the random variable y, and this estimate is based on data/measurements; (2.12) is the empirical counterpart of (2.9). It holds that $\hat{x}' = \hat{x} - \vartheta$.

The expectation (or mean) of y, x, is unknown, and will remain unknown forever; we can only come up with an estimate $\hat{x}$ for this parameter.

The second central sample moment (hence second sample moment about the mean) reads

$$\hat{\sigma'}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{x})^2 \tag{2.13}$$

which is an unbiased estimate for the variance, once the mean x is known (a-priori), and substituted for $\hat{x}$ in the above equation. Unbiasedness is generally a desirable property for an estimator, meaning that on average the result is spot-on (see also section 4.3). When the mean is not known (a-priori), estimate (2.13), with $\hat{x}$ from (2.12) inserted, is *not* unbiased, meaning that on average it is 'somewhat off'. If the mean is unknown, we generally use the following estimate for the variance instead

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{x})^2 \tag{2.14}$$

which is unbiased. Generally (2.14) is referred to as the sample variance; it is the empirical counterpart of (2.10). The square root of estimate (2.14) will be our default estimate, the **sample standard deviation**.

The difference of N and $N - 1$ in the denominator in (2.13) and (2.14) is negligible in practice, when N is large.

Note that in the second case with (hypothetically) only one observation y_1 ($N = 1$), the mean (2.12) becomes $\hat{x} = y_1$ and the variance (2.14) is undefined — from one sample it is not possible to estimate both mean and variance. Estimate $\hat{\sigma'}^2$ would give zero in this case.

The difference $y_i - \hat{x}$ in the above equations will later be denoted by $\hat{e}$, and referred to as the **measurement residual**

$$\hat{e}_i = y_i - \hat{x} \tag{2.15}$$

see section 6.1. Where e is the unknown measurement error, cf. (2.1), $\hat{e}$ is the *estimated* error. Residual $\hat{e} = y - \hat{x}$ equals the difference of the observation y and the *estimated* mean $\hat{x}$ (or, the average fitted to the observations).

In the context of section 6.1.1, the second sample moment about the mean (2.13) can be regarded as the mean of the squared residuals, see also section 6.1.4, as $\sum_{i=1}^{N}(y_i - \hat{x})^2$ is the sum of squared residuals.

2.5.3. Empirical moments: precision [*]

The estimates for the mean and variance (2.12) and (2.13) can be shown to result as Maximum Likelihood (ML) estimates from the model $E(\underline{y}) = Ax$, with $A = (1, \ldots, 1)^T$, and $D(\underline{y}) = \sigma^2 I_N$, with $\underline{y}$ normally distributed, cf. chapter 4 (4.2). Also, it can be shown that the estimator $\underline{\hat{\sigma}}^2$ (2.13)/(2.14) is not correlated with $\underline{\hat{x}}$ (2.12).

As the estimates $\hat{x}$ for the mean, and $\hat{\sigma}^2$ for the variance are based on the measurements (which are not perfect), we can expect these estimates to be not perfect either. The estimators are unbiased, meaning that $E(\underline{\hat{x}}) = x$ and $E(\underline{\hat{\sigma}}^2) = \sigma^2$. The variances of the estimators in (2.12) and (2.14) are given (without proof) by

$$\sigma_{\hat{x}}^2 = \frac{\sigma^2}{N} \tag{2.16}$$

and

$$\sigma_{\hat{\sigma}^2}^2 = \frac{2\sigma^4}{N-1} \tag{2.17}$$

The more measurements we take (bigger N) the smaller these variances get, which is intuitively appealing. The more data you use, the more precise these estimators (for mean and variance) get. The expression for $\sigma_{\hat{\sigma}^2}^2$ holds only for normally distributed observables.

Typically one is interested in the standard deviation rather than the variance, hence with $\underline{\hat{\sigma}} = G(\underline{\hat{\sigma}}^2) = \sqrt{\underline{\hat{\sigma}}^2}$ (the standard deviation is a non-linear function of the variance, through the square-root) and through the first-order approximation $\frac{\partial G}{\partial \hat{\sigma}^2} = \frac{1}{2\sqrt{\hat{\sigma}^2}}$, one can use (3.12), and obtain $\sigma_{\hat{\sigma}} \approx \frac{\sigma_{\hat{\sigma}^2}}{2\sqrt{\hat{\sigma}^2}} = \frac{\sigma_{\hat{\sigma}^2}}{2\hat{\sigma}}$ (the standard deviation of the estimator for the standard deviation), and use in this case $\sigma_{\hat{\sigma}^2} = \frac{\sqrt{2}\sigma^2}{\sqrt{N-1}}$.

2.6. Mean square error: accuracy

In practice, a measurement may be biased. This means that on average it does *not* deliver (the value of) the parameter which we hope it delivers. The bias is another (typically undesired) parameter ϑ, which enters the equation: $\underline{y} = x + \vartheta + \underline{e}$. Here, x is the unknown parameter in which our interest lies, $\underline{y}$ is the observable (which is a random variable), $\underline{e}$ is the random measurement error (for which we assume that it has zero mean, hence, it will cause individual measurements (samples) to be off from the true value, but, taking the average over a large number of (repeated) measurements, will provide an outcome close to the true and wanted value), and ϑ represents the **bias**, or (constant) offset in the measurement, which is a *systematic* effect.

In terms of the laser-disto-meter, one can think of an unwanted time delay of the signal in the electronic circuitry of the device, which translates into a certain fixed error (offset) in the travel-time of the laser-pulse, and hence into the measured distance — the effect is there all the time (also in repeated measurements).

Systematic errors also include scaling effects, e.g. $\underline{y} = \lambda x + \underline{e}$, with λ as a scale factor, but this is beyond the scope of this primer.

In this section, we present a measure of the spread in the uncertain outcome of the measurement, which includes also the unwanted bias-part. In the end, we would like to have a measure of how close our observation is to the true (wanted) distance. Instead of the variance, we consider the mean squared error (MSE), cf. also section 20.3 in [1].

In the previous section there was no bias, and the expectation of the observable was $E(\underline{y}) = x$, as $\underline{y} = x + \underline{e}$. Now, with $\underline{y} = x + \vartheta + \underline{e}$, we have instead $E(\underline{y}) = x + \vartheta$, where ϑ is a bias. The variance (2.10) is the second moment about the mean of the observable $E(\underline{y}) = x + \vartheta$.

But — as an all-in measure — we are now interested in the second moment about $E'(\underline{y}) = x$, (2.8), namely about the true distance x. Therefore one defines the Mean Squared Error (MSE)

$$D'(\underline{y}) = \int_{-\infty}^{+\infty} (y - E'(\underline{y}))^2 f(y)\, dy \quad = E((\underline{y} - E'(\underline{y}))^2)$$

instead of the variance (2.10), and the mean $E(\underline{y})$ (2.9) has been replaced by the **true** distance $E'(\underline{y}) = x$ (2.8).

We will show that the MSE can be written as $D'(\underline{y}) = \sigma^2 + \vartheta^2$. Therefore we develop the MSE into

$$D'(\underline{y}) = E((\underline{y} - E'(\underline{y}))^2) = E((\underline{y} - E(\underline{y}) + E(\underline{y}) - E'(\underline{y}))^2)$$

$$= E((\underline{y} - E(\underline{y}))^2 + (E(\underline{y}) - E'(\underline{y}))^2 + 2(\underline{y} - E(\underline{y}))(E(\underline{y}) - E'(\underline{y})))$$

$$= E((\underline{y} - E(\underline{y}))^2 + (E(\underline{y}) - E'(\underline{y}))^2)$$

as the factor $(E(\underline{y}) - E'(\underline{y}))$ is just a constant, and taking it out of the expectation operator in the cross-term, we are left with $2(E(\underline{y}) - E'(\underline{y}))E(\underline{y} - E(\underline{y}))$, which is just zero as $E(E(\underline{y})) = E(\underline{y})$. Hence,

$$D'(\underline{y}) = E((\underline{y} - E(\underline{y}))^2 + (E(\underline{y}) - E'(\underline{y}))^2) = \sigma^2 + \vartheta^2$$

the MSE equals the variance plus the squared bias. The MSE accounts for the spread in the result, as well as a bias, when present. When there is no bias $\vartheta = 0$, the MSE simply equals the variance σ^2.

The name of Mean Squared Error (MSE) explains by recognizing that we take the error (in our case, by how much $\underline{y}$ deviates from the true value $E'(\underline{y}) = x$), square it, and eventually take the mean (expectation), as the observable $\underline{y}$ is a random variable (in general we have no knowledge or control about the random error $\underline{e}$ included in $\underline{y}$). The MSE provides us with a general, overall measure of the deviation we can expect in the outcome; the MSE measures *accuracy*.

According to the New Oxford Dictionary of English **accuracy** is 'the quality or state of being correct or precise', with the specialized technical subsense as 'the degree to which the results of a measurement, calculation, or specification conforms to the correct value or a standard'; it represents the degree of closeness of a measurement of a certain quantity to the actual true, or reference value of that quantity.

The spread in the outcome of a repeated experiment is referred to as **repeatability**; the degree to which repeated measurements, under unchanged conditions, show same results.

The formal notion of repeatability is **precision**. The standard deviation σ, or variance σ^2, is a measure of precision. For a common uni-modal Probability Density Function (PDF), the standard deviation measures the width of this formal function.

Loosely spoken, one could say that: 'accuracy equals precision plus bias'.

2.6.1. Empirical MSE

The Mean Squared Error (MSE) can be used in practice for instance in a calibration or verification campaign. With the laser-disto-meter, the manufacturer may have a calibration test-range available, for which actual distances are known already (known with a much higher accuracy, better by orders of magnitude than what the laser-disto-meter will deliver, for instance by using different equipment). Then, the laser-disto-meter is employed on the test-range, and the spread of the (repeated) measurements $y_1, y_2, \ldots, y_N$ is quantified by the second moment about the true distance x, which is known in this case (and not about the mean). If — unexpectedly — a bias is present in the measurements, it will be reflected in the resulting mean squared error (MSE).

Also, one could correct the obtained measurements beforehand for the known distance, therefore be dealing with samples of $(\underline{y} - x) = \vartheta + \underline{e}$, which are assumed to present a zero mean error, as we are initially not aware of the bias ϑ. Next, taking the second moment about zero yields the Mean Squared Error (MSE).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - x)^2 \tag{2.18}$$

In practice one often takes the square root of the MSE, leading to the Root Mean Squared Error (RMS) or (RMSE), which is in units of the observed quantity.

The above expression for the empirical MSE looks very much like the second sample moment about the mean (2.13), and (2.14), but carefully note that in (2.18) the true value x is involved, whereas in (2.13), and (2.14), it is the *estimated* mean $\hat{x}$.

2.6.2. Example on bias, precision and accuracy

With a large bias, the observable can be very precise (small random error), but it will not be accurate. With a big spread (large standard deviation), the observable is neither precise, nor accurate.

Suppose a laser-disto-meter is being tested on an accurately known distance, with x=7.452 m. Five measurements are taken (N=5), and the MSE is computed, based on the known distance x, hence (2.18) can be rewritten (by subtracting and adding the term $\hat{x}$ in between the brackets) into

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{x})^2 + \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - x)^2$$

in a way much similar to the variance plus bias decomposition of the MSE. The first term is the second sample moment $\hat{\sigma'}^2$ (2.13), which is approximately equal to the (estimated) variance (when N is large), and the second term, with $\hat{x}$ in (2.12), equal to just $(\hat{x} - x)^2$, is the square of the estimated bias, as $\hat{x}$ is the (ordinary) mean of the observations, which now includes the bias ϑ, and x is the true distance.

We consider three different cases. They are listed in table 2.1. In the first case the distance observable is obviously both precise (small standard deviation) and accurate (small MSE). In the second case the observable is not precise (big standard deviation) and not accurate (big

case	y_1	y_2	y_3	y_4	y_5	$\hat{x}$	$\hat{\sigma}'$	$RMSE$
1	7.452	7.454	7.450	7.453	7.451	7.452	0.001	0.001
2	7.452	7.462	7.442	7.457	7.447	7.452	0.007	0.007
3	7.459	7.461	7.457	7.460	7.458	7.459	0.001	0.007

Table 2.1: Three different cases of testing a laser-disto-meter on a calibrated distance. The true distance is x=7.452 m. Each time N=5 measurements have been taken. The sample mean (2.12), the (square root of the) second order sample moment (2.13), as an approximation to the sample standard deviation, and the Root Mean Squared Error $\sqrt{MSE}$ (2.18) have been computed (in the latter, x has been replaced by $\hat{x}'$). All values are in [m].

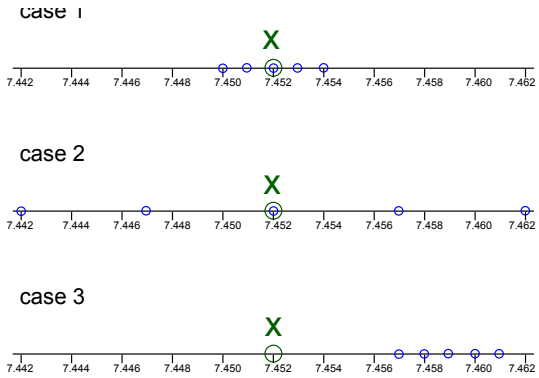


Figure 2.4: Three different cases of testing a laser-disto-meter on a calibrated distance. The true distance is $x = 7.452$ m, indicated in green. Each time $N = 5$ measurements have been taken, shown by blue circles.

MSE). In the first two cases there is no bias. In the third case, the observable is again precise (small standard deviation), but it is not accurate (big MSE), due to a bias in the measurements, the sample mean deviates a lot from the known distance (7.459 m versus 7.452 m).

Figure 2.4 displays the three cases, and table 2.2 presents the summary in terms of the distance observable being precise or not, and accurate or not. In some sources, you may find that case 2 is not precise, but, surprisingly, is accurate. The cause of this confusion may lie in the fact that in case 2 the distance observable is *not biased*, and thereby averaging can be used to reduce the uncertainty. According to (2.16), the standard deviation of the mean is smaller by a factor $\sqrt{N}$, when the mean is taken over N samples. In case 2, a *single* distance observable is not precise, but the *mean* over all five together could be, and hence, all together they could provide a result which is close to the truth (and thereby — confusingly — rated accurate).

In case 3, the measurements are biased. This bias is preserved, when taking the average over all five measurements. The result will stay (far) off from the true value.

case	precise	accurate
1	yes	yes
2	no	no
3	yes	no

Table 2.2: Summary on the precision and accuracy of the distance observable in three different cases of testing a laser-disto-meter on a calibrated distance.

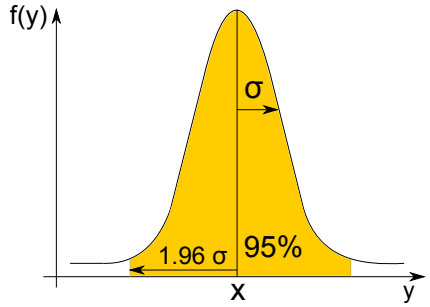


Figure 2.5: Normal probability density function (PDF) $f(y)$. The width of this function represents the uncertainty in measured outcomes of the observable $\underline{y}$. When the experiment could be repeated many times, the width of the PDF would reflect the spread in the outcome, for instance the observed distances. When the distance observable is normally distributed, the 1-sigma (σ) interval to both sides about the mean x contains about 68% of the samples (it is referred to as the 68% interval), and 95% of the samples (in yellow) will lie in the interval [-1.96 σ, 1.96 σ] about the mean.

2.7. Probabilities and intervals

The Probability Density Function (and hence, also the Cumulative Distribution Function (CDF)) are mathematical concepts, which are very useful in practice, once one would like to compute probabilities, specifically, probabilities that a certain random variable lies in a certain interval.

$$P[a \leq \underline{y} \leq b] = \int_a^b f(y)\,dy = F(b) - F(a) \tag{2.19}$$

A normal probability density function $\underline{y} \sim N(x, \sigma^2)$ is shown in Figure 2.5.

Through the PDF (and CDF), probabilities and interval-bounds are intimately related. The p-th quantile (with p a number between 0 and 1), or the 100p-th percentile of the distribution of $\underline{y}$ is the smallest number q_p such that

$$F(q_p) = P[\underline{y} \leq q_p] = p \tag{2.20}$$

Quantiles and percentiles can also be obtained empirically, from the observed data. Suppose that we have N samples of a certain random variable, then we order the samples ascendingly, and each sample basically represents a probability of $\frac{1}{N}$. When a proportion p is less than a certain number k (i.e. 100p% of the samples have all values less than k), and a proportion $1-p$ is greater than this number, this number k is the 100p-th empirical quantile, or the 100p-th sample percentile.

For realizations of a random variable (error), which is (or can be) assumed to have zero mean, one is generally interested in just the *magnitude* of the error. Then typically the absolute value is taken of the samples, they are ordered ascendingly, and then the empirical percentile is determined. In that case, the corresponding formal 100p-th percentile is defined as

$$P[|\underline{y}| \leq q_p] = p$$

2.8. Exercises and worked examples

This section presents a number of problems and worked answers.

Question 1 A random variable is normally distributed, and has zero mean, and standard deviation equal to 1. You could think of this random variable as the measurement error $\underline{e}$, for instance of an observable with the laser-disto-meter, and units in millimeters. Compute the

probability that a single sample of this random variable will have a value of 1.27, or less. This is, compute the probability $P[\underline{e} < 1.27]$.

Answer 1 The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix B. We need $P[\underline{e} < r_\alpha] = \Phi(r_\alpha)$, with $r_\alpha = 1.27$. The table gives probabilities α, as the right tail probabilities $\alpha = 1-\Phi(r_\alpha)$, rather than left tail. With $r_\alpha = 1.27$, we obtain $\alpha = 0.1020$, hence the requested probability is $P[\underline{e} < r_\alpha] = \Phi(r_\alpha) = 1-\alpha = 0.8980$.

Question 2 A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to question 1. Compute the probability that a single sample of this random variable will have a value of -1.27, or less. This is, compute the probability $P[\underline{e} < -1.27]$.

Answer 2 The random variable has again a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix B. We need $P[\underline{e} < -r_\alpha] = \Phi(-r_\alpha)$, with $r_\alpha = 1.27$. The table gives probabilities α, as the right tail probabilities $\alpha = 1 - \Phi(r_\alpha)$, but only for positive arguments. However, the normal distribution is symmetric about its mean. Hence, when the mean is zero, we have $\Phi(-r_\alpha) = P[\underline{e} < -r_\alpha] = P[\underline{e} > r_\alpha] = 1 - P[\underline{e} < r_\alpha] = 1 - \Phi(r_\alpha)$. With $r_\alpha = 1.27$, we obtain $\alpha = 0.1020$, hence the requested probability is $P[\underline{e} < -r_\alpha] = \alpha = 0.1020$.

Question 3 A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to question 1. Compute the probability that a single sample of this random variable will have a value of -1.23, or more. This is, compute the probability $P[\underline{e} > -1.23]$.

Answer 3 The random variable has again a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix B. We need $P[\underline{e} > -r_\alpha] = 1 - \Phi(-r_\alpha)$, which is $1 - \Phi(-r_\alpha) = 1 - (1 - \Phi(r_\alpha)) = \Phi(r_\alpha))$, with $r_\alpha = 1.23$. The table gives probabilities α, as the right tail probabilities $\alpha = 1 - \Phi(r_\alpha)$. With $r_\alpha = 1.23$ we obtain $\alpha = 0.1093$, hence the requested probability is $\Phi(r_\alpha) = 1 - \alpha = 0.8907$.

Question 4 A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to question 1. Compute the probability that a single sample of this random variable will have a value lying in between -2.00 and 1.50. This is, compute the probability $P[-2.00 < \underline{e} < 1.50]$.

Answer 4 The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix B. We need $P[-r_{\alpha,1} < \underline{e} < r_{\alpha,2}] = \Phi(r_{\alpha,2}) - \Phi(-r_{\alpha,1}) = \Phi(r_{\alpha,2}) - (1 - \Phi(r_{\alpha,1}))$, with $r_{\alpha,1} = 2.00$ and $r_{\alpha,2} = 1.50$. With the table we obtain $P[-2.00 < \underline{e} < 1.50] = (1 - 0.0668) - (1 - (1 - 0.0228)) = 0.9104$.

Question 5 A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to question 1. For what boundary value r_α holds that the probability that a single sample of this random variable will have a value of r_α, or less, equals 0.975? This is, solve the probability statement $P[\underline{e} < r_\alpha] = 0.975$, for r_α.

Answer 5 The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix B. We need $P[\underline{e} < r_\alpha] = \Phi(r_\alpha) = 0.975$, or $\alpha = 0.025$, given in the table. This yields $r_\alpha = 1.96$.

Question 6 A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to question 1. For what boundary value r_α holds that the probability that a single sample of this random variable will have a value in between $-r_\alpha$ and r_α, equals 0.95? This is, solve the probability statement $P[-r_\alpha < \underline{e} < r_\alpha] = 0.95$ for r_α.

Answer 6 The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix B. We need $P[-r_\alpha < \underline{e} < r_\alpha] = \Phi(r_\alpha) - \Phi(-r_\alpha) = \Phi(r_\alpha) - (1 - \Phi(r_\alpha)) = 0.95$. This equals $\Phi(r_\alpha) - (1 - \Phi(r_\alpha)) = 2\Phi(r_\alpha) - 1 = 0.95$, or $\Phi(r_\alpha) = \frac{1.95}{2}$. With $\alpha = 1 - \frac{1.95}{2}$, given in the table, this yields $r_\alpha = 1.96$.

Question 7 A random variable is normally distributed, and has mean equal to 2 (unlike

previous questions), and standard deviation equal to 1. Compute the probability that a single sample of this random variable will have a value in between 1.27 and 2.00. This is, compute the probability $P[1.27 < \underline{e} < 2.00]$.

Answer 7 The random variable now does not have a standard normal distribution. Though, it can be turned into a standard normally distributed variable by subtracting the mean (this is a linear operation and the new random variable is normally distributed as well), $\underline{e} \sim N(2,1)$, and $(\underline{e}-2) \sim N(0,1)$, which is tabulated in Appendix B. We need $P[r_{\alpha,1} < \underline{e} < r_{\alpha,2}] = P[(r_{\alpha,1}-2) < (\underline{e}-2) < (r_{\alpha,2}-2)] = \Phi(r_{\alpha,2}-2) - \Phi(r_{\alpha,1}-2)$, with $r_{\alpha,1} = 1.27$ and $r_{\alpha,2} = 2.00$. With the table, we obtain $P[1.27 < \underline{e} < 2.00] = (1-0.5000) - (1-(1-0.2327)) = 0.2673$.

Question 8 Random variable $\underline{e}$ is normally distributed, and has mean equal to 1, and standard deviation equal to 2. Compute the probability that a single sample of this random variable will have a value in between -1.00 and 1.00. This is, compute the probability $P[-1.00 < \underline{e} < 1.00]$.

Answer 8 The random variable now does not have a standard normal distribution. Though, it can be turned into a standard normally distributed variable by subtracting the mean, and dividing by the standard deviation (these are both linear operations and the new random variable is normally distributed as well), $\underline{e} \sim N(1,4)$, and $\frac{\underline{e}-1}{2} \sim N(0,1)$, which is tabulated in Appendix B. We need $P[-r_{\alpha,1} < \underline{e} < r_{\alpha,2}] = P[\frac{-r_{\alpha,1}-1}{2} < \frac{\underline{e}-1}{2} < \frac{r_{\alpha,2}-1}{2}] = \Phi(\frac{r_{\alpha,2}-1}{2}) - \Phi(\frac{-r_{\alpha,1}-1}{2})$, with $r_{\alpha,1} = 1.00$ and $r_{\alpha,2} = 1.00$. With the table, we obtain $P[-1.00 < \underline{e} < 1.00] = (1-0.5000) - (1-(1-0.1587)) = 0.3413$.

Question 9 For the error $\underline{e}$ in a distance observable $\underline{y}$, with $\underline{y} = x + \underline{e}$, and x the unknown true distance, is given that it is distributed as $\underline{e} \sim N(0,\sigma^2)$, with standard deviation σ = 3 mm. What is the probability that — when we take a distance measurement in practice — the *magnitude* of the measurement error is bigger than 6 mm?

Answer 9 The required probability is $P[|\underline{e}| > 6]$. The absolute sign (because of the word 'magnitude') can be removed through: $P[|\underline{e}| > 6] = P[\underline{e} < -6] + P[\underline{e} > 6]$, which equals $2P[\underline{e} > 6]$, as the normal distribution is symmetric here about zero (given zero mean). Then we convert into a *standard* normal distribution through $2P[\frac{\underline{e}}{\sigma} > \frac{6}{\sigma}]$, which, with the table in Appendix B ($r_\alpha = 2.00$), yields $2P[\frac{\underline{e}}{\sigma} > \frac{6}{3}] = 2 \cdot 0.0228$ = 0.0456.

Question 10 A laser disto meter is deployed on an accurately calibrated test-range (hence the true distance x is known). Suppose that the distance observable $\underline{y}$ has a standard deviation of 2 mm, and that the instrument has a bias ϑ of 1 cm (hence all measured distances are systematically too long by 1 cm). The distance observable is normally distributed. The manufacturer analyses, by measuring the known distance repeatedly, the *magnitude* of the error $\underline{y} - x$, where he presumes the error to be zero mean (as a-priori he is not aware of the presence of the bias). Can you give — based on the above given data — a good estimate for the 95th-percentile that the manufacturer is going to find?

Answer 10 The deviation in the observable from the known distance $\underline{y} - x = \vartheta + \underline{e}$ is distributed as $\vartheta + \underline{e} \sim N(\vartheta,\sigma^2)$, in this case with ϑ=10 mm, and σ=2 mm. We need to find the value for q_p in $P[-q_p < \vartheta + \underline{e} < q_p] = p$, which is an interval symmetric about zero, with p=0.95. Transformation of the random variable yields $P[\frac{-q_p-\vartheta}{\sigma} < \frac{\underline{e}}{\sigma} < \frac{q_p-\vartheta}{\sigma}] = p$, or $\Phi(\frac{q_p-\vartheta}{\sigma}) - \Phi(\frac{-q_p-\vartheta}{\sigma}) = p$, as the random variable $\frac{\underline{e}}{\sigma}$ has a standard normal distribution. This has to be solved iteratively, by numerical search, i.e. trying different values for q_p until we obtain the desired $p = 0.95$. Specifically, we start with value $q_p = 0$, and increase it each step by 0.01, until we reach the desired 95% probability, i.e. find the first occurrence where the above equation yields a probability of 95% or more. The result is q_p=13.29. In this case the left tail does actually not contribute to the exceedance probability, it is below 10^{-30}; the 5% of the samples beyond the bounds $[-q_p, q_p]$ will typically all lie at the right hand side, that is,

beyond q_p. Suppose the bias would be ϑ=1 mm, then the bound is found to be q_p=4.37. In the left tail we have an exceedance probability of 0.0036, and in the right tail 0.0460. Please verify these figures yourself, with the table in Appendix B.

Question 11 With a laser-disto-meter four times the same distance has been measured on a calibration test-range. The distance is known, with very high accuracy, and equals 2.894 m. The four observed distances are: y_1 = 2.890 m, y_2 = 2.899 m, y_3 = 2.875 m, and y_4 = 2.886 m. Compute the (empirical) Mean Squared Error (MSE).

Answer 11 The empirical MSE follows from equation (2.18). In this case x = 2.894 m, there are four observations, hence $N = 4$, and the four observations y_1, y_2, y_3 and y_4 are given. Substituting this in equation (2.18) yields 116.5 mm^2.

3

Multi-variate: random vector

So far, we have been dealing with a single random variable. With two laser-disto-meters in place, there are two random variables in parallel, denoted by $\underline{y}_1$ and $\underline{y}_2$ respectively, as we have two separate processes of taking measurements; a sample of $\underline{y}_1$ is denoted by y_1, and a sample of $\underline{y}_2$ by y_2. In this chapter we treat the multi-variate case.

From now on, the notion $\underline{y}$ will refer to a random vector. By default, we assume that this vector has m elements, hence

$$\underline{y} = \begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} \tag{3.1}$$

and a sample of this vector is a set of one sample of each of the m random variables in this vector

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

In the second part of this chapter we address the question of what happens to random variables upon mathematical manipulation. Suppose the two random variables $\underline{y}_1$ and $\underline{y}_2$ are independent, and have variances $\sigma^2_{y_1}$ and $\sigma^2_{y_2}$ respectively, what will be the variance of the sum of the two $\underline{y}_1 + \underline{y}_2$? Propagation laws are a crucial concept to be able to assess the uncertainty in a result, which is computed from a set of measurements.

3.1. Probability density function and moments

The mean of $\underline{y}$, the vector with random variables, is obviously a vector as well. Similar to (2.9) we have

$$E(\underline{y}) = \int_{-\infty}^{+\infty} y f(y)\, dy \tag{3.2}$$

which now is a multiple integral expression; for instance for the element i of this vector we have

$$E(\underline{y}_i) = \int_{y_1=-\infty}^{+\infty} \int_{y_2=-\infty}^{+\infty} \dots \int_{y_i=-\infty}^{+\infty} \dots \int_{y_m=-\infty}^{+\infty} y_i f(y_1, y_2, \dots, y_i, \dots, y_m)\, dy_1 dy_2 \dots dy_i \dots dy_m$$

The mean, or expectation of vector $\underline{y}$ is

$$E(\underline{y}) = \begin{pmatrix} E(\underline{y}_1) \\ E(\underline{y}_2) \\ \vdots \\ E(\underline{y}_m) \end{pmatrix} \tag{3.3}$$

Instead of a single variance σ^2, we now get a full $m \times m$ variance matrix Q_{yy}. Formally the second central moment of the vector $\underline{y}$ reads

$$D(\underline{y}) = E((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T) = \int_{-\infty}^{+\infty} (y - E(\underline{y}))(y - E(\underline{y}))^T f(y)\, dy \tag{3.4}$$

which is a multiple integral expression (the superscript $(\cdot)^T$ denotes the transpose of a vector or matrix). In case vector $\underline{y}$ consists of just a single random variable (m=1), the above form reduces again to (2.10). The variance matrix is

$$D(\underline{y}) = Q_{yy} = \begin{pmatrix} \sigma_{y_1}^2 & \sigma_{y_1 y_2} & \cdots & \sigma_{y_1 y_m} \\ \sigma_{y_2 y_1} & \sigma_{y_2}^2 & \cdots & \sigma_{y_2 y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_m y_1} & \sigma_{y_m y_2} & \cdots & \sigma_{y_m}^2 \end{pmatrix} \tag{3.5}$$

On the diagonal we find the variances for the individual random variables $\underline{y}_1$ through $\underline{y}_m$, and on the off-diagonals, we find the covariances, each time pertaining to a pair of random variables, hence $\sigma_{y_i y_j}$ is the covariance between $\underline{y}_i$ and $\underline{y}_j$, in textbooks on statistics often denoted as $Cov(\underline{y}_i, \underline{y}_j)$. The covariance is

$$\sigma_{y_i y_j} = E((\underline{y}_i - E(\underline{y}_i))(\underline{y}_j - E(\underline{y}_j)))$$

and the correlation coefficient is defined as

$$\rho_{y_i y_j} = \frac{\sigma_{y_i y_j}}{\sigma_{y_i} \sigma_{y_j}}$$

and can be regarded as a normalized covariance, as $|\rho_{y_i y_j}| \leq 1$, [1]. When $\sigma_{y_i y_j} = 0$, and thereby $\rho_{y_i y_j} = 0$, the two random variables $\underline{y}_i$ and $\underline{y}_j$ are said to be uncorrelated.

One could regard the variance $\sigma_{y_i}^2$ as the covariance of $\underline{y}_i$ with itself ($\sigma_{y_i y_i} = \sigma_{y_i}^2$); the square denotes the variance, in order to distinguish it from the standard deviation σ_{y_i}.

As $\sigma_{y_i y_j} = \sigma_{y_j y_i}$, matrix Q_{yy} (3.5) is symmetric. The variance matrix Q_{yy} is a positive (semi) definite matrix, just like the variance (2.10) by definition is a non-negative quantity, meaning that for any vector $u \in R^m$, it holds that the quadratic form $u^T Q_{yy} u \geq 0$ in case it is positive semi-definite, and $u^T Q_{yy} u > 0$ in case it is positive definite.

3.1.1. Multi-variate normal distribution

In case all random variables in vector $\underline{y}$ are normally distributed, their joint distribution is a multi-dimensional, or multi-variate normal distribution, and the PDF $f(y) = f(y_1, y_2, \ldots, y_m)$ is given by

$$f(y) = \frac{1}{\sqrt{|2\pi Q_{yy}|}} e^{-\frac{1}{2}(y-x)^T Q_{yy}^{-1}(y-x)} \tag{3.6}$$

$$= \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{|Q_{yy}|}} e^{-\frac{1}{2}(y-x)^T Q_{yy}^{-1}(y-x)}$$

where $|Q|$ denotes the determinant of matrix Q, and vector x denotes the mean of $\underline{y}$, hence $x = E(\underline{y})$.

3.2. Mean and variance propagation laws

We often transform one random vector into another one. When this transformation is linear, the mean and variance matrix of the new random vector can be fairly easily computed, once the mean and variance matrix of the original random vector are available; this takes place through the so-called propagation laws.

We consider the following linear transformation

$$\underline{v} = R\underline{y} + s \tag{3.7}$$

where vector $\underline{v}$ has n elements, and consequently matrix R has n rows and m columns, and vector s is an n-vector.

The mean of $\underline{v}$ is easily obtained through

$$E(\underline{v}) = E(R\underline{y} + s) = RE(\underline{y}) + s \tag{3.8}$$

where $E(\underline{v})$ is an n-vector. The $n \times n$ variance-matrix of $\underline{v}$ follows as

$$Q_{vv} = RQ_{yy}R^T \tag{3.9}$$

In practice, (3.9) is also referred to as the error propagation law.

Proofs of the above two propagation laws can be found in Appendix A.2.

Finally we state that when vector $\underline{y}$ is normally distributed, then through a linear transformation as (3.7), vector $\underline{v}$ is also normally distributed. Hence, if $\underline{y} \sim N$ then through (3.7) also $\underline{v} \sim N$.

3.3. Example

The height x_1 of a benchmark — monumented in the outer wall of a church tower — has been previously surveyed. The surveyed height of this point 1 is available and denoted as observable $\underline{y}_1$, with standard deviation $\sigma_{y_1} = \sigma$ (e.g. with σ = 3 mm); the surveyed height will be a good estimate for the unknown height x_1, but not be perfect.

Next, we level from point 1 to 2, and eventually from point 2 to 3, see Figure 3.1. The measured height difference $y_{1,2}$ equals the difference of the height of point 2 and the height of point 1, hence $y_{1,2} = x_2 - x_1$, apart of course, from a measurement error; and $y_{2,3} = x_3 - x_2$. In order to properly account for random errors in these measurements, they are regarded as random variables: observables $\underline{y}_{1,2}$ and $\underline{y}_{2,3}$, with standard deviations $\sigma_{y_{1,2}} = \sigma_{y_{2,3}} = \sigma$ (and

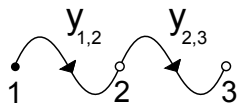


Figure 3.1: Levelling from point 1, via point 2 to point 3.

zero mean measurement error $\underline{y}_{1,2} = x_2 - x_1 + \underline{e}_{1,2}$ and $E(\underline{y}_{1,2}) = x_2 - x_1$, as $E(\underline{e}_{1,2}) = 0$ is assumed zero).

The surveyed height $\underline{y}_1$, and the height difference observables $\underline{y}_{1,2}$ and $\underline{y}_{2,3}$ are uncorrelated.

With the given surveyed height and the measured height differences we can determine the heights of points 2 and 3. The question now is, what uncertainty can be expected in these figures? So, what will be the variability in the determined height for point 3 for instance? Therefore, we need to compute the variance (or standard deviation) of the height (parameter) we determine for this point.

We handle this problem in a structured and systematic way. First the height of point 1 can be estimated; the estimator is trivial

$$\underline{\hat{x}}_1 = \underline{y}_1$$

From the second paragraph of text above we can deduce that

$$\underline{\hat{x}}_2 = \underline{y}_1 + \underline{y}_{1,2}$$

and that

$$\underline{\hat{x}}_3 = \underline{y}_1 + \underline{y}_{1,2} + \underline{y}_{2,3}$$

We see that there is only one straightforward way to get to know the heights of points 2 and 3. The above three equations can be cast in matrix-vector form

$$\begin{pmatrix} \underline{\hat{x}}_1 \\ \underline{\hat{x}}_2 \\ \underline{\hat{x}}_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}}_{M} \begin{pmatrix} \underline{y}_1 \\ \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix}$$

and now resembles (3.7) with the R-matrix as the above 3×3-matrix M. We basically have $\underline{\hat{x}} = M\underline{y}$. We are concerned here with a *linear* transformation.

In order to determine the variances (or standard deviations) of the height estimators (we want to compute $Q_{\hat{x}\hat{x}}$), we would like to apply (3.9). Therefore, we still need the variance matrix of the observables Q_{yy}. The third paragraph of text says that all y's are uncorrelated, and we know that they all have standard deviation equal to σ. Hence variance matrix Q_{yy} is just an identity matrix, scaled by σ^2. Applying (3.9) yields

$$Q_{\hat{x}\hat{x}} = MQ_{yy}M^T = \sigma^2 \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

and the requested variances can be obtained from the diagonal, hence $\sigma^2_{\hat{x}_1} = \sigma^2$, $\sigma^2_{\hat{x}_2} = 2\sigma^2$, and $\sigma^2_{\hat{x}_3} = 3\sigma^2$.

What we see here is an *accumulation* of uncertainty, when you add together observables each with associated uncertainties. The estimator for the height of point 2 has a variance which is double the one for point 1, where we started. And for point 3, this is even a factor of three. We return to this phenomenon with Figure 5.14 on the so-called open levelling line. You can also see that the three height-estimators are correlated, the off-diagonal elements are not equal to zero; this also makes sense as they share part of the information they are based on; for instance, the surveyed height $\underline{y}_1$ appears in all three equations (for $\underline{\hat{x}}_1$, $\underline{\hat{x}}_2$, and $\underline{\hat{x}}_3$).

3.4. Non-linear mean and variance propagation laws

In section 3.2 we considered a *linear* relation in equation (3.7). In practice, we may also face a *non-linear* relation:

$$\underline{v} = G(\underline{y}) \tag{3.10}$$

where G is a mapping from R^m to R^n; vector $\underline{v}$ has n elements (random variables), and vector $\underline{y}$ has m elements (random variables).

Theory is available to propagate the probability density function of $\underline{y}$, $f(y)$, into the one of $\underline{v}$, $f(v)$. In this section we restrict to just the expectation and the variance matrix of random vector $\underline{v}$.

An approximation for the expectation of $\underline{v}$ is given by

$$E(\underline{v}_i) \approx G_i(E(\underline{y})) + \frac{1}{2}\mathrm{trace}\left(\left.\frac{\partial^2 G_i}{\partial y y^T}\right|_{E(\underline{y})} Q_{yy}\right) \tag{3.11}$$

for $i = 1, \ldots, n$, and where trace means taking the sum of the diagonal elements, in this case of the matrix product $\left.\frac{\partial^2 G_i}{\partial y y^T}\right|_{E(\underline{y})} Q_{yy}$, which is an $m \times m$ matrix. Matrix $\frac{\partial^2 G_i(y)}{\partial y y^T}$ is the so-called Hessian matrix (in this case with dimensions $m \times m$), and contains the second order partial derivatives of the non-linear function $G_i(y)$. This equation shows that $E(\underline{v}_i) \neq G_i(E(\underline{y}))$; supplying the mean or expectation of $\underline{y}$ in the non-linear function G, does *not* yield the mean/expectation of $\underline{v}$! With approximation (3.11), there is already an extra term, which depends on the variance matrix of $\underline{y}$. The proof of (3.11) can be found in Appendix A.3.

An approximation for the variance matrix of $\underline{v}$ is given by

$$Q_{vv} \approx \left.\frac{\partial G}{\partial y^T}\right|_{E(\underline{y})} Q_{yy} \left.\frac{\partial G}{\partial y^T}\right|^T_{E(\underline{y})} \tag{3.12}$$

where $\frac{\partial G(y)}{\partial y^T}$ is an $n \times m$ matrix, containing, as rows, the gradient vectors of non-linear functions $G_i(y)$, with $i = 1, \ldots, n$, all evaluated at $E(\underline{y})$. Defining the matrix $M = \left.\frac{\partial G}{\partial y^T}\right|_{E(\underline{y})}$, the above variance propagation law becomes $Q_{vv} \approx M Q_{yy} M^T$, which is then very similar to (3.9), though (3.12) is an approximation. The proof of (3.12) can be found in Appendix A.3.

In practice, the expectation of the observable vector $\underline{y}$ may not be known, hence the derivatives $\frac{\partial G(y)}{\partial y^T}$ and $\frac{\partial^2 G_i(y)}{\partial y y^T}$ are evaluated at a sample value instead.

3.5. Exercises and worked examples

This section presents a problem with a worked answer on error propagation.

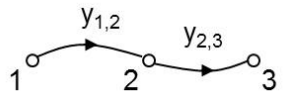


Figure 3.2: Levelling from point 1, via point 2 to point 3.

Question 1 As shown in Figure 3.2, the height-difference between point 1 and point 2 has been leveled, and the height-difference between points 2 and 3 has been leveled. Both observables have a standard deviation of $\frac{3}{2}$ mm. The two observables are uncorrelated. Based on these two observed height-differences, the height-difference between point 1 and point 3 is determined. What is the standard deviation of this height-difference?

Answer 1 The height-difference between point 1 and point 3 follows as $y_{1,3} = y_{1,2} + y_{2,3}$. Or, in terms of a matrix and a vector (and random variables)

$$\underline{y}_{1,3} = \underbrace{\begin{pmatrix} 1 & 1 \end{pmatrix}}_{R} \begin{pmatrix} \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix}$$

With the variance matrix of $\underline{y}_{1,2}$ and $\underline{y}_{2,3}$ being

$$D\begin{pmatrix} \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix} = \begin{pmatrix} \frac{9}{4} & 0 \\ 0 & \frac{9}{4} \end{pmatrix}$$

application of equation (3.9) yields

$$\sigma^2_{y_{1,3}} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{9}{4} & 0 \\ 0 & \frac{9}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{18}{4}$$

Hence, the requested standard deviation becomes $\sigma_{y_{1,3}} = \frac{3}{2}\sqrt{2}$ mm. By adding two equally precise observables, the result has a standard deviation which is worse by a factor of $\sqrt{2}$.

Question 2 Given is a random variable $\underline{y}$ with mean equal to $E(\underline{y}) = 40$ mm, and a standard deviation of $\sigma_y = 3$ mm. Compute the expectation and standard deviation of $\underline{v}$, with $\underline{v} = \underline{y}^2$.

Answer 2 The given transformation $\underline{v} = \underline{y}^2$ is *non-linear*, hence the propagation laws for a linear transformation can not be applied. We have to use (3.11) and (3.12) instead. With $G(y) = y^2$, we obtain $\frac{\partial G(y)}{\partial y} = 2y$ and $\frac{\partial^2 G(y)}{\partial y^2} = 2$ (which have to be evaluated at $E(\underline{y}) = 40$ mm), and this results into

$$E(\underline{v}) \approx 40^2 + \frac{1}{2} 2 \cdot 9 = 1609 \text{ mm}^2$$

$$\sigma^2_v \approx 80 \cdot 9 \cdot 80 = 57600 \text{ mm}^4$$

hence $\sigma_v \approx 240$ mm^2. Mind that the expectation $E(\underline{v})$ deviates, though slightly, from $(E(\underline{y}))^2 = 1600$ mm^2. The larger the uncertainty in observation y, the larger its variance, and hence the larger the effect of the second term of (3.11).

4

Observation modelling and parameter estimation

4.1. Introduction

In this chapter we propose a structured way of dealing with measurements. For each measurement we will formulate an equation, which expresses the observed quantity in terms of parameters of interest. In chapter 2, we used a very simple example: we had a distance observable $\underline{y}$, and related it to the unknown (true) distance x. The observation equation reads:

$$\underline{y} = x + \underline{e} \tag{4.1}$$

This equation says that distance observable $\underline{y}$ is equal to the unknown, true distance, plus a random error term $\underline{e}$ representing the measurement error. When the instrument is fine, and the measurement is carried out properly, we can expect that the measurement (sample) y is not perfect, hence we will *not* have $y = x$, though y should be close to x. The symbol e accounts for the measurement error, which, due to uncontrollable effects, will be positive one time, negative a next time, small, somewhat bigger etc. But, on average, the measurement is expected to be spot-on, that is, on average the measurement error is zero, $E(\underline{e}) = 0$, and therefore $E(\underline{y}) = x$. The spread we can expect once we would take many repeated measurements, and the uncertainty present in a single measurement, is quantified by standard deviation σ. With a laser-disto-meter for instance, a value of σ=0.002 m is a fair number. Then we have $\sigma_y = \sigma_e = \sigma$.

In this simple example, the unknown parameter of interest is the distance. In practice survey-problems can be much more complicated. Measurements can be angles and distances, and the unknown parameters are (position) coordinates of certain points of interest. An observation equation is an equation, expressing the observed quantity in terms of the unknown parameters. When you express a distance in terms of position coordinates (in a two or three dimensional space), this equation may even be non-linear. Later we return to non-linear observation equations, but first we address the linear model of observation equations.

4.2. Observation modelling

The model of observation equations is given by

$$E(\underline{y}) = Ax \; ; \; D(\underline{y}) = Q_{yy} \tag{4.2}$$

where y is the m-vector of observations, x the n-vector of unknown parameters, A the $m \times n$ design matrix (of full rank equal n) containing a total of mn *known* coefficients, and Q_{yy} the $m \times m$ variance-matrix (rank equal m). The observation equations are assumed to be linear here, so (4.2) is a linear system. The system can also be written with the measurement errors occuring explicitly

$$\underline{y} = Ax + \underline{e}\ ;\ D(\underline{y}) = D(\underline{e}) = Q_{yy} \tag{4.3}$$

with $E(\underline{e}) = 0$.

The linear regression model in chapter 17 of [1], with offset α and slope β is indeed a simple example of the above linear model of observation equations; these two parameters α and β would together constitute vector x.

4.2.1. Example

In the example on levelling in the previous chapter (Figure 3.1), the unknown parameters of interest are the heights of certain points (as x_2 and x_3), and through levelling we measure height-*differences*: e.g. $y_{2,3} = x_3 - x_2$. The model of observation equations reads

$$\underbrace{E\begin{pmatrix} \underline{y}_{1} \\ \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix}}_{E(\underline{y})} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{x}$$

In this special case, $m = n = 3$, and matrix A is a square and invertible matrix, and indeed we had $\hat{\underline{x}} = M\underline{y}$ which, with $M = A^{-1}$, equals $\hat{\underline{x}} = A^{-1}\underline{y}$. The above system of equations is solved for by simply an inversion of matrix A, but this can be done *only* in the case with a square and invertible matrix A.

4.2.2. Redundancy

When there are n unknown parameters (in vector x) to determine, we need at least $m = n$ observations (in vector y). In practice, typically (some) more measurements are done than strictly necessary (for good reasons), and in this primer we will always consider the case $m \geq n$. The excess of m over n is referred to as redundancy.

4.3. Parameter estimation

Given a vector of observations y we have to determine the values for the entries in vector x. In the simple example above (in section 4.1), computing the estimator for the unknown distance is very simple: $\hat{\underline{x}} = \underline{y}$, once we have a sample for y, we simply equate the result for x to this observation: $\hat{x} = y$. By the hat-symbol, we denote that this is an estimate for the unknown distance. We will never know the actual true distance, but we can make a guess, or estimate of it, based on our observation. The true distance x is still unknown, but, the 'best to our knowledge' guess of it is $\hat{x} = y$.

When, with multiple measurements and unknown parameters, we have $m = n$, then $\hat{x} = A^{-1}y$, as A was assumed a full rank matrix before (and with $m = n$ it is also a square matrix). This is still a fairly simple case to solve.

As said before, in practice we typically deal with the case $m > n$, hence there are 'too many' measurements, and for the reason of measurement errors, the system $y = Ax$ will not be a consistent system; given the measurements, there will not be a solution for x which satisfies all equations in the system $y = Ax$.

4.3.1. Example

Suppose we measure the same distance twice. One time, we get y_1=7.452 m, and next we get y_2=7.454 m. Though, these two measurements are close together, we can not make them fit perfectly in the assumed measurement model. For both observables holds that they are related to the same, unknown distance, hence

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (x)$$

but we can never find a value for x which satisfies y_1=7.452 m $= x$, and at the same time y_2=7.454 m $= x$. Therefore $y \neq Ax$.

4.3.2. Least-squares estimate

The least-squares principle provides a solution to a system of observation equations which is redundant ($m > n$) and inconsistent ($y \neq Ax$ due to measurement errors). The estimate for the unknown parameter vector x shall be computed according to

$$\hat{x} = (A^T A)^{-1} A^T y \tag{4.4}$$

The name least-squares explains, as we are trying to make the system $y = Ax$ consistent, by using — instead of the vector of observations y — a vector of estimated, 'slightly adapted', observation values $\hat{y}$, that is $\hat{y} = A\hat{x}$ (properly said, $\hat{y}$ is the estimate for the mean of the observable y). Of course, one could choose very weird values for $\hat{x}$ and consequently for $\hat{y}$, and arrive at a consistent system, but, we would like to have $\hat{y}$ **close** to y, as afterall, we expect the measured values to be close to the true values (the measurements do contain useful information). So, we are going to 'change', or 'adjust' the observed values not too much.

The underlying criterion is, given vector y, to find x such that the length (norm) of the vector $(y - Ax)$ is smallest

$$\min_x \|y - Ax\|^2 \tag{4.5}$$

The solution $\hat{x}$ to this minimization problem yields the smallest length, hence the vector $(y - A\hat{x}) = (y - \hat{y})$ is shortest (the norm is obtained by squaring all entries of the vector and summing them, and this should be at minimum, hence the term least-squares). A proof that solution (4.4) results from this minimization is provided in Appendix A.4.

4.3.3. Example

When we observe the same distance x twice, then the least-squares estimator for the unknown distance equals just the average of the two measurements:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{A} x$$

and using (4.4) we find that $\hat{x} = \frac{1}{2}(y_1 + y_2)$.

In section 2.5.2 we actually used a model $y = Ax$ with $y = (y_1, \dots, y_N)^T$ and $A = (1, \dots, 1)^T$ with $m = N$ and $n = 1$, just a single unknown parameter.

4.3.4. Minimum variance estimator

With least-squares we have not used yet all available information. We used just the functional relations between observations and unknown parameters cast in $y = Ax$. In (4.2) we have also available, the variance matrix of the observables Q_{yy} (and it does not occur at all in (4.4)). The parameter estimation process should use this information, to be optimal. Observables with small variances should get more weight in the solution than observables with larger variances. The first ones are more precise than the latter, and the resulting estimator should reflect this. The information contained in the variance matrix Q_{yy} is taken into account in an optimal way, by the following estimator, which we state without proof:

$$\underline{\hat{x}} = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} \underline{y} \tag{4.6}$$

You can easily see that, when the variance matrix Q_{yy} is a (scaled) identity matrix, we are back at (4.4). Note that the above estimator $\underline{\hat{x}}$ for the vector of unknown parameters is a random vector, as it is a function of the vector of observables $\underline{y}$. The least-squares principle is not concerned with statistical aspects, and therefore no 'underscores' are used in equation (4.4).

The above estimator (4.6) has three distinct properties. The first one is that $\underline{\hat{x}}$ is a **linear** function of the observables $\underline{y}$, through matrix $n \times m$-matrix $(A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1}$. Second, the estimator is **unbiased**, that is, on average it delivers values which agree with the unknown true values; taking the expectation of (4.6) and using (3.8) we get $E(\underline{\hat{x}}) = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} E(\underline{y})$, which, with $E(\underline{y}) = Ax$, yields $E(\underline{\hat{x}}) = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} Ax = x$, thus indeed $E(\underline{\hat{x}}) = x$. And finally the estimators in $\underline{\hat{x}}$ have **minimum variance**. The vector $\underline{\hat{x}}$ is a (linear) function of vector $\underline{y}$, which is a vector with random variables, hence all entries are subject to (random) measurement errors. Therefore, the entries of vector $\underline{\hat{x}}$ will also be subject to uncertainty. By using (3.9) on

$$\underline{\hat{x}} = \underbrace{(A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1}}_{H} \underline{y}$$

we obtain the $n \times n$ variance matrix for the estimator $\underline{\hat{x}}$ as

$$Q_{\hat{x}\hat{x}} = H Q_{yy} H^T = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} Q_{yy} Q_{yy}^{-1} A (A^T Q_{yy}^{-1} A)^{-1}$$

which simplifies into

$$Q_{\hat{x}\hat{x}} = (A^T Q_{yy}^{-1} A)^{-1} \tag{4.7}$$

It can be shown that this matrix — among the variance matrices of all possible linear and unbiased estimators — has minimum trace, that is, this estimator is best in the sense that the sum of all n variances together is smallest. The least-squares solution (4.4) only shares the first two properties with the minimum variance solution (4.6), hence being linear and unbiased.

The estimator (4.6) is unbiased, and therefore minimum variance implies best accuracy, cf. section 2.6. The estimator (4.6) is also known as the Best Linear Unbiased Estimator (BLUE). It provides a generalization of the least-squares estimation of offset α and slope β in chapter 22 of [1]. An example on regression, or line-fitting is presented in chapter 6. With the above BLUE one can compute a properly weighted least-squares solution (for vector x) to any proper linear problem. As the *inverse* of matrix Q_{yy} is involved in (4.6), precise observables (small variances) receive larger weights, and less precise observables (large variances) receive smaller weights.

4.3.5. Example

We repeat the example of observing the same distance twice. However in this case the first measurement is made with a better instrument, and the standard deviation of $\underline{y}_1$ is equal to $\frac{1}{2}$. The second measurement is carried out with the default instrument, as before, and the standard deviation of $\underline{y}_2$ equals 1. We have

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{A} x \; ; \; D\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} = Q_{yy} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{pmatrix}$$

and using (4.6) we find that $\underline{\hat{x}} = \frac{4}{5}\underline{y}_1 + \frac{1}{5}\underline{y}_2$, hence we arrive at the weighted mean, rather than the ordinary mean. Observable $\underline{y}_1$ has a variance which is four times smaller than the variance of observable $\underline{y}_2$, and therefore the coefficient in the final estimator is four times bigger, $\frac{4}{5}$ versus $\frac{1}{5}$, and because $\frac{4}{5} + \frac{1}{5} = 1$ all information is taken into account (total weight equals 1, and the estimator is unbiased).

Finally, with (4.7), one can see that the variance of the estimator is $\sigma_{\hat{x}}^2 = \frac{1}{5}$, which is better, and smaller than any of the two observables at the input!

Collecting redundant measurements typically leads to inconsistent systems of equations, but eventually improves the precision of the estimator!

4.4. Non-linear observation equations

The chapter started off from a *linear* model of observation equations (4.2), and proposed the least-squares estimate and the Best Linear Unbiased Estimator (BLUE). It provides a nice theoretical framework. But, in practice there are hardly any measurements which carry a linear relation with the unknown parameters. Indeed, a levelled height-difference is linear in the (unknown) height-parameters of the two points. But a distance is clearly a non-linear function of the coordinate differences, see (5.4).

The approach to systems of *non-linear* observation equations will be to *approximate* them by linear equations. The originally non-linear equations will be linearized with respect to the unknown parameters and the resulting system of linear(ized) equations will be treated using (4.4) or (4.6) on (4.2). Of course, we need to make sure that the approximation we make, is sufficiently good.

The model of observation equations

$$E(\underline{y}) = F(x) \; ; \; D(\underline{y}) = Q_{yy} \tag{4.8}$$

where matrix-vector product Ax has been replaced by $F(x)$, a non-linear function, or mapping from R^n to R^m; it is a collection of m non-linear functions of n parameters.

The mapping $F(x)$ can be detailed as

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2, \dots, x_n) \\ f_2(x_1, x_2, \dots, x_n) \\ \vdots \\ f_m(x_1, x_2, \dots, x_n) \end{pmatrix}$$

or in words, (the expectation of) each observable $\underline{y}_i$ is a (scalar) function f_i of n parameters, namely $x_1, x_2, \dots, x_n$ (f_i is a mapping from R^n to R), and row-by-row we do have m such functions, $i = 1, \dots, m$. Examples of function f_i are given by (5.2) and (5.4).

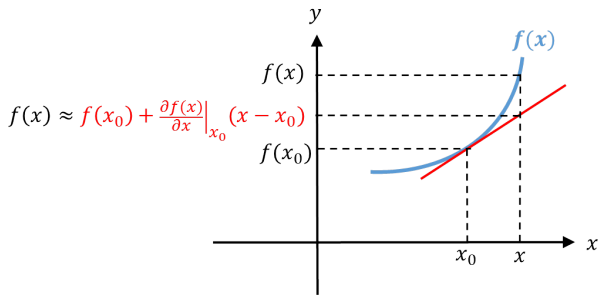


Figure 4.1: Linearization of the non-linear function $\underline{y} = f(x) + \underline{e}$ $(m = n = 1)$. The red line represents the first-degree Taylor polynomial of $f(x)$, shown by the blue curve, centered at x_o.

4.4.1. Linearization

The function $F(x)$ is approximated by the zero-order and first-order terms. Higher-order terms are neglected. We rely on the **Taylor series**, using a point x_o, which is presumed to be reasonably close to the actual/true x; vector $x_o = (x_{1,o}, x_{2,o}, \dots, x_{n,o})^T$ contains approximate values for all n unknown parameters. So,

$$F(x) \approx F(x_o) + \frac{\partial F(x)}{\partial x^T}\bigg|_{x_o}(x - x_o)$$

where the zero order term and the first derivative are evaluated at x_o. To compute the first order derivative, all m non-linear functions, one by one, are differentiated with respect to x_1, x_2, until x_n, hence this turns into an $m \times n$ matrix. The first row of this matrix reads $\frac{\partial f_1}{\partial x_1}\frac{\partial f_1}{\partial x_2}\cdots\frac{\partial f_1}{\partial x_n}$, with the partial derivatives evaluated at x_o.

Substituting this first-degree (linear) approximation of $F(x)$ into (4.8) yields

$$E(\underline{y}) \approx F(x_o) + \frac{\partial F(x)}{\partial x^T}\bigg|_{x_o}(x - x_o)\ ;\ D(\underline{y}) = Q_{yy}$$

or

$$E(\underbrace{\underline{y} - F(x_o)}_{\Delta\underline{y}}) \approx \underbrace{\frac{\partial F(x)}{\partial x^T}\bigg|_{x_o}}_{A}\underbrace{(x - x_o)}_{\Delta x}\ ;\ D(\underline{y} - F(x_o)) = Q_{yy} \tag{4.9}$$

Here, the first-order derivative (of dimensions $m \times n$) takes the role of the design matrix A. The vector of observations y is replaced by $y - F(x_o)$, that is, the observations minus the observations as they are computed based on just the approximate value x_o for the unknown parameters: $y_o = F(x_o)$. And through least-squares we will not be estimating the vector of unknown parameters x, but $(x - x_o)$ instead, the differences of x with respect to the approximate values x_o that we already introduced.

4.4.2. Estimation

Accepting the approximation made in (4.9), the estimator for Δx follows as

$$\underline{\Delta\hat{x}} = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1}(\underline{y} - F(x_o))$$

with

$$Q_{\Delta\hat{x}\Delta\hat{x}} = (A^T Q_{yy}^{-1} A)^{-1}$$

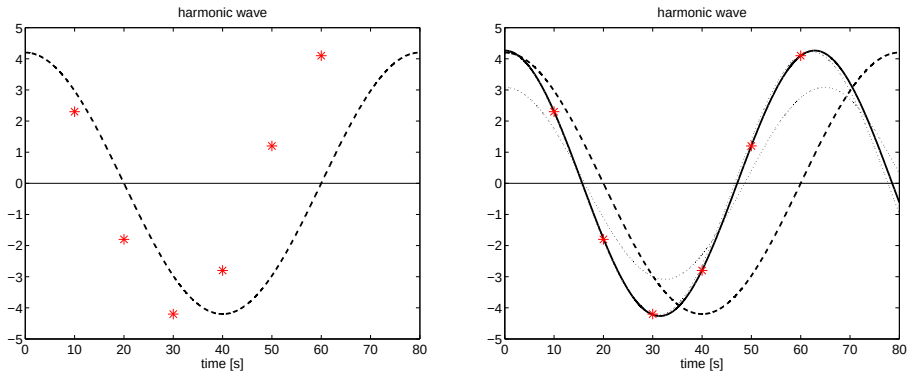


Figure 4.2: Example of non-linear estimation problem, fitting a cosine with unknown amplitude and frequency through a given set of data points, indicated by red asterisks (at left), and results of Gauss-Newton iterated estimation (at right).

and the eventual estimator for the vector of unknown parameter is obtained as

$$\underline{\hat{x}} = x_o + \underline{\Delta\hat{x}}$$

with

$$Q_{\hat{x}\hat{x}} = Q_{\Delta\hat{x}\Delta\hat{x}}$$

The model (4.9) is only an approximation of the actual non-linear model. For the approximation to be good and valid, the approximate value x_o should be close to the true unknown value x. Therefore, the above procedure is repeated (iterated). One starts off with as good as possible guess for x_o, next one determines the estimate $\hat{x}$, and then takes this estimate as a new approximate value, as likely it is closer to the true, unknown x than x_o was, and repeats the above procedure (and on and on, if necessary). This iterative procedure is known as the Gauss-Newton method. A further discussion of this method, its properties and convergence behaviour is beyond the scope of this primer (as well as alternatives to the Gauss-Newton method).

Concerns with regard to non linear estimation are briefly mentioned in Appendix A.5.

4.4.3. Example

A certain object moves harmonically, for instance a tall rise building vibrating under the load of wind. The motion in one dimension, as a function of time, can be described by a cosine, with zero phase offset (our assumption/simplification in this example), but with unknown amplitude and frequency. The positions of the object are observed at times $t_1 = 10$, $t_2 = 20$, $t_3 = 30$, $t_4 = 40$, $t_5 = 50$, and $t_6 = 60$ seconds (timing is assumed to be perfect), and the measurements are denoted by y_1, y_2, y_3, y_4, y_5 and y_6 (and these position measurements are subject to errors). The observation model reads

$$E(\underline{y}_i) = x_1 \cos(2\pi x_2 t_i) \tag{4.10}$$

for $i = 1, \ldots, 6$. Unknown parameters are the amplitude x_1, and the frequency x_2 in Hertz.

The observation values are $y_1 = 2.3$, $y_2 = -1.8$, $y_3 = -4.2$, $y_4 = -2.8$, $y_5 = 1.2$ and $y_6 = 4.1$, and they are shown in Figure 4.2 by the red asterisks. The observation model is non-linear in the unknown paramaters x_1 and x_2, and will now be approximated by zero and first order terms. In order to do so, one needs approximate values for the unknown

	iteration	0	1	2	3	4	5
$y_1 = 2.3$		2.9698	1.7446	2.2474	2.3027	2.3046	2.3046
$y_2 =$-1.8		0.0000	-1.1066	-1.8431	-1.7699	-1.7729	-1.7729
$y_3 =$-4.2	$F(x_o)$	-2.9698	-2.9975	-4.2056	-4.2163	-4.2210	-4.2210
$y_4 =$-2.8		-4.2000	-2.2871	-2.6249	-2.7887	-2.7898	-2.7898
$y_5 = 1.2$		-2.9698	0.4080	1.4168	1.2011	1.2054	1.2054
$y_6 = 4.1$		0.0000	2.7491	4.1302	4.0874	4.0928	4.0928
		-0.6698	0.5554	0.0526	-0.0027	-0.0046	-0.0046
		-1.8000	-0.6934	0.0431	-0.0301	-0.0271	-0.0271
	$y - F(x_o)$	-1.2302	-1.2025	0.0056	0.0163	0.0210	0.0210
		1.4000	-0.5129	-0.1751	-0.0113	-0.0102	-0.0102
		4.1698	0.7920	-0.2168	-0.0011	-0.0054	-0.0054
		4.1000	1.3509	-0.0302	0.0126	0.0072	0.0072
$(x_1)_o =$4.2000	$\hat{x}_1$	3.0818	4.2308	4.2595	4.2641	4.2641	4.2641
$(x_2)_o =$0.0125	$\hat{x}_2$	0.0154	0.0161	0.0159	0.0159	0.0159	0.0159

Table 4.1: Gauss-Newton iteration: the left column shows the observed values y_1, y_2, y_3, y_4, y_5 and y_6, and the two approximate values for the unknown parameters at bottom $(x_1)_o$ and $(x_2)_o$. The columns 0 through 5 show the results of the 5 iteration steps; on top the $F(x_o)$, $y - F(x_o)$ and the resulting estimate $\hat{x}$ at bottom.

parameters. Our choice is to simply set the amplitude $(x_1)_o$ equal to the largest observed value (in absolute sense) $(x_1)_o = |y_4|$, and the data points in Figure 4.2 seems to show roughly one period of the harmonic motion in a time span of 80 seconds, hence the frequency is set to $(x_2)_o = \frac{1}{80}$ (Hz). The dotted line in the graph on the left presents the harmonic wave defined by these approximate values, hence $y_o(t) = (x_1)_o \cos(2\pi(x_2)_o t)$.

The graph on the right in Figure 4.2 shows again the harmonic wave defined by the initial approximate values, and also the results (estimates) of the first and second iteration (in thin dotted lines), and eventually the result of the last step, the fifth iteration, as a solid line: $\hat{y}(t) = \hat{x}_1 \cos(2\pi\hat{x}_2 t)$. One can clearly see that with the initial approximate values the harmonic wave was a bit off, and that a few iterations quickly make the wave fit to the observated data points. Numerical results are reported in table 4.1. The values $F(x_o)$ computed for the observations in each iteration, directly follow from using the values for $\hat{x} = (\hat{x}_1, \hat{x}_2)^T$ obtained in the step before.

It should be noted that the approximate values for the unknown amplitude and frequency should be carefully chosen — not any value will do. The problem is in fact highly non-linear — the cosinus is a highly curved function, in particular for high(er) frequencies. The approximate values should be sufficiently close already to the true, but unknown values for the parameters to be estimated.

5

Land-surveying

Land-surveying is in support of Civil Engineering activities, as there is need for knowledge about shape, attitude and location (position) of objects, topography and the Earth's surface. Measurements are done to gather this knowledge, and the knowledge is eventually presented — most often — in terms of position coordinates. In this chapter we present the most common types of measurements in land-surveying. The two basic measurement types are angle (or azimuth, or direction), and distance. And they are actually closely related, as the angle — expressed in radians — equals the quotient of the distance along the arc of a circle, and its radius.

As Civil Engineering activities mostly concern local and regional areas, we use a two dimensional (Euclidean) geometry in the (local) horizontal plane. The height is used separately as a third dimension. Azimuths, directions, angles, and distances are parametrized in terms of two-dimensional coordinates. The coordinates of a certain point i read $((x_1)_i, (x_2)_i)$.

Next, in section 5.5 we introduce the theory necessary to evaluate the precision of the resulting position coordinate estimators. The last section of this chapter provides a review of elementary measurement set-ups, together with an analysis of precision of the resulting position coordinates.

5.1. Levelled height-difference

A height-difference y_{ij} obtained through levelling is simply the difference of the heights of the two points involved:

$$E(\underline{y}_{ij}) = x_j - x_i \tag{5.1}$$

This is a linear equation in the unknown parameters x_i and x_j, and has been dealt with before, see section 3.3.

5.2. Azimuth and angle measurements

The azimuth a_{ij} is the argument of the line-segment from point i to (target) point j.

$$E(\underline{a}_{ij}) = \arctan \frac{(x_1)_j - (x_1)_i}{(x_2)_j - (x_2)_i} \tag{5.2}$$

see Figure 5.1. The azimuth provides the angle of the line-segment from point i to j, with respect to a fixed reference direction, typically the x_2-axis. For measurements of azimuth we use the symbol a rather than y; for distance we use l in the next section, rather than y.

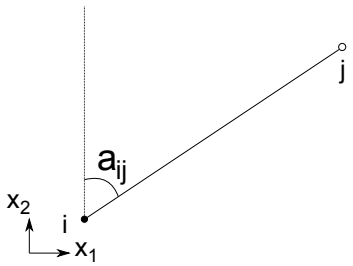


Figure 5.1: Azimuth measurement a_{ij}.

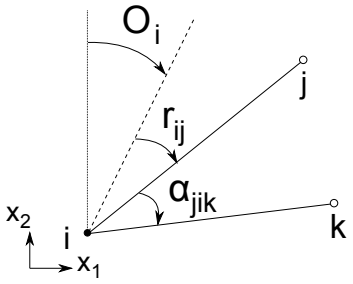


Figure 5.2: Measurement of direction r_{ij}, and measurement of angle α_{jik}.

For an observation of direction r_{ij}, the zero-reference — indicated by the dashed line in Figure 5.2 — has an arbitrary offset (angle) with respect to the second coordinate axis x_2. This unknown offset parameter enters the observation equation (5.2); it should be subtracted from the right-hand side. All observations of direction taken with a single set up of the instrument, share the same orientation offset.

$$E(\underline{r}_{ij}) = \arctan \frac{(x_1)_j - (x_1)_i}{(x_2)_j - (x_2)_i} - O_i$$

An angle observation α_{jik} is just the difference of two azimuths, hence the angle α_{jik} at point i from point k to point j, is just $a_{ik} - a_{ij}$, see also Figure 5.2.

With section 4.4, linearization of (5.2) with azimuth a_{ij} expressed in radians, yields

$$E(\underline{\Delta a}_{ij}) = -\frac{(x_2)_{ij,o}}{l_{ij,o}^2}(\Delta x_1)_i + \frac{(x_1)_{ij,o}}{l_{ij,o}^2}(\Delta x_2)_i + \frac{(x_2)_{ij,o}}{l_{ij,o}^2}(\Delta x_1)_j - \frac{(x_1)_{ij,o}}{l_{ij,o}^2}(\Delta x_2)_j \tag{5.3}$$

where $(x_1)_{ij,o} = (x_1)_{j,o} - (x_1)_{i,o}$, $(x_2)_{ij,o} = (x_2)_{j,o} - (x_2)_{i,o}$, and $l_{ij,o}^2 = (x_1)_{ij,o}^2 + (x_2)_{ij,o}^2$. The above equation follows by using that $d \arctan(x)/dx = 1/(1 + x^2)$.

5.3. Distance measurements

The Euclidean distance between points i and j is the length of the line-segment from point i to point j.

$$E(\underline{l}_{ij}) = \sqrt{((x_1)_j - (x_1)_i)^2 + ((x_2)_j - (x_2)_i)^2} \tag{5.4}$$

see Figure 5.3.

Linearization of (5.4) yields

$$E(\underline{\Delta l}_{ij}) = -\frac{(x_1)_{ij,o}}{l_{ij,o}}(\Delta x_1)_i - \frac{(x_2)_{ij,o}}{l_{ij,o}}(\Delta x_2)_i + \frac{(x_1)_{ij,o}}{l_{ij,o}}(\Delta x_1)_j + \frac{(x_2)_{ij,o}}{l_{ij,o}}(\Delta x_2)_j \tag{5.5}$$

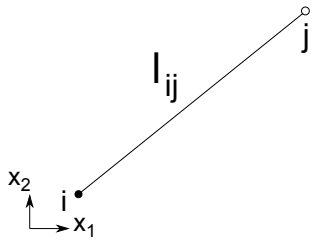


Figure 5.3: Measurement of distance in two dimensions between points i and j.

where $(x_1)_{ij,o} = (x_1)_{j,o} - (x_1)_{i,o}$, $(x_2)_{ij,o} = (x_2)_{j,o} - (x_2)_{i,o}$, and the approximate value for the distance $l_{ij,o} = \sqrt{(x_1)^2_{ij,o} + (x_2)^2_{ij,o}}$.

In practice distance measurements can be subject to a scale factor and/or to an offset. These aspects are beyond the scope of this primer.

5.4. Idealisation

In the mathematical model of observation equations $\underline{y} = Ax + \underline{e}$ (chapter 4), the parameters x, for instance representing position coordinates of a point, are deterministic quantities. Each parameter represents a single — though unknown — numerical value.

Next, we should realize that — for the purpose of surveying and making maps — we are *modeling* the Earth's surface and its topography and objects by means of basic geometric entities. The real world is reduced to points, lines and polygons/areas. For this reducation the surveyor relies on his insight and experience.

Nevertheless we should realize that there is a random component involved in this process. The corner of a brick-wall building can be identified quite well, though, looking at the millimeter scale, you will see that not all bricks in a wall are perfectly aligned. For the center or border line of a ditch, the reduction will be much more difficult; how can we identify the (shape of the) ditch in rough terrain? Where does the ditch start and where does it end? The (additional) uncertainty may increase to the centimeter or even decimeter level in this case. The additional random component is referred to as the *idealisation error*. Figure 5.4 shows an example of reducing reality to points and lines.

Idealisation is about the question how well we can *identify* what we actually measure. Upon concluding the surveying exercise, we can make for instance a statement about the distance between two objects on Earth and when stating the precision of this distance, we should account for the *measurement precision*, and also for the *idealisation precision*, in order to present realistic figures. In this primer however, we will *not* account for this additional random component in surveying. To say, our assumption is that we can identify the objects of interest infinitely precise.

5.5. Analysis of measurement set-up: confidence ellipse

In this section we present the confidence ellipse as a concept for evaluating the precision of a random vector, and this is instrumental to (performance) analysis of a survey measurement set-up. The outcome of surveying (that is, taking measurements and carrying out their consequent processing) is generally a set of *position coordinates*. We take measurements, and we want to know where (a certain identified point on) an object is located. Observables are turned into estimators for position coordinates, and the quality of these estimators needs to meet requirements. An important aspect of quality is *precision*, and we present it in terms

Figure 5.4: Example of idealisation and generalisation. This country-road you find on the map typically as a (series of) straight line-segment(s). In reality the edge of the road, shown in orange, which separates the asphalt-surface and the grass along the road, is really not a straight line. Geometric entities in maps, like points and lines have infinitesimally small dimensions. A line connecting two points for instance is infinitesimally thin, whereas objects in reality, like the edge of the road, have real dimensions. This reduction, or approximation of dimensions is referred to as idealisation and generalisation. Accurately surveying the actual edge with twists and turns every centimeter would be an unrealistic (and very costly) job.

of a confidence ellipse (for the estimators of the two position coordinates of the point under consideration).

The confidence ellipse is an area (in the two dimensional space) centered at the estimated position, which contains the true, but unknown position, with a certain stated probability. In case the position is described with just a single coordinate (for instance when we are only interested in the height) the confidence ellipse turns into an interval. And for a three-dimensional position coordinates vector, it is an ellipsoid. One can also use the general term confidence region, rather than confidence ellipse, to suit any dimension. In the sequel we work with two-dimensional position coordinate vectors. One can for instance think of the indicated center of a pillar to support a future bridge-deck, and be concerned with the (horizontal) coordinates of this point of interest: $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. We are about to analyze the precision of the obtained position coordinate estimators $\underline{\hat{x}}$, which typically is a function of the observables $\underline{\hat{x}} = G(\underline{y})$ (non-linear or linear), with variance matrix $Q_{\hat{x}\hat{x}}$. The estimators are assumed to be *normally distributed* here (remember that normally distributed observables yield normally distributed estimators, with section 3.2 and equation (4.6)).

We repeat equation (3.6), but now applied to $\underline{\hat{x}}$, rather than $\underline{y}$ in (3.6).

$$f(\hat{x}) = \frac{1}{\sqrt{|2\pi Q_{\hat{x}\hat{x}}|}} e^{-\frac{1}{2}(\hat{x}-E(\underline{\hat{x}}))^T Q_{\hat{x}\hat{x}}^{-1}(\hat{x}-E(\underline{\hat{x}}))} \tag{5.6}$$

The peak of the PDF is centered at $E(\underline{\hat{x}}) = x$, the true value (as we assume an unbiased estimator here); $\underline{\hat{x}} \sim N(x, Q_{\hat{x}\hat{x}})$. Mind, that in practice, we generally do *not* know the true value .... What is actually of interest for precision analysis, is **how close** position estimates can expected to be to the true value, therefore we consider the PDF of the *difference* $\underline{z} = \underline{\hat{x}} - x$, which we temporarily denote by z; $\underline{z} \sim N(0, Q_{zz})$, which is easily obtained using section 3.2 (subtracting a constant vector does not change the variance matrix, so $Q_{zz} = Q_{\hat{x}\hat{x}}$)). Such a

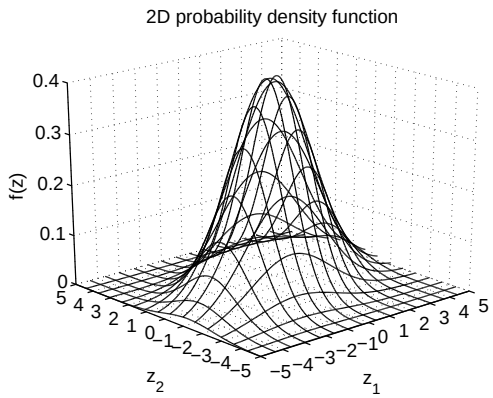


Figure 5.5: Joint probability density function (PDF) of $\underline{z}_1$ and $\underline{z}_2$.

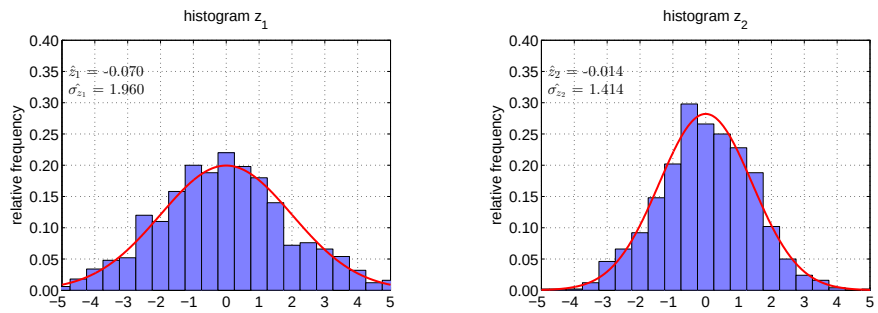


Figure 5.6: Histogram of z_1 on the left, and of z_2 on the right. The (theoretical) normal probability density function is imposed: for $\underline{z}_1$ with standard deviation $\sigma_{z_1} = 2$, and for $\underline{z}_2$ with standard deviation $\sigma_{z_2} = \sqrt{2}$, and both $\underline{z}_1$ and $\underline{z}_2$ with zero mean.

PDF is shown in Figure 5.5. The PDF is specified by the mean vector and the variance matrix, in this two-dimensional example:

$$E\begin{pmatrix} \underline{z}_1 \\ \underline{z}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad D\begin{pmatrix} \underline{z}_1 \\ \underline{z}_2 \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$$

Figure 5.6 shows the histograms for 1000 samples, separately of $\hat{x}_1$ and $\hat{x}_2$, and Figure 5.7 (on the left) shows the samples $(\hat{x}_1,\hat{x}_2)$ in a two-dimensional scatter plot; in both cases they have been corrected for their known true values (x_1,x_2). So, we had to know the true values for the two coordinates, and we do so here for demonstration purpose.

In practice it is not very convenient to draw — for each pair of coordinate estimators — the PDF in a full three dimensional image, as done in Figure 5.5. Though one would like to present the main feature of this function. This is done by considering a cross-sectional cut of the bell-shaped curve, horizontally, at a certain height k; this cut provides an iso contour line.

$$f(z) = \frac{1}{\sqrt{|2\pi Q_{zz}|}} e^{-\frac{1}{2} z^T Q_{zz}^{-1} z} = k$$

hence

$$z^T Q_{zz}^{-1} z = -2\ln(k\sqrt{|2\pi Q_{zz}|}) = k' \tag{5.7}$$

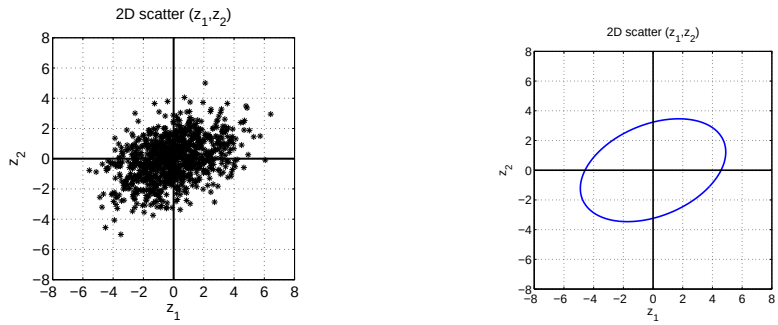


Figure 5.7: On the left: scatter plot of 1000 samples (z_1,z_2); that is, all samples $(\hat{x}_1,\hat{x}_2)$ have been corrected for their known true value (x_1,x_2). On the right: contour of PDF, ellipse of concentration, of position estimator, corrected for true position (x_1,x_2), hence, centered at the origin; an estimated position $(\hat{x}_1,\hat{x}_2)$, also corrected for the true position, is inside the ellipse with probability $1-\alpha = 0.95$ and $k' = \chi^2_\alpha(2,0) = 5.9915$.

where $\ln$ is the natural logarithm, and vector $z \in R^n$. By doing this, $f(z) = k$, we obtain an ellipse. The border of this ellipse represents all points z (end-points of vectors z) with equal probability density. Optionally it is mentioned that the axes of the ellipse can be found through eigenvalue-decomposition of the variance matrix [1] .

The ellipse nicely captures the shape of the PDF, but still does not tell us much about how much probability is actually contained. For this, we need another type of probability density function, namely the Chi-squared distribution.

If random vector $\underline{z}$, consisting of n random variables, is distributed as $\underline{z} \sim N(0, Q_{zz})$, then $\underline{z}^T Q_{zz}^{-1} \underline{z} \sim \chi^2(n,0)$. The PDF is shown in Figure 5.8, and a table of the Chi-squared distribution can be found in appendix C. Hence we have for the quadratic form

$$\underline{T} = \underline{z}^T Q_{zz}^{-1} \underline{z} \sim \chi^2(n,0)$$

in practice typically with $n = 2$.

In terms of $\hat{x}$ we have (using $\underline{z} = \underline{\hat{x}} - x$)

$$(\underline{\hat{x}} - x)^T Q_{\hat{x}\hat{x}}^{-1} (\underline{\hat{x}} - x) \sim \chi^2(n,0)$$

Hence, the above constant k' simply follows from the Chi-squared distribution. In the table in appendix C values can be found for the one-minus-the Cumulative Distribution Function, or to say, the exceedence probability α. The area $P[(\underline{\hat{x}} - x)^T Q_{\hat{x}\hat{x}}^{-1} (\underline{\hat{x}} - x) \leq \chi^2_\alpha(n,0)] = 1-\alpha$ contains probability $1-\alpha$, with the values for α on the top-row in the table, and the $\chi^2_\alpha(n,0)$-values tabulated. For the points in this area holds that $f(\hat{x} - x) \geq k$.

Above we considered the probability that the difference of the position estimator with the true position is inside an ellipse, which equals the probability that the true (and in practice, unknown) position is inside the ellips, but centered at the estimate $\hat{x}$. The ellipse of identical shape as in Figure 5.7 on the right, but centered at the obtained estimate for the position coordinate vector, that is at $(\hat{x}_1, \hat{x}_2)$, is the **confidence ellipse**. It shows the area, which

[1] optional: the orientation (principal axes) of the ellipse, or ellipsoid, can be found through eigenvalue-decomposition of variance matrix Q_{zz}:

$$Q_{zz} = U \Lambda U^T$$

with orthogonal matrix $U = (u_1, \ldots, u_n)$ containing the orthonormal eigenvectors u_i, and diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with positive eigenvalues λ_i, with $i = 1, \ldots, n$, see also chapter 7 on symmetric matrices and quadratic forms in [2]. The directions of the principal axes of the ellipse/ellipsoid are given by the vectors u_i in matrix U, and the length of the i-th principal axis by $\sqrt{\lambda_i k'}$.

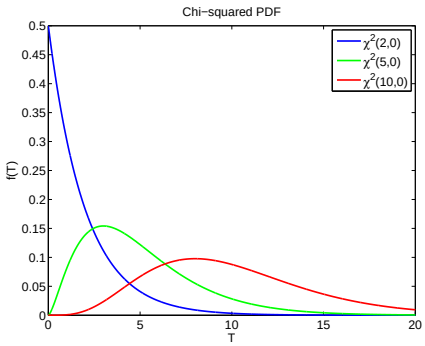


Figure 5.8: Probability density function of a Chi-squared distributed random variable $\underline{T} = \underline{z}^T Q_{zz}^{-1} \underline{z}$; shown are the central Chi-squared distributions for 2, 5 and 10 degrees of freedom respectively. The quadratic form T ranges from zero to infinity.

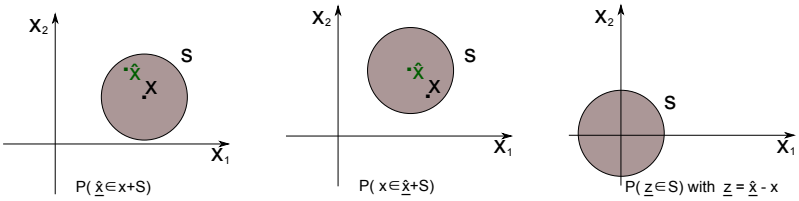


Figure 5.9: The diagram on the left shows, centered at the true position x, the area S which contains the position estimator $\underline{\hat{x}}$ with a certain probability (contour of the PDF). The diagram in the middle shows the confidence region for the true position x, centered at the position estimate $\hat{x}$. The diagram on the right shows the area S for the difference of the position estimator and the true position $\underline{z} = \underline{\hat{x}} - x$, it is centered at the origin (sometimes referred to as the error-region). In these diagrams the area S has simply been shown as a circle.

contains the true (but unknown) position, with a certain, specified probability. After all, the goal of surveying is to determine the position of a point of interest. We will never know the actual, true value, but instead we come up with an estimate for it, and then we would like to know, how close our estimate is to the actual, true position.

In the sequel, in section 5.7, we use the PDF contour ellipse to demonstrate the quality of position solutions using several different measurement set-ups, for instance using solely distance measurements, using solely azimuth measurements, and using a combination of them. Such an analysis is typically done during the design-phase of the survey.

Finally we note that with an estimator for a one-dimensional quantity, we are considering a single random variable, and the error region is just an interval, namely the $\sqrt{k'}$-times-sigma interval. In (5.7), the variance matrix then consists of just a single variance, $(\hat{x} - x)^T Q_{\hat{x}\hat{x}}^{-1} (\hat{x} - x) = k'$, and we have

$$\frac{(\hat{x} - x)^2}{\sigma_{\hat{x}}^2} = k'$$

or

$$|\hat{x} - x| = \sqrt{k'} \sigma_{\hat{x}}$$

The corresponding confidence interval is centered at $\hat{x}$, and extends to both sides by $\sqrt{k'}\sigma_{\hat{x}}$ and contains the true x with a probability $1-\alpha$, where values for k' are tabulated in appendix C, as a function of α, with $n = 1$.

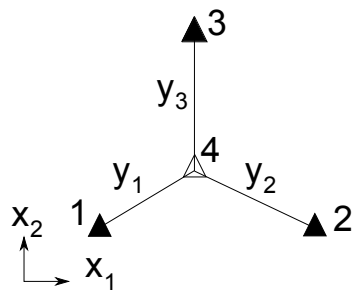


Figure 5.10: Geometric construction to determine position coordinates of new/unknown point 4 based on observed distances to three neighbouring known points.

5.6. Example: resection with distances

For a simple example of measuring three distances (from three known points) to determine the position of a fourth point, we will analyse quality (precision) of the coordinate estimators of this fourth point. It is actually a 'strength'-analysis of the geometric construction (or surveying network in general).

As shown in Figure 5.10 distances are measured from points 1, 2 and 3, to the new/unknown point 4. The coordinates of the points 1, 2 and 3 are known and indicated in the figure: point 1=(0,0), point 2=(100$\sqrt{3}$, 0), point 3=(50$\sqrt{3}$, 150). Approximate values for the coordinates of point 4 are (50$\sqrt{3}$, 50). Units can be assumed to be in meters. The three distance observables are uncorrelated and all have variance σ^2 (for instance σ = 5 m; quite a large value, but done for convenience here).

In the sequel we set up, and compute the design matrix as it occurs in the model of linearized observation equations (using the approach of section 4.4). Next, we compute the variance matrix of the coordinate estimators for point 4.

The three non-linear observation equations read:

$$E(\underline{l}_{14}) = \sqrt{((x_1)_4 - (x_1)_1)^2 + ((x_2)_4 - (x_2)_1)^2}$$

$$E(\underline{l}_{24}) = \sqrt{((x_1)_4 - (x_1)_2)^2 + ((x_2)_4 - (x_2)_2)^2}$$

$$E(\underline{l}_{34}) = \sqrt{((x_1)_4 - (x_1)_3)^2 + ((x_2)_4 - (x_2)_3)^2}$$

There are $m = 3$ observations, and only $n = 2$ unknown parameters, namely $(x_1)_4$ and $(x_2)_4$; the other coordinates are known. With equation (5.5), the given coordinates of points 1, 2 and 3, and $(x_1)_{4,o}$ and $(x_2)_{4,o}$ as the approximate values for $(x_1)_4$ and $(x_2)_4$, the 3×2 design matrix A of the model of linearized observation equations (cf. equation (4.9)) becomes

$$A = \begin{pmatrix} \frac{1}{2}\sqrt{3} & \frac{1}{2} \\ -\frac{1}{2}\sqrt{3} & \frac{1}{2} \\ 0 & -1 \end{pmatrix}$$

The variance matrix is simply $Q_{yy} = \sigma^2 I_3$, a scaled identity matrix.

With (4.7) the variance matrix is obtained as

$$Q_{\hat{x}\hat{x}} = \sigma^2 \begin{pmatrix} \frac{2}{3} & 0 \\ 0 & \frac{2}{3} \end{pmatrix}$$

With this particular geometry (a re-section of point 4 with three angles all nicely of 60 degrees), the two coordinate estimators are uncorrelated, and also happen to have equal

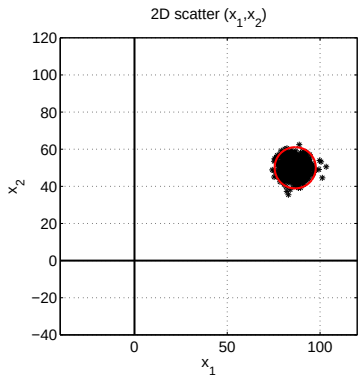


Figure 5.11: Coordinates estimator of point 4 based on distance observables to three neighbouring known points. Shown is the contour ellipse of the PDF, with 97.5% confidence, $k' = 7.3778$, as well as the outcome of carrying out this experiment $N = 1000$ times (each time measuring three distances and determining estimates for $(x_1)_4$ and $(x_2)_4$).

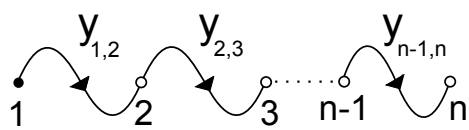


Figure 5.12: Open levelling line from (known) point 1 to point n.

precision $\sigma_{(\hat{x}_1)_4} = \sigma_{(\hat{x}_2)_4}$. The contour ellipse of the PDF turns into just a circle, see Figure 5.11. This figure also shows a scatter of $N = 1000$ trials of this experiment; 975 of these trials should lie inside the ellipse, and 25 should lie outside.

The above analysis of geometry and precision can be done once approximate values for the unknown parameters are available. Actual measurements are not (yet) needed. As there are no measuremements, we can not iterate as done in the example of section 4.4, hence we obtain only a first approximation of the measures of precision.

5.7. Elementary measurement set-up

In this section we present six of the most elementary local measurement set-ups, three on levelling, which is about positioning in just one-dimension, and three on tachymetry (with angles and distances) in two-dimensional geometry. For these six set-ups we analyse the precision of the coordinate estimators for the unknown points.

5.7.1. Levelling

A height-difference y_{ij}, observed with an optical levelling-instrument in a local context, equals the difference of two (orthometric) heights x_i and x_j (apart from a measurement error). The observation equation $E(\underline{y}_{ij}) = x_j - x_i$ is linear in the unknown parameters (coordinates) x_i and x_j. The coordinate system is one dimensional, and the direction of the axis (up) is determined by gravity. The scale is provided by the marks on the levelling rod. With optical levelling, the measurement precision is typically in the order of a few millimeter.

Figure 5.12 presents a so-called *open levelling line*. The height of point 1 is known from an earlier survey, and one levels from 1 to 2, from 2 to 3 etc., until point n. The line consists of $(n-1)$ stretches, also called level-runs.

In an open levelling line, the heights for the unknown points simply follow by adding the observed height-differences to the height of the known point, just like we did in the

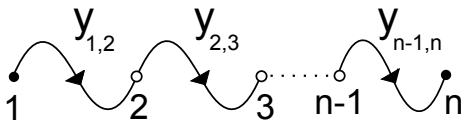


Figure 5.13: Levelling line which terminates at (known) points, point 1 and point n — connected levelling line.

example with three points in chapter 3 (section 3.3). This set-up does not allow for an internal consistency check. In case a gross error is present in one of the levelled height-differences, it will go unnoticed (and spoil part of the results). Hence, such a set-up is not recommended for critical applications!

When the height-difference observables $\underline{y}_{ij}$ have standard deviation equal to $\sigma_{y_{ij}} = \sigma$, and also the given height of point 1 has standard deviation equal to $\sigma_{y_1} = \sigma$ (and all observables are uncorrelated), then the variances for the points along the line are

$$\sigma^2_{\hat{x}_i} = i\sigma^2 \text{ for } i = 1, \dots, n \tag{5.8}$$

This behaviour is shown as a straight line in Figure 5.14. The variance of the height increases linearly with the number of the stretches (accumulation of errors).

What would be the variance of point 1, i.e. $\sigma^2_{\hat{x}_1}$ in case point n would be the known point (instead of point 1)?

Figure 5.13 shows a so-called *connected levelling line* (or, closed level traverse). The heights of both the first and last point are known from an earlier survey. The line connects known points 1 and n, through a series of level runs, visiting unknown points 2, 3, ..., n-1.

This set-up is more secure than an open levelling line, as there is now a check at the end (the line is connected). There is now one more measurement than strictly needed (redundancy equals one).

When the height-difference observables $\underline{y}_{ij}$ have standard deviation equal to $\sigma_{y_{ij}} = \sigma$, and also the given heights of points 1 and n have standard deviation equal to $\sigma_{y_1} = \sigma_{y_n} = \sigma$ (and all observables are uncorrelated), then the variances for the points along the line are

$$\sigma^2_{\hat{x}_i} = i\sigma^2[1 - \frac{i\sigma^2}{(n+1)\sigma^2}] \text{ for } i = 1, \dots, n \tag{5.9}$$

Compared with the variance expression for the open levelling line, an additional factor shows up, between the square brackets, and it is smaller than 1, hence the variances of the heights in the connected line are smaller than those in the open line. This behaviour is shown as a curved line in Figure 5.14. The curve (as a function of i) is actually determined by the parabola $-i^2 + (n+1)i$, which has its top at $i = \frac{n+1}{2}$. The curve starts at a (slightly) lower value for point 1 than the straight line for the open line, and it is symmetric about $i = 5\frac{1}{2}$.

The above expression for the variance can be checked for instance by considering an example with 4 points, with points 1 and 4 given, and unknown points 2 and 3, and we have three levelled height-differences. The model of observation equations (4.2) reads

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_{12} \\ \underline{y}_{23} \\ \underline{y}_{34} \\ \underline{y}_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

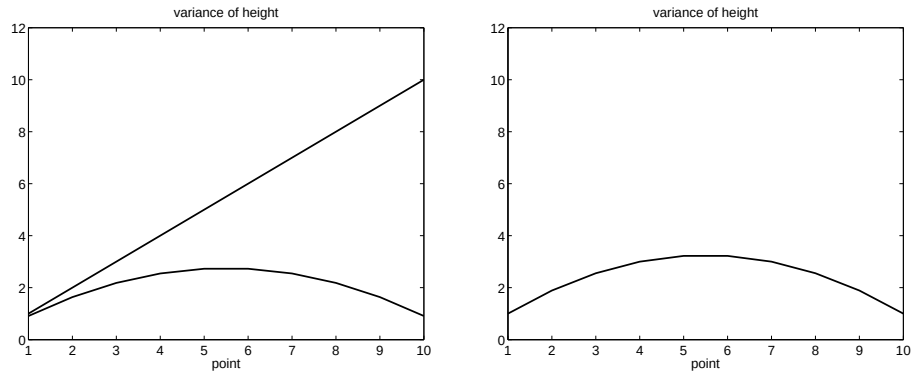


Figure 5.14: Variance of heights determined for points along open and connected levelling line of n=10 points (on the left), and along (closed) levelling loop of 9 points, where point 10 equals point 1 again, (on the right). Variance σ^2 was simply set to 1.

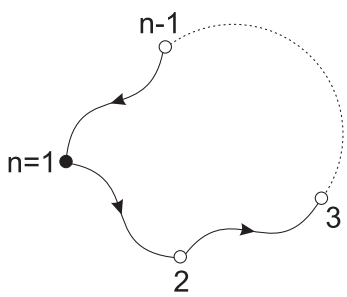


Figure 5.15: Levelling loop of $(n-1)$ points, where (known) point n equals point 1.

and the variance matrix is just a 5×5 identity matrix, scaled by σ^2. The expression for $\sigma^2_{\hat{x}_i}$ can be checked by computing the $Q_{\hat{x}\hat{x}}$ matrix, cf. (4.7).

Figure 5.15 shows a *levelling loop*, (or, closed loop level traverse). The levelling 'line' now returns to the starting point, and the height of the first point is known from an earlier survey.

Also this set-up is more secure than an open levelling line, as there is a check: all observed height differences added together need to yield just zero (return to the same point); this is the loop-closure condition. There is now one more measurement than strictly needed (redundancy equals one).

Suppose the levelling loop of Figure 5.15 consists of just four points, hence $n-1=4$. Then there are $m=5$ observations: y_{12}, y_{23}, y_{34} and y_{41}, and the height of point 1 is given and fixed - this provides the fifth observation y_1. There are three points with unknown heights: x_2, x_3, x_4, they are to be determined, and x_1 is included in order to be able to relate 'observation' y_1 (the given height). The model of observation equations (4.2) reads:

$$E\begin{pmatrix} \underline{y}_{1} \\ \underline{y}_{12} \\ \underline{y}_{23} \\ \underline{y}_{34} \\ \underline{y}_{41} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

With $m=5$ observations and $n=4$ unknown parameters (mind, n now describes the number of unknown parameters, no longer the number of points), there is a redundancy of 1; to say,

there is one measurement too much. This redundancy can be made explicit. The (expectation of the) five observables can be described by four unknown parameters. This implies that there should hold a *condition* ($m - n = 1$ condition in this case) on the observations.

We multiply both left and right side by the same matrix

$$\underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \end{pmatrix}}_{B^T} E \begin{pmatrix} \underline{y}_{1} \\ \underline{y}_{12} \\ \underline{y}_{23} \\ \underline{y}_{34} \\ \underline{y}_{41} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

resulting in

$$E(\underline{y}_{12} + \underline{y}_{23} + \underline{y}_{34} + \underline{y}_{41}) = 0$$

and this is the above mentioned loop closure condition (the expectation operator is linear). It implies that the sum of the levelled height-differences is (expected to be) equal to zero. We need to return to the same point, and this provides us with a check. In practice, the sum of the height differences $y_{12} + y_{23} + y_{34} + y_{41} \neq 0$ (though hopefully close to zero), and this is referred to as the *misclosure*. For the $(m-n)$xm matrix B^T describing the condition(s), holds that $B^T A = 0$.

When the height-difference observables $\underline{y}_{ij}$ have standard deviation equal to $\sigma_{y_{ij}} = \sigma$, and also the given height of point 1 has standard deviation equal to $\sigma_{y_1} = \sigma$ (and all observables are uncorrelated), then the variances for the points along the line are

$$\sigma^2_{\hat{x}_i} = \sigma^2 + \frac{\sigma^2}{n-1}(i-1)(n-i) \quad \text{for } i = 1, \dots, n \tag{5.10}$$

where n is indicating the number of points, as in Figure 5.15. The variance shows a behaviour similar as for the connected line, see Figure 5.14 on the right, driven by a similar parabola, but at a higher level than for the connected line (as there is only one known point involved in the loop, versus two with the connected line). Obviously the variance of point 10 equals the variance of point 1.

5.7.2. Intersection with azimuths

In Figure 5.16 the position of an unknown point is determined (in two dimensions) by observing the azimuths at two known points 1 and 2. For four possible locations of the unknown point (all on the line $x_1 = 5$) the PDF contour ellipse is given.

When the unknown point is close to the line connecting the points 1 and 2 (dashed line, which parallels the x_1-axis here), the precision is good in the x_2-direction (perpendicular to the connection line), but very poor in the x_1 direction (along the connection line). This situation is reversed when the point is located far away (to the top or bottom of the graph). An intersection of the two azimuth-lines at the unknown point, at a right angle, delivers a homogeneous precision (with the ellipse of concentration being close to a circle).

5.7.3. Polar coordinates

In two dimensions the position of an unknown point can be also be determined using one measurement of azimuth and one measurement of distance, taken at a single known point. This is the typical way of working with a tachymeter or a total station: measuring distances

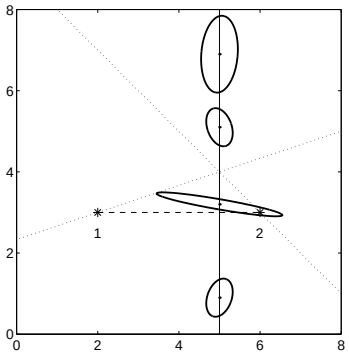


Figure 5.16: Precision of positioning in two dimensions using azimuths, measured at (known) points 1 and 2; PDF contour ellipses (P=0.3935) for four possible locations of the unknown point. The standard deviation was set as σ_a = 0.13 radian.

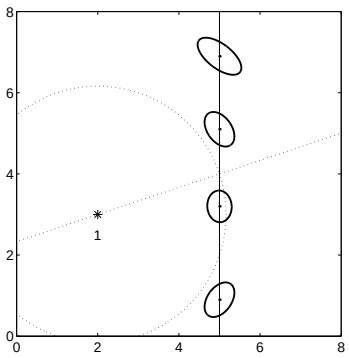


Figure 5.17: Precision of positioning in two dimensions using azimuths and distances (polar coordinates at point 1); PDF contour ellipses for four possible locations of the unknown point. The standard deviations were set as σ_a = 0.13 radian, and σ_l = 0.30.

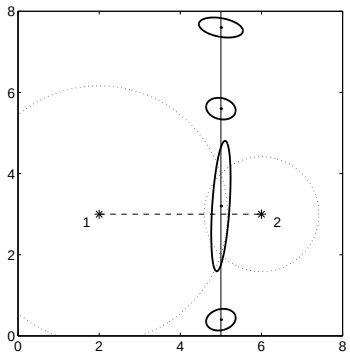


Figure 5.18: Precision of positioning in two dimensions using distance measurements at points 1 and 2; PDF contour ellipses for four possible locations of the unknown point. The standard deviation was set as σ_l = 0.30.

and azimuths (or directions) to points to be determined. In Figure 5.17 PDF contour ellipses are given again for the same four possible locations of the unknown point.

With the particular choice for the azimuth and distance observables' standard deviations as noted in Figure 5.17, a homogeneous precision is achieved for the two coordinates of the unknown point (ellipse close to a circle), provided that the unknown point is not too far away from the known station 1. For points far away, one has to realize that a fixed uncertainty in the measurement of angle translates — at a larger distance — into a larger uncertainty in the position, in the direction perpendicular to the line connecting the unknown point and the instrument.

5.7.4. Intersection with distances

Finally we determine the position of an unknown point using two measurements of distance, a set up which is pretty much similar to the example with azimuths in one of the previous sections. From each of the two known points a distance is measured to the unknown point. In Figure 5.18 PDF contour ellipses are given for the same four possible locations of the unknown point. The standard deviation of the distance observable was kept fixed, whereas in practice it may vary with distance (typically, larger standard deviation for larger distances).

When the unknown point is close to the line connecting the points 1 and 2 (dashed line, which parallels the x_1-axis), the precision is good in the x_1-direction (the direction along the connection line), but very poor in the other direction (perpendicular). This situation is reversed when the point is located far away (to the top or bottom of the graph). An intersection of the two distance-lines at the unknown point, at a right angle, delivers a homogeneous precision (with the ellipse of concentration being close to a circle).

6

Validation

In chapter 4 we introduced the mathematical model $\underline{y} = Ax + \underline{e}$, with $m \geq n$, where m is the dimension of vector y, and n the dimension of vector x. The system $y = Ax$ is generally not consistent, hence, $y \neq Ax$. The least-squares estimate (4.4), and the minimum variance estimator (4.6) were introduced, sections 4.3.2 and 4.3.4. Once an estimate $\hat{x}$ for the unknown parameters is available, one can compute an estimate for the observations: $\hat{y} = A\hat{x}$, and this system is *consistent*. One could regard this parameter estimation as 'changing' (or adjusting) the observed values from y to $\hat{y}$ in order to turn a non-consistent system into a consistent one. As outlined with the least-squares criterion (4.5), one keeps vector $y - \hat{y}$ as short as possible, the observed values should be 'changed only as little as possible'.

In this chapter we introduce the least-squares residuals, and show how they can be used in an overall *consistency check*, to answer the question whether the collected measurements y and the assumed model Ax can be deemed to be mutually consistent. Next we present a worked example of line-fitting (regression). Eventually we briefly introduce the more advanced (optional) subject of observation testing, and present the final results of our data processing and analysis.

6.1. Least-squares residuals

Once an estimator $\underline{\hat{y}}$ is available for the vector of observables, the least-squares residuals follow as

$$\underline{\hat{e}} = \underline{y} - \underline{\hat{y}} \tag{6.1}$$

The least-squares residuals measure the difference between the observations (as measured) y, and the estimated, or adapted ones $\hat{y}$ (see section 4.3.2, $\underline{\hat{y}} = A\underline{\hat{x}}$). The least-squares residuals provide an estimate for the (unknown) measurement error e (that is why also the residuals are denoted with a hat-symbol). They carry important diagnostic information about the parameter estimation process. When the residuals are small, the situation is looking good. One does not need to 'change' the observed values by much, in order to make them fit into the model Ax. However, when they are large, this might be a reason for reconsideration. It could be that there are large outliers or faults present in the measurements (e.g. entering an observed height-difference as 0.413 m, instead of 0.143 m), or that the assumed model is not appropriate for the case at hand (in a dynamic system, with a moving object, we may assume that the object is moving with constant velocity, but this may turn out not to be the case). Small residuals tend to be OK — large ones are not. But what is small, and what is big? Fortunately, we can devise an objective criterion to judge on their size.

6.1.1. Overall consistency test

When the data are normally distributed, $\underline{y} \sim N(Ax, Q_{yy})$, under the working model (the null-hypothesis), then also the residuals will be normally distributed (with zero mean). It can be shown — though the proof is omitted here — that $\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}}$ has — under the working model (4.2) — a central Chi-square distribution with $m-n$ degrees of freedom, hence $\chi^2(m-n, 0)$ (from the m-vector of observations, n unknown parameters in x are estimated, and hence only $(m-n)$ degrees of freedom are 'left' for the residuals). The Chi-squared distribution was introduced in section 5.5, and shown in Figure 5.8. The **squared norm of the residuals-vector** $\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}}$ is an *overall* consistency measure[1]. It provides an objective criterion on judging the amount by which we need to 'change' the observed values, to fit the assumed or supposed model (Ax).

$$\underline{T} = \underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}} \sim \chi^2(m-n, 0) \tag{6.2}$$

Now one can set a level of significance (probability) α, see chapter 26 in [1], e.g. 5%, and not accept the residual vector $\hat{e}$, once its squared norm is located in the upper 5% of the distribution (right tail); occurrence of such a value is deemed to be too unlikely to be true under the working model. When there are large outliers or faults present in the observations, or when the assumed model is not appropriate, the residuals tend to be larger, leading to a *larger* value for the test-statistic T, and hence we set a right-sided critical region.

In practice, one computes $T = \hat{e}^T Q_{yy}^{-1} \hat{e}$, and retrieves threshold $k' = \chi^2_\alpha(m-n, 0)$ from the table and concludes (decides):

- model and data are consistent when $T < k'$
- model and data are *not* consistent when $T > k'$

For example with m=5 and n=2, $m-n$=3, and with a 5% level of significance, the threshold (or critical) value k' for the above squared norm of the least-squares residuals, T, is 7.8147, see the table in Appendix C ($\chi^2_\alpha(3, 0) = 7.8147$).

6.1.2. Simplification

In the simple case that the observables' variance matrix is a diagonal matrix $Q_{yy} = \mathrm{diag}(\sigma^2_{y_1}, \sigma^2_{y_2}, \ldots, \sigma^2_{y_m})$ (all observables are uncorrelated), the above overall consistency measure can be given a simple and straightforward interpretation. Namely

$$\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}} = \sum_{i=1}^{m} \frac{\underline{\hat{e}}_i^2}{\sigma^2_{y_i}} = \sum_{i=1}^{m} \left(\frac{\underline{\hat{e}}_i}{\sigma_{y_i}} \right)^2 \tag{6.3}$$

and it compares — per observation — the residual $\hat{e}_i$ with the standard deviation σ_{y_i} of, i.e. the expected uncertainty in, the observable $\underline{y}_i$.

The overall consistency measure equals the sum of all those m squared ratios. The overall consistency test, as the name says, aims to *detect*, in general sense, any inconsistency between model and data.

6.1.3. Discussion

As an example we consider the case in which the same unknown distance is measured twice, cf. section 4.3.1. If one of these measurements is biased by a large amount (for instance 10 m), and the other is not, then the overall consistency test will likely detect that the two

[1] actually a measure of in-consistency

measurements are not consistent with the model, namely, according to the model (measuring the same distance twice) the numerical values of two measurements should be the same, or close together. Intuitively, as the measurements are not close together, this gives rise to suspicion (there might be something wrong with the measurements). Anomalies in the measurements which cause the data to be (still) consistent with the assumed model can *not* be detected by this test. In our example, if both observations are biased in the same way, let us say by 10 m, then the two observations are still consistent (with each other in the model; they are both in error). Intuitively, as their values are the same or close together, this does not raise any suspicion. Being able to detect all relevant anomaly scenarios is part of designing a good measurement set-up.

6.1.4. Example: repeated measurements

Suppose m measurements of the same unknown quantity are made. Then the general model of observation equations $E(\underline{y}) = Ax;\ D(\underline{y}) = Q_{yy}$ (4.2) reads:

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} x \ ; \ D\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \sigma^2 I_m$$

where we assumed that all observables have equal precision (all variances equal to σ^2), and there is no correlation. There are m observations, and there is $n = 1$ unknown parameter.

In this case, the minimum variance estimate $\hat{x}$ for the unknown parameter x, equals just the mean of the observations, similar to (2.12).

In this simple example $\hat{y}_i = \hat{x}$, and hence, the least-squares residuals $\hat{e}_i = y_i - \hat{x}$. The squared norm of the residuals vector becomes

$$T = \hat{e}^T Q_{yy}^{-1} \hat{e} = \frac{1}{\sigma^2} \sum_{i=1}^{m} \hat{e}_i^2 = \frac{1}{\sigma^2} \sum_{i=1}^{m} (y_i - \hat{x})^2$$

which shows that, in this case, this measure of consistency is closely related to the sample variance (2.14), which you determine based on measurements taken. We have

$$\frac{T}{m-1} = \frac{1}{\sigma^2} \frac{1}{m-1} \sum_{i=1}^{m} (y_i - \hat{x})^2 = \frac{\hat{\sigma}^2}{\sigma^2}$$

where we use m instead of N as in (2.14). The measure of consistency equals, in this case, the ratio of the sample variance and the formal, a-priori variance.

6.2. Example

In this example we consider the classical problem of *line fitting*. This problem is frequently encountered in science and engineering. One can think of measuring the extension of a certain object (e.g. of steel) due to temperature, for which one assumes a linear behaviour, so the length of the object is measured, at different temperatures, and next one would like to determine the length of the object at some reference temperature and the coefficient of extension, i.e. by how much it extends for every increase of one degree in temperature. In literature, this subject is often referred to as regression analysis, see e.g. chapter 6 on orthogonality and least-squares in [2], in particular section 6 with applications to linear models.

The subject can be seen much broader however, namely as *curve fitting*. The principle is not restricted to just straight lines, one can also use parabolas and higher order polynomials for instance.

In this example we consider a vehicle which is driving along a straight line, and we are interested in the position along the road. Therefore, a laser-tracker is used, and this device measures/reports the position of the vehicle every second. For convenience, the laser-tracker is at the origin of this one dimensional coordinate system.

Over a period of four seconds, we take measurements: $y = (y_1, y_2, y_3, y_4)^T$, hence the vector of observations has dimension $m = 4$. Correspondingly at times t_1, t_2, t_3 and t_4, the unknown positions are $x(t_1)$, $x(t_2)$, $x(t_3)$ and $x(t_4)$. The measurements equal the unknown positions, apart from measurement errors, i.e. $y_i = x(t_i) + e_i$ for $i = 1, \dots, 4$.

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \underline{y}_3 \\ \underline{y}_4 \end{pmatrix} = \begin{pmatrix} x(t_1) \\ x(t_2) \\ x(t_3) \\ x(t_4) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x(t_1) \\ x(t_2) \\ x(t_3) \\ x(t_4) \end{pmatrix}$$

In case we determine these $n = 4$ four unknown positions from the $m = 4$ four observed positions, estimation is pretty trivial, namely $\hat{x}(t_i) = y_i$ for $i = 1, \dots, 4$.

We have reasons however, to assume that the vehicle is driving at constant speed, and thereby we can model the unknown motion of the vehicle, by just an unknown initial position $x(t_0)$, and its velocity $\dot{x}$.

$$\begin{pmatrix} x(t_1) \\ x(t_2) \\ x(t_3) \\ x(t_4) \end{pmatrix} = \begin{pmatrix} 1 & (t_1 - t_0) \\ 1 & (t_2 - t_0) \\ 1 & (t_3 - t_0) \\ 1 & (t_4 - t_0) \end{pmatrix} \begin{pmatrix} x(t_0) \\ \dot{x} \end{pmatrix}$$

which we can substitute in the above system of observation equations, and hence

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \underline{y}_3 \\ \underline{y}_4 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & (t_1 - t_0) \\ 1 & (t_2 - t_0) \\ 1 & (t_3 - t_0) \\ 1 & (t_4 - t_0) \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x(t_0) \\ \dot{x} \end{pmatrix}}_{x}$$

Now there are still $m = 4$ observations, but just $n = 2$ unknown parameters in vector x. The distance observables by the laser-tracker are all uncorrelated, and all have the same variance σ^2. Hence $Q_{yy} = \sigma^2 I_4$. Standard deviation σ can be taken here as $\sigma = 1$.

In the sequel we develop the example into a real numerical example. The observation times are $t_1 = 1$, $t_2 = 2$, $t_3 = 3$ and $t_4 = 4$ seconds, and $t_0 = 0$ (and timing is assumed here to be perfect — no errors; all coefficients of matrix A are known — when this assumption cannot be made, refer to Appendix A.6).

As can be seen in Figure 6.1, we try to fit a straight line through the observed data points. Therefore we estimate the offset/intercept of the line $x(t_0)$, and its slope $\dot{x}$; in terms of regression, they are the (unknown) regression coefficients. In this example time t is the explanatory or independent variable (also called the regressor); the (observed) position depends on time t. And, the observation y is the dependent (or response) variable. Least-squares estimation will yield a best fit of the line with the observed data points, through minimizing (4.5).

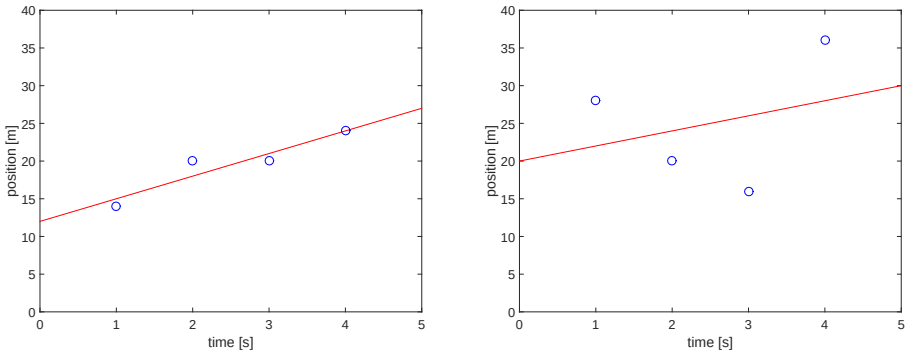


Figure 6.1: Line fitting on day 1 (left) and day 2 (right). Shown are the observed data points as blue dots, and the straight line in red, fitted by means of least-squares.

The experiment was repeated the next day, and in the sequel we consider — simultaneously — the outcomes of *both* experiments. The measurements on day 1 were $y = (14, 20, 20, 24)^T$, and on day 2 $y = (28, 20, 16, 36)^T$. In Figure 6.1, the observed distances (in meters) are shown, together with the fitted line, based on the least-squares estimates. Verify yourself that, with (4.6), one obtains

$$\hat{x} = \begin{pmatrix} 12 \\ 3 \end{pmatrix} \text{for day 1, and } \hat{x} = \begin{pmatrix} 20 \\ 2 \end{pmatrix} \text{for day 2}$$

with units in meters and meters per second respectively.

With $\hat{x}(t_0) = 12$ and $\hat{\dot{x}} = 3$ on day 1, one can determine $\hat{y}$ through $\hat{y} = A\hat{x}$, and eventually obtain $\hat{e} = y - \hat{y}$. And do this for day 2 as well.

$$\hat{e} = \begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \end{pmatrix} \text{ for day 1, and } \hat{e} = \begin{pmatrix} 6 \\ -4 \\ -10 \\ 8 \end{pmatrix} \text{ for day 2}$$

From these numbers, and also Figure 6.1, one can already conclude that for day 1 (on the left) we have a pretty good fit, while on day 2 (on the right) the fit is pretty poor, likely a second order polynomial would do here much better. It indicates that the motion of the vehicle on day 2 has not really been at constant speed. For an objective criterion in judging a good and poor fit, we use the squared norm of the residual vector.

$$T = \hat{e}^T Q_{yy}^{-1} \hat{e} = 6 \text{ for day 1, and } T = \hat{e}^T Q_{yy}^{-1} \hat{e} = 216 \text{ for day 2}$$

With the table in appendix C, we find that with $m - n = 2$ and $\alpha = 0.01$ the threshold value equals $\chi^2_\alpha = 9.2103$, and hence the line fit of day 1 is not rejected ($6 < 9.2103$), but the line fit of day 2 is rejected ($216 > 9.2103$)!

At this stage we recall that least-squares estimation is driven, see (4.5), by minimization of the squared norm of the vector $y - Ax$, see Figure 6.2. Least-squares minimizes the objective function $\min_x \|y - Ax\|^2$. Vector $y - Ax$ contains the differences between the actual observations y and the modelled observations by Ax. One should choose values for the elements in vector x, such that the squared norm of th vector $y - Ax$ is at minimum. Figure 6.2 shows at left, for the observations of day 1, along the vertical axis, the squared norm of the vector $y - Ax$, which is just a single number, as a function of $x(t_0)$ and $\dot{x}$ in the horizontal plane, with offset parameter $x(t_0)$ along the axis in front, and slope parameter $\dot{x}$ along the axis to the back.

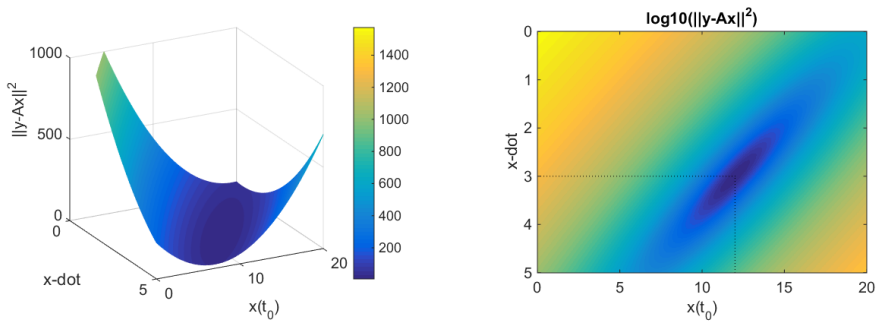


Figure 6.2: The sum of squared residuals as a function of $x(t_0)$ and $\dot{x}$. At right, actually the $\log 10$-value is shown. The minimum, which is equal to 6, is attained for $x(t_0)$=12 and $\dot{x}$=3, indicated by the dotted lines, hence $\min_x \|y - Ax\|^2 = 6$ (for day 1).

We have to keep in mind that the true motion of the vehicle is unknown. It is *our assumption* that we can mathematically describe the motion through a constant velocity model. For day 1 there is no indication that this is not working. But for day 2, a constant velocity model, most likely, is not providing a proper description of the actual motion. Maybe a second order model (constant acceleration) gives a better/acceptable fit.

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \underline{y}_3 \\ \underline{y}_4 \end{pmatrix} = \begin{pmatrix} 1 & (t_1 - t_0) & \frac{1}{2}(t_1 - t_0)^2 \\ 1 & (t_2 - t_0) & \frac{1}{2}(t_2 - t_0)^2 \\ 1 & (t_3 - t_0) & \frac{1}{2}(t_3 - t_0)^2 \\ 1 & (t_4 - t_0) & \frac{1}{2}(t_4 - t_0)^2 \end{pmatrix} \begin{pmatrix} x(t_0) \\ \dot{x}(t_0) \\ \ddot{x} \end{pmatrix}$$

There are now $n = 3$ unknown parameters, with still $m = 4$ observations.

Mind that if we go for a third order polynomial, we will have $n = 4$ unknown parameters, and all information contained in the $m = 4$ observations, will go into the unknown parameters, and nothing will be left for the least-squares residuals, as $m - n = 0$. Without redundancy, the overall consistency check gets void. In that case it looks like the observations perfectly fit the assumed model, but you actually do not have any means to verify this.

6.3. Observation testing and outlook [*]

The squared norm of the residuals is an overall measure of consistency between measurements and assumed mathematical model. The statistical test of $\hat{e}^T Q_{yy}^{-1} \hat{e}$ being smaller or larger than the critical value k' from the Chi-squared distribution is referred to as the overall consistency test, see section 6.1.1. In literature you may also encounter it as the F-test, then $\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}}/(m-n) \sim F(m-n, \infty, 0)$ under the working model (null-hypothesis), where F represents the F-distribution.

The test can be derived from the principle of statistical hypothesis testing. More specific statistical hypothesis tests exist, for instance tests to identify outliers, blunders, and anomalies in single measurements. A true coverage of this subject is beyond the scope of this primer, but we will introduce — without any derivation, or proof of optimality — a simple test which aims to identify an **outlier** in a set of measurements (then only *one* measurement from this set is affected by the outlier). Typically this test is used multiple times, namely to test *each* of the m observations in vector $y = (y_1, \ldots, y_m)^T$ separately. It is based again on the least-squares residuals $\underline{\hat{e}} = \underline{y} - \underline{\hat{y}}$. Using the error propagation law (3.9), with (4.6), and $\underline{\hat{y}} = A\underline{\hat{x}}$, one can

derive that

$$Q_{\hat{e}\hat{e}} = Q_{yy} - A(A^T Q_{yy}^{-1} A)^{-1} A^T = Q_{yy} - A Q_{\hat{x}\hat{x}} A^T$$

where $Q_{\hat{x}\hat{x}}$ was given by (4.7).

Under the working model the least-squares residuals have zero mean, hence $\underline{\hat{e}} \sim N(0, Q_{\hat{e}\hat{e}})$, and the vector $\underline{\hat{e}}$ is normally distributed as it is a linear function of $\underline{y}$, which is also taken to be normally distributed.

The least-squares residuals $\underline{\hat{e}}$ is a vector with m random variables $\underline{\hat{e}} = (\underline{\hat{e}}_1, \underline{\hat{e}}_2, \ldots, \underline{\hat{e}}_m)^T$. For each of the residuals we have $\underline{\hat{e}}_i \sim N(0, \sigma^2_{\hat{e}_i})$ with $i = 1, \ldots, m$. Usually, if (just) observation y_i contains a large error, the corresponding residual $\hat{e}_i$ will deviate (substantially) from zero. We use this to propose — valid only for the case when the observables have a diagonal variance matrix Q_{yy} — the w-teststatistic as

$$\underline{w}_i = \frac{\underline{\hat{e}}_i}{\sigma_{\hat{e}_i}} \tag{6.4}$$

and check whether it deviates from zero.

Division of the residual by the standard deviation $\sigma_{\hat{e}_i}$ (you may use (3.9) again) causes the w-teststatistic to be *standard normally distributed*: $\underline{w}_i \sim N(0, 1)$. In fact, it is the *normalized* residual. Setting a level of significance α, one can find the critical value. Mind that deviations both in positive and negative direction may occur. Hence, the level of significance α is split into $\frac{\alpha}{2}$ for the right tail, and $\frac{\alpha}{2}$ for the left tail. The critical value $\tilde{k} = N_{\frac{\alpha}{2}}(0, 1)$ follows from the table in Appendix B. The hypothesis 'no outlier present in measurement y_i' is rejected if $|w_i| > \tilde{k}$, i.e. if $w_i < -\tilde{k}$ or $w_i > \tilde{k}$. The test is typically done for all measurements $i = 1, \ldots, m$.

In practice the largest of the w-teststatistics (in absolute sense) indicates the most likely faulty measurement. This measurement is removed from the data, and estimation and validation is repeated with one less observation, until no more measurements are rejected, or until redundancy runs low. This procedure of 'data checking' can be carried out automatedly, without human intervention, and allows for a robust processing of the measurements.

There are many alternative approaches to robust estimation, delivering robust estimators, which are insensitive to outliers, but these are beyond the scope of this primer.

6.4. Example — continued

Eventually we wrap up, and present the final results of our data processing and analysis. We have computed the estimates for the parameters of interest in the example of section 6.2, namely the initial position $x(t_0)$ and the velocity $\dot{x}$, and we have verified the consistency of the model and the data.

In Figure 6.3 we consider the data of day 1, hence giving a good fit, see Figure 6.1 at left. In Figure 6.3 at left, we present again the four measurements, and, we indicate with them the measurement uncertainty. The standard deviation of each measurement was $\sigma = 1$. The graph presents the so-called error-bars for a confidence level of 95%. When the observables are normally distributed, one can verify with the table in Appendix B, that the error-bar extends by $r_\alpha \sigma$ to either side of the observed value, with $r_\alpha = 1.96$.

The graph at right presents again the fitted line, based on the least-squares estimate $\hat{x}$. As outlined in the introduction of this chapter, estimates for the observations can be computed as $\hat{y} = A\hat{x}$. With this, you can check yourself that while the observation $y_2 = 20$, $\hat{y}_2 = 18$ (and the latter is located on the red line).

The corresponding variance matrix of $\underline{\hat{y}}$ can be found, using (3.9) on $\underline{\hat{y}} = A\underline{\hat{x}}$, as $Q_{\hat{y}\hat{y}} = A Q_{\hat{x}\hat{x}} A^T$, with $Q_{\hat{x}\hat{x}}$ from (4.7). So, for the estimated observation at t_2 holds that $\sigma_{\hat{y}_2} \approx 0.55$,

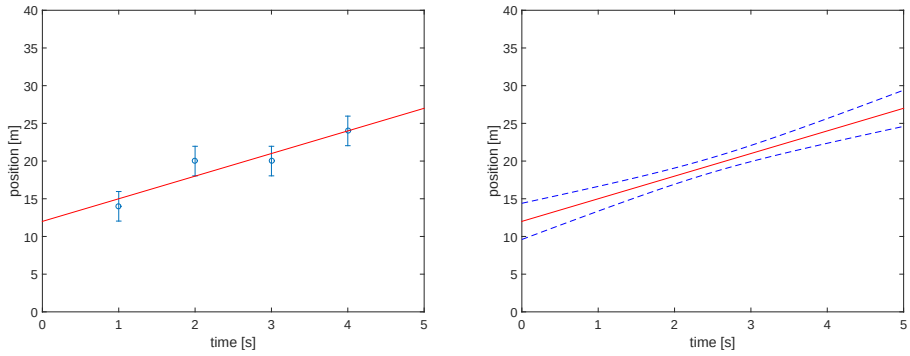


Figure 6.3: Line fitting on day 1: observed data points as blue dots, with error-bars, also in blue, at the left, and the straight line fitted by means of least-squares with confidence interval at the right.

whereas $\sigma_{y_2} = 1$ (hence, it is much smaller). Then the 95% confidence interval for the true y_2 is centered at $\hat{y}_2$, and extends to $r_\alpha \sigma_{\hat{y}_2} \approx 1.07$ on either side (and hence it is much smaller than with the individual observation, in the graph at left). Suppose the experiment is repeated, then 95% of the realizations of this (random) interval will actually contain the true position, see also chapter 23 in [1]. In fact, the confidence interval can be presented for the position at any time t, see the dashed blue lines on either side of the fitted line, in the graph at right; it is the confidence interval for $y(t)$ based on $\hat{y}(t)$, for any time t.

6.5. Exercises and worked examples

This section presents a problem with a worked answer on parameter estimation and validation.

Question 1 To determine the sea-level height in the year 2000, and also its rate of change over time, five observations of the sea-level are available, for instance from a tide gauge station. The observations are:

- observation y_1 = 25 mm, in the year 1998
- observation y_2 = 24 mm, in the year 1999
- observation y_3 = 27 mm, in the year 2000
- observation y_4 = 26 mm, in the year 2001
- observation y_5 = 28 mm, in the year 2002

Based on these five observations, compute the *least-squares* estimates for the sea-level in the year 2000 and the rate of change. All quantities are referenced to the start of the year. The rate of change can be considered constant over the time span considered.

Answer 1 Similar to the example of section 6.2, the model reads

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \underline{y}_3 \\ \underline{y}_4 \\ \underline{y}_5 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x(t_0) \\ \dot{x} \end{pmatrix}}_{x}$$

and basically we need to fit a straight line through five data points. In the above $y = Ax$ model, $x(t_0)$ is the *unknown* sea-level height in the year 2000 (expressed in [mm]), and $\dot{x}$ is

the (assumed constant) rate of change of the sea-level (and also unknown, and expressed in [mm/year]). On the left hand side we have the five observations, of which the third, y_3, is the *observed* sea-level height in the year 2000 (in [mm]). For example, the last observation, y_5, is related to the two unknown parameters as: y_5 equals (on average) the sum of the sea-level in the year 2000 $x(t_0)$, plus twice the yearly change $\dot{x}$. These two coefficients, 1 and 2, show up, as the last row, in the A-matrix. There is no information given with regard to the precision of the observables (no matrix Q_{yy}), hence we use the basic least-squares equation (4.4) $\hat{x} = (A^T A)^{-1} A^T y$.

$$A^T A = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix} \quad (A^T A)^{-1} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{pmatrix}$$

and $\hat{x} = (A^T A)^{-1} A^T y$ becomes

$$\hat{x} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 25 \\ 24 \\ 27 \\ 26 \\ 28 \end{pmatrix} = \begin{pmatrix} 26 \\ \frac{8}{10} \end{pmatrix}$$

The least-squares estimate for the sea-level height in the year 2000 is 26 mm, and the rate of change is 0.8 mm/year. Note that the least-squares estimate for the height in the year 2000 does *not* equal the observed height (y_3). The least-squares estimate is determined based on *all* available observations.

Question 2 With the model and observations of the previous question, determine whether the overall consistency test is passed, when the level of significance is set to 10%. The observables can be assumed to be all uncorrelated, and have a standard deviation of 1 mm.

Answer 2 The variance matrix of the observables reads $Q_{yy} = I_5$, and this does not change anything to the computed least-squares estimates (see equation (4.6)). The overall consistency test is $T = \hat{e}^T Q_{yy}^{-1} \hat{e}$, hence we need the vector of least-squares residuals (6.1) $\hat{e} = y - \hat{y} = y - A\hat{x}$.

$$\begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \\ \hat{e}_4 \\ \hat{e}_5 \end{pmatrix} = \begin{pmatrix} 25 \\ 24 \\ 27 \\ 26 \\ 28 \end{pmatrix} - \begin{pmatrix} 24.4 \\ 25.2 \\ 26.0 \\ 26.8 \\ 27.6 \end{pmatrix} = \begin{pmatrix} 0.6 \\ -1.2 \\ 1.0 \\ -0.8 \\ 0.4 \end{pmatrix}$$

Then the value for the overall consistency test becomes $T = \hat{e}^T Q_{yy}^{-1} \hat{e} = 3.6$. The threshold, with $\alpha = 0.1$, is $k' = \chi^2_\alpha(m-n, 0)$, and $m = 5$ and $n = 2$, hence $m - n = 3$. With the table in Appendix C we obtain $k' = \chi^2_{0.1}(3, 0) = 6.2514$, and hence $T < k'$, and the overall consistency test is accepted. There is no reason to suspect that something is wrong; the assumed model and the made observations seem to be in agreement with each other; the fit is good.

7

Interpolation [*]

In this chapter we cover the subject of interpolation. After the introduction we cover deterministic interpolation, and next stochastic interpolation, thereby focussing in particular on Kriging.

7.1. Introduction

Interpolation is about determining the value of an attribute, like the height or water-depth, at a certain position within a spatial domain, from a set of observations of that attribute in that domain. For instance, with a hydrographic survey, by means of echo-sounding, the attribute water-depth is measured, and the measurements $y_1, \dots, y_m$ are taken at specific discrete positions, for instance while the vessel is sailing along a regular pattern (forth and back) over the water-way. Next, one is interested to know the water-depth z at *another* position, where *no* specific *measurement* is available, see Figure 7.1.

One of the most common techniques is *linear* interpolation. The value of the attribute z_0 at position p_0 (position coordinate vector) is estimated as a linear combination

$$\hat{z}_0 = w_1 y_1 + \cdots + w_m y_m \tag{7.1}$$

of observations $y_1, \dots, y_m$ at positions $p_1, \dots, p_m$ in the neighbourhood of p_0. The coefficients w_1 to w_m indicate the weights given to each of the observations. By stacking the observations in vector $y = (y_1, \dots, y_m)^T$, and the weights in vector $w = (w_1, \dots, w_m)^T$, equation (7.1) can be summarized into

$$\hat{z}_0 = w^T y \tag{7.2}$$

Rudimentary interpolation could imply to copy just the nearest observation, in this case $w_i = 1$, with p_i being the observation position closest to p_0, and all other $w_{j \neq i} = 0$. This is referred to as nearest neighbour interpolation.

Another simple alternative is to take the mean of all available observations in the domain. Then we have $w_1 = w_2 = \cdots = w_m = \frac{1}{m}$.

Figure 7.2 shows two interpolation configurations in a two-dimensional spatial domain. Observation positions are given in blue, and the desired interpolation position is indicated in red. In the configuration on the left (gridded data-set), many observations are available in all directions in a regular way around the desired interpolation position, while in the configuration on the right, observations are sparse and irregular — in some directions observations are available, whereas not in other directions, while in addition, some observations have their

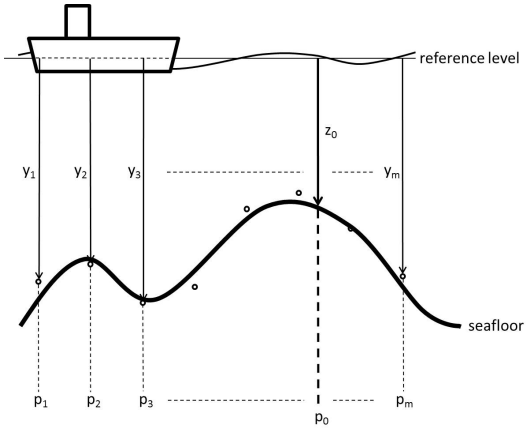


Figure 7.1: The water-depth z_0 at position p_0 has *not* been measured. The water-depth z_0 is interpolated, based on measurements of water-depth $y_1, \ldots, y_m$ at positions p_1 through p_m. Note that the measurements — the open dots — are *not* exactly on the sea-floor, as they are subject to (small) measurement errors. Generally, the resulting estimate or prediction $\hat{z}_0$ (interpolated value) will not lie exactly on the sea-floor either, though hopefully be close to it.

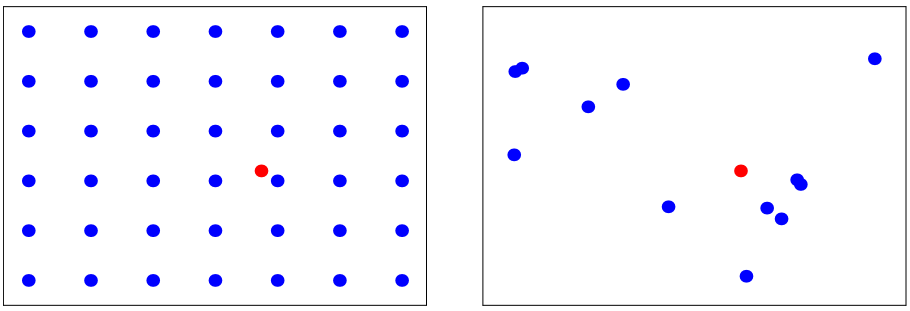


Figure 7.2: Observations of the attribute in a regular grid, in a two-dimensional domain, in blue (left), and at arbitrary positions (right). The intended interpolation position p_0 is indicated by the red dot.

positions almost coinciding. In the first case, a simple but fast method will probably work well, while in the latter case, one should be more careful on how the method handles clusters of data and directional variability.

If the desired position (red dot) is within the convex hull of surrounding observation positions, then one speaks of interpolation, otherwise of extrapolation. Two classes of interpolation methods are distinguished here. In deterministic methods, the uncertainty of, and correlation between observations is ignored, while stochastic methods do take this uncertainty and correlation into account. The next two sections present the basics of deterministic interpolation and stochastic interpolation.

7.2. Deterministic interpolation

Deterministic interpolation methods are distinguished by the way in which they distribute weights w_i with $i = 1, \ldots, m$ over the available observations (with $w_i \geq 0$).

As mentioned before, two very basic ways of interpolation are nearest neighbour interpolation, in which all weight is given to the closest observation, and averaging all available observations by distributing the weights equally over all observations. Most common, slightly more advanced, methods are inverse distance interpolation and triangular interpolation.

7.2.1. Inverse distance interpolation

It is intuitively appealing that observations close-by (in the spatial domain) show large similarity with the attribute at the target interpolation position. The observations are weighted by their spatial distance to the interpolation position. Let d_i be the distance between points p_i and p_0, with $i = 1, \dots, m$. The *inverse distance interpolation* of (for instance) the height at position p_0 is given by

$$\hat{z}_0 = \frac{1}{\sum_{i=1}^{m} \frac{1}{d_i}} \left(\frac{1}{d_1} y_1 + \cdots + \frac{1}{d_m} y_m \right) \tag{7.3}$$

based on the height observations $y_1, \dots, y_m$. Therefore, the weight given to the j-th observation equals

$$w_j = \frac{\frac{1}{d_j}}{\sum_{i=1}^{m} \frac{1}{d_i}} \tag{7.4}$$

Note that the sum of all weights equals one.

More or less weight to closeby observations can be given by incorporating a power p in equation (7.3), respectively by power $p > 1$ (more) and $p < 1$ (less). Inverse distance interpolation of power p reads:

$$\hat{z}_0 = \frac{1}{\sum_{i=1}^{m} \frac{1}{d_i^p}} \left(\frac{1}{d_1^p} y_1 + \cdots + \frac{1}{d_m^p} y_m \right) \tag{7.5}$$

Inverse distance interpolation works well for dense and regularly space data.

Often only observations within some threshold distance R to the interpolation location are incorporated ($d_i \leq R$); the neighbourhood of position p_0 is restricted to those positions within a distance of R.

7.2.2. Triangular interpolation

Rather than using a single nearby observation for interpolation, as done with nearest neighbour interpolation, one may want to use the three observations, which are, in a two-dimensional spatial domain, directly surrounding the target interpolation position p_0. Triangular interpolation is often used for *dense* data, like airborne laser scanning data (for creating 3D terrain models, for instance). This method consists of two steps. Assume m height observations $y_1, \dots, y_m$ are available at positions $p_1, \dots, p_m$, with $p_1 = ((x_1)_1, (x_2)_1), \dots, p_m = ((x_1)_m, (x_2)_m)$, and that we want to obtain a height estimate at a 2D position p_0.

The first step consists of determining the *Delaunay triangulation* of the observation positions $p_1, \dots, p_m$, see Figure 7.3 on the left.

In the second step one finds the (smallest) triangle consisting of three nodes (positions p_i, p_j, and p_k) which contains p_0, see Figure 7.3 on the right. Only the three observations y_i, y_j, y_k at positions p_i, p_j, and p_k are used. All other observations get a weight equal to zero for interpolation at p_0, hence $w_l = 0$ for $l \neq i, j, k$. Positive weights for y_i, y_j and y_k are obtained from the triangular weight equation

$$\hat{z}_0 = \frac{A_{0jk}}{A_{ijk}} y_i + \frac{A_{i0k}}{A_{ijk}} y_j + \frac{A_{ij0}}{A_{ijk}} y_k = w_i y_i + w_j y_j + w_k y_k \tag{7.6}$$

where A_{ijk} denotes the area of the triangle with vertices i, j, k, and noting that $A_{0jk} + A_{i0k} + A_{ij0} = A_{ijk}$. So, the closer p_0 is to vertex i, the larger the triangle on the other side (A_{0jk}) and the more weight y_i gets.

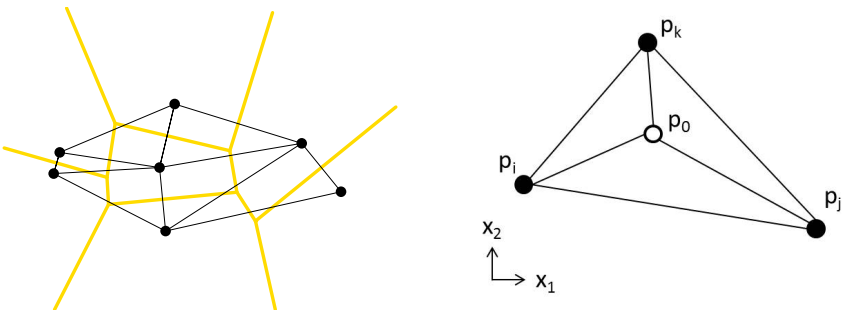


Figure 7.3: Delaunay triangulation in black of a set of observation points/positions (black dots) in two-dimensional $\mathbb{R}^2$, and Voronoi diagram in yellow, on the left. In triangular interpolation the weights w_i, w_j and w_k of the attribute at positions p_i, p_j and p_k are proportional to their relative distance to the center point p_0.

An example of a Delaunay triangulation is given in black in Figure 7.3 (on the left), while in yellow its dual structure, the Voronoi diagram is given. To create the Voronoi diagram, one starts from the perpendicular bisectors between any two positions (black dots), and uses these to construct the smallest possible convex polygon around each position (practically spoken, the smallest possible area (cell) around each position). This yields the Voronoi cells in yellow, Figure 7.3 (on the left). Triangular interpolation is relatively fast, as efficient algorithms exist for creating a Delaunay triangulation, cf. [3].

Once the Delaunay triangulation has been done for the full set of observation positions, typically irregularly distributed, triangular interpolation is often used to create a regular interpolation grid; then the attribute is determined (interpolated) at each grid point.

7.3. Stochastic interpolation

Though deterministic interpolation is intuitively appealing at first sight, there are a few problems associated.

In stochastic interpolation we consider the *spatial variation* of the attribute in a statistical way. The attribute to-be-interpolated, which is subject to variations, is modelled as a *random function* $\underline{z}(p)$, which depends on position p. As an example we can assume that the seafloor is flat (deterministic trend), but quite naturally, small variations may occur from place to place, see Figure 7.4. These random variations are referred to as the *signal* in the quantity of interest, in this case the seafloor-depth; the seafloor naturally goes up and down, though smoothly. Considering the observation positions, the random function $\underline{z}(p)$ actually consists of a set of *random variables* $\underline{z}(p_1), \ldots, \underline{z}(p_m)$. We note that examples of *spatial* interpolation are shown here, but the approach presented in this section equally applies to *temporal* interpolation (with the attribute, for instance temperature, as a function of time).

Secondly, as we know already from chapter 4, observables $\underline{y}_1, \ldots, \underline{y}_m$ are *random* variables, and the *measurement* uncertainty, or noise, should be taken into account in the interpolation, like we did with parameter estimation. For *optimal* interpolation results, the measurement uncertainty should be reflected in the weights in (7.1).

Finally one should also propagate the quality of the observables into measures of quality of the interpolation result, taking into account also the natural variability of the signal, so that one is able, to not only present the result, but also in a realistic way to evaluate its quality (uncertainty). What counts in the end is, how far off the interpolated value is from the actual sea-floor depth. The result of interpolation, estimate $\hat{z}_0$ at position p_0 is shown in Figure 7.5.

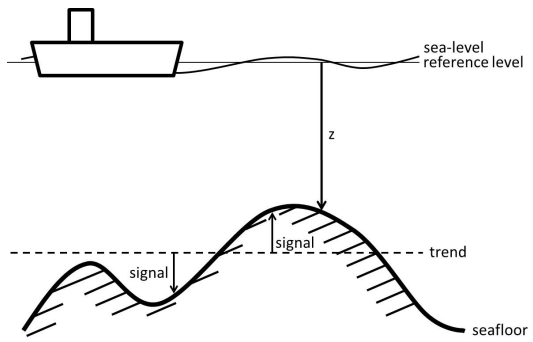


Figure 7.4: The to-be-interpolated attribute z (e.g. the seafloor-depth) is modelled (simply) as a trend, in this case the flat dashed horizontal line (average seafloor-level), supplemented by a small variation signal, the solid line. The signal behaves smoothly in the spatial domain, there are no sudden changes or jumps. The trend and signal together represent the actual seafloor-depth z. By including a *random* signal in the model, the interpolation can accomodate small local variations of the attribute (e.g. the depth) with respect to the trend. Water-depth is measured with respect to the thin line, which represents the reference or zero height/depth level (for instance average sea-level).

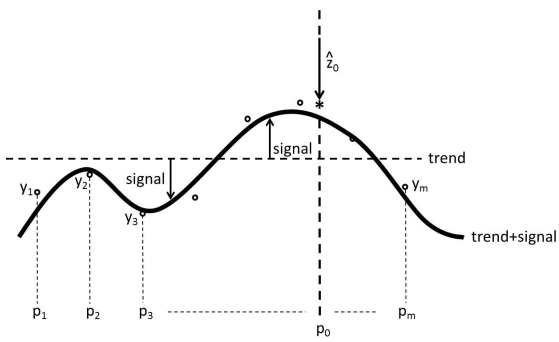


Figure 7.5: The result of interpolation: water-depth $\hat{z}_0$. The open dots indicate water-depth observations $y_1, \ldots, y_m$. Because of measurement errors, they do not perfectly coincide with the actual sea-floor. The actual sea-floor is not known to us - we just got the observations at positions $p_1, \ldots, p_m$. Interpolation is about determining the water-depth at another location p_0, and do this in a best way. The interpolated water-depth $\hat{z}_0$, indicated by the asterisk, is hopefully close to the actual sea-floor.

7.3.1. Kriging

A class of methods that takes care of the above issues is called Kriging. Kriging is a common technique in geology and environmental sciences, and fits in the theory of linear *prediction* developed in (geo-)statistics and geodesy.

Kriging takes as input, the observation values, and also takes into account stochastic properties of the attribute and the observable. In this section we restrict ourselves to working with the first two central moments of a random variables' distribution: the mean and the variance.

The spatial variation of the attribute with respect to the trend is captured by the signal, denoted by $\underline{s}$, and it depends on the position p, i.e. $\underline{s}(p)$. The signal is assumed here to have zero mean, i.e. on average, the variations of the seafloor with respect to the assumed flat plane equal zero. The variation is described by a *covariance function*, cf. Figure 7.7 on the right; this implies that signal values at two nearby positions are largely correlated (dependent), whereas signal values of two positions far away from each other will have little, or no dependence. In this way physical smoothness of the variation is translated into statistical correlation. The covariance is assumed to depend on the separation vector between two positions, and in practice often only on its norm, hence the Euclidean *distance* between two positions, and not on the particular positions themselves.

Secondly we take into account the uncertainty in the observables $\underline{y}$. As one can see in Figure 7.5, random deviations in the observations from the trend consist of firstly, the signal variation (the solid line hoovers around the dashed line), and secondly, the measurement uncertainty (the open dots are not exactly on the solid line). With respect to model (4.3) in chapter 4, the random signal $\underline{s}$ is added, and we have

$$\underline{y} = Ax + \underline{s} + \underline{e} \tag{7.7}$$

with vector y holding the observations $y_1, \dots, y_m$ at positions $p_1, \dots, p_m$, with Ax the so-called trend, vector s the signal at positions $p_1, \dots, p_m$, i.e. $s = (s(p_1), \dots, s(p_m))^T$, and vector e the measurement errors $e = (e_1, \dots, e_m)^T$. The full mxm variance matrix Q_{yy} of all observables, cf. (3.5), and section 4.3.4, takes both effects into account (error in the measurement system and signal in the observed attribute). It starts from the variance matrix Q_{yy} in chapter 4, representing the measurement noise, and adding now the mxm variance matrix Q_{ss} of vector $\underline{s}$, constructed from the covariance function, with elements $\sigma_{s(p_i)s(p_j)}$ for $i = 1, \dots, m$ and $j = 1, \dots, m$. Positions p_i and p_j are separated by a certain distance d_{ij}, and given this distance, matrix entry $\sigma_{s(p_i)s(p_j)}$ is simply read from Figure 7.7, on the right. In short-hand notation we have $Q_{yy} := Q_{yy} + Q_{ss}$ (assuming that measurement noise and signal are not correlated). In many applications of Ordinary Kriging, the measurement noise — when compared to the signal variation — can often be, or is often neglected. In that case we simply have $Q_{yy} := Q_{ss}$.

In order to derive the *best* possible interpolator, the interpolation error is defined as $\underline{\hat{\epsilon}} = \underline{z}_0 - \underline{\hat{z}}_0$, the difference between the actual water-depth z_0 at position p_0 and the interpolated value $\hat{z}_0$. We start from the Mean Squared Error (MSE), cf. section 2.6 and impose conditions on the solution like that the interpolation is a *linear* combination of the observations, cf. equation (7.2), and that the interpolation is *unbiased*, thus $E(\underline{\hat{z}}) = E(\underline{z})$, and this leads to an expression for the error variance of the interpolator. Minimizing the error variance (implying minimum MSE) results in a solution that in practice is obtained by solving a system of linear equations. Kriging not only provides a numerical interpolation result that is optimal in the sense of minimum error variance $\sigma^2_{\hat{\epsilon}}$ (least uncertainty), but also provides this variance as a measure of the uncertainty of the interpolation result.

Finally we note that in this chapter we always work with a given covariance function. In practice, the covariance function will not be known, and has to be estimated. In particular

selecting an appropriate type or shape of covariance function is important. These subjects are beyond the scope of this primer.

7.3.2. Ordinary Kriging

The most common way of Kriging is so-called Ordinary Kriging. Ordinary Kriging assumes, first, that the attribute under consideration has a constant mean over the entire spatial domain (i.e. the seafloor globally is assumed to be flat and level, as in Figure 7.4), but the mean water-depth is *unknown*. The attribute $\underline{z}$ at position p_i equals the unknown trend x, plus the signal $\underline{s}$ at that position, cf. Figure 7.4

$$\underline{z}(p_i) = x + \underline{s}(p_i) \tag{7.8}$$

Stating that the attribute has a constant unknown mean over the entire spatial domain implies that matrix A in (7.7) reduces to $A = l$, with l a vector of all ones, i.e. $l = (1, \dots, 1)^T$, of length m.

Secondly, it assumes that the covariance is the same over the entire spatial domain under consideration (typically a decaying function of distance, as the example in Figure 7.7 on the right).

The above assumptions, in combination with requirements on unbiasedness and linearity as above, see Appendix A.7, lead to the following Ordinary Kriging system:

$$\begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix} \begin{pmatrix} w \\ \nu \end{pmatrix} = \begin{pmatrix} Q_{yz_0} \\ 1 \end{pmatrix} \tag{7.9}$$

The last row of this matrix-vector system, $l^T w = 1$, ensures that the sum of all weights equals one (and represents the unbiasedness condition). Matrix Q_{yy} is the variance matrix of $\underline{y}$ in (7.7). Vector Q_{yz_0} contains the covariances $\sigma_{s(p_i)s(p_0)}$ between the signal at observation position p_i and the signal at interpolation position p_0, and this for $i = 1, \dots, m$; the covariances depend, according to the covariance function, on the distances d_{i0}.

Equation (7.9) is a square system with $m + 1$ unknowns, namely $w = (w_1, \dots, w_m)^T$ and Lagrange multiplier λ, with $\nu = \frac{\lambda}{2}$, cf. Appendix A.7, and can be solved by inverting the $(m + 1)$x$(m + 1)$ matrix, resulting in

$$\begin{pmatrix} w \\ \nu \end{pmatrix} = \begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} Q_{yz_0} \\ 1 \end{pmatrix} \tag{7.10}$$

Finally, once vector w has been obtained, the interpolated value is computed with (7.2). Ordinary Kriging is based on Best Linear Unbiased Prediction (BLUP), similar to BLUE in section 4.3.4. With the weights in w resulting from the above equation, it can be shown, cf. Appendix A.7, that the interpolator (7.2) can be rewritten into

$$\underline{\hat{z}}_0 = \underline{\hat{x}} + Q_{z_0 y} Q_{yy}^{-1} (\underline{y} - l\underline{\hat{x}}) \tag{7.11}$$

with estimator $\underline{\hat{x}}$ as

$$\underline{\hat{x}} = (l^T Q_{yy}^{-1} l)^{-1} l^T Q_{yy}^{-1} \underline{y} \tag{7.12}$$

similar to (4.6) with $A = l$, and Q_{yy} the variance matrix of $\underline{y}$ in (7.7). The estimator $\underline{\hat{x}}$ is also a linear function of $\underline{y}$. Matrix $Q_{z_0 y}$ is a row-vector, containing the covariances $\sigma_{s(p_0)s(p_i)}$.

In Figure 7.6 we show an example of Ordinary Kriging. On the left $m = 5$ observations of terrain height are shown in a two-dimensional spatial domain. The graph on the right shows

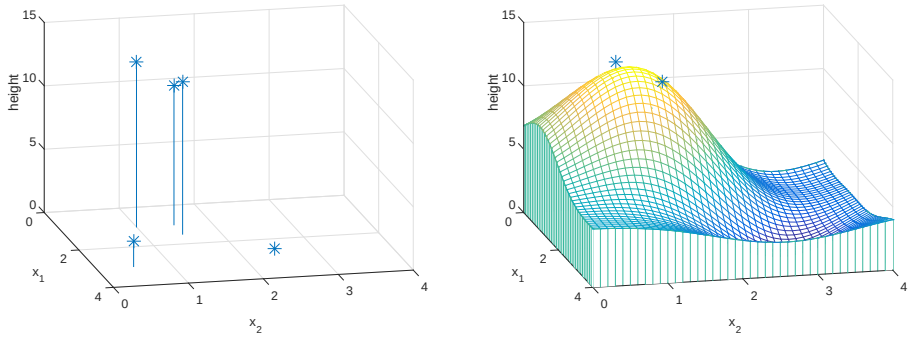


Figure 7.6: Example of Ordinary Kriging. On the left, five observations of attribute height, indicated by asterisks, in a two-dimensional spatial domain (x_1, x_2). The observed heights are 0, 2, 13, 11, 12, at positions $(2\frac{1}{2}, 2\frac{1}{2})$, $(3, \frac{1}{2})$, $(1, 1)$, $(1, 1\frac{1}{2})$, $(1\frac{1}{2}, 1\frac{1}{2})$. On the right the interpolated terrain model - two of the observations are above the interpolated surface, and three are below (and thereby not visible). The used covariance function was $\sigma_{s(p_i)s(p_j)} = ae^{-bd_{ij}^2}$, with $a = 1$ and $b = \frac{1}{2}$, and d_{ij} the Euclidean distance between positions p_i and p_j. The measurement noise $\underline{e}$ was set to have a variance of $\sigma_e^2 = \frac{1}{10}$.

the interpolation result, computed for a 0.1×0.1 grid of positions. Using (7.12), the trend x (mean terrain height) is estimated as $\hat{x} = 4.7$. This is quite remarkable, as the mean of the five observations is 7.6. This is caused by the fact that variance matrix Q_{yy} of $\underline{y}$ in (7.7) now contains measurement noise *and* signal noise. The last three observations are closely together (clustered in the spatial domain) and hence, according to a covariance function, like in Figure 7.7 on the right, highly correlated (dependent). In computing the mean height for the full spatial domain, they therefore 'count for' one observation (their mean equals 12, and counts effectively as one observation, rather than three).

With universal Kriging one can use a general polynomial trend model (for instance a sloped line or plane), rather than just a constant, as with Ordinary Kriging.

In the next sub-section we cover a variant of Kriging, with an additional restriction.

7.3.3. Simple Kriging

Simple Kriging, as the name suggests, implies a further simplification of Ordinary Kriging. Instead of an unknown trend, it is assumed that the mean value of the attribute is *known*. Parameter x in (7.8) is known.

Simple Kriging will be covered by means of an example. The temperature in a room is kept constant, at a pre-set, known (scalar) value x. If this were all available information, predicting the temperature would simply be $\hat{z}_0 = x$ (for any time t_0); it equals the known mean, obviously. In practice however, small variations of temperature occur over time. The random signal $\underline{s}$ now represents the deviations from the nominal temperature x. This signal has a constant, zero mean, as we had with (7.8) as well.

The temperature in the room is *observed* at regular time instants $t_1, \ldots, t_m$, and the observations are denoted as $y_1, \ldots, y_m$. Compared to the assumed variation in the temperature, the measurement uncertainty is negligible in this example (though measurement noise can also be accomodated with Simple Kriging, similar as with Ordinary Kriging).

Now we would like to interpolate the temperature to some time instant t_0 in between the observation time instants, based on all available information, i.e. the known mean x, and the observations in vector $y = (y_1, \ldots, y_m)^T$. The result reads

$$\underline{\hat{z}}_0 = x + w^T(\underline{y} - lx) \tag{7.13}$$

where l is again a vector with all ones, $l = (1, \ldots, 1)^T$, and $w = Q_{yy}^{-1} Q_{yz_0}$, similar to (7.10),

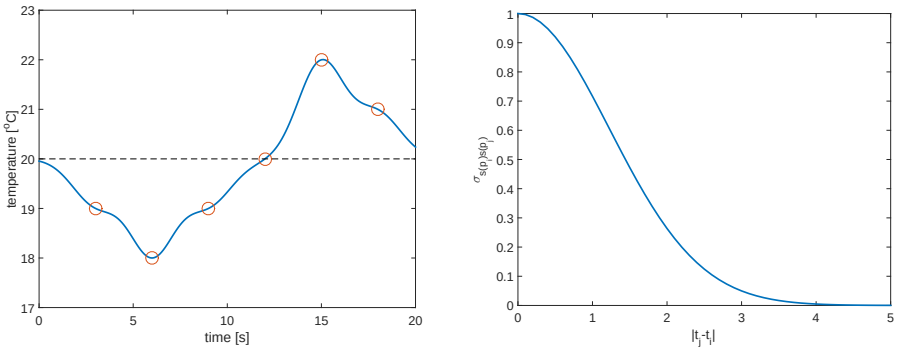


Figure 7.7: Example of Simple Kriging. On the left, six observations (red circles) of attribute temperature (at time instants 3, 6, 9, 12, 15 and 18), and the interpolation result (blue line). On the right the covariance function of the signal $\sigma_{s(t_i)s(t_j)} = ae^{-b|t_j-t_i|^2}$ with $a = 1$ and $b = \frac{1}{3}$; the amount of covariance between the signal at two positions or instants depends on their distance, shown along the horizontal axis, i.e. the covariance depends on how far the positions or instants are apart.

though omitting the last row (this result is given here without proof/derivation).

This equation can — as a side note — be shown to deliver the interpolation value, as a function of both the known mean and the observations, through rewriting it as

$$\underline{\hat{z}}_0 = (1 - w^T l)x + w^T \underline{y} = (1 - \sum_{i=1}^{m} w_i)x + \sum_{i=1}^{m} w_i y_i$$

Substituting the expression for weight-vector w in (7.13) yields

$$\underline{\hat{z}}_0 = x + Q_{z_0 y} Q_{yy}^{-1}(\underline{y} - lx) \tag{7.14}$$

This result is based on Best Linear Prediction (BLP). The result actually looks very similar to (7.11), but in (7.14) the known mean x is used, whereas estimator $\underline{\hat{x}}$ is used in (7.11).

Figure 7.7 shows an example of Simple Kriging, specifically the temperature as a function of time t. The known mean equals $x = 20$ and is shown by the dashed line. The actual observations are shown by red circles, $y = (19, 18, 19, 20, 22, 21)^T$. The blue line presents the interpolation result. One can see that in between observations, the interpolation has the tendency of going back to the known mean (dashed line) — the interpolation is a combination of the known mean and the observations cf. (7.13).

The graph on the right shows the covariance function of random signal $\underline{s}$. The amount of covariance between the signal at two positions p_i and p_j depends on the mutual distance d_{ij} between them, or, in this example, the difference in time between instants t_i and t_j: $d_{ij} = |t_j - t_i|$. Typically, an exponentially decaying function is used, as $\sigma_{s(p_i)s(p_j)} = ae^{-bd_{ij}^2}$. Parameter b governs the width of the covariance function; a larger value for b yields a function which is more narrow, and a smaller value yields a wider covariance function. The width of the covariance function measures the *smoothness* of the signal; a wide function corresponds to a smooth signal, a narrow function allows for rapid variations. Parameter a represents the signal variance, i.e. $\sigma_{s(p_i)s(p_i)} = a$, and equals σ_z^2; it is a measure of the *magnitude* of the signal variations with respect to the trend.

The interpolation result in Figure 7.7 on the left, is computed using equation (7.14). The

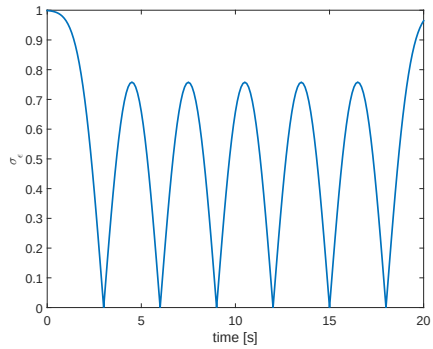


Figure 7.8: Example of Simple Kriging: interpolation error standard deviation $\sigma_{\hat{\epsilon}}$ as a function of the interpolation time instant.

6x6-variance matrix Q_{yy} reads

$$Q_{yy} = \begin{pmatrix} 1 & 0.05 & & & & \\ 0.05 & 1 & 0.05 & & & \\ & 0.05 & 1 & 0.05 & & \\ & & 0.05 & 1 & 0.05 & \\ & & & 0.05 & 1 & 0.05 \\ & & & & 0.05 & 1 \end{pmatrix}$$

For the interpolation $\hat{z}_0$ at time instant $t_0 = 4$, matrix $Q_{z_0 y}$ reads

$$Q_{z_0 y} = \begin{pmatrix} 0.72 & 0.26 & 0 & 0 & 0 & 0 \end{pmatrix}$$

which has only two non-zero values. In line with the covariance function in Figure 7.7 on the right and equation (7.14), interpolation at $t_0 = 4$ depends on the known mean, and on the two neighbouring observations, at $t_i = 3$ we have $|t_i - t_0| = 1$, and at $t_i = 6$ we have $|t_i - t_0| = 2$. The interpolated value at $t_0 = 4$ becomes $\hat{z}_0 = 18.85$, well in between the two neighbouring observations of 19 at $t = 3$, and 18 at $t = 6$.

Figure 7.8 shows the interpolation error standard deviation $\sigma_{\hat{\epsilon}}$ from (A.5), as a function of the interpolation position, in this case, interpolation time instant. At $t_0 = 4$ the standard deviation reads $\sigma_{\hat{\epsilon}(t_0=4)} = 0.66$. Note that the interpolated signal passes exactly through the observations (red circles), cf. Figure 7.7. At the observation positions (time instants), the error standard deviation is zero cf. Figure 7.8, as at these positions there is no uncertainty on the interpolated height/depth (as actually no interpolation is needed at these positions). When there is measurement noise present, the interpolated signal will *not* pass exactly through the observations. In between observation instants, the interpolation error standard deviation will increase. The interpolation error standard deviation as shown in Figure 7.8 initially increases with increasing distance to the observation positions, but levels off at a maximum value at the so-called range distance $\sigma_{\hat{\epsilon}} = \sqrt{a}$ (the level when there would be no observation), the distance beyond which there is no correlation anymore between observables, according to the covariance function used. For more information see e.g. [4].

7.4. Exercises and worked examples

This section presents a derivation-exercise and a simple problem with a worked answer on interpolation.

Question 1 Piecewise linear interpolation of an attribute, in a one-dimensional spatial, or temporal domain implies just connecting two succesive observations by a straight line,

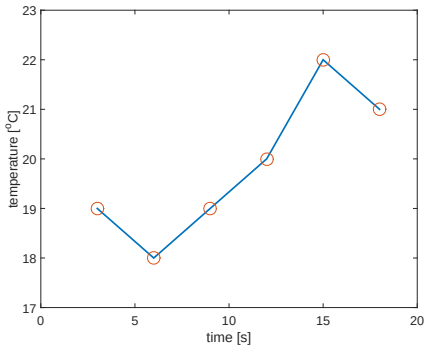


Figure 7.9: Piecewise linear interpolation of observations in the example of Figure 7.7, with question 1.

see Figure 7.9. Start from equation (7.3) for inverse distance interpolation, and restrict it to $m = 2$, for instance interpolating z for position p_0 with $p_0 \in [p_1, p_2]$, using only y_1 and y_2 in the interpolation. Show that in this case one obtains

$$\hat{z}_0 = \frac{p_2 - p_0}{p_2 - p_1} y_1 + \frac{p_0 - p_1}{p_2 - p_1} y_2 = y_1 + \frac{p_0 - p_1}{p_2 - p_1}(y_2 - y_1)$$

Answer 1 Restricting $m = 2$ and using solely y_1 and y_2 in (7.3) yields

$$\hat{z}_0 = \frac{1}{\frac{1}{d_{10}} + \frac{1}{d_{20}}}\left(\frac{1}{d_{10}} y_1 + \frac{1}{d_{20}} y_2\right)$$

with d_{10} the distance between p_1 and p_0, and d_{20} the distance between p_2 and p_0. This can be rewritten as

$$\hat{z}_0 = \frac{d_{10} d_{20}}{d_{10} + d_{20}}\left(\frac{1}{d_{10}} y_1 + \frac{1}{d_{20}} y_2\right) = \frac{d_{20}}{d_{10} + d_{20}} y_1 + \frac{d_{10}}{d_{10} + d_{20}} y_2$$

which shows the given equation, when the horizontal coordinate p is introduced (e.g. $d_{20} = p_2 - p_0$).

Question 2 Two observations of terrain height are given: $y_1 = 3$ at position $p_1 = 1$, and $y_2 = 4$ at position $p_2 = 3$, see Figure 7.10 on the left. The measurement noise of the observables has a variance of $\sigma^2_{y_1} = \sigma^2_{y_2} = \frac{1}{2}$, and is uncorrelated across the observables. In

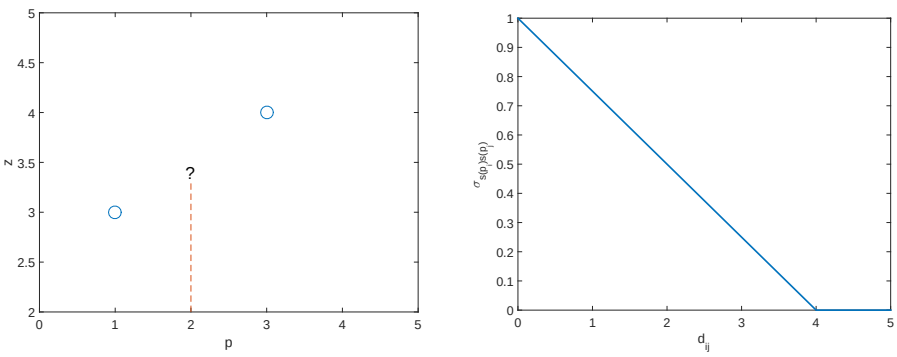


Figure 7.10: On the left, problem statement for question 2. Two observations of height are given (shown by the circles), $y_1 = 3$ and $y_2 = 4$, at positions $p_1 = 1$ and $p_2 = 3$ respectively, and one is asked to interpolate the height for position $p_0 = 2$. On the right, covariance function $\sigma_{s(p_i)s(p_j)}$ for question 2.

this spatial domain from $p = 0$ to $p = 5$, the terrain can be assumed flat, though the mean terrain-height is unknown. On the right in Figure 7.10 the covariance function is given, it describes the 'smoothness' of the terrain. It is mathematically given as $\sigma_{s(p_i)s(p_j)} = -\frac{1}{4}d_{ij} + 1$ for $0 \leq d_{ij} \leq 4$, and zero otherwise, with $d_{ij} = |p_j - p_i|$. With all this information, one can perform interpolation using Ordinary Kriging. Compute the interpolated terrain-height for $p = 2$. Also do this for $p = 1\frac{1}{2}$.

Answer 2 The terrain-height exactly halfway the two observation positions can, logically, be expected to be the average of the two observations. Kriging will also provide this answer. The variance matrix for the signal at the two observation positions p_1 and p_2, using the covariance function in Figure 7.10 on the right, reads

$$Q_{ss} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

Together with the measurement noise $\sigma^2_{e_1} = \sigma^2_{e_2} = \frac{1}{2}$, the variance matrix of vector $\underline{y}$, as outlined with (7.7), becomes

$$Q_{yy} = \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} \end{pmatrix}$$

and its inverse reads

$$Q_{yy}^{-1} = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

With (7.12) and $l = (1, 1)^T$, the estimated mean terrain height becomes $\hat{x} = 3\frac{1}{2}$. Then with (7.11), and matrix, or actually vector Q_{z_0y} as $Q_{z_0y} = \begin{pmatrix} \frac{3}{4} & \frac{3}{4} \end{pmatrix}$, based on the distance from p_0 to p_1, and the distance from p_0 to p_2, and the covariance function of Figure 7.10, we obtain, as expected

$$\hat{z}_0 = 3\frac{1}{2} + \begin{pmatrix} \frac{3}{4} & \frac{3}{4} \end{pmatrix} \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} \end{pmatrix} [\begin{pmatrix} 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} 3\frac{1}{2}] = 3\frac{1}{2}$$

Repeating the exercise for $p_0 = 1\frac{1}{2}$ yields $\hat{z}_0 = 3\frac{3}{8}$. The full interpolation result is shown in Figure 7.11. Apparently, interpolation with Ordinary Kriging, does *not* yield a straight line passing through the two observation points. The observations y_1 and y_2 are subject to measurement error. And also mind that in Ordinary Kriging, one assumes that the attribute under consideration has a constant mean across the entire spatial domain. That is why the result is an in-between the straight line and a flat horizontal line, and does not pass through the two observations. As a final exercise, verify yourself that for $p_0 = 1$ you obtain $\hat{z}_0 = 3$, and $\hat{z}_0 = 4$ for $p_0 = 3$, when the measurement noise can be neglected, i.e. when $\sigma^2_{e_1} = \sigma^2_{e_2} = 0$. In this case the interpolated result is exactly matching the observations y_1 and y_2.

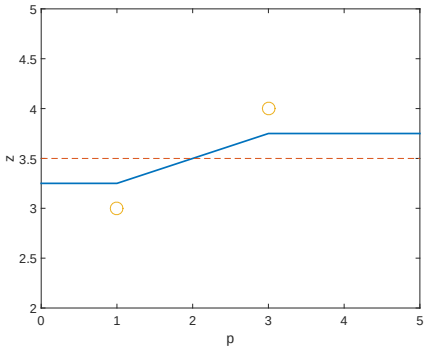


Figure 7.11: Interpolation result using Ordinary Kriging for question 1, $\hat{x}$ is given by the dashed line, and the interpolation $\hat{z}_0$ by the solid line in blue; the two circles give the observations.

8

Conclusion

In land surveying it is good practice to take more measurements than strictly necessary. Using redundant measurements serves the purpose of increasing the precision of the estimators for the unknown parameters of interest, and it allows for checking the mutual and internal consistency of the set of measurements.

Using the variance matrix of chapter 4, one can see that when the measurement set-up is doubled, the variances of the estimators go down by a factor of 2. And in chapter 6, we have seen that redundant measurements can be used to check the validity of the assumed mathematical model, and to diagnose whether big outliers or anomalies are present in the data-set. This check provides a safeguard against producing, unknowingly, largely incorrect results.

A

Several proofs [*]

A.1. Mean and variance of normal distribution

The Probability Density Function (PDF) of a normally distributed random variable $\underline{y}$ reads

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}} \tag{A.1}$$

and in this section we prove that x is the mean, and σ the standard deviation (σ^2 the variance).

Using (2.9) the mean reads

$$E(\underline{y}) = \int_{-\infty}^{+\infty} y \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy \tag{A.2}$$

which can be split into

$$E(\underline{y}) = x \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy + \int_{-\infty}^{+\infty} \frac{(y-x)}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy$$

The first part yields just the mean x, and in the second part we apply the change of variable $z = \frac{y-x}{\sigma}$

$$E(\underline{y}) = x + \int_{-\infty}^{+\infty} \sigma \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

where the factor σ enters in the second term as the change of variable implies $\sigma dz = dy$.

$$E(\underline{y}) = x + \frac{\sigma}{\sqrt{2\pi}} \left[-e^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} = x$$

as the second term yields zero. The mean of the random variable $\underline{y}$ reads x.

Using (2.10) the variance reads

$$D(\underline{y}) = \int_{-\infty}^{+\infty} (y-x)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy \tag{A.3}$$

where we used already $E(\underline{y}) = x$. Again we apply the change of variable $z = \frac{y-x}{\sigma}$

$$D(\underline{y}) = \int_{-\infty}^{+\infty} \sigma^2 \frac{z^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

which can be split into

$$D(\underline{y}) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (ze^{-\frac{z^2}{2}})(z)\,dz$$

where we apply integration by parts

$$D(\underline{y}) = \frac{\sigma^2}{\sqrt{2\pi}} \left(-\left[ze^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}\,dz \right) = \sigma^2$$

where we used the definite integral $\int_0^{\infty} e^{-\frac{z^2}{(\sqrt{2})^2}}\,dz = \frac{\sqrt{2\pi}}{2}$. The variance of the random variable $\underline{y}$ reads σ^2.

A.2. Mean and variance propagation laws

We consider the following linear transformation, as given by (3.7), of m-vector $\underline{y}$

$$\underline{v} = R\underline{y} + s$$

where vector $\underline{v}$ has n elements, and consequently matrix R has n rows and m columns, and vector s is an n-vector.

In this section we prove that $E(\underline{v}) = RE(\underline{y}) + s$ (3.8) and that $Q_{vv} = RQ_{yy}R^T$ (3.9).

Using

$$E(G(\underline{y})) = \int_{-\infty}^{\infty} G(y)f(y)\,dy$$

we compute the mean of $\underline{v}$

$$E(\underline{v}) = E(R\underline{y} + s) = \int_{-\infty}^{\infty} (Ry + s)f(y)\,dy$$

$$= R\int_{-\infty}^{\infty} yf(y)\,dy + s\int_{-\infty}^{\infty} f(y)\,dy = RE(\underline{y}) + s$$

which proves (3.8) in section 3.2.

With (3.4) applied to vector $\underline{v}$

$$D(\underline{v}) = E((\underline{v} - E(\underline{v}))(\underline{v} - E(\underline{v}))^T) = E((R\underline{y} + s - RE(\underline{y}) - s)(R\underline{y} + s - RE(\underline{y}) - s)^T)$$

$$= RE((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T)R^T = RD(\underline{y})R^T = RQ_{yy}R^T = Q_{vv}$$

which proves (3.9).

A.3. Non-linear mean and variance propagation laws

In section 3.4 the mean and variance propagation laws were given for the non-linear transformation $\underline{v} = G(\underline{y})$ (3.10), where $G()$ represents n non-linear functions of each time m random variables (the elements of vector $\underline{y}$).

The Taylor series, up to second order term, of one of the n non-linear functions $\underline{v}_i = G_i(\underline{y})$ with $i = 1, \ldots, n$, at $E(\underline{y})$, reads

$$G_i(\underline{y}) \approx G_i(E(\underline{y})) + \left.\frac{\partial G_i}{\partial y^T}\right|_{E(\underline{y})} (\underline{y} - E(\underline{y})) + \frac{1}{2}(\underline{y} - E(\underline{y}))^T \left.\frac{\partial^2 G_i}{\partial y y^T}\right|_{E(\underline{y})} (\underline{y} - E(\underline{y}))$$

where $\frac{\partial G_i(y)}{\partial y}$ is the $m \times 1$ gradient vector of $G_i(y)$, and $\frac{\partial^2 G_i(y)}{\partial y y^T}$ is the $m \times m$ Hessian matrix of $G_i(y)$ (matrix with second order partial derivatives).

The expectation of $\underline{v}_i$ in terms of $\underline{y}$ becomes

$$E(\underline{v}_i) = E(G_i(\underline{y})) \approx E(G_i(E(\underline{y}))) + E(\frac{\partial G_i}{\partial y^T}\bigg|_{E(\underline{y})} (\underline{y} - E(\underline{y}))) + E(\frac{1}{2}(\underline{y} - E(\underline{y}))^T \frac{\partial^2 G_i}{\partial y y^T}\bigg|_{E(\underline{y})} (\underline{y} - E(\underline{y})))$$

$$= G_i(E(\underline{y})) + \frac{1}{2}\text{trace}(\frac{\partial^2 G_i}{\partial y y^T}\bigg|_{E(\underline{y})} E((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T))$$

$$= G_i(E(\underline{y})) + \frac{1}{2}\text{trace}(\frac{\partial^2 G_i}{\partial y y^T}\bigg|_{E(\underline{y})} Q_{yy})$$

as $E(\underline{y} - E(\underline{y})) = E(\underline{y}) - E(E(\underline{y})) = E(\underline{y}) - E(\underline{y}) = 0$, and where trace means taking the sum of the diagonal elements of the matrix, and, we used definitions (3.4) and (3.5) for Q_{yy}. This completes the proof of (3.11).

The $n \times n$ variance matrix of $\underline{v}$ is formally given by

$$Q_{vv} = E((\underline{v} - E(\underline{v}))(\underline{v} - E(\underline{v}))^T)$$

see (3.4) and (3.5). We consider one element (i, j), which is $Q_{v_i v_j}$

$$Q_{v_i v_j} = E((\underline{v}_i - E(\underline{v}_i))(\underline{v}_j - E(\underline{v}_j))^T)$$

with $i = 1, \ldots, n$ and $j = 1, \ldots, n$. We supply the above Taylor series, up to second order term, of $\underline{v}_i = G_i(\underline{y})$, and of $\underline{v}_j = G_j(\underline{y})$, and we substitute the just obtained approximation for $E(\underline{v}_i)$, and for $E(\underline{v}_j)$. In the expansion we neglect all terms with cross products of first and second order derivatives of $G(y)$, as well as products of second order derivatives, and we are left with just one term, namely

$$Q_{v_i v_j} \approx E(\frac{\partial G_i}{\partial y^T}\bigg|_{E(\underline{y})})(\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T \frac{\partial G_j}{\partial y^T}\bigg|_{E(\underline{y})}^T$$

$$= \frac{\partial G_i}{\partial y^T}\bigg|_{E(\underline{y})} E((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T) \frac{\partial G_j}{\partial y^T}\bigg|_{E(\underline{y})}^T = \frac{\partial G_i}{\partial y^T}\bigg|_{E(\underline{y})} Q_{yy} \frac{\partial G_j}{\partial y^T}\bigg|_{E(\underline{y})}^T$$

where we used again definition (3.4). Finally the above expression can be used for any element $Q_{v_i v_j}$, hence, the full $n \times n$ variance matrix Q_{vv} results as

$$Q_{vv} \approx \frac{\partial G}{\partial y^T}\bigg|_{E(\underline{y})} Q_{yy} \frac{\partial G}{\partial y^T}\bigg|_{E(\underline{y})}^T$$

This completes the proof of (3.12). The term $\frac{\partial G(y)}{\partial y^T}$ is an $n \times m$ matrix, containing, as rows, the gradient vectors of non-linear functions $G_i(y)$, with $i = 1, \ldots, n$, all evaluated at $E(\underline{y})$.

A.4. Least-squares

In this section we prove that $\hat{x} = (A^T A)^{-1} A^T y$ (4.4) is the solution to $\min_x \|y - Ax\|^2$ (4.5).

For the least-squares solution we minimize, for x, the function $g(x) = \|y - Ax\|^2 = (y - Ax)^T (y - Ax)$. The gradient, the first order derivative, is set equal to zero, and the Hessian, the second order derivative, should be larger than zero (be a positive definite matrix), in order for the found (single) extremum to be a global minimizer.

$$g(x) = y^T y - y^T Ax - x^T A^T y + x^T A^T Ax = y^T y - 2x^T A^T y + x^T A^T Ax$$

as the innerproduct $y^T b = b^T y$. Setting $\frac{\partial g(x)}{\partial x} = 0$ yields

$$-2A^T y + 2A^T Ax = 0$$

where we used $\frac{\partial (x^T b)}{\partial x} = b$ and $\frac{\partial (x^T Mx)}{\partial x} = (M + M^T)x$ (which both can be verified by expansion), and recognize that $A^T A$ is a symmetric $n \times n$ matrix. Solving this for x yields

$$A^T A\hat{x} = A^T y$$

which is referred to as the system of normal equations (it is a system with n unknowns in n equations), and inverting matrix $A^T A$ yields indeed (4.4). The Hessian reads

$$\frac{\partial^2 g(x)}{\partial x x^T} = 2A^T A > 0$$

which is indeed a positive definite matrix, and where we used $\frac{\partial (b^T x)}{\partial x^T} = b^T$.

A.5. Concerns on non-linear estimation

For model (4.8) in section 4.4, application of the least-squares criterion (4.5), taking into account the variance matrix Q_{yy}, yields

$$\underline{\hat{x}}' = \arg \min_{x \in R^n} \|y - F(x)\|^2_{Q_{yy}^{-1}}$$

the non-linear estimator $\underline{\hat{x}}' = G(\underline{y})$, which in most cases in practice can not be found in an analytical form. Also we mention that non-linear estimation is not a trivial subject. The propagation of the random characteristics of $\underline{y}$ into those of $\underline{\hat{x}}'$ is difficult. For instance the mean and variance can not be propagated through a non-linear relation in a straightforward way, e.g. $E(\underline{\hat{x}}') = E(G(\underline{y})) \neq G(E(\underline{y}))$, the non-linear estimator is biased $E(\underline{\hat{x}}') \neq x$.

Through the iterative procedure in section 4.4 the estimate $\hat{x}$ is a numerical approximation of the realization of the non-linear estimator $\underline{\hat{x}}'$.

A.6. Line-fitting with observed independent variable

Equation (4.2) $E(\underline{y}) = Ax; D(\underline{y}) = Q_{yy}$ presented the (linear) model of observation equations, with observations y, unknown parameters x, and known coefficients in mxn-matrix A. Example (6.2) was on line fitting, and positions were observed, with error, but timing, the independent variable, was assumed to be perfect (error-free), yielding the coefficients of matrix A.

If this assumption (on the independent variable) can*not* be made, the model for line fitting has to be set up differently. As an example we consider the measurement of the expansion of a steel bar due to a rise in temperature. Both the dependent variable (the length, observed, in vector y; sometimes called the response), and the independent variable (temperature, occuring in matrix A), are now subject to observational error.

The length of the bar is measured m times: $y_1, y_2, \ldots, y_m$ (and inevitably measurement errors are being made), and, at the same time also the (corresponding) temperature is *measured*: $T_1, T_2, \ldots, T_m$ (and also here inevitably with some error). The functional model (in absence of any error) reads $y_i = T_i x_1 + x_2$, for $i = 1, \ldots, m$, where x_2 is the length of the bar at reference temperature (e.g. $T = 0$), and x_1 is the coefficient of thermal expansion (in this case the increase in length per degree, for instance for steel approximately 10 μm per degree, for a 1 meter bar).

Trying to formulate the model as before yields

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \underbrace{\begin{pmatrix} T_1 & 1 \\ T_2 & 1 \\ \vdots & \vdots \\ T_m & 1 \end{pmatrix}}_{A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

but, this model now has observational error also in the independent variable T_i, and is referred to as an Error-in-Variables (EIV) model. This is an example of a so-called *total least squares* problem. There exist various approaches to solve such a problem, e.g. using singular value decomposition. In the sequel we describe another way.

Earlier, only y was a random vector, namely $\underline{y}$, and T was not, but now also T is a random vector, namely $\underline{T}$, and with keeping just x_1 and x_2 as unknown parameters, we would have random variables in the design matrix A. Therefore the model is set up differently, namely by introducing also the temperatures as unknown parameters, next to x_1 and x_2. The actual temperatures are not known — they have been measured, but with error. The full model of observation equations becomes:

$$\begin{aligned} E(\underline{T}_i) &= T_i \\ E(\underline{y}_i) &= T_i x_1 + x_2 \quad \text{for} \quad i = 1, 2, \ldots, m \end{aligned} \quad ; \; D\begin{pmatrix} \underline{T} \\ \underline{y} \end{pmatrix} = \begin{pmatrix} Q_{TT} & 0 \\ 0 & Q_{yy} \end{pmatrix}$$

with in total $2m$ observations (which is exactly all we measured), m temperatures and m lengths, and $m + 2$ unknown parameters, namely x_1, x_2, and $T_1, T_2, \ldots, T_m$. The total $2m$ x $2m$ variance matrix will typically be a block-diagonal matrix, as there is generally no correlation between temperature and length readings.

The above extended model of observation equations is *non-linear*, notably by the product of T_i and x_1, cf. section 4.4 on linearization and iteration.

Figure A.1 shows an example *without* (at left), and *with* (at right) observational error in the independent variable. In the former case, the least-squares residuals (indicated by the dashed lines) are exactly along the vertical axis, whereas in the latter case both observed values T_i and y_i get 'corrected'.

A.7. Ordinary Kriging

In this section we prove the key equation for Ordinary Kriging (7.9).

The requirement of *linear* interpolation is already implied in (7.2), $\hat{\underline{z}}_0 = w^T \underline{y}$.

The interpolation is also required to be *unbiased*. With just a constant unknown mean (7.7), with $A = l$, yields $E(\underline{y}) = lx$, where l is a vector of all ones. Then with (7.2) and (3.8) $E(\hat{\underline{z}}_0) = w^T E(\underline{y}) = w^T lx$. Now with (7.8) for $\underline{z}(p_0) = \underline{z}_0$ we know that $E(\underline{z}_0) = x$, hence $E(\hat{\underline{z}}_0) = E(\underline{z}_0)$ yields the constraint $w^T l = 1$, or $\sum_{i=1}^{m} w_i = 1$.

Eventually, to achieve the *best* interpolator $\hat{\underline{z}}_0$, we require minimum error variance with the error as $\hat{\underline{\epsilon}} = \underline{z}_0 - \hat{\underline{z}}_0$. Variance $\sigma^2_{\hat{\epsilon}}$ should be as small as possible, implying the largest

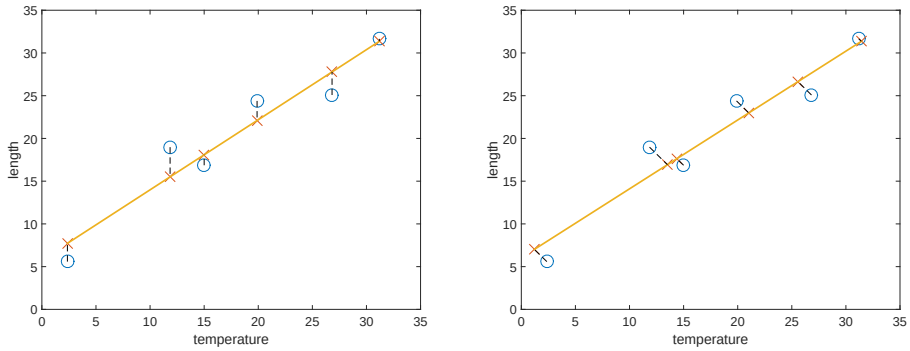


Figure A.1: Line fitting: the circles show the (same) original observations, pairs (T_i, y_i), on the left the (ordinary) least-squares solution when temperature is error-free, and at right, when both length and temperature measurement are subject to observational error, and the alternative (non-linear) approach with temperatures as unknown parameters is used. The crosses show pairs $(T_i, \hat{y}_i)$ at left, for the line $\hat{x}_1$=0.82 and $\hat{x}_2$=5.79, and pairs $(\hat{T}_{i,NL}, \hat{y}_{i,NL})$ at right, for the line $\hat{x}_{1,NL}$=0.81 and $\hat{x}_{2,NL}$=6.05; crosses are all on the fitted line in both cases.

probability on a small error at any position, and practically, the interpolated value being as close as possible to the actual water-depth. Hence, the goal is to determine the elements of vector w, such that

$$\min_w \sigma_{\hat{\epsilon}}^2 \quad \text{subject to} \quad w^T l - 1 = 0 \tag{A.4}$$

The error variance is obtained through noting, with (7.2), that

$$\underline{\hat{\epsilon}} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \underline{z}_0 \\ \underline{\hat{z}}_0 \end{pmatrix} = \begin{pmatrix} 1 & -w^T \end{pmatrix} \begin{pmatrix} \underline{z}_0 \\ \underline{y} \end{pmatrix}$$

with the variance matrix of the vector on the right hand side as

$$\begin{pmatrix} Q_{z_0 z_0} & Q_{z_0 y} \\ Q_{y z_0} & Q_{yy} \end{pmatrix}$$

and $Q_{z_0 z_0} = \sigma_{z_0}^2$. Applying the variance propagation law (3.9) yields

$$\sigma_{\hat{\epsilon}}^2 = \sigma_{z_0}^2 + w^T Q_{yy} w - Q_{z_0 y} w - w^T Q_{y z_0} = \sigma_z^2 + w^T Q_{yy} w - 2 w^T Q_{y z_0} \tag{A.5}$$

and the error variance depends on the weights in vector w.

The minimization (A.4), including the constraint, is solved using the Lagrange multiplier rule. The Lagrange function becomes

$$L(w, \lambda) = \sigma_{\hat{\epsilon}}^2 + \lambda(w^T l - 1) = \sigma_z^2 + w^T Q_{yy} w - 2 w^T Q_{y z_0} + \lambda(w^T l - 1)$$

where λ is the Lagrange multiplier, and (A.5) has been substituted. Setting the partial derivatives of $L(w, \lambda)$ to zero yields

$$\frac{\partial L(w, \lambda)}{\partial w} = 2 Q_{yy} w - 2 Q_{y z_0} + \lambda l = 0$$

which is an m-vector, and with Q_{yy} a symmetric matrix, and

$$\frac{\partial L(w, \lambda)}{\partial \lambda} = w^T l - 1 = 0$$

These $m+1$ equations, with $m+1$ unknowns can be cast in

$$\begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix} \begin{pmatrix} w \\ \nu \end{pmatrix} = \begin{pmatrix} Q_{yz_0} \\ 1 \end{pmatrix} \tag{A.6}$$

with $\nu = \frac{\lambda}{2}$, see (7.9), from which the weights can be obtained through (7.10).

The inverse in (7.10) can be shown to read

$$\begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix}^{-1} = \begin{pmatrix} Q_{yy}^{-1} - Q_{yy}^{-1} l (l^T Q_{yy}^{-1} l)^{-1} l^T Q_{yy}^{-1} & Q_{yy}^{-1} l (l^T Q_{yy}^{-1} l)^{-1} \\ (l^T Q_{yy}^{-1} l)^{-1} l^T Q_{yy}^{-1} & -(l^T Q_{yy}^{-1} l)^{-1} \end{pmatrix}$$

Once the values for vector w have been obtained, the interpolation error variance can be evaluated using (A.5).

B

Normal distribution: table

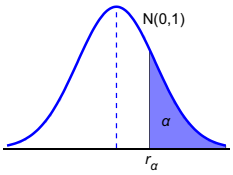


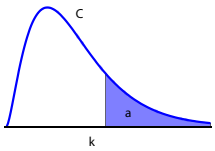
Figure B.1: Standard normal distribution $\mathrm{N}(0,1)$: one-sided level of significance α as function of the critical value r_α, i.e. $\alpha = 1 - \Phi(r_\alpha)$.

r_α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002

Table B.1: Standard normal distribution $\mathrm{N}(0,1)$: one-sided level of significance α as function of the critical value r_α, i.e. $\alpha = 1 - \Phi(r_\alpha)$. Values of r_α are given up to the first decimal in the first column, the second decimal in the first row. Example: $\alpha = 0.0250$ for $r_\alpha = 1.96$.

C

Chi-square distribution: table



α	0.9990	0.9950	0.9900	0.9750	0.9500	0.9000	0.7500	0.5000	0.2500	0.1000	0.0500	0.0250	0.0100	0.0050	0.0010
n															
1	0.0000	0.0000	0.0002	0.0010	0.0039	0.0158	0.1015	0.4549	1.3233	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	0.0020	0.0100	0.0201	0.0506	0.1026	0.2107	0.5754	1.3863	2.7726	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	0.0243	0.0717	0.1148	0.2158	0.3518	0.5844	1.2125	2.3660	4.1083	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	0.0908	0.2070	0.2971	0.4844	0.7107	1.0636	1.9226	3.3567	5.3853	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	0.2102	0.4117	0.5543	0.8312	1.1455	1.6103	2.6746	4.3515	6.6257	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150
6	0.3811	0.6757	0.8721	1.2373	1.6354	2.2041	3.4546	5.3481	7.8408	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577
7	0.5985	0.9893	1.2390	1.6899	2.1673	2.8331	4.2549	6.3458	9.0371	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219
8	0.8571	1.3444	1.6465	2.1797	2.7326	3.4895	5.0706	7.3441	10.2189	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245
9	1.1519	1.7349	2.0879	2.7004	3.3251	4.1682	5.8988	8.3428	11.3888	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772
10	1.4787	2.1559	2.5582	3.2470	3.9403	4.8652	6.7372	9.3418	12.5489	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883
11	1.8339	2.6032	3.0535	3.8157	4.5748	5.5778	7.5841	10.3410	13.7007	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641
12	2.2142	3.0738	3.5706	4.4038	5.2260	6.3038	8.4384	11.3403	14.8454	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095
13	2.6172	3.5650	4.1069	5.0088	5.8919	7.0415	9.2991	12.3398	15.9839	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282
14	3.0407	4.0747	4.6604	5.6287	6.5706	7.7895	10.1653	13.3393	17.1169	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233
15	3.4827	4.6009	5.2293	6.2621	7.2609	8.5468	11.0365	14.3389	18.2451	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973
16	3.9416	5.1422	5.8122	6.9077	7.9616	9.3122	11.9122	15.3385	19.3689	23.5418	26.2962	28.8454	31.9999	34.2672	39.2524
17	4.4161	5.6972	6.4078	7.5642	8.6718	10.0852	12.7919	16.3382	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185	40.7902
18	4.9048	6.2648	7.0149	8.2307	9.3905	10.8649	13.6753	17.3379	21.6049	25.9894	28.8693	31.5264	34.8053	37.1565	42.3124
19	5.4068	6.8440	7.6327	8.9065	10.1170	11.6509	14.5620	18.3377	22.7178	27.2036	30.1435	32.8523	36.1909	38.5823	43.8202
20	5.9210	7.4338	8.2604	9.5908	10.8508	12.4426	15.4518	19.3374	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968	45.3147
21	6.4467	8.0337	8.8972	10.2829	11.5913	13.2396	16.3444	20.3372	24.9348	29.6151	32.6706	35.4789	38.9322	41.4011	46.7970
22	6.9830	8.6427	9.5425	10.9823	12.3380	14.0415	17.2396	21.3370	26.0393	30.8133	33.9244	36.7807	40.2894	42.7957	48.2679
23	7.5292	9.2604	10.1957	11.6886	13.0905	14.8480	18.1373	22.3369	27.1413	32.0069	35.1725	38.0756	41.6384	44.1813	49.7282
24	8.0849	9.8862	10.8564	12.4012	13.8484	15.6587	19.0373	23.3367	28.2412	33.1962	36.4150	39.3641	42.9798	45.5585	51.1786
25	8.6493	10.5197	11.5240	13.1197	14.6114	16.4734	19.9393	24.3366	29.3389	34.3816	37.6525	40.6465	44.3141	46.9279	52.6197
26	9.2221	11.1602	12.1981	13.8439	15.3792	17.2919	20.8434	25.3365	30.4346	35.5632	38.8851	41.9232	45.6417	48.2899	54.0520
27	9.8028	11.8076	12.8785	14.5734	16.1514	18.1139	21.7494	26.3363	31.5284	36.7412	40.1133	43.1945	46.9629	49.6449	55.4760
28	10.3909	12.4613	13.5647	15.3079	16.9279	18.9392	22.6572	27.3362	32.6205	37.9159	41.3371	44.4608	48.2782	50.9934	56.8923
29	10.9861	13.1211	14.2565	16.0471	17.7084	19.7677	23.5666	28.3361	33.7109	39.0875	42.5570	45.7223	49.5879	52.3356	58.3012
30	11.5880	13.7867	14.9535	16.7908	18.4927	20.5992	24.4776	29.3360	34.7997	40.2560	43.7730	46.9792	50.8922	53.6720	59.7031
40	17.9164	20.7065	22.1643	24.4330	26.5093	29.0505	33.6603	39.3353	45.6160	51.8051	55.7585	59.3417	63.6907	66.7660	73.4020
50	24.6739	27.9907	29.7067	32.3574	34.7643	37.6886	42.9421	49.3349	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900	86.6608
60	31.7383	35.5345	37.4849	40.4817	43.1880	46.4589	52.2938	59.3347	66.9815	74.3970	79.0819	83.2977	88.3794	91.9517	99.6072
70	39.0364	43.2752	45.4417	48.7576	51.7393	55.3289	61.6983	69.3345	77.5767	85.5270	90.5312	95.0232	100.4252	104.2149	112.3169
80	46.5199	51.1719	53.5401	57.1532	60.3915	64.2778	71.1445	79.3343	88.1303	96.5782	101.8795	106.6286	112.3288	116.3211	124.8392
90	54.1552	59.1963	61.7541	65.6466	69.1260	73.2911	80.6247	89.3342	98.6499	107.5650	113.1453	118.1359	124.1163	128.2989	137.2084
100	61.9179	67.3276	70.0649	74.2219	77.9295	82.3581	90.1332	99.3341	109.1412	118.4980	124.3421	129.5612	135.8067	140.1695	149.4493

Table C.1: Central χ^2 distribution: critical value $\chi^2_\alpha(n,0)$ as function of one-sided level of significance α (top row) and degrees of freedom n (left column). Given is one-minus-the CDF, hence α represents the right-tail probability. Example: α = 0.010 and n = 10 yield $\chi^2_\alpha(n,0)$ = 23.2093.

Concluding remarks

This primer on Mathematical Geodesy addresses the following two learning objectives: 'knowledge, understanding, application, analysis and evaluation in the area of calculus of observation (in Dutch: waarnemingsrekening), and the assessment and presentation of (measures of) quality', and 'being able to convert a surveying problem into a mathematical model, subsequently determining parameters of interest, and eventually inserting results back into the application in reality', and addressing these objectives from the perspective of Surveying and Mapping applications in the field of Civil Engineering.

Bibliography

[1] F. Dekking, C. Kraaikamp, H. Lopuhaa, and L. Meester, *A modern introduction to probability and statistics — understanding why and how* (Springer-Verlag, 2005).

[2] D. Lay, *Linear algebra and its applications*, 4th ed. (Pearson, 2014).

[3] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Application*, 3rd ed. (Springer Verlag, 2008).

[4] M. Oliver and R. Webster, *Basic Steps in Geostatistics: The Variogram and Kriging* (SpringerBriefs in Agriculture, 2015).