Vrije Universiteit Brussel

Faculty of Engineering

System Identification

prof. Johan Schoukens

29th April 2013

Acknowledgment

Parts of this course are based on the books:

- System Identification. A Frequency Domain Approach. Rik Pintelon and Johan Schoukens (2001). IEEE Press, Piscataway, NJ 08855-1331.
- System Identification. A Frequency Domain Approach. Second Edition. Rik Pintelon and Johan Schoukens (2012). IEEE Press-Wiley.
- Mastering System Identification in 100 Exercises. Johan Schoukens, Rik Pintelon and Yves Rolain (2012). IEEE Press-Wiley

Contents

List of Figures

Chapter 1

An Introduction to identification

Chapter 1

An introduction to identification

In this chapter a brief, intuitive introduction to the identification theory is given. By means of a simple example the reader is made aware of a number of pitfalls associated with a model built from noisy measurements. Next an overview of the identification process is given. Eventually, a statistical characterization of the parameters is introduced.

1.1 What is identification?

From the beginning of our lives, when we grew up as babies we interacted with our environment. Intuitively, we learned to control our actions by predicting their effect. These predictions are based on an inborn model fitted to reality, using our past experiences. Starting from very simple actions (if I push a ball, it rolls), we soon became very able to deal with much more complicated challenges (walking, running, biking, playing ping-pong). Finally, this process culminates in the design of very complicated systems like radios, air-planes, mobile phones, ... etc. to satisfy our needs. We even build models just to get a better understanding of our observations of the universe: what does the life cycle of the sun

look like? Can we predict the weather of this afternoon, tomorrow, next week, next month? From all these examples it is seen that we never deal with the whole of nature at once: we always focus on those aspects we are interested in and don't try to describe all of reality using one coherent model. The job is split up, and efforts are concentrated on just one part of reality at a time. This part is called the system, the rest of nature being referred to as the environment of the system. Interactions between the system and its environment are described by input and output ports. For a very long time in the history of mankind the models were qualitative, and even nowadays we describe most real life situations using this "simple" approach: e.g. a ball will roll downhill; temperature will rise if the heating has been switched on; it seems it will rain since the sky looks very dark. In the last centuries this qualitative approach was complemented with quantitative models based on advanced mathematics, and until the last decade this seemed to be the most successful approach in many fields of science. Most physical laws are quantitative models describing some part of our impression of reality. However, it also became clear, very soon, that it can be very difficult to match a mathematical model to the available observations and experiences. Consequently, qualitative logical methods typified by fuzzy modeling became more popular, once more. In this book we deal with the mathematical, quantitative modeling approach. Fitting these models to our observations creates new problems. We look at the world through "dirty" glasses: when we measure a length, the weight of a mass, the current or voltage, ... etc. we always make errors since the instruments we use are not perfect. Also the models are imperfect, reality is far more complex than the rules we apply. Many systems are not deterministic. They also show a stochastic behavior which makes it impossible to predict exactly their output. Noise in a radio receiver, Brownian motion of small particles, variation of the wind speed in a thunder storm are all illustrations of this nature. Usually we split the model into a deterministic part and a stochastic part. The deterministic aspects are captured by the mathematical system model, while the stochastic behavior is modeled as a noise distortion.

The aim of identification theory is to provide a systematic approach to fit the mathematical model, as well as possible, to the deterministic part, eliminating the noise distortions as much as possible.

Later in this book the meaning of terms like "system" and "goodness of fit" will be precisely described. Before formalizing the discussion we want to motivate the reader by analyzing a very simple example, illustrating many of the aspects and problems that appear in identification theory.

1.2 Identification: a simple resistance example

1.2.1 Estimation of the value of a resistor

Two groups of students had to measure a resistance. Their measurement setup is shown in 1.1. They passed a constant but unknown current through the resistor. The voltage u_0 across the resistor and the current i_0 through it were measured using a voltmeter and an ampere meter. The input impedance of the voltmeter is very large compared with the unknown resistor so that all the measured current is assumed to pass through the resistor. A set of voltage and current measurements, respectively, $u(k)$, $i(k)$ with $k = 1, 2, \ldots, N$ is made. The measurement results of each group are shown in 1.2. Since the measurements were very noisy the groups decided to average their results. Following a lengthy discussion, 3 estimators for the resistance were proposed:

$$\hat{R}_{\mathrm{SA}}(N) = \frac{1}{N}\sum_{k=1}^{N}\frac{u(k)}{i(k)} \tag{1.1}$$

$$\hat{R}_{\mathrm{LS}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k)}{\frac{1}{N}\sum_{k=1}^{N} i^2(k)} \tag{1.2}$$

$$\hat{R}_{\mathrm{EV}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)}{\frac{1}{N}\sum_{k=1}^{N} i(k)} \tag{1.3}$$

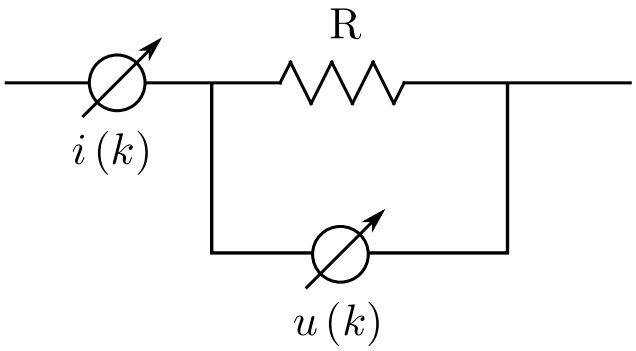
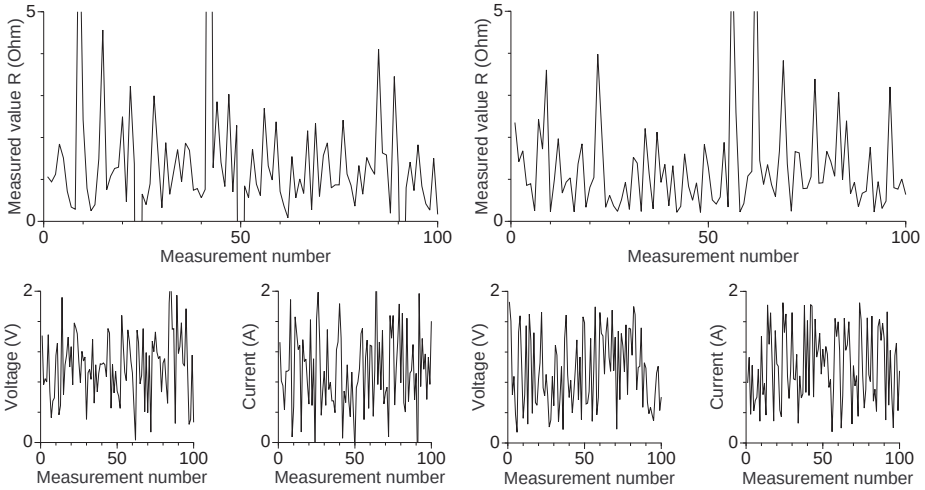


Figure 1.1: Measurement of a resistor.



(a) Group A (b) Group B

Figure 1.2: Measurement results $u(k), i(k)$ for groups A and B.
The plotted value $R(k)$ is obtained by direct division of the voltage by the current: $R(k) = u(k)/i(k)$.

The index N indicates that the estimate is based on N observations. Note that the three estimators result in the same estimate on noiseless data. Both groups processed their measurements and their results are given in Figure 1.3. From this figure a number of interesting observations can be made:

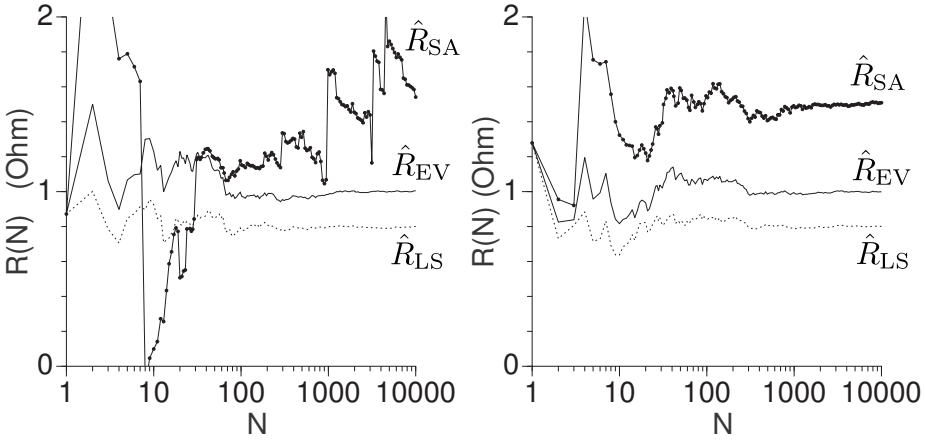


Figure 1.3: Estimated resistance values $\hat{R}(N)$ for both groups as a function of the number of processed data N.
Full dotted line: $\hat{R}_{\mathrm{SA}}$, dotted line: $\hat{R}_{\mathrm{LS}}$, full line: $\hat{R}_{\mathrm{EV}}$.

- All estimators have large variations for small values of N, and seem to converge to an asymptotic value for large values of N, except $\hat{R}_{\mathrm{SA}}(N)$ of group A. This corresponds to the intuitively expected behavior: if a large number of data points are processed we should be able to eliminate the noise influence due to the averaging effect.

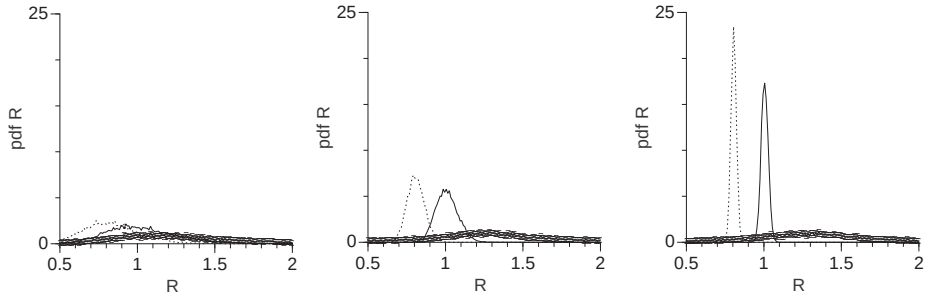
- The asymptotic values of the estimators depend on the kind of averaging technique that is used. This shows that there is a serious problem: at least 2 out of the 3 methods converge to a wrong value. It is not even certain that any one of the estimators is doing well. This is quite catastrophic: even an infinite amount of measurements does not guarantee that the exact value is found.

- The $\hat{R}_{\mathrm{SA}}(N)$ of group A behaves very strangely. Instead of converging to a fixed value, it jumps irregularly up and down before convergence is reached.

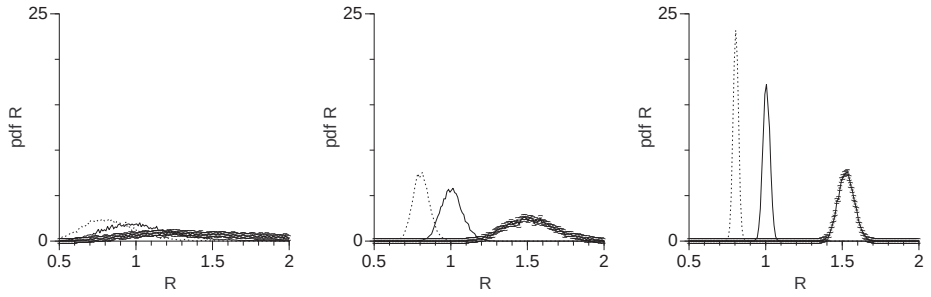
These observations prove very clearly that a good theory is needed to explain and understand the behavior of candidate estimators. This will allows us to make a sound selection out of many possibilities and to indicate in advance,

before running expensive experiments, if the selected method is prone to serious shortcomings.

In order to get a better understanding of their results the students repeated their experiments many times and looked to the histogram of $\hat{R}(N)$ for $N = 10, 100$ and 1000. Normalizing these histograms gives an estimate of the pdf (probability density function) of $\hat{R}(N)$ as shown in 1.4. Again the students could learn a lot from these figures:



(a) Group A



(b) Group B

Figure 1.4: Observed pdf of $\hat{R}(N)$.
From the left to the right 10, 100 and 1000: full dotted line: $\hat{R}_{\mathrm{SA}}(N)$, dotted line: $\hat{R}_{\mathrm{LS}}(N)$, full line: $\hat{R}_{\mathrm{EV}}(N)$.

- For small values of N the estimates are widely scattered. As the number of processed measurements increases, the pdf becomes more concentrated.
- The estimates $\hat{R}_{\mathrm{LS}}(N)$ are less scattered than $\hat{R}_{\mathrm{EV}}(N)$, while for $\hat{R}_{\mathrm{SA}}(N)$ the odd behavior in the results of group A appears again. The distribution of this estimate does not contract for growing values of N for group A, while it does for group B.

- Again it is clearly visible that the distributions are concentrated around different values.

At this point in the exercise, the students could still not decide which estimator is the best. Moreover, there seems to be a serious problem with the measurements of group A because $\hat{R}_{\mathrm{SA}}(N)$ behaves very oddly. Firstly they decided to focus on the scattering of the different estimators, trying to get more insight into the dependency on N. In order to quantify the scattering of the estimates, their standard deviation is calculated and plotted as a function of N in 1.5.

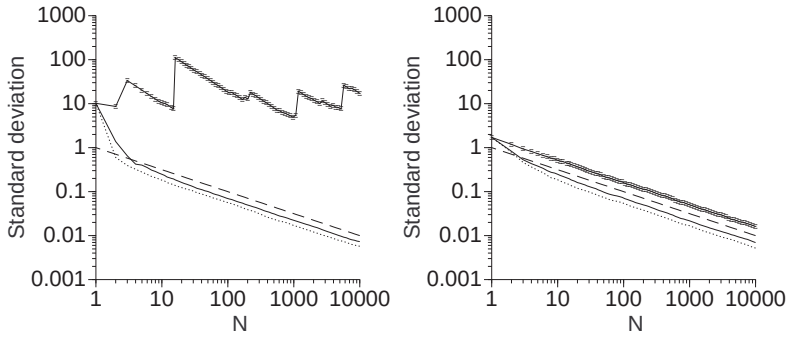


Figure 1.5: Standard deviation of $\hat{R}(N)$ for the different estimators, and comparison with $1/\sqrt{N}$.
Full dotted line: $\hat{R}_{\mathrm{SA}}(N)$, dotted line: $\hat{R}_{\mathrm{LS}}(N)$, full line: $\hat{R}_{\mathrm{EV}}(N)$, dashed line $1/\sqrt{N}$.

- The standard deviation of $\hat{R}(N)$ decreases monotonically with N, except for the pathological case, $\hat{R}_{\mathrm{SA}}(N)$, of group A. Moreover, it can be concluded by comparing with the broken line that the standard deviation is proportional to $1/\sqrt{N}$. This is in agreement with the rule of thumb which states that the uncertainty on an averaged quantity obtained from independent measurements decreases as $1/\sqrt{N}$.

- The uncertainty in this experiment depends on the estimator. Moreover, the proportionality to $1/\sqrt{N}$ is only obtained for sufficiently large values of N for $\hat{R}_{\mathrm{LS}}(N)$ and $\hat{R}_{\mathrm{EV}}(N)$.

Since both groups of students used the same programs to process their measurements, they concluded that the strange behavior of $\hat{R}_{\mathrm{SA}}(N)$ in group A should

be due to a difference in the raw data. For that reason they took a closer look at the time records given in Figure 1.2. Here it can be seen that the measurements of group A are a bit more scattered than those of group B. Moreover, group A measured some negative values for the current while group B did not. In order to get a better understanding, they made a histogram of the raw current data as shown in Figure 1.6.

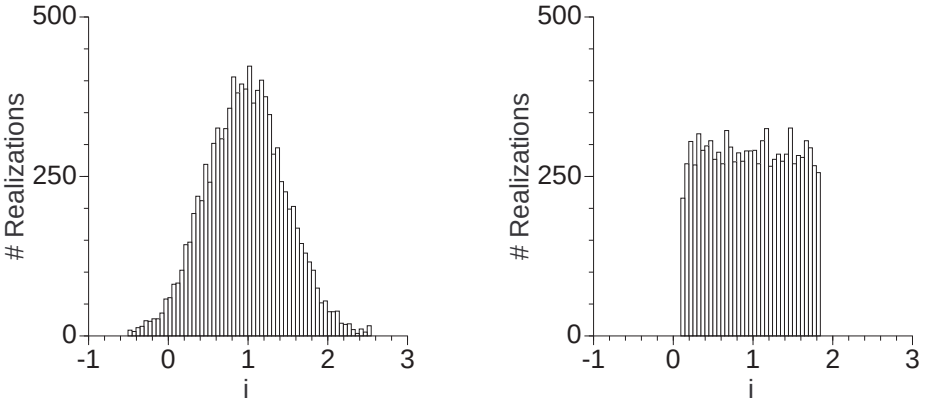


Figure 1.6: Histogram of the current measurements.

These histograms clarify the strange behavior of $\hat{R}_{\mathrm{SA}}$ of group A. The noise on the measurements of group A looks completely different from that of group B. Due to the noise on the current measurements, there is a significant risk of getting current values that are very close to zero for group A, while this is not so for group B. These small current measurements blow up the estimate $\hat{R}_{\mathrm{SA}}(k) = \frac{u(k)}{i(k)}$ for some k, so that the running average $\hat{R}_{\mathrm{SA}}$ cannot converge, or more precisely: the expected value $\mathrm{E}\left\{\frac{u(k)}{i(k)}\right\}$ does not exist. This will be discussed in more detail later in this chapter. This example shows very clearly that there is a strong need for methods which can generate and select between different estimators. Before setting up a general framework, the resistance problem is further elaborated.

It is also remarkable to note that although the noise on the measurements is completely differently distributed, the distribution of the estimated resistance values $\hat{R}_{\mathrm{LS}}$ and $\hat{R}_{\mathrm{EV}}$ seems to be the same in Figure 1.4 for both groups.

1.2.2 Simplified analysis of the estimators

With the knowledge they got from the previous series of experiments, the students eliminated $\hat{R}_{\mathrm{SA}}$, but they were still not able to decide whether $\hat{R}_{\mathrm{LS}}$ or $\hat{R}_{\mathrm{EV}}$ was the best. More advanced analysis techniques are needed to solve this problem. As the estimates are based on a combination of a finite number of noisy measurements, there are bound to be stochastic variables. Therefore, an analysis of the stochastic behavior is needed to select between both estimators. This is done by calculating the limiting values and making series expansions of the estimators. In order to keep the example simple, we will use some of the limit concepts quite loosely. Precise definitions are postponed till Section 1.3. Three observed problems are analyzed below:

- Why do the asymptotic values depend on the estimator?
- Can we explain the behavior of the variance?
- Why does the $\hat{R}_{\mathrm{SA}}$ estimator behave strangely for group A?

To do this it is necessary to specify the stochastic framework: how are the measurements disturbed with the noise (multiplicative, additive), and how is the noise distributed? For simplicity we assume that the current and voltage measurements are disturbed by additive zero mean, independently and identically distributed noise, formally formulated as:

$$i(k) = i_0 + n_i(k) \qquad \text{and} \qquad u(k) = u_0 + n_u(k). \tag{1.4}$$

where i_0 and u_0 are the exact but unknown values of the current and the voltage, $n_i(k)$ and $n_u(k)$ are the noise on the measurements.

Assumption 1: Disturbing noise

$n_i(k)$ and $n_u(k)$ are mutually independent, zero mean, independent and identically distributed (iid) random variables with a symmetric distribution and with variance σ_u^2 and σ_i^2.

1.2.2.1 Asymptotic value of the estimators

In this section the limiting value of the estimates for $N \to \infty$ is calculated. The calculations are based on the observation that the sample mean of iid random variables $x(k)$, $k = 1, \ldots, N$ converges to its expected value, $\mathrm{E}\{x\}$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x(k) = \mathrm{E}\{x\}. \tag{1.5}$$

Moreover, if $x(k)$ and $y(k)$ obey Assumption 1, then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x(k)\, y(k) = 0 \tag{1.6}$$

Since we are dealing here with stochastic variables, the meaning of this statement should be defined more precisely, but in this section we will just use this formal notation and make the calculations straightforwardly (see Section1.3) for a formal definition).

The first estimator we analyze is $\hat{R}_{\mathrm{LS}}(N)$. Taking the limit of (1.2), gives

$$\begin{aligned} \lim_{N\to\infty} \hat{R}_{\mathrm{LS}}(N) &= \lim_{N\to\infty} \frac{\sum_{k=1}^{N} u(k)i(k)}{\sum_{k=1}^{N} i^2(k)} && (1.7)\\ &= \frac{\lim_{N\to\infty} \sum_{k=1}^{N} (u_0 + n_u(k))(i_0 + n_i(k))}{\lim_{N\to\infty} \sum_{k=1}^{N} (i_0 + n_i(k))^2} && (1.8) \end{aligned}$$

Or, after dividing the numerator and denominator by N

$$\lim_{N\to\infty} \hat{R}_{\mathrm{LS}}(N) = \frac{\lim_{N\to\infty} \left[u_0 i_0 + \frac{u_0}{N} \sum_{k=1}^{N} n_i(k) + \frac{i_0}{N} \sum_{k=1}^{N} n_u(k) + \frac{1}{N} \sum_{k=1}^{N} n_u(k) n_i(k)\right]}{\lim_{N\to\infty} \left[i_0^2 + \frac{1}{N} \sum_{k=1}^{N} n_i^2(k) + \frac{2i_0}{N} \sum_{k=1}^{N} n_i(k)\right]} \tag{1.9}$$

Since n_i and n_u are zero mean iid, it follows from (1.5) and (1.6) that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} n_u(k) = 0,$$
$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} n_i(k) = 0 \text{ and}$$
$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} n_u(k) n_i(k) = 0.$$

However, the sum of the squared current noise distributions does not converge to zero, but to a constant value different from zero

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} n_i^2(k) = \sigma_i^2$$

so that the asymptotic value becomes:

$$\lim_{N\to\infty} \hat{R}_{\mathrm{LS}}(N) = \frac{u_0 i_0}{i_0^2 + \sigma_i^2} = \frac{R_0}{1 + \sigma_i^2 / i_0^2}. \tag{1.10}$$

This simple analysis gives a lot of insight into the behavior of the $\hat{R}_{\mathrm{LS}}(N)$ estimator. Asymptotically, this estimator underestimates the value of the resistance due to quadratic noise contributions in the denominator. Although the noise disappears in the averaging process of the numerator, it contributes systematically in the denominator. This results in a systematic error (called bias) that depends on the signal-to-noise ratio (SNR) of the current measurements: i_0/σ_i.

The analysis of the second estimator $\hat{R}_{\mathrm{EV}}(N)$ is completely similar. Using

(1.3), we get

$$\lim_{N\to\infty} \hat{R}_{\mathrm{EV}}(N) = \lim_{N\to\infty} \frac{\sum_{k=1}^{N} u(k)}{\sum_{k=1}^{N} i(k)} \tag{1.11}$$

$$= \frac{\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} (u_0 + n_u(k))}{\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} (i_0 + n_i(k))} \tag{1.12}$$

or

$$\lim_{N\to\infty} \hat{R}_{\mathrm{EV}}(N) = \frac{u_0 + \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} n_u(k)}{i_0 + \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} n_i(k)} \tag{1.13}$$

$$= \frac{u_0}{i_0} = R_0 \tag{1.14}$$

so that we can conclude now that $\hat{R}_{\mathrm{EV}}(N)$ converges to the true value and should be preferred over $\hat{R}_{\mathrm{LS}}(N)$. These conclusions are also confirmed by the students' results in 1.3, where it is seen that the asymptotic value of $\hat{R}_{\mathrm{LS}}(N)$ is much smaller than that of $\hat{R}_{\mathrm{EV}}(N)$.

1.2.2.2 Strange behavior of the "simple approach"

Finally, we have to analyze $\hat{R}_{\mathrm{SA}}(N)$ in order to understand its strange behavior. Can't we repeat the previous analysis here? Consider

$$\hat{R}_{\mathrm{SA}}(N) = \frac{1}{N} \sum_{k=0}^{N} \frac{u(k)}{i(k)} = \frac{1}{N} \sum_{k=0}^{N} \frac{u_0 + n_u(k)}{i_0 + n_i(k)}. \tag{1.15}$$

A major difference with the previous estimators is the order of summing and dividing: here the measurements are first divided and then summed together, while for the other estimators we first summed the measurements together before making the division. In other words, for $\hat{R}_{\mathrm{LS}}(N)$ and $\hat{R}_{\mathrm{EV}}(N)$ we first applied an averaging process (summing over the measurements) before making the division. This makes an important difference.

$$\hat{R}_{\mathrm{SA}}(N) = \frac{1}{N} \frac{u_0}{i_0} \sum_{k=0}^{N} \frac{1 + n_u(k)/u_0}{1 + n_i(k)/i_0} \tag{1.16}$$

In order to process $\hat{R}_{\mathrm{SA}}(N)$ along the same lines as the other estimators, we should get rid of the division, for example by making a Taylor series expansion:

$$\frac{1}{1+x} = \sum_{l=0}^{\infty}(-1)^l x^l \text{ for } |x| < 1. \tag{1.17}$$

with $x = \frac{n_i(k)}{i_0}$. Since the terms $n_i^{2l+1}(k)$ and $n_u^l(k)n_i^l(k)$ disappear in the averaging process (the pdf is symmetric), the limiting value becomes

$$\lim_{N\to\infty} \hat{R}_{\mathrm{SA}}(N) = R_0\left(1 + \frac{1}{N}\sum_{k=1}^{N}\left(\frac{n_i(k)}{i_0}\right)^2 + \frac{1}{N}\sum_{k=1}^{N}\left(\frac{n_i(k)}{i_0}\right)^4 + ...\right) \tag{1.18}$$

with $\left|\frac{n_i(k)}{i_0}\right| < 1$. If we neglect all terms of order 4 or more, the final result becomes

$$\lim_{N\to\infty} \hat{R}_{\mathrm{SA}}(N) = R_0\left(1 + \frac{\sigma_i^2}{i_0^2}\right) \tag{1.19}$$

if $\left|\frac{n_i(k)}{i_0}\right| < 1, \forall\, k$.

From this analysis we can make two important conclusions

- The asymptotic value only exists if the following condition on the measurements is met: the series expansion must exist otherwise (1.19) is NOT valid. The measurements of group A violate the condition that is given in (1.18) while those of group B obey it (see Figure 1.6). A more detailed analysis shows that this condition is too rigorous. In practice it is enough that the expected value $\mathrm{E}\left\{\hat{R}_{\mathrm{SA}}(N)\right\}$ exists. Since this value depends on the pdf of the noise, a more detailed analysis of the measurement noise would be required. For some noise distributions the expected value exists even if the Taylor expansion does not!

- If the asymptotic value exists, 1.19 shows that it will be too large. This is also seen in the results of Group B in Figure 1.3. We know already that $\hat{R}_{\mathrm{EV}}(N)$ converges to the exact value, and $\hat{R}_{\mathrm{SA}}(N)$ is clearly significantly larger.

1.2.2.3 Variance analysis

In order to get a better understanding of the sensitivity of the different estimators to the measurement noise, the students made a variance analysis using first order Taylor series approximations. Again they began with the $\hat{R}_{\mathrm{LS}}(N)$. Starting from(1.8), and neglecting all second order contributions like $n_u(k)n_i(k)$, or $n_i^2(k)$ it is found that

$$\hat{R}_{\mathrm{LS}}(N) \approx R_0 \left(1 + \frac{1}{N}\sum_{k=1}^{N}(\frac{n_u(k)}{u_0} - \frac{n_i(k)}{i_0})\right) = R_0 + \Delta R. \tag{1.20}$$

The approximated variance $\mathrm{Var}\left\{\hat{R}_{\mathrm{LS}}(N)\right\}$is (using Assumption 1)

$$\mathrm{Var}\left\{\hat{R}_{\mathrm{LS}}(N)\right\} = \mathrm{E}\left\{(\Delta R)^2\right\} = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right) \tag{1.21}$$

with $\mathrm{E}\{\cdot\}$ the expected value. Note that during the calculation of the variance, the shift of the mean value of $\hat{R}_{\mathrm{LS}}(N)$ is not considered since it is a second order contribution.

For the other two estimators, exactly the same results are found:

$$\mathrm{Var}\left\{\hat{R}_{\mathrm{EV}}(N)\right\} = \mathrm{Var}\left\{\hat{R}_{\mathrm{SA}}(N)\right\} = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right). \tag{1.22}$$

The result $\mathrm{Var}\left\{\hat{R}_{\mathrm{SA}}(N)\right\}$ is only valid if the expected values exist.

Again, a number of interesting conclusions can be made from this result

- The standard deviation is proportional to $1/\sqrt{N}$ as was found before in Figure 1.5.
- Although it is possible to reduce the variance by averaging over repeated measurements, this is no excuse for sloppy experiments since the uncertainty is inversely proportional to the SNR of the measurements. Increasing the SNR requires many more measurements in order to get the same final uncertainty on the estimates.

- The variance of the three estimators should be the same. This seems to conflict with the results of Figure 1.5. However, the theoretical expressions are based on first order approximations. If the SNR drops to values that are too small, the second order moments are no longer negligible. In order to check this, the students set up a simulation and tuned the noise parameters so that they got the same behavior as they had observed in their measurements. These values were: $i_0 = 1\text{A}$, $u_0 = 1\text{V}$, $\sigma_i = 1\text{A}$, $\sigma_u = 1\text{V}$. The noise of group A is normally distributed and uniformly distributed for group B. Next they varied the standard deviations and plotted the results in 1.7 for $\hat{R}_{\text{EV}}(N)$ and $\hat{R}_{\text{LS}}(N)$. Here it is clear that for higher SNR the uncertainties coincide while they differ significantly for the lower SNR. To give closed form mathematical expressions for this behavior, it is not enough any more to specify the first and second order moments of the noise (mean, variance) but the higher order moments or the pdf of the noise are also required.

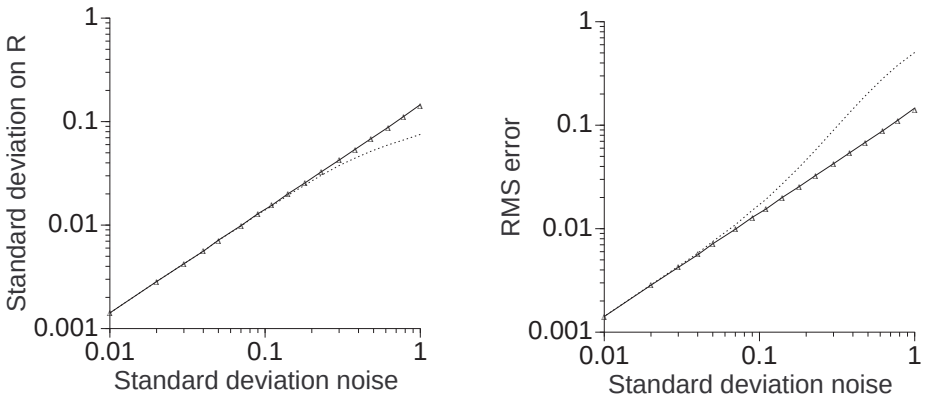


Figure 1.7: Evolution of the standard deviation and the RMS error on the estimated resistance value as a function of the standard deviation of the noise ($\sigma_u = \sigma_i$).
——— : $\hat{R}_{\text{EV}}(N)$, $\cdots\cdots$: $\hat{R}_{\text{LS}}(N)$, $\triangle\triangle\triangle$: theoretical value σ_R.

- Although $\hat{R}_{\text{LS}}(N)$ has a smaller variance than $\hat{R}_{\text{EV}}(N)$ for low SNR, its total root mean square (RMS) error (difference with respect to the true value) is significantly larger due to its systematic error. The following is

quite a typical observation: many estimators reduce the stochastic error at the cost of systematic errors. For the $\hat{R}_{\mathrm{EV}}$ the RMS error is completely due to the variability of the estimator since the RMS error coincides completely with the theoretical curve of the standard deviation.

1.2.3 Interpretation of the estimators: a cost function based approach

The previous section showed that there is not just one single estimator for each problem. Moreover, the properties of the estimators can vary quite a lot. This raises two questions: how can we generate good estimators and how can we evaluate their properties? The answers are given in this and the following sections. In order to recognize good estimators it is necessary to specify what a good estimator is. This is done in the next section. First we will deal with the question of how estimators are generated. Again there exist different approaches. A first group of methods starts from a deterministic approach. A typical example is the observation that the noiseless data should obey some model equations. The system parameters are then extracted by intelligent manipulation of these equations, usually inspired by numerical or algebraic techniques. Next, the same procedure is used on noisy data. The major disadvantage of this approach is that it does not guarantee, at all, that the resulting estimator has a good noise behavior. The estimates can be extremely sensitive to disturbing noise. The alternative is to embed the problem in a stochastic framework. A typical question to be answered is: where does the disturbing noise sneak into my problem and how does it behave? To answer this question, it is necessary to make a careful analysis of the measurement setup. Next, the best parameters are selected using statistical considerations. In most cases these methods lead to a cost function interpretation and the estimates are found as the arguments that minimize the cost function. The estimates of the previous section can be found as the minimizers of the following cost functions:

- $\hat{R}_{\mathrm{SA}}(N)$: Consider the successive resistance estimates $R(k) = u(k)/i(k)$. The overall estimate after N measurements is then the argument minimizing the following cost function:

$$\hat{R}_{\mathrm{SA}}(N) = \arg\min_{R} V_{\mathrm{SA}}(R, N) \text{ with } V_{\mathrm{SA}}(R, N) = \sum_{k=1}^{N} \left(R(k) - R\right)^2 . \tag{1.23}$$

  This is the most simple approach ("SA" stands for simple approach) of the estimation problem. As seen before it has very poor properties.

- $\hat{R}_{\mathrm{LS}}(N)$: A second possibility is to minimize the equation errors in the model equation $u(k) - Ri(k) = e(k, R)$ in least squares (LS) sense. For noiseless measurements $e(k, R_0) = 0$, with R_0 the true resistance value.

$$\hat{R}_{\mathrm{LS}}(N) = \arg\min_{R} V_{\mathrm{LS}}(R, N) \text{ with } V_{\mathrm{LS}}(R, N) = \sum_{k=1}^{N} e^2(k, R). \tag{1.24}$$

- $\hat{R}_{\mathrm{EV}}(N)$: The basic idea of the last approach is to express that the current as well as the voltage measurements are disturbed by noise. This is called the errors-in-variables (EV) approach. The idea is to estimate the exact current and voltage (i_0, u_0), parametrised as $(i_{\mathrm{p}}, u_{\mathrm{p}})$ keeping in mind the model equation $u_0 = Ri_0$.

$$\hat{R}_{\mathrm{EV}}(N) = \arg\min_{R, i_{\mathrm{p}}, u_{\mathrm{p}}} V_{\mathrm{EV}}\left(R, i_p, u_p, N\right) \text{ subject to } u_{\mathrm{p}} = Ri_{\mathrm{p}} \tag{1.25}$$

$$\text{with } V_{\mathrm{EV}}\left(R, i_p, u_p, N\right) = \sum_{k=1}^{N} \left(u(k) - u_{\mathrm{p}}\right)^2 + \sum_{k=1}^{N} \left(i(k) - i_{\mathrm{p}}\right)^2 \tag{1.26}$$

This wide variety of possible solutions and motivations illustrates very well the need for a more systematic approach. In this book we put the emphasis on a stochastic embedding approach, selecting a cost function on the basis of a noise analysis of the general measurement setup that is used.

All the cost functions that we presented are of the 'least squares' type. Again there exist many other possibilities, for example, the sum of the absolute

values. There are two reasons for choosing for a quadratic cost: firstly it is easier to minimize than other functions, and secondly we will show that normally distributed disturbing noise leads to a quadratic criterion. This does not imply that it is the best choice from all points of view. If it is known that some outliers in the measurements can appear (due to exceptionally large errors, a temporary sensor failure or a transmission error, etc.), it can be better to select a least absolute values cost function (sum of the absolute values) because these outliers are strongly emphasized in a least squares concept (Huber, 1981; Van den Bos, 1985). Sometimes a mixed criterion is used, e.g. the small errors are quadratically weighted while the large errors only appear linear in the cost to reduce the impact of outliers (Ljung, 1995).

1.3 Description of the stochastic asymptotic behavior of estimators

Since the estimates are obtained as a function of a finite number of noisy measurements, they are stochastic variables as well. Their pdf is needed in order to characterize them completely. However, in practice it is usually very hard to derive it, so that the behavior of the estimates is described by a few numbers only, such as their mean value (as a description of the location) and the covariance matrix (to describe the dispersion). Both aspects are discussed below.

1.3.1 Location properties: unbiased and consistent estimates

The choice for the mean value is not obvious at all from a theoretical point of view. Other location parameters like the median or the mode (Stuart and Ord, 1987) could be used too, but the latter are much more difficult to analyze in most cases. Since it can be shown that many estimates are asymptotically normally distributed under weak conditions, this choice is not so important

because in the normal case, these location parameters coincide. It seems very natural to require that the mean value equals the true value, but it turns out to be impractical. What are the true parameters of a system? We can only speak about true parameters if an exact model exists. It is clear that this is a purely imaginary situation since in practice we always stumble on model errors so that only excitation dependent approximations can be made. For theoretical reasons it still makes sense to consider the concept of "true parameters", but it is clear at this point that we have to generalize to more realistic situations. One possible generalization is to consider the estimator evaluated in the noiseless situation as the "best" approximation. These parameters are then used as reference value to compare the results obtained from noisy measurements. The goal is then to remove the influence of the disturbing noise so that the estimator converges to this reference value.

Definition 1.3.1: unbiasedness
An estimator $\hat{\theta}$ of the parameters θ_0 is unbiased if $\mathrm{E}\left\{\hat{\theta}\right\} = \theta_0$, for all true parameters θ_0. Otherwise it is a biased estimator.

If the expected value only equals the true value for an infinite number of measurements, then the estimator is called asymptotically unbiased. In practice it turns out that (asymptotic) unbiasedness is a hard requirement to deal with.

Example 1.3.1: Unbiased and biased estimators
At the end of their experiments the students wanted to estimate the value of the voltage over the resistor. Starting from the measurements (1.4), they first carry out a noise analysis of their measurements by calculating the sample mean value and the sample variance:

$$\hat{u}(N) = \frac{1}{N}\sum_{k=1}^{N} u(k) \text{ and } \hat{\sigma}_u^2(N) = \frac{1}{N}\sum_{k=1}^{N} \left(u(k) - \hat{u}(N)\right)^2 . \tag{1.27}$$

Applying the previous definition it is readily seen that

$$\mathrm{E}\{\hat{u}(N)\} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{E}\{u(k)\} = \frac{1}{N}\sum_{k=1}^{N}u_0 = u_0, \tag{1.28}$$

since the noise is zero mean, so that their voltage estimate is unbiased. The same can be done for the variance estimate:

$$\mathrm{E}\left\{\hat{\sigma}_u^2(N)\right\} = \frac{N-1}{N}\sigma_u^2. \tag{1.29}$$

This estimator shows a systematic error of σ_u^2/N and is thus biased. However, as $N \to \infty$ the bias disappears and following the definitions it is asymptotically unbiased. It is clear that a better estimate would be $\frac{1}{N-1}\sum_{k=1}^{N}\left(u(k)-\hat{u}(N)\right)^2$ which is the expression that is found in the handbooks on statistics.

For many estimators, it is very difficult or even impossible to find the expected value analytically. Sometimes it does not even exist as it was the case for $\hat{R}_{\mathrm{SA}}(N)$ of group A. Moreover, unbiased estimators can still have a bad distribution, e.g. the pdf of the estimator is symmetrically distributed around its mean value, with a minimum at the mean value. Consequently, a more handy tool (e.g. consistency) is needed.

Definition 1.3.2: consistency
An estimator $\hat{\theta}(N)$ of the parameters θ_0 is weakly consistent, if it converges in probability to θ_0: $\mathrm{plim}_{N\to\infty}\hat{\theta}(N) = \theta_0$, and strongly consistent if it converges with probability one (almost surely) to θ_0: $\mathrm{a.\,s.}\lim_{N\to\infty}\hat{\theta}(N) = \theta_0$.

Loosely spoken it means that the pdf of $\hat{\theta}(N)$ contracts around the true value θ_0, or $\lim_{N\to\infty}\mathrm{P}\left(\left|\hat{\theta}(N)-\theta_0\right| > \delta > 0\right) = 0$. Other convergence concepts exists, but these will not be discussed here. For the interested reader a short summary is given in (1.A) till (1.C). The major advantage of the consistency concept is purely mathematical: it is much easier to prove consistency than unbiasedness using probabilistic theories starting from the cost function interpretation. Another nice property of the plim is that it can be interchanged with

a continuous function: $\text{plim} f(a) = f(\text{plim}(a))$ if both limits exist. In fact it was this property that we applied during the calculations of the limit values of $\hat{R}_{\text{LS}}$ and $\hat{R}_{\text{EV}}$, for example

$$\begin{aligned} \operatorname*{plim}_{N\to\infty} \hat{R}_{\text{EV}}(N) &= \operatorname*{plim}_{N\to\infty} \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)}{\frac{1}{N}\sum_{k=1}^{N} i(k)} && (1.30)\\ &= \frac{\text{plim}_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} u(k)}{\text{plim}_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} i(k)} && (1.31)\\ &= \frac{u_0}{i_0} = R_0. && (1.32) \end{aligned}$$

Consequently $\hat{R}_{\text{EV}}(N)$ is a weakly consistent estimator. Calculating the expected value is much more involved in this case due to the division. Therefore, consistency is a better suited concept than (asymptotic) unbiasedness to study it.

1.3.2 Dispersion properties: efficient estimators

In this book the covariance matrix is used to measure the dispersion of an estimator, i.e. to ascertain how much the actual estimator is scattered around its limiting value? Again this choice, among other possibilities (like, for example, percentiles), is highly motivated from a mathematical point of view. Within the stochastic framework used it will be quite easy to calculate the covariance matrix while it is much more involved to obtain the other measures. For normal distributions, all dispersion measures are obtainable from the covariance matrix so that for most estimators this choice is not too restrictive because their distribution converges to a normal one.

As users we are highly interested in estimators with minimal errors. However, since we can collect only a finite number of noisy measurements it is clear that there are limits on the accuracy and precision we can reach. This is precisely quantified in the Cramér-Rao inequality. This inequality provides a lower bound on the covariance matrix of a(n) (un)biased estimator starting from the likelihood function. First we introduce the likelihood function, next we present

the Cramér-Rao lower bound.

Consider the measurements $z \in \mathbb{R}^N$ obtained from a system described by a hypothetical, exact model that is parameterised in θ. These measurements are disturbed by noise and are hence stochastic variables that are characterized by a probability density function $f(z|\,\theta_0)$ that depends on the exact model parameters θ_0 with $\int_{z\in\mathbb{R}^N} f(z|\,\theta_0)\,\mathrm{d}\,z = 1$. Next we can interpret this relation conversely, viz.: how likely is it that a specific set of measurements $z = z_m$ are generated by a system with parameters θ? In other words, we consider now a given set of measurements and view the model parameters as the free variables:

$$L\,(z_m|\,\theta) = f\,(z = z_m|\,\theta) \tag{1.33}$$

with θ the free variables. $L\,(z_m|\,\theta)$ is called the likelihood function. In many calculations the log likelihood function $l(z|\,\theta) = \ln L\,(z|\,\theta)$ is used. In 1.33 we used z_m to indicate explicitly that we use the numerical values of the measurements that were obtained from the experiments. From here on we just use z as a symbol because it will be clear from the context what interpretation should be given to z. The reader should be aware that $L(z|\,\theta)$ is not a probability density function with respect to θ since $\int_\theta L(z|\,\theta)\,\mathrm{d}\,\theta \neq 1$. Notice the subtle difference in terminology, i.e. probability is replaced by likeliness.

The Cramér-Rao lower bound gives a lower limit on the covariance matrix of parameters.

Under quite general conditions (see 1.D), this limit is universal and independent of the selected estimator: no estimator that violates this bound can be found. It is given by

$$\mathrm{CR}\,(\theta_0) = \left(\mathrm{I}_{n_\theta} + \frac{\partial b_\theta}{\partial \theta}\right)^{\mathrm{T}} \mathrm{Fi}^{-1}\,(\theta_0)\left(\mathrm{I}_{n_\theta} + \frac{\partial b_\theta}{\partial \theta}\right) \tag{1.34}$$

$$\mathrm{Fi}\,(\theta_0) = \mathrm{E}\left\{\left(\frac{\partial l\,(z|\,\theta)}{\partial \theta}\right)^{\mathrm{T}}\left(\frac{\partial l\,(z|\,\theta)}{\partial \theta}\right)\right\} = -\,\mathrm{E}\left\{\frac{\partial^2 l\,(z|\,\theta)}{\partial \theta^2}\right\} \tag{1.35}$$

The derivatives are calculated in $\theta = \theta_0$, and $b_\theta = \mathrm{E}\left\{\hat{\theta}\right\} - \theta_0$ is the bias on the estimator. For unbiased estimators (1.34) reduces to

$$\mathrm{CR}(\theta_0) = \mathrm{Fi}^{-1}(\theta_0). \tag{1.36}$$

$\mathrm{Fi}(\theta)$ is called the Fisher information matrix: it is a measure for the information in an experiment: the larger the matrix the more information there is. In (1.35) it is assumed that the first and second derivatives of the log likelihood function exist with respect to θ.

Example 1.3.2: Influence of the number of parameters on the Cramér-Rao lower bound

A group of students wanted to determine the flow of tap water by measuring the height $h_0(t)$ of the water in a measuring jug as a function of time t. However, their work was not precise and in the end they were not sure about the exact starting time of their experiment. They included it in the model as an additional parameter: $h_0(t) = a(t - t_{\mathrm{start}}) = at + b$, and $\theta = [a, b]^{\mathrm{T}}$. Assume that the noise $n_h(k)$ on the height measurements is iid zero mean normally distributed $N(0, \sigma^2)$, and the noise on the time instances is negligible $h(k) = at_k + b + n_h(k)$, then the following stochastic model can be used

$$\mathrm{P}\left(h(k), t_k\right) = \mathrm{P}\left(h(k) - (at_k + b)\right) = \mathrm{P}\left(n_h(k)\right)$$

where $\mathrm{P}\left(h(k), t_k\right)$ is the probability to make the measurements $h(k)$ at t_k. The likelihood function for the set of measurements $h = \{(h(1), t_1), ..., (h(N), t_N)\}$ is

$$L(\,h|\,a, b) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(h(k) - at_k - b)^2}, \tag{1.37}$$

and the loglikelihood function becomes

$$l(\,h|\,a, b) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}\left(h(k) - at_k - b\right)^2. \tag{1.38}$$

The Fisher information matrix and the Cramér-Rao lower bound are found using (1.35):

$$\mathrm{Fi}(a,b) = \frac{N}{\sigma^2}\begin{bmatrix} s^2 & \mu \\ \mu & 1 \end{bmatrix} \tag{1.39}$$

$$\Downarrow$$

$$\mathrm{CR}(a,b) = \mathrm{Fi}^{-1}(a,b) = \frac{\sigma^2}{N(s^2-\mu^2)}\begin{bmatrix} 1 & -\mu \\ -\mu & s^2 \end{bmatrix} \tag{1.40}$$

with $\mu = \frac{1}{N}\sum_{k=1}^{N} t_k$ and $s^2 = \frac{1}{N}\sum_{k=1}^{N} t_k^2$. These expressions are very informative. First of all we can note that the attainable uncertainty is proportional to the standard deviation of the noise. This means that inaccurate measurements result in poor estimates, or identification is no excuse for sloppy measurements. The uncertainty decreases as $\sqrt{N}$, which can be used as a rule of thumb whenever independent measurements are processed. Finally it can also be noted that the uncertainty depends on the actual time instances used in the experiment. In other words, by making a proper design of the experiment, it is possible to influence the uncertainty on the estimates. Another question we can answer now is what price is paid to include the additional model parameter b to account for the unknown starting time. By comparing $\mathrm{Fi}^{-1}(a,b)$ to $\mathrm{Fi}^{-1}(a)$ (assuming that b is known) it is found that

$$\sigma_a^2(a,b) = \frac{\sigma^2}{N(s^2-\mu^2)} \geq \frac{\sigma^2}{Ns^2} = \sigma_a^2(a) \tag{1.41}$$

where $\sigma_a^2(a,b)$ is the lower bound on the variance of a if both parameters are estimated, else $\sigma_a^2(a)$ is the lower bound if only a is estimated. This shows that adding additional parameters to a model increases the minimum attainable uncertainty on it. Of course these parameters may be needed to remove systematic errors so that a balance between stochastic errors and systematic errors is achieved. This is further elaborated in the chapter on model validation.

The Cramér-Rao lower bound is a conservative estimate of the smallest possible covariance matrix that is not always attainable (the values may be too small). Tighter bounds exist (Abel, 1993) but these are more involved to calculate. Consequently, the Cramér-Rao bound is the most used criterion to verify the efficiency of an estimator.

Definition 1.3.3: efficiency
An unbiased estimator is called efficient if its covariance matrix is smaller than that of any other unbiased estimator. An unbiased estimator that reaches the Cramér-Rao lower bound is also an efficient estimator.

1.4 Basic steps in the identification process

Each identification session consists of a series of basic steps. Some of them may be hidden or selected without the user being aware of his choice. Clearly, this can result in poor or sub optimal results. In each session the following actions should be taken:

- Collect information about the system;
- Select a model structure to represent the system;
- Choose the model parameters to fit the model as well as possible to the measurements: selection of a "goodness of fit" criterion;
- Validate the selected model.

Each of these points is discussed in more detail below.

1.4.1 Collect information about the system

If we want to build a model for a system we should get information about it. This can be done by just watching the natural fluctuations (e.g. vibration analysis of a bridge that is excited by normal traffic), but most often it is more efficient to set up dedicated experiments that actively excite the system (e.g. controlled

excitation of a mechanical structure using a shaker). In the latter case the user has to select an excitation that optimizes his own goal (for example, minimum cost, minimum time or minimum power consumption for a given measurement accuracy) within the operator constraints (e.g. the excitation should remain below a maximum allowable level). The quality of the final result can heavily depend on the choices that are made.

1.4.2 Select a model structure to represent the system

A choice should be made within all the possible mathematical models that can be used to represent the system. Again a wide variety of possibilities exist such as

Parametric versus nonparametric models: In a parametric model, the system is described using a limited number of characteristic quantities called the parameters of the model, while in a nonparametric model the system is characterized by measurements of a system function at a large number of points. Examples of parametric models are the transfer function of a filter described by its poles and zeros, the motion equations of a piston, etc. An example of a nonparametric model is the description of a filter by its impulse response at a large number of points.

Usually it is simpler to create a non-parametric model than a parametric one because the modeler needs less knowledge about the system itself in the former case. However, physical insight and concentration of information is more substantial for parametric models than for nonparametric ones. In this book we will concentrate on transfer function models (parametric models), but also the problem of frequency response function measurements (nonparametric model) will be elaborated.

White box models versus black box models: In the construction of a model, physical laws whose availability and applicability depend on the insight

and skills of the experimenter can be used (Kirchhoff's laws, Newton's laws, etc.). Specialized knowledge relating to different scientific fields may be brought into this phase of the identification process. The modelling of a loudspeaker, for example, requires extensive understanding of mechanical, electrical and acoustical phenomena. The result may be a physical model, based on comprehensive knowledge of the internal functioning of the system. Such a model is called a white box model.

Another approach is to extract a black box model from the data. Instead of making a detailed study, and developing a model based upon physical insight and knowledge, a mathematical model is proposed which allows sufficient description of any observed input and output measurements. This reduces the modelling effort significantly. For example, instead of modelling the loudspeaker using physical laws, an input-output relation, taking the form of a high-order transfer function, could be proposed.

The choice between the different methods depends on the aim of the study: the white box approach is better for gaining insight into the working principles of a system, but a black box model may be sufficient if the model will only be used for prediction of the output.

Although, as a rule of thumb, it is advisable to include as much prior knowledge as possible during the modelling process, it is not always easy to do so. If we know, for example, that a system is stable, it is not simple to express this information if the polynomial coefficients are used as parameters.

Linear models versus nonlinear models: In real life almost every system is nonlinear. Because the theory of nonlinear systems is very involved, these are mostly approximated by linear models, assuming that in the operation region the behavior can be linearised. This kind of approximation makes it possible to use simple models without jeopardizing properties which are of importance to the modeler. This choice depends strongly on the intended use of the model. For example, a nonlinear model is needed to describe the distortion of an amplifier,

but a linear model will be sufficient to represent its transfer characteristics if the linear behavior is dominant and is of the only interest.

Linear-in-the-parameters versus nonlinear-in-the-parameters: A model is called linear-in-the-parameters if there exists a linear relation between these parameters and the error that is minimized. This does not imply that the system itself is linear. For example $\varepsilon = y - (a_1 u + a_2 u^2)$ is linear in the parameters a_1 and a_2 but describes a non-linear system. On the other hand

$$\varepsilon(j\omega) = Y(j\omega) - \frac{a_0 + a_1 j\omega}{b_0 + b_1 j\omega} U(j\omega)$$

describes a linear system but the model is non-linear in the b_0 and b_1 parameters. Linearity in the parameters is a very important aspect of models since it has a strong impact on the complexity of the estimators if a (weighted) least squares cost function is used. In that case the problem can be solved analytically for models that are linear in the parameters so that an iterative optimization problem is avoided.

1.4.3 Match the selected model structure to the measurements

Once a model structure is chosen (e.g. a parametric transfer function model) it should be matched as well as possible with the available information about the system. Mostly, this is done by minimizing a criterion that measures a goodness of the fit. The choice of this criterion is extremely important since it determines the stochastic properties of the final estimator. As seen from the resistance example, many choices are possible and each of them can lead to a different estimator with its own properties. Usually, the cost function defines a distance between the experimental data and the model. The cost function can be chosen on an ad hoc basis using intuitive insight, but there exists also a more systematic approach based on stochastic arguments as explained in Chapter 2. Simple tests

on the cost function exist (necessary conditions) to check even before deriving the estimator if it can be consistent, but this outside the scope of this book.

1.4.4 Validate the selected model

Finally, the validity of the selected model should be tested: does this model describe the available data properly or are there still indications that some of the data are not well modelled, indicating remaining model errors? In practice the best model (= the smallest errors) is not always preferred. Often a simpler model that describes the system within user-specified error bounds is preferred. Tools will be provided that guide the user through this process by separating the remaining errors into different classes, for example unmodelled linear dynamics and non-linear distortions. From this information further improvements of the model can be proposed, if necessary.

During the validation tests it is always important to keep the application in mind. The model should be tested under the same conditions as it will be used later. Extrapolation should be avoided as much as possible. The application also determines what properties are critical.

1.4.5 Conclusion

This brief overview of the identification process shows that it is a complex task with a number of interacting choices. It is important to pay attention to all aspects of this procedure, from the experiment design till the model validation, in order to get the best results. The reader should be aware that besides this list of actions other aspects are also important. A short inspection of the measurement setup can reveal important shortcomings that can jeopardize a lot of information. Good understanding of the intended applications helps to setup good experiments and is very important to make the proper simplifications during the model building process. Many times, choices are made that are not based on complicated theories but are dictated by the practical circumstances. In these cases a good theoretical understanding of the applied methods will help

the user to be aware of the sensitive aspects of his techniques. This will enable him to put all his effort on the most critical decisions. Moreover, he will become aware of the weak points of the final model.

Appendices

1.A Definitions of stochastic limits

Let $x(N)$, $N = 1, 2, \ldots$ be a scalar random sequence. There are several ways in which the sequence might converge to a (random) number x as $N \to \infty$. We will define four modes of stochastic convergence.

Convergence 1: in mean square

The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *in mean square* if, $\mathrm{E}\left\{|x|^2\right\} < \infty$, $\mathrm{E}\left\{|x(N)|^2\right\} < \infty$ for all N, and $\lim_{N\to\infty} \mathrm{E}\left\{|x(N) - x|^2\right\} = 0$. We write

$$\operatorname*{l.i.m.}_{N\to\infty} x(N) = x \quad \Leftrightarrow \quad \lim_{N\to\infty} \mathrm{E}\left\{|x(N) - x|^2\right\} = 0 \tag{1.42}$$

Convergence 2: with probability 1

The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *with probability 1* (w.p. 1) or *almost surely* if, $\lim_{N\to\infty} x^{[\omega]}(N) = x^{[\omega]}$ for almost all realizations ω, except those $\omega \in A$ such that $\mathrm{P}\left(() A\right) = 0$. We write

$$\operatorname*{a.\,s.\,lim}_{N\to\infty} x(N) = x \quad \Leftrightarrow \quad \mathrm{P}\left(\lim_{N\to\infty} x(N) = x\right) = 1 \tag{1.43}$$

This definition is equivalent to (Theorem 2.1.2 of Lukacs, 1975)

$$\operatorname*{a.\,s.\,lim}_{N\to\infty} x(N) = x \quad \Leftrightarrow \quad \forall \varepsilon > 0\colon \lim_{N\to\infty} \mathrm{P}\left(\sup_{k\geq N} \{|x(k) - x|\} \leq \varepsilon\right) = 1 \tag{1.44}$$

Convergence 3: in probability

The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *in probability* if, for every $\varepsilon, \delta > 0$ there exists a N_0 such that for every $N > N_0$: $\mathrm{P}\left(|x(N) - x| \leq \varepsilon\right) > 1 - \delta$.

We write

$$\operatorname*{plim}_{N\to\infty} x(N) = x \quad \Leftrightarrow \quad \forall \varepsilon > 0 \colon \lim_{N\to\infty} \mathrm{P}\left(|x(N) - x| \leq \varepsilon\right) = 1 \tag{1.45}$$

Convergence 4: in distribution

Let $F_N(x)$ and $F(x)$ be the distribution functions of, respectively, $x(N)$ and x. The sequence $x(N)$, $N = 1, 2, \ldots$ converges to x *in law* or *in distribution* if $F_N(x)$ converges weakly[1] to $F(x)$. We write

$$\operatorname*{Lim}_{N\to\infty} x(N) = x \quad \Leftrightarrow \quad \operatorname*{Lim}_{N\to\infty} F_N(x) = F(x)$$

1.B Interrelations between stochastic limits

In the previous section we defined several modes of stochastic convergence. The connections between these concepts are

Interrelation 1:

Almost sure convergence implies convergence in probability, the converse is not true (Theorem 2.2.1 of Lukacs, 1975)

Interrelation 2:

Convergence in mean square implies convergence in probability, the converse is not true (Theorem 2.2.2 of Lukacs, 1975)

Interrelation 3:

Convergence in probability implies convergence in law, the converse is not true (Theorem 2.2.3 of Lukacs, 1975)

Interrelation 4:

There is no implication between almost sure and mean square convergence

Interrelation 5:

a sequence $x(N)$ converges in probability to x if and only if every subsequence

[1] This means at all continuity points of the limiting function and is denoted by "*Lim*"

$x(N_k)$ contains a sub-subsequence $x(N_{k_i})$ which converges ($i \to \infty$) almost surely to x (Theorem 2.4.4 of Lukacs, 1975)

Interrelation 6:

A sequence converges in probability to a constant if and only if it converges in law to a degenerate distribution[2] (Corollary to Theorem 2.2.3 of Lukacs, 1975)

A graphical representation of the convergence area of the different stochastic limits is given in Figure1.8. The interrelations between the concepts are summarized in Figure1.9. Since these allow a better understanding of the stochastic limits, some proofs are given in appendix. The importance of Interrelation 5 is that any theorem proven for the almost sure limit is also valid for the limit in probability. Before illustrating some of the interrelations by (counter) examples, we cite the Borel-Cantelli and the Fréchet-Shohat lemmas which are useful to establish, respectively, convergence w.p. 1 and convergence in distribution. The Borel-Cantelli lemma roughly says that if the convergence in probability or in mean square is sufficiently fast, then this implies convergence with probability 1.

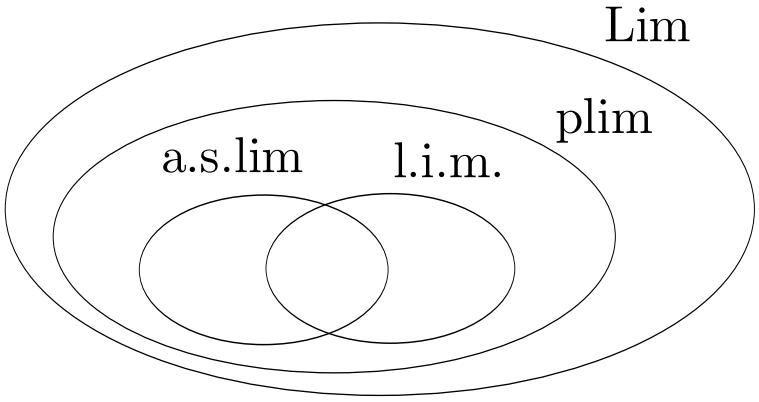


Figure 1.8: Convergence area of the stochastic limits.

[2] $F(x)$ is degenerate if there exists a x_0 such that $F(x) = 0$ for $x < x_0$ and $F(x) = 1$ for $x \geq x_0$.

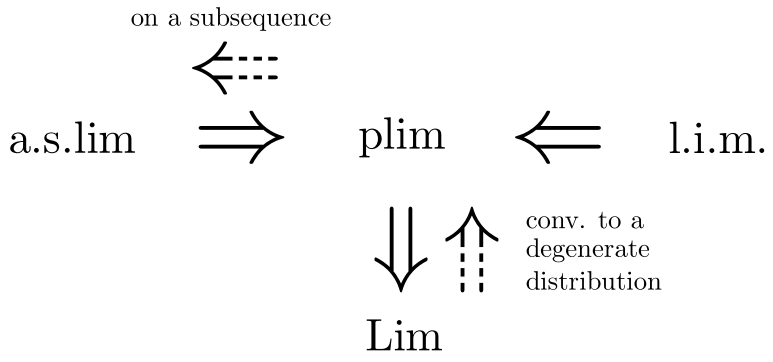


Figure 1.9: Interrelations between the stochastic limits.

Lemma 1: Borel-Cantelli

if

$$\sum_{N=1}^{\infty} \mathrm{P}\left(|x(N)-x|>\varepsilon\right)<\infty \text{ or } \sum_{N=1}^{\infty} \mathrm{E}\left\{|x(N)-x|^2\right\}<\infty \tag{1.46}$$

then $x(N)$ converges to x w.p. 1.

Lemma 2: Fréchet-Shohat

let x have a distribution function $F(x)$ that is uniquely determined by its moments (cumulants). If the moments (cumulants) of the sequence $x(N)$ converge for $N \to \infty$ to the moments (cumulants) of x, then $x(N)$ converges in distribution to x.

Example 1.B.1:

Convergence w.p. 1 and convergence in probability do not imply convergence in mean square (Example 2.1.1 of Stout). Take ω to be uniform in $[0,1]$, and build the sequence $x(N)$ such that

$$x^{[\omega]}(N)=\begin{cases} N & \forall\,\omega\in\left[0,\frac{1}{N}\right) \\ 0 & \forall\,\omega\in\left[\frac{1}{N},1\right] \end{cases}$$

Two realizations of the sequence are, for example,

$$\left\{x^{[0.3]}(N)\right\}=\{1,2,3,0,0,0,0,0,\ldots\}$$
$$\left\{x^{[0.15]}(N)\right\}=\{1,2,3,4,5,6,0,0,\ldots\}$$

We see that $x^{[\omega]}(N)$ is zero for N sufficiently large, which suggests that it will converge to zero. Formally, $\text{plim}_{N\to\infty} x(N) = \text{a. s. lim}_{N\to\infty}\ x(N) = 0$ since

$$\mathrm{P}\left(\sup_{k\geq N}|x(k)|\leq\varepsilon\right)=\mathrm{P}\left(|x(N)|\leq\varepsilon\right)=\mathrm{P}\left(x(N)=0\right)=1-\frac{1}{N}$$

is arbitrarily close to 1 for N sufficiently large. There is just one sequence, $x^{[0]}(N)$, which does not converge. This is not in contradiction with the previous results because the probability to get this particular realization is zero: $\mathrm{P}(\omega = 0) = 0$. The mean square limit $\text{l. i. m.}_{N\to\infty} x(N)$ does not exist since $\mathrm{E}\left\{x^2(N)\right\} = N$ is unbounded. Note that the Borel-Cantelli lemma cannot be used in this example to establish the almost sure convergence from the convergence in probability. Indeed,

$$\sum_{N=1}^{\infty}\mathrm{P}\left(|x(N)|>\varepsilon\right)=\sum_{N=1}^{\infty}\frac{1}{N}=\infty.$$

Example 1.B.2: Convergence in probability and convergence in mean square do not imply convergence w.p. 1 (Example 2.1.2 of Stout, 1974)

Take ω to be uniform in $[0, 1)$, and build the sequence $T(n, k)$ such that

$$T^{[\omega]}(n,k)=\begin{cases}1 & \forall\,\omega\in\left[\frac{k-1}{n},\frac{k}{n}\right)\\ 0 & \text{elsewhere}\end{cases}$$

for $k = 1, 2, \ldots, n$ and $n \geq 1$. Let

$$\{x(N)\}=\{\{T(1,k)\},\{T(2,k)\},\{T(3,k)\},\ldots\}$$

with $\{T(n,k)\} = \{T(n,1), T(n,2), \ldots, T(n,n)\}$ and $N = \frac{n(n-1)}{2} + k$. Two realizations of the sequence are, for example,

$$\begin{aligned}\left\{x^{[0.27]}(N)\right\}&=\{\{1\},\{1,0\},\{1,0,0\},\{0,1,0,0\},\ldots\}\\ \left\{x^{[0.85]}(N)\right\}&=\{\{1\},\{0,1\},\{0,0,1\},\{0,0,0,1\},\ldots\}\end{aligned}$$

We see that the length of each subsequence $\{T(n,k)\}$ of $\{x(N)\}$ increases with n and that it contains exactly one non-zero term. This suggests that $x(N)$ will converge in probability (the probability to get a 1 goes to zero), but not w.p. 1 (the supremum is 1 for any value of N). Formally, $\text{plim}_{N\to\infty} x(N) = 0$ since

$$\lim_{N\to\infty} \mathrm{P}\left(|x(N)| \leq \varepsilon\right) = \lim_{N\to\infty} \mathrm{P}\left(T(n,k) = 0\right) = \lim_{N\to\infty} \left(1 - \frac{1}{n}\right) = 1$$

and $\text{l. i. m.}_{N\to\infty} x(N) = 0$ because

$$\lim_{N\to\infty} \mathrm{E}\left\{x^2(N)\right\} = \lim_{N\to\infty} \mathrm{E}\left\{T^2(n,k)\right\} = \lim_{N\to\infty} \frac{1}{n} = 0$$

The almost sure limit $\text{a. s. lim}_{N\to\infty} x(N)$ does not exist since $\mathrm{P}\left(\sup_{r\geq N}\{|x(r)|\} > \varepsilon\right) = 1$. Note that the subsequence $T(n,k)$, with k fixed and $n \geq 1$, converges w.p. 1 to zero. This is an illustration of interrelation 5

Example 1.B.3: Convergence in mean square and convergence w.p. 1 are compatible (Example 2.2.3 of Lukacs, 1975)

Let $x(N)$ be a random variable which assumes only the values $\frac{1}{N}$ and $-\frac{1}{N}$ with equal probability. We find $\text{l. i. m.}_{N\to\infty} \mathrm{E}\left\{x^2(N)\right\} = 0$ since

$$\lim_{N\to\infty} \mathrm{E}\left\{x^2(N)\right\} = \lim_{N\to\infty} \frac{1}{N^2} = 0$$

Also $\text{a. s. lim}_{N\to\infty} x(N) = 0$ because $|x(k)| < |x(N)|$ for any $k > N$ so that $\mathrm{P}\left(\sup_{k\geq N}\{|x(k)| \leq \varepsilon\}\right) = \mathrm{P}\left(|x(N)| \leq \varepsilon\right)|_{N>1/\varepsilon} = 1$

Example 1.B.4: Convergence in distribution does not imply convergence in probability (Example 2.2.4 of Lukacs, 1975)

Let x be a random variable that can only take the values 0 and 1 with equal probability. Next, construct the sequence $x(N) = 1 - x$. We have $\text{Lim}_{N\to\infty} x(N) = x$ since $x(N)$ and x have the same distribution functions $F_N(x) = F(x)$. However, the limit in probability $\text{plim}_{N\to\infty} x(N)$ does not exist since $|x(N) - x| = 1$ so that $\mathrm{P}\left(|x(N) - x| \leq \varepsilon\right) = 0$.

1.C Properties of stochastic limits

The properties of the stochastic limits are similar to those of the classical (deterministic) limit, but there are some subtle differences. The general properties are

Property 1:

A continuous function and the almost sure limit may be interchanged

$$\text{a. s. lim}_{N\to\infty} f\left(x(N)\right) = f(x) \text{ with } x = \text{a. s. lim}_{N\to\infty} x(N) \tag{1.47}$$

Property 2:

The almost sure limit and the expected value may be interchanged for uniformly bounded sequences (Theorem 5.4 of Billingsley, 1995)

$$\lim_{N\to\infty} \mathrm{E}\left\{x(N)\right\} = \mathrm{E}\left\{\text{a. s. } \lim_{N\to\infty} x(N)\right\} \tag{1.48}$$

A direct consequence of Property 2 is that

$$\mathrm{E}\left\{O_{\text{a.s.}}\left(N^{-k}\right)\right\} = O\left(N^{-k}\right) \tag{1.49}$$

Property 3:

A continuous function and the limit in probability may be interchanged (Theorem 2.3.3 of Lukacs, 1975)

$$\operatorname*{plim}_{N\to\infty} f\left(x(N)\right) = f(x) \text{ with } x = \operatorname*{plim}_{N\to\infty} x(N) \tag{1.50}$$

Property 4:

The limit in probability and the expected value may be interchanged for uniformly bounded sequences (Theorem 5.4 of Billingsley, 1995)

$$\lim_{N\to\infty} \mathrm{E}\left\{x(N)\right\} = \mathrm{E}\left\{\operatorname*{plim}_{N\to\infty} x(N)\right\} \tag{1.51}$$

A direct consequence of Property4 is that

$$\mathrm{E}\left\{O_{\mathrm{p}}(N^{-k})\right\} = O(N^{-k}) \tag{1.52}$$

Property 5:

The mean square limit is linear (Theorem 3.1 of Jazwinski, 1970)

$$\underset{N\to\infty}{\mathrm{l.\,i.\,m.}}\,(ax(N) + by(N)) = a\ \underset{N\to\infty}{\mathrm{l.\,i.\,m.}}\ x(N) + b\ \underset{N\to\infty}{\mathrm{l.\,i.\,m.}}\ y(N) \tag{1.53}$$

where a and b are deterministic (non-random) numbers.

Property 6:

The mean square limit and the expected value may be interchanged (Theorem 3.1 of Jazwinski, 1970),

$$\lim_{N\to\infty} E\left\{x(N)\right\} = E\underset{N\to\infty}{\mathrm{l.\,i.\,m.}}\, x(N) \tag{1.54}$$

A direct consequence of 6 is that

$$\mathrm{E}\left\{O_{\mathrm{m.s.}}\left(N^{-k}\right)\right\} = O\left(N^{-k}\right) \tag{1.55}$$

Property 7:

if $\mathrm{l.\,i.\,m.}_{N\to\infty}\, x(N) = x$ and $E\left\{(x(N) - x)^2\right\} = O(N^{-k})$, with $k > 0$, then

$$x(N) = x + O_{\mathrm{m.s.}}(N^{-k/2}) \text{ and } x(N) = x + O_{\mathrm{p}}(N^{-k/2}) \tag{1.56}$$

This is a direct consequence of (1.55) and Interrelation 2.

Property 8:

If the sequence $x(n)$ is deterministic (non-random), then the limit in mean square, the limit w.p. 1 and the limit in probability reduce to the deterministic limits.

Property 1 follows directly from the definition of convergence 2 with proba-

bility one while property 3 follows from Interrelation 5 and property 1. Properties 1 and 3 require the continuity of the function at ALL values of the limit random variable x. If x is a constant (non-random), then continuity in a closed neighborhood of x is sufficient. Note that the limit in mean square and a continuous function may, in general, NOT be interchanged. Note also that the almost sure limit and the limit in probability, in general, do NOT commute with the expected value.

1.D Cramér-Rao lower bound

Consider the identification of the parameter vector $\theta \in \mathbb{R}^{n_\theta}$ using noisy measurements $z \in \mathbb{R}^N$. The quality of the estimator $\hat{\theta}(z)$ can be represented by its mean square error matrix

$$\mathrm{MSE}\left(\hat{\theta}(z)\right) = \mathrm{Cov}\left\{\hat{\theta}(z)\right\} + b_\theta b_\theta^T \tag{1.57}$$

where θ_0 and b_θ denote, respectively, the true value and the bias on the estimates. We may wonder whether there exists a lower limit on the value of the mean square error (1.57) that can be obtained with various estimators. The answer is given by the *generalized Cramér-Rao lower bound.*

Theorem 1: Generalized Cramér-Rao lower bound

Let $f_z(z, \theta_0)$ be the probability density function of the measurements $z \in \mathbb{R}^N$. Assume that $f_z(z, \theta_0)$ and its first and second order derivatives w.r.t. $\theta \in \mathbb{R}^{n_\theta}$ exist for all θ_0-values. Assume, furthermore, that the boundaries of the domain of $f_z(z, \theta_0)$ w.r.t. z are θ_0-independent. Then, the *generalized Cramér-Rao lower bound* on the mean square error of any estimator $\hat{G}(z)$ of the function $G(\theta) \in \mathbb{C}^r$ of θ is

$$\mathrm{MSE}\left(\hat{G}\left(\hat{\theta}\left(z\right)\right)\right) \geq \left(\frac{\partial G\left(\theta_0\right)}{\partial \theta_0} + \frac{\partial b_G}{\partial \theta_0}\right) \mathrm{Fi}^{+}\left(\theta_0\right) \left(\frac{\partial G\left(\theta_0\right)}{\partial \theta_0} + \frac{\partial b_G}{\partial \theta_0}\right)^{\mathrm{H}} + b_G b_G^{\mathrm{H}} \tag{1.58}$$

with $b_G = \mathrm{E}\left\{\hat{G}(z)\right\} - G(\theta_0)$ the bias that might be present in the estimate, and $\mathrm{Fi}(\theta_0)$ the *Fisher information matrix* of the parameters θ_0

$$\mathrm{Fi}(\theta_0) = \mathrm{E}\left\{\left(\frac{\partial \ln f_z(z,\theta_0)}{\partial \theta_0}\right)^{\mathrm{T}}\left(\frac{\partial \ln f_z(z,\theta_0)}{\partial \theta_0}\right)\right\} = -\mathrm{E}\left\{\frac{\partial^2 \ln f_z(z,\theta_0)}{\partial \theta_0^2}\right\} \tag{1.59}$$

Equality holds in (1.58) if and only if there exists a non-random matrix Γ such that

$$\hat{G}\left(\hat{\theta}(z)\right) - \mathrm{E}\left\{\hat{G}\left(\hat{\theta}(z)\right)\right\} = \Gamma\left(\frac{\partial \ln f_z(z,\theta_0)}{\partial \theta_0}\right)^{\mathrm{T}} \tag{1.60}$$

The expectations in (1.58) and (1.59) are taken w.r.t. the measurements z.

Note that the calculation of the Cramér-Rao lower bound requires knowledge of the true parameters θ_0 which is often not available (except in simulations). An approximation can be calculated by replacing θ_0 by its estimated value $\hat{\theta}$ in (1.58). Two special cases of the Cramér-Rao inequality are worthwhile mentioning.

If $G(\theta) = \theta$, $b_G = 0$ and $\mathrm{Fi}(\theta_0)$ is regular, then we obtain the *Cramér-Rao lower bound for unbiased estimators* (abbreviated as UCRB)

$$\mathrm{Cov}\left\{\hat{\theta}(\mathrm{z})\right\} \geq \mathrm{Fi}^{-1}(\theta_0) \tag{1.61}$$

If condition (1.60) is not satisfied, $\hat{\theta}(z) - \theta_0 \neq \Gamma\left(\frac{\partial \ln f_z(z,\theta_0)}{\partial \theta_0}\right)^{\mathrm{T}}$, then the lower bound (1.61) is too conservative, and there may still be an unbiased estimator which has smaller variance than any other unbiased estimator. Better (larger) bounds exist when (1.61) is not attainable, but they are often (extremely) difficult to compute. An overview of tighter bounds can be found in Abel (1993).

If $G(\theta) = \theta$, $b_G \neq 0$ and $Fi(\theta_0)$ is regular, then we find the *Cramér-Rao lower bound on the mean square error of biased estimators* (abbreviated as CRB)

$$\mathrm{MSE}\left(\hat{\theta}(z)\right) \geq \left(\mathrm{I}_{n_\theta} + \frac{\partial b_\theta}{\partial \theta_0}\right)\mathrm{Fi}^{-1}(\theta_0)\left(\mathrm{I}_{n_\theta} + \frac{\partial b_\theta}{\partial \theta_0}\right)^{\mathrm{T}} + b_\theta b_\theta^{\mathrm{T}} \tag{1.62}$$

It follows that the Cramér-Rao lower bound for asymptotically unbiased estima-

tors ($b_\theta \to 0$ as $N \to \infty$) is asymptotically given by (1.61), only if the derivative of the bias w.r.t. θ_0 is asymptotically zero. Likewise the unbiased case, the lower bound (1.62) may be too conservative and tighter bounds exist (Abel, 1993). Note that the first term in the right hand side of (1.62) can be zero for biased estimators.

In general, it is impossible to show that the bias (and its derivative w.r.t. θ) of a weakly or strongly consistent estimator converges to zero as $N \to \infty$. However, the moments of the limiting random variable often exist. The (asymptotic) covariance matrix or mean square error of the limiting random variable is then compared to the UCRB. In this context, the concept of asymptotic efficiency is also used for weakly or strongly consistent estimators.

Chapter 2

A statistical approach to the estimation problem

Chapter 2

A statistical approach to the estimation problem

In this chapter a systematic approach to the parameter estimation problem is made: what criterion should be used to match the model to the data? A statistical approach to select a criterion to measure the 'goodness' of the fit is made. Basic concepts of statistics such as the expected value, the covariance matrix, probability density functions are assumed to be known.

2.1 Introduction

In the previous sections it was shown that an intuitive approach to a parameter estimation problem can cause serious errors without even being noticed. To avoid severe mistakes, a theoretical framework is needed. Here a statistical development of the parameter estimation theory is made. Four related estimators are studied: the least squares (LS) estimator, weighted least squares (WLS) estimator, maximum likelihood (ML) estimator and, finally, the Bayes estimator. It should be clear that as mentioned before, it is still possible to use other estimators, like the least absolute values. However, a comprehensive overview of all possible techniques is beyond the scope of this book.

To use the Bayes estimator, the a priori probability density function (pdf) of the unknown parameters and the pdf of the noise on the measurements is required. Although it seems, at first, quite strange that the parameters have a pdf, we will illustrate in the next section that we use this concept regularly in daily life. The ML estimator only requires knowledge of the pdf of the noise on the measurements, and the WLS estimator can be applied optimally if the covariance matrix of the noise is known. Even if this information is lacking, the LS method is usable. Each of these estimators will be explained in more detail and illustrated in the following sections.

2.2 Least Squares estimation

Consider a multiple input, single output system modelled by

$$y_0(k) = g\left(u_0(k), \theta_0\right) \tag{2.1}$$

with k the measurement index, $y(k) \in R$, $u_0(k) \in R^{1\times n_u}$, and $\theta_0 \in R^{n_\theta}$ the true parameter vector. The aim is to estimate the parameters from noisy observations at the output of the system:

$$y(k) = y_0(k) + n_y(k). \tag{2.2}$$

This is done by minimizing the errors

$$e(k,\theta) = y(k) - y(k,\theta), \tag{2.3}$$

with $y(k,\theta)$ the modelled output. To do this, a scalar criterion that expressed the 'goodness' of the fit is needed. Many possibilities exist, for example:

minimizing the sum of the absolute values

$$\hat{\theta}_{\mathrm{NLA}}(N) = \arg\min_{\theta} V_{\mathrm{NLA}}(\theta, N), \text{ with } V_{\mathrm{NLA}}(\theta, N) = \sum_{k=1}^{N} \left|e(k,\theta)\right|, \tag{2.4}$$

where NLA stands for nonlinear least absolute values. An alternative is to minimize the sum of the squared values, leading to the nonlinear least squares (NLS):

$$\hat{\theta}_{\mathrm{NLS}}(N) = \arg\min_{\theta} V_{\mathrm{NLS}}(\theta, N), \text{ with } V_{\mathrm{NLS}}(\theta, N) = \frac{1}{2}\sum_{k=1}^{N} e^2(k, \theta). \tag{2.5}$$

The least squares estimator is the most popular. Since this is an arbitrary choice, initially, it is clear that the result is not necessarily optimal. Some of the other loss functions with their related properties are studied explicitly in the literature. In this book we concentrate on least squares, a choice strongly motivated by numerical aspects: minimizing a least squares cost function is usually less involved than the alternative cost functions. Later on, this choice will also be shown to be motivated from stochastic point of view. Normally distributed noise leads, naturally, to least squares estimation. On the other hand $\hat{\theta}_{\mathrm{NLA}}(N)$ will be less sensitive to outliers in the data. A full treatment of the problem is beyond the scope of this book.

As seen in the resistance example, even within the class of least squares estimators, there are different possibilities resulting in completely different estimators with different properties. As a rule of thumb, it is important to see where the noise enters into the raw data. Next a cost function should be selected that explicitly accounts for these errors. For the major part of this chapter, we assume that the inputs $u_0(k)$ are exactly known. Only the output observations of the system is disturbed by noise:

$$y(k) = y_0(k) + n_y(k). \tag{2.6}$$

This leads in a natural way to the loss function (2.5). Later in this chapter, we will come back to the situation where also the input observations are disturbed by noise ($u(k) = u_0(k) + n_u(k)$) and one possibility to deal with this generalized problem will be presented.

2.2.1 Nonlinear least squares

In general the analytical solution of the non-linear least squares problem (2.5) is not known because $e(k, \theta)$ is nonlinear in the parameters. Numerical methods must be used to find the parameters minimize $V_{\mathrm{NLS}}(\theta, N)$. A whole bunch of techniques are described in the literature (Fletcher, 1991), and many of them are available in mathematical packages that are commercially available. They vary from very simple techniques like simplex methods that require no derivatives at all, through gradient or steepest descent methods (based on first order derivatives), to Newton methods that make use of second order derivatives. The optimal choice strongly depends on the specific problem. A short discussion of some of these methods is given in the next chapter. Here we will reduce the problem first to the linear least squares where it is much easier to make general statements. Eventually, we come back to the general problem and make some brief comments.

2.2.2 The linear-in-the-parameters least squares estimator

If the model is linear-in-the-parameters,

$$y_0 = K(u_0)\theta_0, \qquad (K \in \mathbb{R}^{N \times n_\theta}) \tag{2.7}$$

and

$$e(\theta) = y - K(u_0)\theta, \tag{2.8}$$

then 2.5 reduces to a linear least squares cost function.

$$V_{\mathrm{LS}}(\theta, N) = \frac{1}{2}\sum_{k=1}^{N} e^2(k, \theta) = \frac{1}{2}\sum_{k=1}^{N} \left(y(k) - K(u_0(k))\theta\right)^2 \tag{2.9}$$

or using matrix notations

$$V_{\mathrm{LS}}(\theta, N) = \frac{1}{2} e^{\mathrm{T}}(\theta) e(\theta) = \frac{1}{2} \left(y - K(u_0)\theta\right)^{\mathrm{T}} \left(y - K(u_0)\theta\right) \tag{2.10}$$

Definition The linear least squares estimate $\hat{\theta}_{\mathrm{LS}}(N)$ is

$$\hat{\theta}_{\mathrm{LS}}(N) = \arg\min_{\theta} V_{\mathrm{LS}}(\theta, N) \tag{2.11}$$

2.2.2.1 Calculation of the explicit solution

The minimizer of this loss function can be explicitly obtained by putting $\frac{\partial V_{\mathrm{LS}}}{\partial \theta} = 0$. In order to keep the expressions compact, we do not include the arguments of K below.

$$\frac{\partial V_{\mathrm{LS}}}{\partial \theta} = e^{\mathrm{T}} \frac{\partial e}{\partial \theta} = 0 \text{ or } \left(\frac{\partial e}{\partial \theta}\right)^{\mathrm{T}} e = 0, \text{ with } \frac{\partial e}{\partial \theta} = -K \tag{2.12}$$

$$-K^{\mathrm{T}} \left(y - K\theta\right) = 0 \tag{2.13}$$

And the explicit solution becomes

$$\hat{\theta}_{\mathrm{LS}}(N) = \left(K^{\mathrm{T}} K\right)^{-1} K^{\mathrm{T}} y \tag{2.14}$$

Although this offers an elegant expression that is very useful for theoretical studies, it is better to calculate $\hat{\theta}_{\mathrm{LS}}$ by solving $\left(K^{\mathrm{T}} K\right) \hat{\theta}_{\mathrm{LS}} = K^{\mathrm{T}} y$ since this requires less numerical operations. Later on, we even will provide a numerical more stable solution that avoids to make the product $K^{\mathrm{T}} K$.

See Chapter 7, Exercise 5 for an illustration of the stable calculation of the linear least squares solution in MATLAB™.

Example 2.2.1: Weighing a loaf of bread

John is asked to estimate the weight of a loaf of bread from N noisy measurements $y(k) = \theta_0 + n_y(k)$ with θ_0 the true but unknown weight, $y(k)$ the weight

measurement and $n_y(k)$ the measurement noise. The model becomes

$$y = K\theta + n_y, \text{ with } K = (1, 1, \ldots, 1)^{\mathrm{T}}. \tag{2.15}$$

Using (2.14) the estimate is

$$\hat{\theta}_{\mathrm{LS}}(N) = \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}y = \frac{1}{N}\sum_{k=1}^{N} y(k) \tag{2.16}$$

2.2.3 Properties of the linear least squares estimator

Note that we did not formulate any assumption on the behavior of the noise n_y to arrive at 2.14. It can be directly calculated from measurements without bothering about the noise behavior. However, in order to make statements about the properties of the estimator, it is necessary to give some specifications on the noise behavior.

2.2.3.1 Expected value of $\hat{\theta}_{\mathrm{LS}}(N)$

The expected value is obtained by

$$\begin{aligned}
\mathrm{E}\left\{\hat{\theta}_{\mathrm{LS}}(N)\right\} &= \mathrm{E}\left\{\left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}y\right\} && (2.17)\\
&= \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}\,\mathrm{E}\left\{y_0 + n_y\right\} && (2.18)\\
&= \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}y_0 + \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}\,\mathrm{E}\left\{n_y\right\} && (2.19)\\
&= \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}K\theta_0 + \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}\,\mathrm{E}\left\{n_y\right\} && (2.20)
\end{aligned}$$

where we made use of 2.7, $y_0 = K(u_0)\theta_0$. Hence the expected value becomes

$$\mathrm{E}\left\{\hat{\theta}_{\mathrm{LS}}(N)\right\} = \theta_0 + \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}\,\mathrm{E}\left\{n_y\right\} \tag{2.21}$$

$\left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}\,\mathrm{E}\left\{n_y\right\}$ equals zero if $\mathrm{E}\left\{n_y\right\} = 0$.

Conclusion: The linear least squares estimate is unbiased if $\mathrm{E}\left\{n_y\right\} = 0$

2.2.3.2 Covariance matrix of $\hat{\theta}_{\mathrm{LS}}(N)$

Also the covariance matrix can be easily obtained.

$$\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}(N)\right\} = \mathrm{E}\left\{\left(\hat{\theta}_{\mathrm{LS}} - \mathrm{E}\left\{\hat{\theta}\right\}\right)\left(\hat{\theta}_{\mathrm{LS}} - \mathrm{E}\left\{\hat{\theta}\right\}\right)^{\mathrm{T}}\right\} \tag{2.22}$$

$$= \mathrm{E}\left\{\left(\left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}n_y\right)\left(\left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}n_y\right)^{\mathrm{T}}\right\} \tag{2.23}$$

$$= \left(\left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}\right)\mathrm{E}\left\{n_y n_y^{\mathrm{T}}\right\}\left(\left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}\right)^{\mathrm{T}} \tag{2.24}$$

$$= \left(\left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}\right)\mathrm{Cov}\left\{n_y\right\}\left(\left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}\right)^{\mathrm{T}} \tag{2.25}$$

with $\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}\right\} = \mathrm{E}\left\{n_y n_y^{\mathrm{T}}\right\}$.

If the disturbing noise n_y is white and uncorrelated, $\mathrm{Cov}\left\{n_y\right\} = \sigma_y^2\,\mathrm{I}_N$ and $\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}\right\}$ becomes $\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}(N)\right\} = \sigma_y^2\left(K^{\mathrm{T}}K\right)^{-1}$.

Conclusion: The covariance matrix of the linear least squares estimator is given by

$$\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}(N)\right\} = \sigma_y^2\left(K^{\mathrm{T}}K\right)^{-1} \text{ for the white noise case,} \tag{2.26}$$

$$\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}(N)\right\} = \left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}\,\mathrm{Cov}\left\{n_y\right\}\,K\left(K^{\mathrm{T}}K\right)^{-1} \tag{2.27}$$

2.2.3.3 Example B continued:

In example 2.2.1 the least squares estimator was obtained from the measurements $y(k) = \theta_0 + n_y(k)$. Assume that the noise $n_y(k)$ is white and uncorrelated: $\mathrm{Cov}\left\{n_y\right\} = \sigma_y^2 I_N$, then 2.15 becomes using $K = (1, 1, \ldots, 1)^{\mathrm{T}}$:

$$\sigma^2_{\hat{\theta}_{\mathrm{LS}}(N)} = \frac{1}{N}\sigma_y^2. \tag{2.28}$$

Observe that this again the $1/N$ rule as discussed in the resistance examples of the previous chapter.

2.2.3.4 Distribution of $\hat{\theta}_{\mathrm{LS}}$

The estimated parameters are given by

$$\hat{\theta}_{\mathrm{LS}}(N) = \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}y \tag{2.29}$$

Consider the dimensions for the matrices:

$$K \in \mathbb{R}^{N\times n_\theta},\ K^{\mathrm{T}}K \in \mathbb{R}^{n_\theta \times n_\theta} \tag{2.30}$$

These matrices do not depend on the noise.

$K^{\mathrm{T}}n_y$ is an $N \times n_\theta$ times $n_\theta \times 1$ matrix, so $K^{\mathrm{T}}n_y \in \mathbb{R}^{n_\theta \times 1}$. The central limit theorem applies under quite general conditions to the sum that is hidden in this matrix product, e.g. $\sum_{j=1}^{N} K_{ij}n_j$, and hence these terms will be asymptotically ($N \to \infty$) Gaussian distributed, EVEN if n_y is not Gaussian distributed.

This brings us to the conclusion that

$$\hat{\theta} \sim \mathcal{N}\left(\mathrm{E}\left\{\hat{\theta}\right\} = \theta_0, C_\theta\right) \tag{2.31}$$

See Chapter 7, Exercise 3.c for an illustration.

2.3 Weighted least squares estimation (Markov estimator)

In (2.9) all measurements are equally weighted. In many problems it is desirable to put more emphasis on one measurement with respect to the other. This can be done to make the difference between measurements and model smaller in some regions, but it can also be motivated by stochastic arguments. If the covariance matrix of the noise is known, then it seems logical to suppress measurements with high uncertainty and to emphasize those with low uncertainty.

In practice it is not always clear what weighting should be used. If it is, for

example, known that model errors are present, then the user may prefer to put in a dedicated weighting in order to keep the model errors small in some specific operation regions, instead of using the weighting dictated by the covariance matrix.

Definition 2.3.1:
In general the weighted linear least squares estimate $\hat{\theta}_{\mathrm{WLS}}(N)$ is

$$\hat{\theta}_{\mathrm{WLS}}(N) = \arg\min_{\theta} V_{\mathrm{WLS}}(\theta, N) \text{ with } V_{\mathrm{WNLS}}(\theta, N) = e^{\mathrm{T}}(\theta) W e(\theta) \qquad (2.32)$$

where $W \in \mathbb{R}^{N \times N}$ is a symmetric positive definite weighting matrix (the asymmetric part does not contribute to a quadratic form).

The explicit solution is found analogously to that of the least squares estimate and is given by

$$\hat{\theta}_{\mathrm{LS}}(N) = \left(K^{\mathrm{T}} W K\right)^{-1} K^{\mathrm{T}} W y \qquad (2.33)$$

Remark: The evaluation of the cost function (2.32) requires $O(N^2)$ operations which might be very time consuming for large N. Consequently, (block) diagonal weighting matrices are preferred in many problems, reducing the number of operations to $O(N)$.

2.3.1 Bias of the weighted linear least squares

The bias condition is not affected by the choice of W. The proof in Section 2.2.3 is directly applicable to the weighted linear least squares problem.

2.3.2 Covariance matrix

The calculation of the covariance matrix in Section 2.2.3 can be repeated here.

For an arbitrary weighting, the covariance matrix is

$$\operatorname{Cov}\left\{\hat{\theta}_{\mathrm{WLS}}(N)\right\} = \left(K^{\mathrm{T}}WK\right)^{-1} K^{\mathrm{T}}W \operatorname{Cov}\{n_y\} WK \left(K^{\mathrm{T}}WK\right)^{-1} \tag{2.34}$$

It is possible to show that among all possible positive definite choices for W, the 'best' one is $W = C_{n_y}^{-1}$ (where $C_{n_y} = \operatorname{Cov}\{n_y\}$) since this minimizes the covariance matrix. In that case the previous expression simplifies to

$$\operatorname{Cov}\left\{\hat{\theta}_{\mathrm{WLS}}(N)\right\} = \left(K^{\mathrm{T}}C_{n_y}^{-1}K\right)^{-1} K^{\mathrm{T}}C_{n_y}^{-1}C_{n_y}C_{n_y}^{-1}K \left(K^{\mathrm{T}}C_{n_y}^{-1}K\right)^{-1} \tag{2.35}$$

$$= \left(K^{\mathrm{T}}C_{n_y}^{-1}K\right)^{-1} \tag{2.36}$$

2.3.3 Properties of the nonlinear least squares estimator

Consider again the full nonlinear model:

$$y_0(k) = g(u_0(k), \theta_0) \tag{2.37}$$

$$y(k) = y_0(k) + n_y(k) \tag{2.38}$$

Because there are no explicit expressions available for the estimator as a function of the measurements, it is not straightforward to study its properties. For this reason special theories are developed to analyze the properties of the estimator by analyzing the cost function. These techniques are covered in detail in the literature.

2.3.3.1 Consistency

Under quite general assumptions on the noise (for example iid noise with finite second and fourth order moments), some regularity conditions on the model $g(u_0(k), \theta)$ and the excitation (choice of $u_0(k)$), consistency of the least squares estimator is proven.

2.3.3.2 Covariance

An approximate expression for the covariance matrix $\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{NLS}}(N)\right\}$ is available:

$$\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{NLS}}(N)\right\} \approx \left(J^{\mathrm{T}}(\theta)J(\theta)\right)^{-1} J^{\mathrm{T}}(\theta)\,\mathrm{Cov}\left\{n_y\right\} J(\theta)\left(J^{\mathrm{T}}(\theta)J(\theta)\right)^{-1}\Big|_{\theta=\hat{\theta}_{\mathrm{LS}}(N)} \tag{2.39}$$

with

$$\mathrm{Cov}\left\{n_y\right\} = \mathrm{E}\left\{n_y n_y^{\mathrm{T}}\right\} \tag{2.40}$$

and $J(\theta)$ the Jacobian matrix $J(\theta) \in \mathbb{R}^{N\times n_\theta}$: $J(\theta) = \dfrac{\partial e(\theta)}{\partial \theta}$ (2.41)

Note that this approximation is still a stochastic variable since it depends on $\hat{\theta}_{\mathrm{NLS}}(N)$, while the exact expression should be in θ_0.

2.4 The Maximum Likelihood estimator

Using the covariance matrix of the noise as weighting matrix allows for prior knowledge about the noise on the measurements. However, a full stochastic characterization requires the pdf of the noise distortions. If this knowledge is available, it may be possible to get better results than those attained with a weighted least squares. Maximum likelihood estimation offers a theoretical framework to incorporate the knowledge about the distribution in the estimator. The pdf f_{n_y} of the noise also determines the conditional pdf $f(\,y|\,\theta_0)$ of the measurements, given the hypothetical exact model,

$$y_0 = G(u_0, \theta_0), \tag{2.42}$$

that describes the system and the inputs that excite the system. Assuming, again, an additive noise model

$$y = y_0 + n_y, \text{ with } y,\, y_0,\, n_y \in \mathbb{R}^N, \tag{2.43}$$

the likelihood function becomes:

$$f(\,y|\,\theta_0, u_0) = f_{n_y}(y - G(u_0, \theta_0)). \tag{2.44}$$

The maximum likelihood procedure consists of two steps.

First the numerical values y_m of the actual measurements are plugged into the expression (2.44) for the variables y.

Next the model parameters θ_0 are considered as the free variables. This results in the so called likelihood function.

The maximum likelihood estimate is then found as the maximizer of the likelihood function

$$\hat{\theta}_{\mathrm{ML}}(N) = \arg\max_{\theta} f(\,y_m|\,\theta, u_0). \tag{2.45}$$

From now on we will no longer explicitly indicate the numerical values y_m but just use the symbol y for the measured values.

Example (*weighing a loaf of bread - continued):* Consider Example 2.2.1 again, but assume that more information about the noise is available. This time John knows that the distribution f_y of n_y is normal with zero mean and standard deviation σ_y. With this information he can build a ML estimator:

$$f(\,y|\,\theta) = f(\,y(1)|\,\theta) = f_{n_y}(y(1) - \theta) \tag{2.46}$$

Plugging in the knowledge of the noise distribution results in:

$$f(\,y|\,\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\theta)^2}{2\sigma_y^2}} \tag{2.47}$$

and the estimated weight becomes $\hat{\theta}_{\mathrm{ML}} = y$. It is therefore not possible to give a better estimate than the measured value itself.

If John makes repeated independent measurements $y(1), \ldots, y(N)$ the likelihood function is

$$f(\,y|\,\theta) = f(\,y(1)|\,\theta) f(\,y(2)|\,\theta) \ldots f(\,y(N)|\,\theta) \tag{2.48}$$

$$= f_{n_y}(y(1) - \theta) f_{n_y}(y(2) - \theta) \ldots f_{n_y}(y(N) - \theta) \tag{2.49}$$

The product of density functions is due to the indepency of the noise. Because the noise is normally distributed, we get eventually:

$$f(\,y|\,\theta) = \frac{1}{\left(2\pi\sigma_y^2\right)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma_y^2}\sum_{k=1}^{N}(y(k)-\theta)^2}. \tag{2.50}$$

The ML estimate is given by the minimizer of

$$\frac{1}{2\sigma_y^2}\sum_{k=1}^{N}(y(k) - \theta)^2 \tag{2.51}$$

(remark that $\left(2\pi\sigma_y^2\right)^{-N/2}$ is parameter independent), and becomes

$$\hat{\theta}_{\mathrm{ML}}(N) = \frac{1}{N}\sum_{k=1}^{N} y(k). \tag{2.52}$$

This is nothing else other than the sample mean of the measurements. It is again easy to check that this estimate is unbiased. Note that in this case the ML estimator and the (weighted) least squares estimator are the same. This is only the case for normally distributed errors.

The unbiased behavior may not be generalized since the MLE can also be biased. This is shown in the next example.

Example 2.4.1: Estimating the sample mean and sample variance of a normal distribution

Consider N samples $y(k)$, $k = 1, \dots N$ drawn from a normally independent distribution. Can we estimate the mean and standard deviation of the distribution from these measurements?

First the likelihood function is formed

$$\begin{aligned} f(\,y|\,\theta) &= f(\,y(1)|\,\theta) f(\,y(2)|\,\theta) \dots f(\,y(N)|\,\theta) \\ &= \frac{1}{\left(2\pi\sigma_y^2\right)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma_y^2}\sum_{k=1}^{N}(y(k)-\mu)^2} \end{aligned} \tag{2.53}$$

The loglikelihood function is:

$$\ln f(\,y|\,\theta) = -\frac{N}{2}\ln\left(2\pi\sigma_y^2\right) - \frac{1}{2\sigma_y^2}\sum_{k=1}^{N}\left(y(k)-\mu\right)^2 \tag{2.54}$$

The parameters to be estimated are the mean value μ and the standard deviation σ_y.

Putting the derivatives with respect to μ, σ_y^2 results in

$$\begin{aligned} &\tfrac{1}{\sigma_y^2}\textstyle\sum_{k=1}^{N}\left(y(k)-\mu\right) = 0 \\ &-\tfrac{N}{2}\tfrac{1}{\sigma_y^2} + \tfrac{1}{2\sigma_y^4}\textstyle\sum_{k=1}^{N}\left(y(k)-\mu\right)^2 = 0 \end{aligned} \tag{2.55}$$

The solution of this set of equations is:

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} y(k), \tag{2.56}$$

$$\hat{\sigma}_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{k=1}^{N}\left(y(k)-\hat{\mu}_{\mathrm{ML}}\right)^2. \tag{2.57}$$

What are the properties of these estimates?

Expected value of the estimated mean It is straight forward that the first one is unbiased:

$$\mathrm{E}\left\{\hat{\mu}_{\mathrm{ML}}\right\} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{E}\left\{y(k)\right\} = \mu \tag{2.58}$$

Expected value of the estimated variance The mean value of the variance is more difficult to obtain:

$$\mathrm{E}\left\{\hat{\sigma}^2_{\mathrm{ML}}\right\} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{E}\left\{\left(y(k)-\hat{\mu}_{\mathrm{ML}}\right)^2\right\} \tag{2.59}$$

To calculate the expected value, the squared term is broken in three parts:

$$\mathrm{E}\left\{\left(y(k)-\hat{\mu}_{\mathrm{ML}}\right)^2\right\} = \mathrm{E}\left\{\left(\left(y(k)-\mu\right)-\left(\hat{\mu}_{\mathrm{ML}}-\mu\right)\right)^2\right\} \tag{2.60}$$

$$= \mathrm{E}\left\{\left(y(k)-\mu\right)^2 - 2\left(y(k)-\mu\right)\left(\hat{\mu}_{\mathrm{ML}}-\mu\right) + \left(\hat{\mu}_{\mathrm{ML}}-\mu\right)^2\right\} \tag{2.61}$$

Each of these terms can be evaluated separately:

$$\mathrm{E}\left\{\left(y(k)-\mu\right)^2\right\} = \sigma_y^2 \tag{2.62}$$

The last term becomes (using the fact that measurement i is independent of j)

$$\mathrm{E}\left\{\left(\hat{\mu}_{\mathrm{ML}}-\mu\right)^2\right\} = \mathrm{E}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\left(y(i)-\mu\right)\right)^2\right\} \tag{2.63}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left\{\left(y(i)-\mu\right)\left(y(j)-\mu\right)\right\} \tag{2.64}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{E}\left\{\left(y(j)-\mu\right)^2\right\} \tag{2.65}$$

$$= \frac{\sigma_y^2}{N} \tag{2.66}$$

The middle term becomes

$$\mathrm{E}\left\{(y(k)-\mu)\left(\hat{\mu}_{\mathrm{ML}}-\mu\right)\right\} = \mathrm{E}\left\{(y(k)-\mu)\left(\frac{1}{N}\sum_{i=1}^{N}(y(i)-\mu)\right)\right\} \tag{2.67}$$

$$= \frac{1}{N}E(y(k)-\mu)^2 = \frac{\sigma_y^2}{N} \tag{2.68}$$

Putting all these results together gives

$$E\hat{\sigma}^2_{\mathrm{ML}} = \frac{\sigma_y^2}{N}\sum_{i=1}^{N}\left(1-\frac{2}{N}+\frac{1}{N}\right) = \sigma_y^2\left(1-\frac{1}{N}\right) \tag{2.69}$$

Conclusion While the first estimate is unbiased, the second one can be shown to be prone to a bias of $\frac{\sigma^2}{N}$ that asymptotically disappears in N: $E\hat{\sigma}^2_{\mathrm{ML}} = \frac{\sigma^2(N-1)}{N}$. This shows that there is a clear need to understand the properties of ML estimator better.

In the literature, a series of important properties is tabled assuming well-defined experimental conditions. Each time these conditions are met, the user knows in advance, before passing through the complete development process, what the properties of the estimator would be. On the other hand, if the conditions are not met, nothing is guaranteed any more and a dedicated analysis is, again, required. In this course we just make a summary statement of the properties; a very precise description can be found in the literature (Goodwin and Payne, 1977; Caines, 1988).

2.4.1 Properties of the ML estimator

Property 9: Principle of invariance

If $\hat{\theta}_{\mathrm{ML}}$ is a ML estimator of $\theta \in R^{n_\theta}$, then $\hat{\theta}_g = g(\hat{\theta}_{\mathrm{ML}})$ is a ML estimator of $g(\theta)$ where g is a function, $\hat{\theta}_g \in R^{n_g}$ and $n_g \leq n_\theta$, with n_θ a finite number.

Property 10: Consistency

If $\hat{\theta}_{\mathrm{ML}}(N)$ is an ML estimator based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ converges to θ_0 almost surely: $\mathrm{a.s.}\lim_{N\to\infty} \hat{\theta}_{\mathrm{ML}}(N) = \theta_0$. If n_θ depends on N the property is no longer valid, and the consistency should be checked again.

See, for example, the errors-in-variables estimator in the previous section where not only is the resistance value estimated, but also the currents $i(1), \ldots, i(N)$ and voltages $u(1), \ldots, u(N)$ (In this case $n_\theta = N + 1$, e.g. the N current values and the unknown resistance value, the voltage is calculated from the estimated current and resistance value).

Property 11: Asymptotic Normality

If $\hat{\theta}_{\mathrm{ML}}(N)$ is a ML estimator based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ converges in law to a normal random variable.The importance of this property is not only that it allows one to calculate uncertainty bounds on the estimates, but that it also guarantees that most of the probability mass gets more and more unimodally concentrated around its limiting value.

Property 12: Asymptotic Efficiency

If $\hat{\theta}_{\mathrm{ML}}(N)$ is a ML estimator based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ is asymptotically efficient ($\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{ML}}(N)\right\}$ asymptotically reaches the Cramér-Rao lower bound).

2.5 The Bayes estimator

As described before, the Bayes estimator requires the most a prior information before it is applicable, namely: the pdf of the noise on the measurements and the pdf of the unknown parameters. The kernel of the Bayes estimator is the conditional pdf of the unknown parameters θ with respect to the measurements y:

$$f(\theta|\, u, y). \tag{2.70}$$

This pdf contains complete information about the parameters θ, given a set of measurements y. This makes it possible for the experimenter to determine the best estimate of θ for the given situation. To select this best value, it is necessary to lay down an objective criterion, for example the minimization of a risk function $C(\theta|\,\theta_0)$ which describes the cost of selecting the parameters θ if θ_0 are the true but unknown parameters. The estimated parameters $\hat{\theta}$ are found as the minimizers of the risk function weighted with the probability $f(\theta|\, u, y)$:

$$\hat{\theta}(N) = \arg\min_{\theta_0} \int_{\theta \in D} C(\theta|\,\theta_0) f(\theta|\, u, y) d\theta \tag{2.71}$$

For some specific choices of $C(\theta|\,\theta_0)$, the solution of expression 2.71 is well known, for example

1. $C(\theta|\,\theta_0) = |\theta - \theta_0|^2$ leads to the mean value,
2. $C(\theta|\,\theta_0) = |\theta - \theta_0|$ results in the median which is less sensitive to outliers since these contribute less to the second criterion than to the first (Eykhoff, 1974).

Another objective criterion is to choose the estimate as

$$\hat{\theta}_{\text{Bayes}}(N) = \arg\max_{\theta} f(\theta|\, u, y) \tag{2.72}$$

The first and second examples are "minimum risk" estimators, the last is the Bayes estimator. In practice, it is very difficult to select the best out of these.

In the next section, we study the Bayes estimator in more detail. To search for the maximizer of 2.72 the Bayes rule is applied:

$$f(\theta|\, u, y) = \frac{f(y|\, \theta, u) f(\theta)}{f(y)} \tag{2.73}$$

In order to maximize the right hand side of this equation it is sufficient to maximize its numerator, because the denominator is independent of the parameters θ, so that the solution is given by looking for the maximum of

$$f(y|\, \theta, u) f(\theta). \tag{2.74}$$

This simple analysis shows that a lot of a priori information is required to use the Bayes estimator: $f(y|\, \theta, u)$ (also appearing in the ML estimator) and $f(\theta)$. In many problems the parameter distribution $f(\theta)$ is unavailable, and this is one of the main reasons why the Bayes estimator is rarely used in practice (Norton, 1986).

Example 2.5.1: Use of the Bayes estimator in daily life

We commonly use some important principles of the Bayes estimator, without being aware of it. This is illustrated in the following story: Joan was walking at night in Belgium and suddenly saw a large animal in the far distance. She decided that it was either a horse or an elephant

$$\mathrm{P}\left(\text{observation}|\text{elephant}\right) = \mathrm{P}\left(\text{observation}|\text{horse}\right). \tag{2.75}$$

However, the probability of seeing an elephant in Belgium is much lower than that of seeing a horse:

$$\mathrm{P}\left(\text{elephant in Belgium}\right) \ll \mathrm{P}\left(\text{horse in Belgium}\right) \tag{2.76}$$

so that from the Bayes principle Joan concludes she was seeing a horse. If she would be on safari in Kenya instead of Belgium the conclusion would be

opposite, because

$$\mathrm{P}\left(\text{elephant in Kenya}\right) \gg \mathrm{P}\left(\text{horse in Kenya}\right). \tag{2.77}$$

Joan continued her walk. When she came closer she saw that the animal had big feet, a small tail, and also a long trunk so that she had to review her previous conclusion on the basis of all this additional information: there was an elephant walking on the street. When she passed the corner she saw that a circus had arrived in town.

From the previous example it is clear that in a Bayes estimator the prior knowledge of the pdf of the estimated parameters is very important. It also illustrates that it balances our prior knowledge with the measurement information. This is more quantitatively illustrated in the next example.

Example 2.5.2: Weighing a loaf of bread - continued

Consider again Example 2.2.1 but assume this time that the baker told John that the bread normally weighs about $w = 800$ g. However, the weight can vary around this mean value due to humidity, the temperature of the oven and so on, in a normal way with a standard deviation σ_w. With all this information John knows enough to build a Bayes estimator. Using normal distributions and noticing that

$$f(\,y|\,\theta) = f_y(n_y) = f_y\left(y-\theta\right), \tag{2.78}$$

the Bayes estimator is found by maximizing with respect to θ

$$f(\,y|\,\theta)f(\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\theta)^2}{2\sigma_y^2}} \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{(\theta-w)^2}{2\sigma_w^2}} \tag{2.79}$$

and the estimated weight becomes

$$\hat{\theta}_{\text{Bayes}} = \frac{y/\sigma_y^2 + w/\sigma_w^2}{1/\sigma_y^2 + 1/\sigma_w^2}. \tag{2.80}$$

In this result, two parts can be distinguished: y the information derived from the

measurement and w the a priori information from the baker. If the quality of the prior information is high compared with that of the measurements ($\sigma_w \ll \sigma_y$), the estimate is determined mainly by the prior information. If the quality of the prior information is very low compared with the measurements ($\sigma_w \gg \sigma_y$), the estimate is determined mainly by the information from the measurements.

After making several independent measurements $y(1), \ldots, y(N)$ the Bayes estimator becomes

$$\hat{\theta}_{\text{Bayes}}(N) = \frac{\sum_{k=1}^{N} \frac{y(k)}{\sigma_y^2} + \frac{w}{\sigma_w^2}}{\frac{N}{\sigma_y^2} + \frac{1}{\sigma_w^2}}. \tag{2.81}$$

The previous conclusions remain valid. However, when the number of measurements increases, the first term dominates the second one, such that the impact of the prior information is reduced (Sörenson, 1980). Finally, when N becomes infinite, the estimate is completely determined by the measurements.

Conclusion: From these examples it is seen that a Bayes estimator combines prior knowledge of the parameters with information from measurements. When the number of measurements is increased, the measurement information becomes more important and the influence of the prior information decreases. If there is no information about the distribution of the parameters, the Bayes estimator reduces to the ML estimator. If the noise is normally distributed, the ML estimator reduces to the weighted least squares. If the noise is white, the weighted least squares boils down to the least squares estimator.

2.6 Identification in the presence of input and output noise

Model

$$y_0(k) = g\left(u_0(k), \theta_0\right) \tag{2.82}$$

Measurements

$$\begin{aligned} u(k) &= u_0(k) + n_u(k) \\ y(k) &= y_0(k) + n_y(k) \end{aligned} \tag{2.83}$$

Note that in this case both the input and the output measurement are disturbed with noise. This is a major difference with the previous situation, where only the output was disturbed by noise.

3 solutions

- MLE formulation / errors-in-variables EIV
- instrumental variables
- total least squares

2.7 Possibility 1: Errors-in-variables (MLE)

Model

$$y_0(k) = g\left(u_0(k), \theta_0\right) \tag{2.84}$$

Measurements

$$\begin{array}{l} u(k) = u_0(k) + n_u(k) \\ y(k) = y_0(k) + n_y(k) \end{array} \text{, pdf of the noise } \begin{array}{l} n_u \to f_{n_u} \\ n_y \to f_{n_y} \end{array} \tag{2.85}$$

for simplicity we assume that:

$$f(n_u, n_y) = f_{n_u} f_{n_u} \tag{2.86}$$

Parameters to be estimated:

- the model parameters θ_0
- the unknown, true input and output: $u_0(k), y_0(k)$

Note that the number of parameters depends on N !!!!!

likelihood function

$$f((y,u)\,|\,(y_0,u_0,\theta_0)) = f_{n_y}(y - y_0|y_0,\theta_0) f_{n_u}(u - u_0|u_0,\theta_0) \tag{2.87}$$

$$\text{with } y_0(k) = g\left(u_0(k), \theta_0\right) \tag{2.88}$$

2.7.1 Example: Estimation of a Resistance

(see Chapter 1)

Model

$$u_0(k) = Ri_0(k) \tag{2.89}$$

Measurements

$$\begin{aligned} u(k) &= u_0(k) + n_u(k) \\ i(k) &= i_0(k) + n_i(k) \end{aligned} \tag{2.90}$$

Noise model

$$n_u(k), n_y(k) \text{ i.i.d. zero mean normally distributed} \tag{2.91}$$

$$n_u(k) \to N(0, \sigma_u^2) \tag{2.92}$$

$$n_y(k) \to N(0, \sigma_y^2) \tag{2.93}$$

Likelihood function

$$f((y,u)\,|\,(y_0,u_0,\theta_0)) = f_{n_y}(y - y_0|y_0,\theta_0) f_{n_u}(u - u_0|u_0,\theta_0) \tag{2.94}$$

$$\text{with } u_0(k) = Ri_0(k) \tag{2.95}$$

or

$$\begin{aligned} &\frac{1}{\left(\sqrt{2\pi\sigma_y^2}\right)^N} \exp\left(-\sum_{k=1}^{N} \frac{(y(k) - y_0(k))^2}{2\sigma_y^2}\right) \\ &\times \frac{1}{\left(\sqrt{2\pi\sigma_u^2}\right)^N} \exp\left(-\sum_{k=1}^{N} \frac{(u(k) - u_0(k))^2}{2\sigma_u^2}\right) \end{aligned} \tag{2.96}$$

The cost function becomes

$$V_{\mathrm{ML}}(y,u,\theta) = \sum_{k=1}^{N} \frac{(y(k) - y_0(k))^2}{2\sigma_y^2} + \frac{(u(k) - u_0(k))^2}{2\sigma_u^2} \tag{2.97}$$

$$\text{with } u_0(k) = Ri_0(k) \tag{2.98}$$

For the rest of the discussion, we refer to Chapter 1.

2.8 Possibility 2: Instrumental Variables

2.8.1 Introduction

In this section we will discuss a final parameter estimation method that is very suitable when both the input and the output are disturbed by noise. Although it does not belong directly to the previous family of estimators we include it in this chapter for use later, to interpret one of the proposed identification schemes.

Example 2.8.1: Measuring a resistance

In the resistance estimation examples, it was shown that the least squares method $\hat{R}_{\mathrm{LS}}(N)$ is biased due to the quadratic noise contributions appearing in the denominator:

$$\hat{R}_{\mathrm{LS}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k)}{\frac{1}{N}\sum_{k=1}^{N} i^2(k)} \tag{2.99}$$

$$\text{with } \lim_{N\to\infty} \hat{R}_{\mathrm{LS}}(N) = R_0 \frac{1}{1+\sigma_i^2/i_0^2} \tag{2.100}$$

This systematic error can be removed by replacing $i(k)$ in the numerator and denominator by $i(k-1)$ so that the new estimate becomes:

$$\hat{R}_{\mathrm{IV}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k-1)}{\frac{1}{N}\sum_{k=1}^{N} i(k)i(k-1)}. \tag{2.101}$$

Making the same analysis as in the previous chapter, it is seen that all quadratic noise contributions are eliminated by this choice, so that

$$\lim_{N\to\infty} \hat{R}_{\mathrm{IV}}(N) = R_0. \tag{2.102}$$

2.8.2 The instrumental variables method

The idea used to generate (2.101) can be generalized as follows. Consider the linear-in-the-parameters model structure

$$y_0 = K(u_0)\theta_0 \tag{2.103}$$

and its least squares estimate

$$\hat{\theta}_{\mathrm{LS}}(N) = \left(K^{\mathrm{T}}K\right)^{-1} K^{\mathrm{T}}y \text{ (see Section 2.2)}, \tag{2.104}$$

Replace K^{T} in (2.14)by G^{T}, to get

$$\hat{\theta}_{\mathrm{IV}}(N) = \left(G^{\mathrm{T}}K(u)\right)^{-1} G^{\mathrm{T}}y. \tag{2.105}$$

The choice of G, a matrix of the same size as $K(u)$, will be set by the conditions that appear in the consistency and the variance analysis. $\hat{\theta}_{\mathrm{IV}}(N)$ is the instrumental variables estimate.

2.8.3 Consistency

Consistency is proven by considering the plim for $N \to \infty$ (Norton, 1986). For simplicity we assume all the plim exists, viz.:

$$\operatorname*{plim}_{N\to\infty} \hat{\theta}_{\mathrm{IV}} = \operatorname*{plim}_{N\to\infty} \left\{ \left(G^{\mathrm{T}}K(u)\right)^{-1} G^{\mathrm{T}}y \right\} \tag{2.106}$$

$$= \left(\operatorname*{plim}_{N\to\infty} G^{\mathrm{T}}K(u_0 + n_u) \right)^{-1} \left(\operatorname*{plim}_{N\to\infty} \left\{ G^{\mathrm{T}}y_0 + G^{\mathrm{T}}n_y \right\} \right) \tag{2.107}$$

$$= \left(\operatorname*{plim}_{N\to\infty} \frac{G^{\mathrm{T}}K(u_0 + n_u)}{N} \right)^{-1} \tag{2.108}$$

$$\cdot \left(\operatorname*{plim}_{N\to\infty} \left\{ \frac{G^{\mathrm{T}}K(u_0)}{N} \right\} \theta_0 + \operatorname*{plim}_{N\to\infty} \frac{G^{\mathrm{T}}n_y}{N} \right) \tag{2.109}$$

If

$$\operatorname*{plim}_{N\to\infty}\left\{\frac{G^{\mathrm{T}}K(u_0+n_u)}{N}\right\} = \operatorname*{plim}_{N\to\infty}\left\{\frac{G^{\mathrm{T}}K(u_0)}{N}\right\} \tag{2.110}$$

$$\operatorname*{plim}_{N\to\infty}\left\{\frac{G^{\mathrm{T}}n_y}{N}\right\} = 0 \tag{2.111}$$

then

$$\operatorname*{plim}_{N\to\infty}\hat{\theta}_{\mathrm{IV}}(N) = \theta_0. \tag{2.112}$$

Equations 2.110 and 2.111 define the necessary conditions for G to get a consistent estimate. Loosely stated, G should not be correlated with the noise on $K(u_0+n_u)$ and the output noise n_y. The variables used for building the entries of G are called the instrumental variables.

2.8.4 Covariance matrix

If the covariance for $C_{n_y} = \sigma^2 I_N$, then an approximate expression for the covariance matrix of the estimates is (Norton, 1986):

$$\operatorname{Cov}\left\{\hat{\theta}_{\mathrm{IV}}(N)\right\} \approx \sigma^2\, R_{GK}^{-1} R_{GG} R_{GK}^{-\mathrm{T}} \tag{2.113}$$

$$\text{with } R_{GK} = \frac{G^{\mathrm{T}}K\,(u)}{N}$$

$$\text{and } R_{GG} = \frac{G^{\mathrm{T}}G}{N}$$

This reveals another condition on the choice of the instrumental variables G: while they should be "uncorrelated" with the noise on the output observation n_y, they should be correlated maximally with K, otherwise R_{GK} tends to zero and $\operatorname{Cov}\left\{\hat{\theta}_{\mathrm{IV}}(N)\right\}$ would become very large.

Example: Measuring a resistance (cont'd) In the case of the resistance estimate, the instrumental variables are the shifted input. Since we used a constant current, no problem arises. In practice this technique can be generalized to varying inputs under the condition that the power spectrum of the noise is

much wider than the power spectrum of the input. In the Exercises below the instrumental variables method is applied to the resistance example.

2.8.5 Conclusion

The Instrumental variables estimate is given by

$$\hat{\theta}_{\mathrm{IV}}(N) = \left(G^{\mathrm{T}}K(u)\right)^{-1} G^{\mathrm{T}}y. \tag{2.114}$$

The choice of G:

- a matrix of the same size as $K(u)$
- $\operatorname{plim}\left\{G^{\mathrm{T}}K(u_0 + n_u)/N\right\} = \operatorname{plim}\left\{G^{\mathrm{T}}K(u_0)/N\right\}$
- $\operatorname{plim}\left\{G^{\mathrm{T}}n_y/N\right\} = 0$
- maximize $R_{GK} = G^{\mathrm{T}}K(u)/N$ to reduce the covariance matrix

2.9 Illustration of the Instrumental Variables and the Errors-In-Variables

It was shown that the presence of disturbing noise on the input measurements creates a systematic error. In this set of exercises more advanced identification methods are illustrated that can deal with this situation. Two methods are studied, the first is called the instrumental variables method (IV), the second is the errors-in-variables (EIV) method. The major advantage of the IV-methods is its simplicity. No additional information is required from the user. The disadvantage is that this method does not always perform well. Both situations are illustrated in the exercises. The EIV performs well in many cases, but in general additional information of the user is required. The covariance matrix of the input-output noise should be known. All methods are illustrated again on the resistance example with measured current and voltage $i(t), u(t)$, $t = 1, 2, \ldots, N$. Both measurements are disturbed by mutually uncorrelated Gaussian noise:

$$\begin{aligned} i(t) &= i_0(t) + n_i(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned} \tag{2.115}$$

The least squares estimate is given by:

$$\hat{R}_{\mathrm{LS}} = \frac{\sum_{t=1}^{N} u(t)i(t)}{\sum_{t=1}^{N} i(t)^2}, \tag{2.116}$$

the instrumental variables estimator (IV) is:

$$\hat{R}_{\mathrm{IV}} = \frac{\sum_{t=1}^{N} u(t)i(t+s)}{\sum_{t=1}^{N} i(t)i(t+s)}, \tag{2.117}$$

with s a user selectable shift parameter. Note that the IV-estimator equals the LS-estimator for $s = 0$.

The EIV estimator is given by

$$\hat{R}_{\text{EIV}} = \frac{\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2} + \sqrt{\left(\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2}\right)^2 + 4\frac{(\sum u(t)i(t))^2}{\sigma_u^2\sigma_i^2}}}{2\frac{\sum u(t)i(t)}{\sigma_u^2}}, \quad (2.118)$$

with σ_u^2, σ_i^2 the variance of the voltage and current noise, the covariance is assumed to be zero in this expression: $\sigma_{ui}^2 = 0$.

Exercise 2.9.1: Noise on input and output: the instrumental variables method

Generate the current $i_0(k)$ from a Gaussian white noise source, filtered by a first order Butterworth filter with cut-off frequency f_{Gen}:

$$i_0 = \text{filter}(b_{\text{Gen}}, a_{Gen}, e_1), \quad (2.119)$$

with `[`b_{Gen}`,`a_{Gen}`]=butter(1,2*`f_{Gen}`)`. Generate the measured current and voltage (2.115), where $n_u(k)$ is white Gaussian noise: $N(0, \sigma_{n_u}^2)$. The current noise $n_i(k)$ is obtained from a Gaussian white noise source filtered by a second order Butterworth filter with cut-off frequency f_{Noise}:i_0`=filter(`$b_{\text{Noise}}, a_{\text{Noise}}, e_2$`)`, with `[`$b_{\text{Noise}}$`,`$a_{\text{Noise}}$`]=butter(2,2*`$f_{\text{Noise}}$`)`, and e_2 white Gaussian noise. Its variance is scaled to $\sigma_{n_u}^2$.

- Experiment 1: Generate three sets of 1000 experiments with $N = 5000$ measurements each, and the following parameter settings:
  - $f_{\text{Gen}} = 0.1$, $f_{\text{Noise}} = [0.999, 0.95, 0.6]$, $\sigma_{i_0} = 0.1$, $\sigma_{n_i} = 0.1$, $\sigma_{n_u} = 1$.
  - Process these measurements with the LS-estimator, and with the IV-estimator with the shift parameter $s = 1$.
- Experiment 2: Generate 1000 experiments with $N = 5000$ measurements each, and the following parameter settings:
  - $f_{\text{Gen}} = 0.1$, $f_{\text{Noise}} = 0.6$, $\sigma_{i_0} = 0.1$, $\sigma_{n_i} = 0.1$, $\sigma_{n_u} = 1$.

– Process these measurements with the LS-estimator, and with the IV-estimator with the shift parameter $s = 1, 2, 5$.

Plot for both experiments:

- the pdf of $\hat{R}_{\mathrm{LS}}$ and $\hat{R}_{\mathrm{IV}}$,
- the auto-correlation function of i_0 and n_i (hint: use the MATLAB™instruction `xcorr`)
- the FRF of the generator and the noise filter.

Observations The results are shown below 2.1 and 2.2. In the first Figure 2.1, the results are shown for a fixed generator filter and a varying noise filter. The shift parameter for the IV is kept constant to 1. From this figure it is clearly seen that the LS are strongly biased. This is due to the noise on the input, the relative bias is in the order of $\sigma_{n_i}^2/\sigma_{i_0}^2$. For the IV-results, the situation is more complicated. For the white noise situation, no bias is visible. However, once the output noise is filtered, a bias becomes visible. The relative bias is proportional to the ratio of the auto correlation functions of the noise and the current $R_{n_i n_i}(s)/R_{u_0 u_0}(s)$.

The same observations can also be made in 2.2. In this figure, the shift parameter is changed while the filters are kept constant. It can be seen that the bias becomes smaller with increasing shift s, because $R_{n_i n_i}(s)/R_{i_0 i_0}(s)$ is getting smaller. At the same time the dispersion is growing, mainly because $R_{i_0 i_0}(s)$ is getting smaller. Observe also that the sign of the bias depends on the sign of $R_{n_i n_i}(s)$. The IV-method works well if the bandwidth of the generator signal is much smaller than that of the noise disturbances.

Exercise 2.9.2: Noise on input and output: the errors-in-variables method

In this exercise the EIV-method is used as an alternative for IV-method to reduce/eliminate the bias of the least squares estimate. This time no constraint

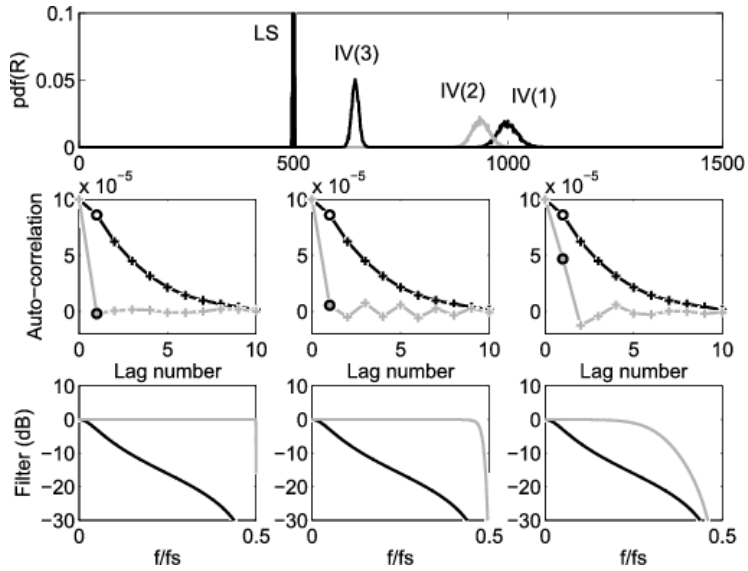


Figure 2.1: Study of the LS- and IV-estimate for a varying noise filter bandwidth and fixed shift $s = 1$.
Top: the LS (black line) and IV estimate (black or gray line). IV(1), IV(2), and IV(3) correspond to the first second, and third filter. Middle: the auto correlation of i_0 (black) and n_i (gray) for the different noise filters. Bottom: the filter characteristics of i_0 (black) and the noise n_i (gray).

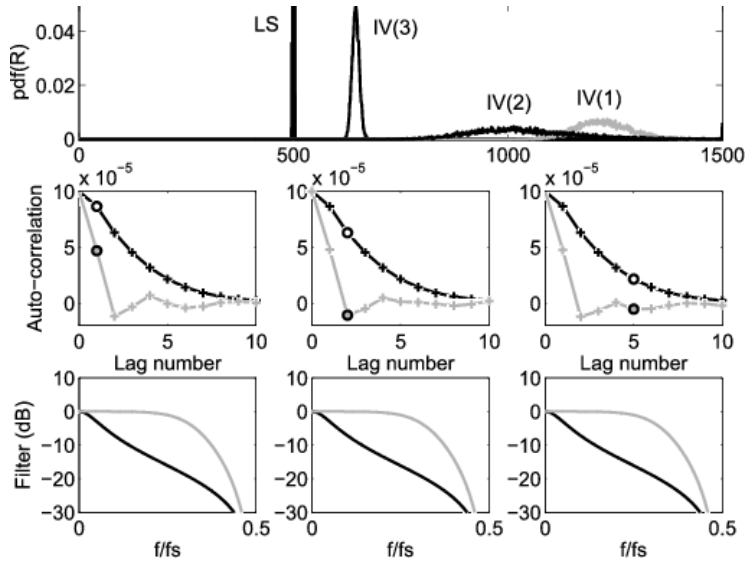


Figure 2.2: Study of the LS- and IV-estimate for a fixed noise filter bandwidth and a varying shift s=1, 2, 5.
Top: the LS (black) and IV (black and gray) estimate. IV(1), IV(2), and IV(3) correspond to a shift of 1,2, and 5 tabs. Middle: the auto correlation of i_0 (black) and n_i (gray). Bottom: the filter characteristics of i_0 (black) and the noise n_i (gray)

is put on the power spectra (bandwidth) of the excitation and the disturbing noise, but instead the variance of the input and output disturbing noise should be priorly given. This is illustrated again on the resistance example with measured current and voltage $i(t), u(t)$, $t = 1, 2, \ldots, N$.The least squares estimate is given by

$$\hat{R}_{\mathrm{LS}} = \frac{\sum_{k=1}^{N} u(k)i(k)}{\sum_{k=1}^{N} i(k)^2}, \tag{2.120}$$

the EIV-estimator is

$$\hat{R}_{\mathrm{EIV}} = \frac{\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2} + \sqrt{\left(\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2}\right)^2 + 4\frac{(\sum u(t)i(t))^2}{\sigma_u^2\sigma_i^2}}}{\frac{\sum u(t)i(t)}{\sigma_u^2}}, \tag{2.121}$$

where the sum runs over $t = 1, \ldots, N$. It is shown to be the minimizer of the following cost function:

$$V_{\mathrm{EIV}} = \frac{1}{N}\sum_{t=1}^{N}\left\{\frac{\left(u(t) - u_0(t)\right)^2}{\sigma_u^2} + \frac{\left(i(t) - i_0(t)\right)^2}{\sigma_i^2}\right\}, \tag{2.122}$$

with respect to u_0, i_0, R_0 under the constraint $u_0(t) = R_0 i_0(t)$.

- Setup: Generate the current $i_0(t)$ from a white zero mean Gaussian noise source $N(0, \sigma_{i_0}^2)$.
  - Generate the measured current and voltage as:

$$\begin{aligned} i(t) &= i_0(t) + n_i(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned}, \tag{2.123}$$

  - $n_u(t)$ and $n_i(t)$ are white Gaussian noise sources with zero mean and vari-ance $\sigma_{n_u}^2$ and $\sigma_{n_i}^2$ respectively
- Generate a set of 1000 experiments with $N = 5000$ measurements each, and the following parameter settings:
  - $R_0 = 1000$, $\sigma_{i_0} = 0.01$, $\sigma_{n_i} = 0.001$, $\sigma_{n_u} = 1$.

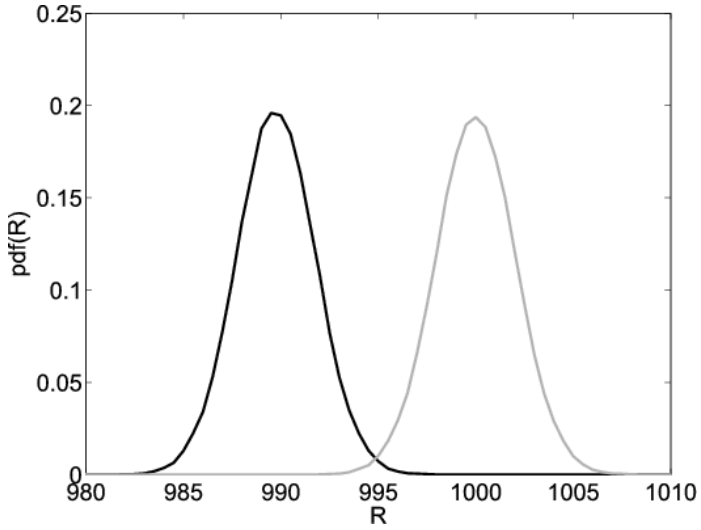


Figure 2.3: Comparison of the pdf of the LS- (black) and the EIV-estimate (gray), calculated with prior known variances.

- Calculate the LS- and EIV-estimate. Plot the histogram with $\hat{R}_{\mathrm{LS}}$ and $\hat{R}_{\mathrm{EIV}}$.

Observations The results are shown in Figure 2.3. From this figure it is clearly seen that the LS are strongly biased (mean value is 990.15). This is due to the noise on the input, the relative bias is in the order of $\sigma_{n_i}^2/\sigma_{i_0}^2$. No systematic error can be observed in the EIV-results (mean value is 999.96). The IV-estimate would fail completely in this situation (why?).

2.10 Possibility 3: Total least squares

$$K\theta \approx Y \tag{2.124}$$

with

$$K = K_0 + N_K, \text{ and } Y = Y_0 + N_Y \tag{2.125}$$

Rearrange the equations

$$\begin{pmatrix} K & Y \end{pmatrix} \begin{pmatrix} \theta \\ -1 \end{pmatrix} \approx 0 \text{ or } L\tilde{\theta} \approx 0 \tag{2.126}$$

Assumption on the noise N_L

$$EN_L^{\mathrm{T}} N_L = \sigma_L^2 I_{n_\theta \times n_\theta} \tag{2.127}$$

The least squares solution

$$L^{\mathrm{T}} L\tilde{\theta}_{\mathrm{LS}} = 0 \tag{2.128}$$

Limit for a large number of samples

$$L^{\mathrm{T}} L\tilde{\theta}_{\mathrm{LS}} \approx 0 \rightarrow \left(L_0^{\mathrm{T}} L_0 + \sigma_L^2 I\right) \tilde{\theta}_{\mathrm{LS}} \approx 0 \tag{2.129}$$

Basic idea: Compensate for the noise

$$L^{\mathrm{T}} L\tilde{\theta}_{\mathrm{LS}} \approx 0 \rightarrow \left(L_0^{\mathrm{T}} L_0 + \sigma_L^2 I\right) \tilde{\theta}_{\mathrm{LS}} \approx 0 \tag{2.130}$$

Hence

$$\left(L^{\mathrm{T}} L - \lambda I\right) \tilde{\theta}_{\mathrm{LS}} = 0 \tag{2.131}$$

leads to the correct solution if $\lambda = \sigma_L^2$

Total least squares idea: Solve the eigenvalue problem

$$\left(L^{\mathrm{T}}L\right)\tilde{\theta}_{\mathrm{TLS}} = \lambda\tilde{\theta}_{\mathrm{TLS}} \tag{2.132}$$

equivalencies

$$\tilde{\theta}_{\mathrm{TLS}} = \arg\min_{\tilde{\theta}} \frac{\left\|A\tilde{\theta}\right\|_2^2}{\left\|\tilde{\theta}\right\|_2^2} \tag{2.133}$$

or

$$\tilde{\theta}_{\mathrm{TLS}} = \arg\min_{\tilde{\theta}} \left\|A\tilde{\theta}\right\|_2^2 \text{ with } \left\|\tilde{\theta}\right\|_2^2 = 1 \tag{2.134}$$

Remarks on TLS:

1. the noise assumption can be relaxed by using a noise weighting
2. TLS is easy to calculate
   - set up the matrix: $L = \begin{bmatrix} K & Y \end{bmatrix}$
   - calculate the SVD: $L = U\Sigma V^{\mathrm{T}}$
   - the solution is given by the last vector in V: $V = \begin{bmatrix} \dots & \tilde{\theta}_{\mathrm{TLS}} \end{bmatrix}$
3. TLS is well suited as a trial to generate starting values

Appendices

2.A Singular value decomposition

For any $A \in \mathbb{C}^{n\times m}$ with $n \geq m$ there exist $U \in \mathbb{C}^{n\times m}$ and $\Sigma, V \in \mathbb{C}^{m\times m}$ such that (Golub and Van Loan, 1996)

$$A = U\Sigma V^{\mathrm{H}} \tag{2.135}$$

where $V^{\mathrm{H}}V = VV^{\mathrm{H}} = U^{\mathrm{H}}U = \mathrm{I}_m$ and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m)$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \geq 0$. The nonnegative real numbers σ_k are the *singular values* of A, and the columns $V_{[:,k]}$ and $U_{[:,k]}$ are the corresponding right and left singular vectors. (2.135) is called the *singular value decomposition* (SVD) of the matrix A. It can be expanded as

$$A = \sum_{k=1}^{m} \sigma_k U_{[:,k]} V_{[:,k]}^{\mathrm{H}} \tag{2.136}$$

A numerically stable calculation of the singular value decomposition is available in standard mathematical software packages.

The singular value decomposition (2.135) contains a lot of information about the structure of the matrix. Indeed, if $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_m = 0$ then

$$\begin{aligned} \mathrm{rank}(A) &= r \\ \mathrm{null}(A) &= \mathrm{span}\left\{V_{[:,r+1]}, V_{[:,r+2]}, \ldots, V_{[:,m]}\right\} \\ \mathrm{range}(A) &= \mathrm{span}\left\{U_{[:,1]}, U_{[:,2]}, \ldots, U_{[:,r]}\right\} \end{aligned}$$

The *condition number* $\kappa(A)$ of a matrix $A \in \mathbb{C}^{n\times m}$ is defined as the ratio of the largest singular value to the smallest singular value $\kappa(A) = \frac{\sigma_1}{\sigma_m}$.

For regular square matrices $m = n$ it is a measure of the sensitivity of the solution of the linear system $Ax = b$, with $b \in \mathbb{C}^n$, to perturbations in A and b.

It can be shown that (Golub and Van Loan, 1996)

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \kappa(A) \left(\frac{\|\Delta A\|_2}{\|A\|_2} + \frac{\|\Delta b\|_2}{\|b\|_2} \right) \tag{2.137}$$

where Δ denotes the perturbation. For rectangular matrices $m > n$ of full rank, it is a measure of the sensitivity of the least squares solution $x_{\mathrm{LS}} = \left(A^{\mathrm{H}}A\right)^{-1} A^{\mathrm{H}}b$ of the overdetermined set of equations $Ax \approx b$, with $b \in \mathbb{C}^m$, to perturbations in A and b. For singular matrices $\kappa(A) = \infty$. If $\kappa(A)$ is large ($\log_{10}(\kappa(A))$ is of the order of the number of significant digits used in the calculations), then A is said to be *ill-conditioned*. Unitary (orthogonal) matrices are perfectly conditioned ($\kappa = 1$), while matrices with small conditions number ($\kappa \approx 1$) are said to be *well-conditioned*.

2.B Moore-Penrose pseudo-inverse

For any matrix $A \in \mathbb{C}^{n \times m}$ there exists a unique generalized inverse $A^+ \in \mathbb{C}^{m \times n}$, also called Moore-Penrose pseudo-inverse, that satisfies the four Moore-Penrose conditions (Ben-Israel and Greville, 1974)

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $\left(AA^+\right)^{\mathrm{H}} = AA^+$
4. $\left(A^+A\right)^{\mathrm{H}} = A^+A$

For regular square matrices it is clear that $A^+ = A^{-1}$. The pseudo-inverse can be constructed using, for example, the singular value decomposition (Golub and Van Loan, 1996). If $\mathrm{rank}(A) = r$ then

$$A^+ = V\Sigma^+U^{\mathrm{H}} \text{ with } \Sigma^+ = \mathrm{diag}\left(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0\right) \tag{2.138}$$

Using (2.138) it can easily be shown that for every matrix A, $\left(A^+\right)^+ = A$, $\left(A^+\right)^{\mathrm{H}} = \left(A^{\mathrm{H}}\right)^+$ and $A^+ = \left(A^{\mathrm{H}}A\right)^+ A^{\mathrm{H}} = A^{\mathrm{H}}\left(AA^{\mathrm{H}}\right)^+$.

Although the properties of the pseudo-inverse very much resemble those of the inverse, in general $(AB)^+ \neq B^+A^+$. If the matrices $A \in \mathbb{C}^{n\times r}$ and $B \in \mathbb{C}^{r\times m}$ with $r \leq \min(n,m)$ are of full rank then $(AB)^+ = B^+A^+$ (Ben-Israel and Greville, 1974).

2.C Solution of the least squares problem using SVD

Consider the least squares solution $\hat{\theta}_{\mathrm{LS}} = \left(K^{\mathrm{T}}K\right)^{-1}K^{\mathrm{T}}y$ as the solution of

$$K^{\mathrm{T}}K\hat{\theta}_{\mathrm{LS}} = K^{\mathrm{T}}y \tag{2.139}$$

Decompose $K = U\Sigma V^{\mathrm{T}}$, then (2.139) becomes

$$\left(V\Sigma U^{\mathrm{T}}\right)\left(U\Sigma V^{\mathrm{T}}\right)\hat{\theta}_{\mathrm{LS}} = \left(V\Sigma U^{\mathrm{T}}\right)y \tag{2.140}$$

$$V\Sigma^2V^{\mathrm{T}}\hat{\theta}_{\mathrm{LS}} = V\Sigma U^{\mathrm{T}}y \tag{2.141}$$

since $U^{\mathrm{T}}U = \mathrm{I}_N$Left multiplication with V^{T} (notice that $V^{\mathrm{T}}V = \mathrm{I}_{n_\theta}$, followed by left multiplication with Σ^{-2}, and eventually again with V leads to

$$\hat{\theta}_{\mathrm{LS}} = V\Sigma^{-1}U^{\mathrm{T}}y = K^+y \tag{2.142}$$

Using the SVD reduces the number of required digits to calculate a stable numerical solution with a factor 2 since the product $K^{\mathrm{T}}K$ is no longer made.

Chapter 3

Model Selection and Validation

Chapter 3

Model selection and validation

A critical step in the identification process is the quality assessment of the identified model. A model without error bounds has no value. For this reason, we need tools to check if all 'input-output' relations in the raw data are captured, and tools to quantify the remaining model errors. Also the validity of the disturbing noise models should be tested.

This chapter provides dedicated tools to test for over- and undermodelling. First the calculation of uncertainty bounds on a model will be recalled. Next it will be shown how overmodelling can be detected, and a tool for selecting an optimal model complexity that balances model errors against noise sensitivity will be introduced and illustrated.

3.1 Introduction

At the end of an identification run, two important questions remain to be answered. What is the quality of the model? Can this model be used to solve my problem? While the first question is an absolute one, the second question shows that in practice the applicability of an identified model strongly depends on the

intended application. Each model is only an approximation of reality and often the existence of a "true" model is only a fiction, in the mind of the experimenter. The deviations between the model and the system that generated the measurements are partitioned in two parts following their nature: systematic errors and stochastic errors. If the experiment is repeated under the same conditions, the systematic errors will be the same, while the stochastic errors vary from one realization to the other. Model validation is directed towards the quantification of the remaining model errors. Once the level of the systematic errors is known, the user should decide whether they are acceptable or not. It is not evident at all that one is looking for the lowest error level, often it is sufficient to push them below a given upper bound. In order to decide if the errors are systematic, it is necessary to know the uncertainty on the estimated model. In this course we use probabilistic uncertainty bounds (e.g. 95% bounds) that describe how the individual realizations are scattered around their mean values. Errors that are outside this bound are considered to be unlikely, so that they are most probably due to systematic deviations.

This short discussion shows, clearly, that model validation starts with the generation of good uncertainty bounds. These bounds can be used in a second step to check for the presence of significant (from statistical point of view) systematic errors. This two step approach is developed in the course of this chapter.

3.2 Assessing the model quality: Quantifying the stochastic errors using uncertainty bounds

As mentioned in the introduction, the first step in the validation process is the partitioning in stochastic and systematic errors. The stochastic error bounds are not only a tool to detect systematic errors, they also are intensively used to describe the overall quality of the model once it is known that systematic errors are no longer dominating.

3.2.1 Covariance matrix of the estimated parameters

The basic "uncertainty" information is delivered under the form of the covariance matrix on the estimated parameters. The actual covariance matrix is mostly too difficult to calculate. But in most cases the Cramér-Rao lower bound can be used for asymptotically efficient estimators. Also for weighted least squares estimators, approximative expressions to calculate the covariance matrix are available. An approximation of both expressions can be calculated easily at the end of the identification process.

3.2.2 Covariance matrix of other model characteristics

However, in many applications the user is not interested in the estimated parameters and their uncertainty, but wants to calculate from these parameters other system characteristics such as the transfer function or the pole positions of this system. The Cramér-Rao lower bound of these derived quantities is generated by simple transformation laws, obtained from the first order derivatives of the actual transformation. The same laws also apply to the approximated covariance matrices:

$$\mathrm{Cov}\left\{f(x)\right\} \approx \left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x} \mathrm{Cov}\left\{f(x)\right\} \left(\left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}\right)^{\mathrm{H}} \tag{3.1}$$

In practice, this works very well as long as the transformations are not heavily nonlinear (e.g. transfer function calculation), but sometimes it fails. A typical example of such a failure is the generation of the uncertainty regions on the estimated poles/zeros. Although the Cramér-Rao bounds (or the approximate covariance matrix) are correct, the actual uncertainties can significantly differ due to the fact that the asymptotic properties on these estimates are not yet reached for practical signal to noise ratios.

Example 3.2.1: Uncertainty bounds on the calculated transfer functions

In this example, we calculate the uncertainty on the amplitude characterisitic of a parametric transfer function model that is identified from a measured FRF (frequency response function).

Consider the measurements:

$$Z = Z_0 + N_Z, \tag{3.2}$$

$$\text{with } Z(k) = G(\Omega_k) = G_0(\Omega_k) + N_G(\Omega_k). \tag{3.3}$$

The noise $N_G(\Omega_k)$ is assumed to be normally, zero mean and independendly distributed over the frequencies. A parametric transfer function model $G(\Omega_k, \hat{\theta})$, together with the covariance matrix C_θ is estimated from these data, and we want to know the reliability of the estimated transfer function as a function of the frequency. Applying eq. 3.1 gives the variance of the transfer function due to the noise sensitivity of the parameter estimates

$$\text{Var}\left\{G(\Omega, \hat{\theta})\right\} \approx \left.\frac{\partial G(\Omega, \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} C_\theta \left(\left.\frac{\partial G(\Omega, \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}}\right)^{\text{H}} \tag{3.4}$$

3.2.3 Uncertainty bounds on the residuals

(**only for information, no questions on the exam**)

A very simple, but popular, validation test is to compare the differences between the measurements and the model

$$e = Z - g(Z, \hat{\theta}(Z)) = f(Z, \hat{\theta}(Z)), \tag{3.5}$$

These differences are called the residuals. For the transfer function example, it are the differences between the measured FRF, $G(\Omega_k)$, and the modelled transfer function, $G(\Omega_k, \hat{\theta})$, $e(k) = G(\Omega_k) - G(\Omega_k, \hat{\theta})$.

In order to decide if these residuals e are significantly different from zero,

their variance should be calculated. Eq. 3.4 of the previous section cannot be applied here directly since $G(\Omega_k) - G(\Omega_k, \hat{\theta})$ depends now not only on $\hat{\theta}(Z)$, but also on the raw data $G(\Omega_k)$, or in general $f(Z, \hat{\theta}(Z))$ depends directly(by Z) and indirectly (by $\hat{\theta}(Z)$) on the raw data Z. Note that $\hat{\theta}(Z)$ and Z are correlated stochastic variables since they depend both on the same noise distortions N_Z. An extended expression for the variance should be used here, taking the correlation into account. It is given here (without proof) for a complex valued $f(Z, \theta)$ and Z:

$$\begin{aligned}\mathrm{Cov}\left\{f(Z,\hat{\theta}(Z))\right\} \approx & \left(\frac{\partial f(Z,\hat{\theta}(Z))}{\partial Z}\right) C_{N_Z} \left(\frac{\partial f(Z,\hat{\theta}(Z))}{\partial Z}\right)^{\mathrm{H}} \\ & + \left(\frac{\partial f(Z,\theta)}{\partial \hat{\theta}(Z)}\right) \mathrm{Cov}\left\{\underline{\hat{\theta}}(Z)\right\} \left(\frac{\partial f(Z,\theta)}{\partial \hat{\theta}(Z)}\right)^{\mathrm{H}} \\ & + 2\,\mathrm{herm}\left(\left(\frac{\partial f(Z,\hat{\theta}(Z))}{\partial Z}\right) \mathrm{Cov}\left\{N_Z, \hat{\theta}(Z) - \tilde{\theta}(Z_0)\right\} \left(\frac{\partial f(Z,\theta)}{\partial \hat{\theta}(Z)}\right)^{\mathrm{H}}\right)\end{aligned} \tag{3.6}$$

$$\mathrm{Cov}\left\{N_Z, \hat{\theta}(Z) - \tilde{\theta}(Z_0)\right\} \approx -C_{N_Z} \left(\frac{\partial \varepsilon(\hat{\theta}(Z), Z)}{\partial Z}\right)^{\mathrm{H}} \left(\frac{\partial \varepsilon(\theta, Z)}{\partial \hat{\theta}(Z)}\right) \tag{3.7}$$

$$\times\, \mathrm{Cov}\left\{\underline{\hat{\theta}}(Z)\right\} \tag{3.8}$$

Example 3.2.2: Variance on the transfer function residuals

Consider the setup of (3.2.1) and assume that the system is excited with a deterministic input (to avoid additional complications to take care for input variations in the case of model errors). It can be shown (without proof in this course) that the variance on the residual $G(\Omega_k) - G(\Omega_k, \hat{\theta})$ becomes:

$$\mathrm{Var}\left\{G(\Omega_k) - G(\Omega_k, \hat{\theta})\right\} = \sigma_G^2(k) - \sigma_G^2(\Omega_k, \hat{\theta}) - \Delta_G(k), \tag{3.9}$$

with $\Delta_G(k)$ a term that is proportional to the model errors. If there are no

model errors, this expression can be further reduced to

$$\mathrm{Var}\left\{G(\Omega_k) - G(\Omega_k, \hat{\theta})\right\} = \sigma_G^2(k) - \sigma_G^2(\Omega_k, \hat{\theta}). \tag{3.10}$$

It is important to remark here that the variance on the residuals is not given by the sum of the measurement and the model variance, but instead the difference should be taken.

Practical application: In general $\sigma_G^2(\Omega_k, \hat{\theta}) \ll \sigma_G^2(k)$ so that the compensation in (3.10) can be neglected. $\sigma_G^2(\Omega_k, \hat{\theta})$ can only become of the same order as $\sigma_G^2(k)$ at those frequencies where the model is very flexible and depends only on a few data points (e.g. very sharp resonances). Since in this situation both terms in (3.10) cancel each other, the result becomes extremely sensitive to model errors. Expression (3.10) can even become negative! In that case the general expression (3.9) should be used. However, since the model errors are not accessible, this is impractical and leads us to the following conclusions: use $\sigma_G^2(k)$ as the uncertainty on the residuals. If $\sigma_G^2(\Omega_k, \hat{\theta}) \approx \sigma_G^2(k)$ the user should accept that in that region he cannot detect the presence of model errors since he has no reliable estimate of the residual uncertainty to decide if they are significantly different from zero or not.

Example 3.2.3: Uncertainty bounds on the simulation poles/zeros (only for information, no questions on the exam)

In this example, we illustrate that the calculation of (co-)variances through linearization of a heavely nonlinear function may fail. We want to know the uncertainty on the poles and zeros of an identified transfer function, once the transfer function parameters θ are estimated. The transfer function is modeled as

$$G(\Omega, \theta) = \frac{B(\Omega, \theta)}{A(\Omega, \theta)} = \frac{\sum_{r=0}^{n_b} b_r \Omega^r}{\sum_{r=0}^{n_a} a_r \Omega^r} \tag{3.11}$$

where $\Omega = s$ for continuous-time models, $\Omega = z^{-1}$ for discrete-time models.

The dispersion of the estimated parameters $\hat{\theta}$ around their mean value is given by the covariance matrix C_θ. Assuming that the estimates are normally distributed, the most compact uncertainty regions are ellipses. Practice has shown that this is a very usable description for realistic signal-to-noise ratios if θ are the coefficients of the numerator and denominator polynomials of the transfer function model. In the previous section it was shown how to calculate the covariance matrix of related system characteristics using linear approximations. However, if the user is interested in the uncertainty of the poles/zeros of the estimated system, it turns out that this linearization may fail. Even for high signal-to-noise ratios, the uncertainty ellipses calculated for the poles and zeros may not cover the true uncertainty regions. This is illustrated in the following simulation example. Consider the system $G(s)$ with zeros $-1.4355 \pm j4.0401$, and poles $1.3010 \pm j4.8553$, $-3.5543 \pm j3.5543$, $-4.8553 \pm j1.3010$. The system has one dominating pole/zero pair and two pole pairs that have a smaller impact on the system. The transfer function is measured in 101 equidistant points between 0 and1.25 rad/s with a signal-to-noise ratio of 40 dB ($\sigma_G(k) = |G(j\omega_k)|/100$). 10000 realizations were generated and for each of them the poles/zeros were calculated and shown in Figure 3.1. Also the "classical" 95% confidence ellipsoids calculated using eq. (3.1) are shown (see Guillaume *et al.,* 1989). In this figure it is clearly seen that not only the shape of the uncertainty regions differs significantly from the ellipsoids (for the non dominating poles), but even for the dominating pole/zeros the uncertainties are significantly underestimated.

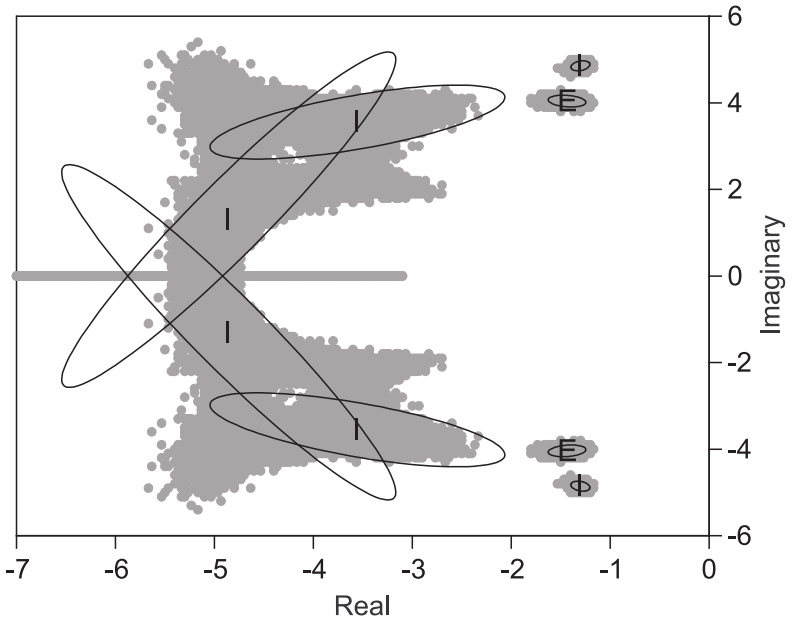


Figure 3.1: 95% confidence ellipsoids compared to the estimated poles and zeros of 10000 simulations.

3.3 Avoiding overmodelling

3.3.1 Introduction: impact of an increasing number of parameters on the uncertainty

(**only for information, no questions on the exam)** In this section we look into the dependency of the model variability on the model complexity. During the modelling process it is often quite difficult to decide whether the introduction of a new parameter is meaningful or not. A simple strategy would be to fit all the parameters which could be of possible interest, but this is not a good idea because the uncertainty on the estimates will then be increased. Consider a model with a partitioned set of parameters $\theta = (\theta_1, \theta_2)$. What is the impact on the model uncertainty if the simple model $G(\theta_1)$ is extended to the more complex one $G(\theta_1, \theta_2)$? In Section 1.3.2 it was already illustrated that the uncertainty will increase. Below, it is shown that this is a general result.

Consider the information matrix of the full model:

$$\mathrm{Fi}(\theta_1, \theta_2) = \begin{pmatrix} \mathrm{Fi}_{11} & \mathrm{Fi}_{12} \\ \mathrm{Fi}_{22} & \mathrm{Fi}_{22} \end{pmatrix}. \tag{3.12}$$

The covariance matrix of the simple model is $C(\theta_1) = \mathrm{Fi}_{11}^{-1}$, while the covariance matrix of the complete model is $C(\theta_1, \theta_2) = \mathrm{Fi}^{-1}$. The covariance matrix C_{θ_1} of the subset θ_1 is related to the covariance matrix $C(\theta_1)$ of the complete set by

$$C_{\theta_1} = \mathrm{Fi}_{11}^{-1} + \mathrm{Fi}_{11}^{-1}\,\mathrm{Fi}_{12}(\mathrm{Fi}_{22} - \mathrm{Fi}_{12}\,\mathrm{Fi}_{22}^{-1}\,\mathrm{Fi}_{12})^{-1}\,\mathrm{Fi}_{21}\,\mathrm{Fi}_{11}^{-1} = C(\theta_1) + \Delta. \tag{3.13}$$

Since $\Delta \geq 0$ it is clear that adding additional parameters to a model increases its uncertainty. A similar result is available for transfer function estimation.

3.3.2 Balancing the model complexity versus the model variability.

In the previous section it was illustrated that the systematic errors reduce with increasing model complexity. However, at the same time the model variability increases as shown in eq. (3.13). In practice the optimal complexity should be selected from the available information. Usually this choice is based on the evolution of the cost function. However, care should be taken. Adding aditional parameters will decrease the minimum value of the cost function (if the optimization routine does not get stuck in a local minimum). But, from a given complexity, the additional parameters no longer reduce the systematic errors, they are only used to follow the actual noise realization on the data. Since these vary from measurement to measurement, they only increase the model variability. Many techniques were proposed to avoid this unwanted behavior. These are based on extending the cost function with a model complexity term that estimates and compensates for the unwanted increasing model variability.

Two popular methods are actually in use, the AIC (Akaike information criterion) and the MDL (minimum description length):

$$\begin{array}{ll} \text{AIC}: & V_{\text{ML}}\left(\hat{\theta}_{\text{ML}}(Z), Z\right) + n_\theta \\ \text{MDL}: & V_{\text{ML}}\left(\hat{\theta}_{\text{ML}}(Z), Z\right) + \frac{n_\theta}{2}\ln\left(2\alpha N\right) \end{array} \tag{3.14}$$

with n_θ the number of identifiable (free) model parameters (= total number of parameters minus the number of constraints), and N is the number of real measurements (one complex measurement counts for 2 real ones). $\alpha = 1$ for output error problems, while $\alpha = 2$ for the errors-in-variables problem. V_{ML} is the non normalized cost function(not devided by the number of data points), e.g.

$$V_{\text{ML}} = \frac{1}{2}\sum_{k=1}^{N} e_k^2(\theta) \tag{3.15}$$

In practice it is better to use the following rules:

$$\begin{array}{ll} \text{AIC}: & V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)\left(1 + \frac{2n_\theta}{N}\right) \\ \text{MDL}: & V_{\text{ML}}(\hat{\theta}_{\text{ML}}(Z), Z)\left(1 + \frac{n_\theta}{N}\ln\left(2\alpha N\right)\right) \end{array} \tag{3.16}$$

The AIC rule is derived in more detail in 3.3.3. The MDL has a much better reputation than AIC. This is also illustrated in the following examples.

3.3.3 Proof of the AIC criterium

In this section, the AIC-model selection rule is proven. The proof is explicitely given (under simplified conditions) because it also gives some additional insight in the cost function behaviour.

Important note: since we will consider limits for N going to infinity, we use normalized cost functions in this section, for example:

$$V_N = \frac{1}{2N}\sum_{k=1}^{N} e_k^2(\theta) \text{ instead of } V_{\text{ML}} = \sum_{k=1}^{N} e_k^2(\theta) \tag{3.17}$$

Consider a (Maximum Likelihood) estimator, with a cost function

$$V_N(\theta, z) \text{ if } N \text{ data points are available.} \tag{3.18}$$

We assume that the standard conditions are met, such that this estimator is consistent and its covariance matrix converges to the Cramér-Rao lower bound (the last condition will only be used in the second part of the proof).

Consider also the limiting situation, for an infinite amount of data:

$$V_*(\theta, z) = \lim_{N\to\infty} V_N(\theta, z) \tag{3.19}$$

The estimates are the minimizers of these cost functions:

$$\hat{\theta}_N(z) = \arg\min_{\theta} V_N(\theta, z), \text{ and } \hat{\theta}_*(z) = \arg\min_{\theta} V_*(\theta, z) \tag{3.20}$$

For notational simplicity, we drop the argument in $\hat{\theta}$ from here on. Note, that θ_* equals the true value of the parameters because it is a consistent estimate by assumption, and no model errors are considered here, we deal with the problem of overmodelling.

In the next step we will first calculate the Taylor series of $V_*(\hat{\theta}_N, z)$ around $\hat{\theta}_*$. Next we setup the Taylor series of $V_N(\hat{\theta}_*, z)$ around $\hat{\theta}_N$. Notice that

$$V'_*(\hat{\theta}_*, z) = \left.\frac{\partial V_*(\theta, z)}{\partial \theta}\right|_{\theta=\hat{\theta}_*} = 0, \text{ and } V'_N(\hat{\theta}_N, z) = \left.\frac{\partial V_N(\theta, z)}{\partial \theta}\right|_{\theta=\hat{\theta}_N} = 0 \tag{3.21}$$

because $\hat{\theta}_N, \hat{\theta}_*$ are the minimizers of there respective costfunctions that are assumed to be derivable in their minimum. So we get that:

$$V_*(\hat{\theta}_N, z) = V_*(\hat{\theta}_*, z) + \frac{1}{2}\left(\hat{\theta}_N - \hat{\theta}_*\right)^{\mathrm{T}} V''_*(\hat{\theta}_*, z)\left(\hat{\theta}_N - \hat{\theta}_*\right) \tag{3.22}$$

$$= V_*(\hat{\theta}_*, z) + \frac{1}{2}\mathrm{trace}\left(V''_*(\hat{\theta}_*, z)\left(\hat{\theta}_N - \hat{\theta}_*\right)\left(\hat{\theta}_N - \hat{\theta}_*\right)^{\mathrm{T}}\right) \tag{3.23}$$

and

$$V_N(\hat{\theta}_*, z) = V_N(\hat{\theta}_N, z) + \frac{1}{2}\left(\hat{\theta}_* - \hat{\theta}_N\right)^{\mathrm{T}} V_N''(\hat{\theta}_N, z)\left(\hat{\theta}_* - \hat{\theta}_N\right) \tag{3.24}$$

$$= V_N(\hat{\theta}_N, z) + \frac{1}{2}\,\mathrm{trace}\left(V_N''(\hat{\theta}_N, z)\left(\hat{\theta}_* - \hat{\theta}_N\right)\left(\hat{\theta}_* - \hat{\theta}_N\right)^{\mathrm{T}}\right) \tag{3.25}$$

$$= V_N(\hat{\theta}_N, z) + \frac{1}{2}\mathrm{trace}\left(V_N''(\hat{\theta}_N, z)\left(\hat{\theta}_N - \hat{\theta}_*\right)\left(\hat{\theta}_N - \hat{\theta}_*\right)^{\mathrm{T}}\right) \tag{3.26}$$

Taking the expectation of (3.22) results in

$$\mathrm{E}\left\{V_*(\hat{\theta}_N, z)\}\right\} = V_*(\hat{\theta}_*, z) + \frac{1}{2}\mathrm{trace}(V_*''(\hat{\theta}_*, z)\mathrm{E}\left\{\left(\hat{\theta}_\mathrm{N} - \hat{\theta}_*\right)\left(\hat{\theta}_\mathrm{N} - \hat{\theta}_*\right)^{\mathrm{T}}\right\}) \tag{3.27}$$

$$= V_*(\hat{\theta}_*, z) + \frac{1}{2}\mathrm{trace}\left(V_*''(\hat{\theta}_*, z)C_{\hat{\theta}_N}\right) \tag{3.28}$$

A similar operation can also be applied to (3.24). Moreover, if we assume (APPROXIMATION!) that

$$\mathrm{E}\left\{\mathrm{V_N}(\hat{\theta}_*, \mathrm{z})\right\} \approx \mathrm{E}\left\{V_*(\hat{\theta}_*, z)\right\} = V_*(\hat{\theta}_*, z) \tag{3.29}$$

and

$$V_N''(\hat{\theta}_N, z) \approx V_*''(\hat{\theta}_*, z), \tag{3.30}$$

we get that

$$V_*(\hat{\theta}_*, z) \approx \mathrm{E}\left\{V_N(\hat{\theta}_N, z)\}\right\} + \mathrm{trace}\left(V_*''(\hat{\theta}_*, z)C_{\hat{\theta}_N}\right) \tag{3.31}$$

So that eventually the following set of relations is obtained:

$$\mathrm{E}\left\{V_*(\hat{\theta}_N, z)\right\} = V_*(\hat{\theta}_*, z) + \frac{1}{2}\,\mathrm{trace}\left(V_*''(\hat{\theta}_*, z)C_{\hat{\theta}_N}\right) \tag{3.32}$$

$$\mathrm{E}\left\{V_N(\hat{\theta}_N, z)\right\} = V_*(\hat{\theta}_*, z) - \frac{1}{2}\,\mathrm{trace}\left(V_*''(\hat{\theta}_*, z)C_{\hat{\theta}_N}\right) \tag{3.33}$$

This is a remarkable result. From the first expression we learn that $\mathrm{E}\left\{V_*(\hat{\theta}_N, z)\right\} > V_*(\hat{\theta}_*, z)$, or in other words, $\hat{\theta}_*$ gives a better description of an infinite number of

data than $\hat{\theta}_N$. This is a quite intuitive result. More surprisingly, at first glance, is that $\mathrm{E}\left\{V_N(\hat{\theta}_N, z)\right\} < V_*(\hat{\theta}_*, z)$. The N data points are better described by their own minimizer θ_N than by the true parameters. This is because in the latter case, a larger part of the noise can be followed by the model parameters. With an infinite number of parameters this becomes impossable at all. Note that this result is obtained without using the 2nd condition (covariance matrix equals the Cramér-Rao lower bound). So this result is also valid for (weighted) least squares estimates that are consitent.

From (3.32) it follows immediately that

$$\mathrm{E}\left\{V_*(\hat{\theta}_N, z)\right\} = \mathrm{E}\left\{V_N(\hat{\theta}_N, z)\right\} + \mathrm{trace}\left(V_*''(\hat{\theta}_*, z) C_{\hat{\theta}_N}\right) \tag{3.34}$$

If V is the maximum likelihood cost function (properly scaled with $1/N$), we get that

$$C_{\mathrm{CR}} = -\left(\frac{\partial^2}{\partial\theta^2}\mathrm{loglikelihood}\right)^{-1} = V_{\mathrm{ML}}''(\theta_*)^{-1} = \left(NV''(\theta_*)\right)^{-1} \tag{3.35}$$

and 3.34 becomes

$$\mathrm{E}\left\{V_*(\hat{\theta}_N, z)\right\} = \mathrm{E}\left\{V_N(\hat{\theta}_N, z)\right\} + \frac{1}{N}\,\mathrm{trace}\left(V_*''(\hat{\theta}_*, z) V_*''(\hat{\theta}_*, z)^{-1}\right) \tag{3.36}$$

$$= \mathrm{E}\left\{V_N(\hat{\theta}_N, z)\right\} + \frac{n_\theta}{N} \tag{3.37}$$

It is a logic choice to consider the value $\mathrm{E}\left\{V_*(\hat{\theta}_N, z)\right\}$ as a measure of the model quality: 'How well would the model describe an infinite amount of data'? The answer to that question is exactly given in 3.36. Of course we do not know in practice $\mathrm{E}\left\{V_N(\hat{\theta}_N, z)\right\}$, only the realized value $V_N(\hat{\theta}_N, z)$. And this becomes the AIC rule as it used in practice: To compare two models with different complexity their respective cost functions are compared after adding a complexity term n_θ/N.

For normaly distributed noise, the weighted least squares cost function is

often used without the normalizing factor.

$$V_{\mathrm{WLS}} = e^{\mathrm{T}} C_e^{-1} e. \tag{3.38}$$

In that case the AIC rule becomes

$$V_{\mathrm{AIC}} = \frac{1}{2} e^{\mathrm{T}} C_e^{-1} e + n_\theta \tag{3.39}$$

3.4 Example of using the AIC-rule

- How to select a 'good' model complexity?
- Effect of the number of parameters on the value of the cost function?
- Optimal order can depend on the signal-to-noise ratio!

The goal of this section is to show how to select an optimal model for a given data set. Too simple models will fail to capture all important aspects of the output, and this will result in too large errors in most cases. Too complex models use too many parameters. As was illustrated in the previous section such models also result in a poor behavior of the modeled output because the model becomes too sensitive to the noise. Hence we need a tool that helps us to select the optimal complexity that balances the model errors against the sensitivity to the noise disturbances. It is clear that this choice will depend on the quality of the data. All these aspects are illustrated in the next exercise where we propose the Akaike information criterion as a tool for model selection. Consider a single input single output linear dynamic system, excited with an input $u_0(t)$ and output $y_0(t) = g_0(t) * u(t)$. The system has an impulse response $g_0(t)$ that is infinitely long (infinite impulse response or IIR-system). For a stable system $g_{0(t)}$ decays exponentially to zero, so that the IIR system can be approximated by a system with a finite length impulse response $g(t)$, $t = 0, 1, \ldots, I$ (finite impulse response or FIR-system). For $t > I$, the remaining contribution can considered to be negligible. The choice of I will depend not only on $g(t)$, but

also on the SNR of the measurements.

$$\hat{y}(t) = \sum_{k=0}^{I} \hat{g}(k) u_0(t-k), \text{ with } u_0(k) = 0 \text{ for } k < 0. \tag{3.40}$$

In (3.40) it is assumed that the system is initially in rest. If this is not the case, transient errors will appear, but these disappear in this model for $t > I$ (why?). The model parameters θ are in this case the values of the impulse response. θ is estimated from the measured data $u_0(t), y(t),\ t = 0, 1, \ldots, N$, with $y(t)$ the output measurement that is disturbed with i.i.d. noise with zero mean and variance σ_v^2:

$$y(t) = y_0(t) + v(t). \tag{3.41}$$

The estimates $\hat{\theta}$ are estimated by minimizing the least squares cost function:

$$V_N(\theta, Z^N) = \frac{1}{2N} \sum_{t=0}^{N} |y(t) - \hat{y}(t)|^2, \text{ with } \hat{y}(t) = \hat{g}(t) * u_0(t) \tag{3.42}$$

Note that this model is linear-in-the-parameters, and solution (**??**) can be used. In order to find the 'best' model, a balance is made between the model errors and the noise errors using a modified cost function that accounts for the complexity of the model. Here we propose to use amongst others the AIC criterion:

$$V_{\text{AIC}} = V_N(\theta) \left(1 + 2\frac{\dim \theta}{N}\right). \tag{3.43}$$

3.4.1 Exercise: Model selection using the AIC criterion

Consider the discrete time system $g_0(t)$ given by its transfer function:

$$G_0(z) = \frac{\sum_{k=0}^{n_b} b_k z^{-k}}{\sum_{k=0}^{n_a} a_k z^{-k}}, \tag{3.44}$$

Generate the filter coefficients a_k, b_k using the MATLAB™-instruction

```
[b, a] = cheby1(3, 0.5, [2 * 0.152 * 0.3]);
```

This is a band pass system with a ripple of 0.5 dB in the pass band. Generate

two data sets D_{est} and D_{val}, the former with length N_e being used to identify the model, the latter with length N_v to validate the estimated model. Note that the initial conditions for both sets are zero! Use the MATLAB™-instructions

```
y0 = filter(b, a, u0);

y = y0 + ny;
```

with u_0 zero mean normally distributed noise with $\sigma_{u_0} = 1$, and v zero mean white Gaussian noise with σ_v equal to 0.5 for a first experiment, and 0.05 for a second experiment. Put $N_{\text{e}} = 1000$, and $N_{\text{val}} = 10000$ in both experiments.

- Use the linear least squares procedure to estimate the model parameters, and this for varying orders from 0 to 100.
- Calculate for each of the models the simulated output $\hat{y} = \text{filter}(\hat{g}, 1, u_0)$, and calculate the cost function(3.42) on D_{est} and on D_{val}.
- Calculate V_{AIC}.
- Calculate $V_0 = \frac{1}{2N} \sum_{t=0}^{N} |y_0(t) - \hat{y}(t)|^2$ on the undisturbed output of the validation set.
- Plot $V_{\text{est}}, V_{AIC}, V_{\text{val}}$ as a function of the model order. Normalize the value of the cost function by σ_v^2 to make an easier comparison of the behavior for different noise levels.
- Plot $\sqrt{V_0/\sigma_v^2}$ as a function of the model order.

Observations: The results are shown in 3.2, the following observations can be made:

1. Increasing the model order results in a monotonic decreasing cost function V_{est}. This result was to be expected because a simpler model is always included by the more complex model, and the linear LS always retrieve the absolute minimum of the cost function, no local minima exist. Hence increasing the complexity of the model should reduce the value of the cost function.

2. On the validation data we observe first a decrease and next an increase of V_{val}. In the beginning, the additional model complexity is mainly used to reduce the model errors, a steep descent of the cost function is observed. From a given order on, the reduction of the model errors is smaller than the increased noise sensitivity due to the larger number of parameters, resulting in a deterioration of the capability of the model to simulate the validation output. As a result the validation cost function V_{val} starts to increase.

3. V_{AIC} gives a good indication, starting from the estimation data only, where V_{val} will be minimum. This reduces the need for long validation records, and it allows to use as much data as possible for the estimation step.

4. The optimal model order increases for a decreasing disturbing noise variance. Since the plant is an IIR system with an infinite long impulse response, it is clear that in the absence of disturbing noise $\sigma_n = 0$, the optimal order would become infinite. In practice this value is never reached due to the presence of calculation errors that act also as a disturbance.

5. A fair idea about the quality of the models is given by V_0. The normalized rms-value $\sqrt{\frac{V_0}{\sigma_v^2}}$ is plotted on the right side of 3.2. This figure shows that a wrong selection of the model can result in much larger simulation errors. The good news is that the selection of the best model order is not so critical, the minimum is quite flat and all model orders in the neighborhood of the minimum result in good estimates.

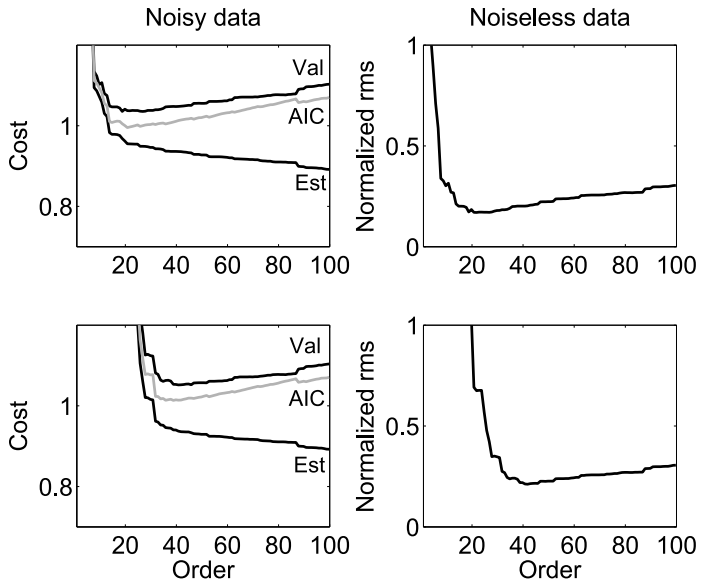


Figure 3.2: Right side: Comparison of the normalized Cost function V_{est}, the AIC-criterion V_{AIC}, and the validation V_{val} for $\sigma_n = 0.5$ (top) and $\sigma_n = 0.05$ (bottom). Left side: evaluation of the model quality on undisturbed (noiseless) data.

Chapter 4

Numerical Optimization Methods

Chapter 4

Numerical Optimization Methods

Abstract: In this chapter, some numerical optimization methods are presented to search the minimum of the cost functions. A vast amount of literature on this topic is available. Here we stick to only a few methods, that are very well suited to take full advantage of the special structure of the cost functions as they appeared in Chapter 2.

4.1 Introduction

If the model is nonlinear-in-the-parameters, it is almost impossible to find an analytical solution for the minimization of the cost function $V(\theta, N)$ (For notational simplicity, we will drop in this chapter the dependency on N, and write $V(\theta)$ instead of $V'(\theta, N)$). As a result, numerical algorithms have to be used in calculating the estimates. The main drawback of this is that insight into the behavior of the solution is partially lost.

The numerical search routines are usually iterative procedures. Starting from an initial value, a better set of parameters is generated, and this process is repeated until it is decided that the process has converged. So, there will be three basic steps in each search routine:

- selection of a set of starting values
- generation of an improved set of parameters
- selection of a stop criterion

4.1.1 Selection of the starting values

Before beginning a non-linear optimization procedure it is necessary to generate starting values. In practice the convergence region of most optimization methods is limited; if the starting values are selected outside this region, the method will diverge. Even if there is convergence, the final result can depend upon the starting values if the cost function has local minima. A priori information can be used to improve the starting values, but usually insufficient information is available. For example, it is very difficult to give reasonable starting values for the coefficients of the transfer function of an unknown system. A more systematic approach is to linearize the original optimization problem, to calculate an analytical solution to the linearized form, and to use this as an approximation of the true values of the parameters.

4.1.2 Generation of an improved set of parameters

The generation of an improved set of parameters is the kernel of each optimization method. For the purposes of this course we will give only a very brief introduction to the different possible techniques. Here we will classify the methods by the highest order of the derivatives they use for the generation of a new value.

Zero order methods These do not require the explicit calculation of derivatives, and are called direct search techniques. Typical examples are the simplex method and Powell's method. They are usable if the derivatives cannot be easily calculated, and converge slowly but monotonically to a local minimum.

First order derivatives These are called gradient methods. In their most simple form they move in a direction opposite to the gradient towards a local minimum. They are also quite robust, but are not always faster than zero order methods as they require calculation of the gradient. Convergence can be speeded up by using more complicated forms, like the conjugated gradient and the Gauss-Newton method. This will be especially important if the minimum is hidden in a long narrow valley; in the neighborhood of such a solution these methods have a quadratic convergence for noiseless cases.

Second order derivatives A function can be approximated around a minimum by a parabolic function which uses the second order derivatives (the first order derivatives are equal to zero) if certain regularity conditions are met. For an exactly parabolic function it is possible to locate the minimum in one step, starting from any arbitrary point where the second order derivatives are known. This principle can also be applied to non-parabolic functions, but then more than one step will be required to reach the minimum. The best known optimization method of this class is the Newton-Raphson algorithm, which will be discussed in more detail later.

4.1.3 Deduction of a stop criterion.

The iteration loop of the optimization method should be stopped at the moment that the stop criterion is met. There are a number of ways of choosing this criterion.

If the results of the identification will be used to predict the output of the system for a given input, then the most important requirement of the model is its prediction capability; the cost function is a measure of this. If there is no longer a significant decrease of the cost function with iteration, then the prediction quality will not be further improved by continuing. We can use this as a stop criterion, even though some parameters may still be changing considerably. The implication of this is that the loss function is insensitive to a change in this group of parameters. This kind of criterion can be extended by tests on the derivatives of the cost function.

If it is the physical interpretation of the parameters and not the prediction ability which is most important, then such a criterion is not acceptable; a check on the variations of the parameters will be better.

4.2 Gradient method.

The gradient method is one of the most simple optimizations. The principle is illustrated in Figure 4.1 where lines of equal cost are shown. The gradient is defined as:

$$\text{grad}V(\theta) = V'(\theta) = \frac{\partial V(\theta)}{\partial \theta} = \left(\frac{\partial V}{\partial \theta_1}, \frac{\partial V}{\partial \theta_2}, \ldots, \frac{\partial V}{\partial \theta_{n_\theta}} \right) \tag{4.1}$$

The gradient is orthogonal to these iso-cost lines and points in the direction of maximum positive slope; the algorithm takes a scaled step in the opposite direction.

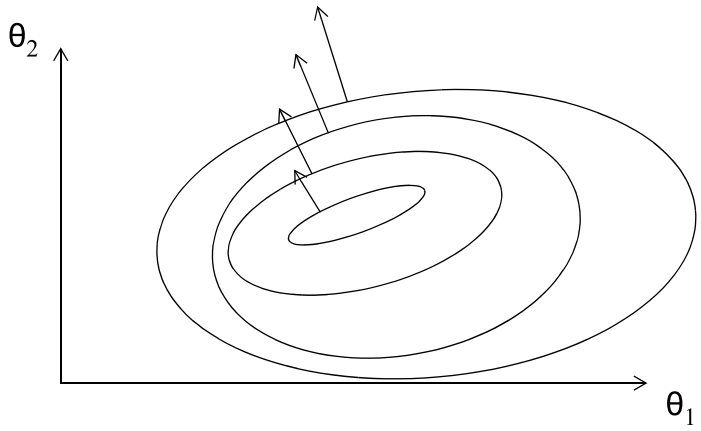


Figure 4.1: Illustration of iso-costlines and the gradient of a cost functions.

In some algorithms, the gradient is numerically approximated by:

$$\text{grad}V(\theta) \approx \left(\frac{V(\theta + \delta_1) - V(\theta)}{|\delta_1|}, \ldots, \frac{V(\theta + \delta_{n_\theta}) - V(\theta)}{|\delta_{n_\theta}|} \right), \tag{4.2}$$

where δ_i is a unit vector with a 1 on the i^{th} element.

Once an estimate of the gradient is available, the basic algorithm is quite simple:

1. Select a set of starting value $\theta^{(0)}$, and choose also a maximum step length $\delta_{\max}$, put the iteration number $i = 0$.

2. Calculate the gradient $V'(\theta^{(i)})$

3. Calculate an updated parameter vector

$$\theta^{(i+1)} = \theta^{(i)} - \delta_{\max} \frac{V'(\theta^{(i)})}{\left|V'(\theta^{(i)})\right|} \tag{4.3}$$

4. Step 4: is this an improvement?

$$V(\theta^{(i+1)}) \leq V(\theta^{(i)})? \tag{4.4}$$

Yes: $i = i + 1$, and jump to Step 2, unless the stop criterion is met

No: reduces $\delta_{\max}$ and go back to Step 3.

Remarks:

1. The variation of the maximum step can be refined. Usually the step size is drastically reduced (e.g. a factor 10), and next it is slowly increased in each successive step (e.g. increase each time with 20%, but avoid that the scale factor is becoming larger than 1).

2. The simple algorithm can get stuck into long, narrow and steep valleys. It starts to jump from one side to the other side of the valley, and moves only very slowly in the desired direction.

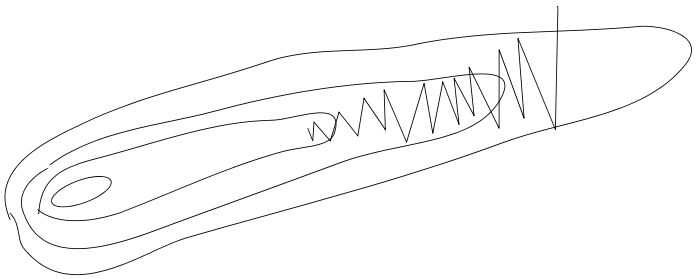


Figure 4.2: Example of a gradient search that gets stuck in a sharp valley.

3. This simple gradient method can be refined in many ways to avoid the above mentioned problem and to speed up the convergence. For example by making a one dimensional search in the direction of the gradient, or by combining the results of successive steps, the convergence rate can be significantly improved.

4.3 Newton-Raphson algorithm

The most important drawback of the gradient method is its slow convergence. This can be speeded up by using higher order derivatives. Consider the Taylor expansion of the cost function:

$$V(\theta + \delta) = V(\theta) + V'(\theta)\delta + \frac{1}{2}\delta^{\mathrm{T}} V''\delta + \dots \tag{4.5}$$

From now on, we neglect the higher order terms. The derivatives have to be taken in θ. A necessary condition for having an extreme for V in $\theta + \delta$ is that the derivative with respect to δ in $\theta + \delta$ should be equal to zero.

$$\frac{\partial V(\theta + \delta)}{\partial \delta} = 0 + V'(\theta) + V''\delta \tag{4.6}$$

The solution of this linear set of equations gives the value of δ:

$$\delta = -V''(\theta)^{-1} V'(\theta) \tag{4.7}$$

The most simple Newton-Raphson algorithm consists of calculation of successive values of δ:

1. Select a set of starting value $\theta^{(0)}$, put $i = 0$
2. Calculate the gradient $V'(\theta^{(i)})$, and the Hessian matrix $V''(\theta^{(i)})$
3. Calculate an updated parameter vector

$$\delta^{(i)} = -V''(\theta^{(i)})^{-1} V'(\theta^{(i)}), \tag{4.8}$$

$$\theta^{(i+1)} = \theta^{(i)} + \delta^{(i)} \tag{4.9}$$

4. If the stop condition is not met, go back to Step 2.

This method will have a quadratic convergence in the neighborhood of the so-

lution. A method is said to converge with order m if

$$\left|\delta^{(i)}\right| = \text{Constant } \left|\delta^{(i-1)}\right|^m \tag{4.10}$$

for i sufficiently large. The biggest difficulty with the Newton-Raphson method is that convergence is not guaranteed. Left multiplication of both sides of equation (4.8) by the gradient $V'(\theta)$ gives:

$$V'(\theta^{(i)})\delta^{(i)} = -V'(\theta^{(i)}) \left(V''(\theta^{(i)})^{-1} V'(\theta^{(i)})\right). \tag{4.11}$$

The left hand side should be negative to guarantee a successful step towards a minimum. However, if the Hessian matrix $V''(\theta^{(i)})$ on the right hand side is not positive definite, there is no guarantee that the expression will be negative, and the procedure can diverge. Only if the cost function is differentiable and the minimum is not a multiple solution (the Hessian matrix is not singular) will it be certain that the second order derivatives will be positive definite in the neighborhood of the solution.

Note that also here it is possible to improve the method by adding for example a line search in each step along the Newton direction. Another simple trick to improve the convergence in case of convergence problems (at a cost of a reduced convergence rate) is to update the parameters with only a fraction of the proposed step.

4.4 Gauss-Newton algorithm

A drawback of the Newton-Raphson method is the need to calculate second order derivatives, which can be quite time consuming. The Gauss-Newton method avoids these calculations by approximating the second order derivatives, making use of the quadratic nature of the cost function. Considering for example the nonlinear least squares cost function (the discussion can be generalized without any problem to weighted nonlinear least squares problems with multiple

outputs):

$$\begin{aligned} V_{\mathrm{LS}}(\theta) &= \tfrac{1}{2} e^{\mathrm{T}}(\theta) e(\theta) = \tfrac{1}{2} \left(y - g(u,\theta)\right)^{\mathrm{T}} \left(y - g(u,\theta)\right) \\ &= \tfrac{1}{2} \sum_{k=1}^{N} \left(y(k) - g(u(k),\theta)\right)^2 \end{aligned} \tag{4.12}$$

the second order derivatives are given by

$$V''_{\mathrm{LS}}(\theta) = \sum_{k=1}^{N} \left(\left(\frac{\partial g(u(k),\theta)}{\partial \theta} \right)^{\mathrm{T}} \left(\frac{\partial g(u(k),\theta)}{\partial \theta} \right) - \left(y(k) - g(u(k),\theta)\right) \frac{\partial^2 g(u(k),\theta)}{\partial \theta^2} \right) \tag{4.13}$$

If the second term in this sum becomes small (i.e. $y - g$ is small) it can be neglected, and the second order derivatives can be approximated by

$$V''_{\mathrm{LS}}(\theta) \approx \sum_{k=1}^{N} \left(\frac{\partial g(u(k),\theta)}{\partial \theta} \right)^{\mathrm{T}} \left(\frac{\partial g(u(k),\theta)}{\partial \theta} \right) = J^{\mathrm{T}} J, \tag{4.14}$$

$$\text{with } J_{[k,i]} = \frac{\partial g(u(k),\theta)}{\partial \theta_i} \text{ (the Jacobian matrix).} \tag{4.15}$$

Substitution of this approximation in the Newton-Raphson algorithm, and replacement of the first derivatives of V by

$$V'(\theta) = -J^{\mathrm{T}} \left(y - g(u,\theta)\right) \tag{4.16}$$

results in the Gauss-Newton method:

$$\delta^{(i)} = \left(J^{(i)\mathrm{T}} J^{(i)} \right)^{-1} J^{(i)\mathrm{T}} \left(y - g(u,\theta^{(i)}) \right), \text{ with } J^{(i)} = J(\theta^{(i)}) \tag{4.17}$$

For most problems the Gauss-Newton method demands less computation time per iteration step than the Newton-Raphson algorithm, and its behavior is quite similar. In the neighborhood of the solution the Newton-Raphson method will converge much faster than Gauss-Newton, except in problems where the term

$$\sum_{k=1}^{N} \left(y(k) - g(u(k),\theta)\right) \frac{\partial^2 g(u(k),\theta)}{\partial \theta^2} \tag{4.18}$$

which we have neglected in the expression for the Hessian goes to zero. But the convergence region of the Gauss-Newton method will be larger because we have replaced the Hessian, which is not guaranteed to be semi-positive definite, by a matrix which is semi-positive definite. A step of the Gauss-Newton algorithm will always be in the proper direction, except at saddle points and singularity points. However, convergence is still not assured; even if a step is in the right direction, the cost function can still increase if change in the parameters is too large.

Another method can be derived which will always converge, at least to a local minimum.

4.5 Method of Levenberg-Marquardt

The method of Levenberg-Marquardt is a combination of the Gauss-Newton method and the gradient method. New parameter updates are generated using the following formula:

$$\delta^{(i)} = \left(J^{(i)\mathrm{T}} J^{(i)} + \lambda \, \mathrm{I}_{n_\theta} \right)^{-1} J^{(i)\mathrm{T}} \left(y - g(u, \theta^{(i)}) \right). \tag{4.19}$$

The Gauss-Newton algorithm is the limit of the expression for λgoing to zero, while the gradient method is the limit for λgoing to $\times$:

$$\begin{aligned} &\lambda \to 0 \qquad \delta^{(i)} = \left(J^{(i)\mathrm{T}} J^{(i)} \right)^{-1} J^{(i)\mathrm{T}} \left(y - g(u, \theta^{(i)}) \right) \\ &\lambda \to \infty \qquad \delta^{(i)} = \tfrac{1}{\lambda} J^{(i)\mathrm{T}} \left(y - g(u, \theta^{(i)}) \right) = -\tfrac{1}{\lambda} \operatorname{grad} V_{\mathrm{LS}}(\theta) \end{aligned} \tag{4.20}$$

The Levenberg-Marquardt method moves progressively from the gradient method to the Gauss-Newton method as λ changes from $\times$ to zero. The addition of the term $\lambda \, \mathrm{I}_{n_\theta}$ has two stabilizing effects:

1. If the algorithm is converging towards a saddle point, then the matrix $J^{(i)\mathrm{T}} J^{(i)}$ which is semi-positive definite will become singular. Consequently, the set of linear equations to be solved is ill-conditioned, and it

will be very difficult to get good numerical solutions. Conditioning of the equations is considerably improved by adding the Levenberg-Marquardt term.

2. A second possible reason for divergence is if the proposed step is too large; even if the direction of the step is correct, the final result may be worse. By increasing λwe also reduce the step length, and move more in the direction of the gradient.

Instead of adding $\lambda \, \mathrm{I}_{n_\theta}$ to the equations, it is also possible to use $\lambda \, \mathrm{diag}\left(J^{(i)\mathrm{T}} J^{(i)}\right)$. The advantage of this approach is that if the original set of equations is diagonally dominant, sensitivity to the individual parameters remains unchanged.

The basic structure of a Levenberg-Marquardt method is in general quite simple:

1. Select a set of starting value $\theta^{(0)}$, and choose a large starting value for $\lambda^{(i)}$, put the iteration number $i = 0$.

2. Calculate the Jacobian matrix $J^{(i)}$

3. Calculate an updated parameter vector

$$\delta^{(i)} = \left(J^{(i)\mathrm{T}} J^{(i)} + \lambda \, \mathrm{I}_{n_\theta}\right)^{-1} J^{(i)\mathrm{T}} \left(y - g(u, \theta^{(i)})\right) \tag{4.21}$$

4. Is this an improvement: $V(\theta^{(i+1)}) \leq V(\theta^{(i)})$?

   Yes: decrease λ, e.g. with a factor 2; $i = i + 1$, and jump to Step 5

   No: increases λ, e.w. with a factor 10, and go back to Step 3, unless the stop criterion is met.

5. Go to Step 2, unless the stop criterion is met

4.6 Summary

In this section we have given a brief introduction to numerical iteration methods of minimizing the value of a cost function. There is an extensive literature on this subject; the few methods we have discussed were selected for their applicability to optimization problems we will meet later on.

Chapter 5

Recursive Identification Methods

Chapter 5

Recursive Identification Methods

In this chapter, recursive parameter estimation schemes are proposed. After each new sample, an update of the estimate is made. This allows online processing of the results. By adding a 'forgetting factor' to the cost function, the least squares least squares is generalized to an adaptive algorithm.

5.1 Introduction

In Chapter 2, a number of cost functions were proposed. The estimates were found as the minimizers of this cost function. Two possibilities exist: in the first case the optimization is postponed till all measurements are available, while in the second case the estimates are calculated each time a new sample is available. In this chapter we focuss on the second class of algorithms. A straightforward solution is to redo all the calculations after each sample. A numerical more efficient solution is to reformulate the problem such that only the newly required calculations are made, recuperating all the previous results. This chapter gives an introduction to this class of methods.

5.1.1 Eample: Recursive calculation of the mean value

The mean value μ_y has to be estimated from a series of measurements $y(k)$, $k = 1, \ldots, N$. These measurements come available one after another. The aim is to find an algorithm that allows to update the mean value estimate, each time a new measurement is available. An estimate for the mean value is:

$$\hat{\mu}_y(N) = \frac{1}{N} \sum_{k=1}^{N} y(k). \tag{5.1}$$

Once the new measurement $y(N+1)$ is available, a new estimate can be calculated:

$$\hat{\mu}_y(N+1) = \frac{1}{N} \sum_{k=1}^{N+1} y(k) \tag{5.2}$$

Instead of summing all the old measurements once more, the previous sum can be recuperated:

$$\begin{aligned} &\hat{\mu}_y(N+1) = \tfrac{1}{N+1} \textstyle\sum_{k=1}^{N} y(k) + \tfrac{1}{N+1} y(N+1) \\ &= \tfrac{N}{N+1} \hat{\mu}_y(N) + \tfrac{1}{N+1} y(N+1) \end{aligned} \tag{5.3}$$

Although this form meets already our requirements, it is possible to rearrange it to a more suitable expression:

$$\hat{\mu}_y(N+1) = \hat{\mu}_y(N) + \frac{1}{N+1} \left(y(N+1) - \hat{\mu}_y(N) \right). \tag{5.4}$$

Although this expression is very simple, it is very informative because almost every recursive algorithm can be reduced to a similar form. The following observations can be made:

- The new estimate $\hat{\mu}_y(N+1)$ equals the old estimate $\hat{\mu}_y(N)$ + a correction term: $\frac{1}{N+1} \left(y(N+1) - \hat{\mu}_y(N) \right)$.
- The correction term consists again of two terms: a gain factor $1/(N+1)$, and an error term.

    1. The gain factor decreases towards zero as more measurements are

already accumulated in the previous estimate. This means that in the beginning of the experiment less importance is given to the old estimate $\hat{\mu}_y(N)$, and more attention is paid to the new incoming measurements. When N starts to grow, the error term becomes small compared to the old estimate. The algorithm relies more and more on the accumulated information in the old estimate $\hat{\mu}_y(N)$, and it does not vary it that much for accidental variations of the new measurements. The additional bit of information in the new measurement becomes small compared with the information that is accumulated in the old estimate.

2. The second term $(y(N+1) - \hat{\mu}_y(N))$ is an error term. It makes the difference between the predicted value of the next measurement on the basis of the model (in this case $\hat{\mu}_y(N)$) and the actual measurement $y(N+1)$

- When properly initiated ($\hat{\mu}_y(1) = y(1)$), this recursive result is exactly equals to the nonrecursive implementation. However, from numerical point of view, it is a very robust procedure, because calculation errors etc. are compensated in each step.

The previous scheme is a special case of a more general class of algorithms: the stochastic approximation methods.

5.1.2 Stochastic approximation algorithms

From the discussion in the previous section, we learned that the recursive algorithm can be written as a combination of an old estimate and a correction term that is properly weighted. This result can be generaly formulated:

$$\hat{\theta}(N+1) = \hat{\theta}(N) - \frac{1}{2}K(N)\text{grad}V(N) \tag{5.5}$$

In this equation is

- $\hat{\theta}(N+1)$: the estimate after $N+1$ measurements
- $\hat{\theta}(N)$: the estimate after N measurements
- $K(N)$: the gain factor, balancing the new information versus the old one
- $V(N)$: the cost function that should be minimized, for example the squared difference between the new measurement and the model based prediction: $V(N) = \Big(y(N+1) - \hat{y}(N+1, \hat{\theta}(N))\Big)^2$

It is clear that not every choice of $K(N)$ will lead to a consistent estimate. It is proven that under quite weak conditions, $\hat{\theta}(N)$ is consistent if the following conditions on the gain are met:

$$K(N) \geq 0, \ \sum_{N=1}^{\infty} K(N) = \infty, \text{ and } \sum_{N=1}^{\infty} K^2(N) \quad \text{is finite.} \tag{5.6}$$

These conditions can be interpreted as follows:

- $K(N) \geq 0$: in order to minimize the cost function, the updates should move against the gradient. This is so if the gain is positive.
- $\sum_{N=1}^{\infty} K(N) = \infty$: This condition expresses that the gain can not go too fast to zero. This is necessary to account sufficiently for the new measurements.
- $\sum_{N=1}^{\infty} K^2(N)$ is finite: This condition imposes that the gain should go fast enough to zero in order to allow the estimate to converge for $N \to \infty$. The variance of the estimates can not converge to zero for $N \to \infty$ if this constraint is not met. The new measurements keep pulling the estimate.

A sequence that meets the previous conditions is

$$K(N) = \frac{1}{N^{\alpha}}, \text{ with } 0.5 < \alpha \leq 1. \tag{5.7}$$

The stochastic approximation algorithm is a very simple scheme. However it is possible to improve it. In the next section, the scalar gain factor $K(N)$ will

be replaced by a vector. This will allow for a faster convergence and more information about the uncertainty on the estimated parameters.

5.2 Recursive least squares with constant parameters

5.2.1 Problem statement

In this section we show how the linear least squares method of Section **??** can be reformulated into a recursive method. These techniques can also be generalized to models that are nonlinear in the parameters, but this is outside the scope of this course.

Consider a multiple input, single output system that is linear in the parameters:

$$y_0(k) = K(u_0)\theta_0, \tag{5.8}$$

with $u_0(k) \in \mathbb{R}^{1\times n_\theta}$, $y_0(k) \in \mathbb{R}^{1\times 1}$.

Noisy output observations are made:

$$y(k) = y_0(k) + n_y(k). \tag{5.9}$$

After collecting N measurements, the equations are under matrix form reformulated:

$$y_N = K_N\theta_0 + n_N, \tag{5.10}$$

with

$$y_N = (y(1), \ldots, y(N))^{\mathrm{T}} \in \mathbb{R}^{N\times 1}, \tag{5.11}$$

$$K_N = \begin{bmatrix} K(u_0, 1) \\ \vdots \\ K(u_0, N) \end{bmatrix} \in \mathbb{R}^{N\times n_\theta} \tag{5.12}$$

$$n_N = (n_y(1), \ldots, n_y(N))^{\mathrm{T}} \in \mathbb{R}^{N\times 1} \tag{5.13}$$

5.2.2 Recursive solution of the least squares problem

The least squares estimate for this problem is given by (2.14). For simplicity we drop the arguments in K.

$$\hat{\theta}_{\mathrm{LS}}(N) = \left(K_N^{\mathrm{T}} K_N\right)^{-1} K_N^{\mathrm{T}} y \tag{5.14}$$

$$= P_N K_N^{\mathrm{T}} y \tag{5.15}$$

with

$$P_N = \left(K_N^{\mathrm{T}} K_N\right)^{-1} \in \mathbb{R}^{n_\theta \times n_\theta}. \tag{5.16}$$

The following recurrence relation follows imediately from the definition:

$$P_N^{-1} = P_{N-1}^{-1} + K^{\mathrm{T}}(u_0, N) K(u_0, N) \tag{5.17}$$

$$K_N^{\mathrm{T}} y_N = K_{N-1}^{\mathrm{T}} y_{N-1} + K^{\mathrm{T}}(u_0, N) y(N) \tag{5.18}$$

Using the matrix inverse lemma[1], eq. (5.17)can be written as:

$$P_N = P_{N-1} - P_{N-1} K^{\mathrm{T}}(u_0, N) \left(K(u_0, N) P_{N-1} K^{\mathrm{T}}(u_0, N) + 1\right)^{-1} K(u_0, N) P_{N-1} \tag{5.19}$$

Note that $\left(K(u_0, N) P_{N-1} K^{\mathrm{T}}(u_0, N) + 1\right)$ is a scalar, so that its inversion requires no time consuming matrix inversions.

Substitution of these results in (5.14) results in

$$\hat{\theta}_{\mathrm{LS}}(N) = \left(K_N^{\mathrm{T}} K_N\right)^{-1} K_N^{\mathrm{T}} \tag{5.20}$$

$$= P_N K_N^{\mathrm{T}} y \tag{5.21}$$

$$= \Big(P_{N-1} - P_{N-1} K^{\mathrm{T}}(u_0, N) \left(K(u_0, N) P_{N-1} K^{\mathrm{T}}(u_0, N) + 1\right)^{-1} K(u_0, N) P_{N-1}\Big) \tag{5.22}$$

$$\times \left(K_{N-1}^{\mathrm{T}} y_{N-1} + K^{\mathrm{T}}(u_0, N) y(N)\right) \tag{5.23}$$

[1] Matrix inversion lemma: If $A^{-1} = B^{-1} + H^{\mathrm{T}} R^{-1} H$, then the following relation is valid: $A = B - BH^{\mathrm{T}} \left(HBH^{\mathrm{T}} + R\right)^{-1} HB$

This results eventually in the following set of recursive equations:

$$\hat{\theta}_{\mathrm{LS}}(N) = \hat{\theta}_{\mathrm{LS}}(N-1) - K_N \left(K(u_0, N)\hat{\theta}_{\mathrm{LS}}(N-1) - y(N) \right) \tag{5.24}$$

$$K_N = \frac{P_{N-1}K^{\mathrm{T}}(u_0, N)}{1 + K(u_0, N)P_{N-1}K^{\mathrm{T}}(u_0, N)} \tag{5.25}$$

$$P_N = P_{N-1} - \frac{P_{N-1}K^{\mathrm{T}}(u_0, N)K(u_0, N)P_{N-1}}{1 + K(u_0, N)P_{N-1}K^{\mathrm{T}}(u_0, N)} \tag{5.26}$$

These expressions allow a recursive implementation of the least squares. If properly initiated (choice of $\hat{\theta}_{\mathrm{LS}}(0), P_0$), they calculate exactly the same solution as the nonrecursive estimated.

5.2.3 Discussion

- The recursive structure, as it was indicated before (see Section 5.1.2) is clearly visible again. The new estimate $\hat{\theta}_{\mathrm{LS}}(N)$ is given by the old estimate $\hat{\theta}_{\mathrm{LS}}(N-1)$ plus an update $K_N \left(K(u_0, N)\hat{\theta}_{\mathrm{LS}}(N-1) - y(N) \right)$. The update consists of the difference between the predicted output measurement $K(u_0, N)\hat{\theta}_{\mathrm{LS}}(N-1)$, and the actual measured value $y(N)$. A vectorial scale factor K_N scales and directs this error signal into a parameter update.

- From Section 2.2.3 we know that for white and uncorrelated disturbing noise $\mathrm{Cov}\{n_y\} = \sigma_y^2 \mathrm{I}_N$ the $\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}\right\}$ is

$$\mathrm{Cov}\left\{\hat{\theta}_{\mathrm{LS}}(N)\right\} = \sigma_y^2 \left(K^{\mathrm{T}}K\right)^{-1} = \sigma_y^2 P_N. \tag{5.27}$$

- So we calculate the covariance matrix at the same time as the updated parameters.

- $P_N = \left(K_N^{\mathrm{T}}K_N\right)^{-1}$ decreases towards zero if the system is persistently excited during the experiment (loosely spoken, persistency means that the experiment is informative, on average, each new sample adds some additional information) because $K_N^{\mathrm{T}}K_N$ grows to infinity. This means

that the uncertainty on the parameters goes to zero.

- Also the gain factor K_N is proportional to P_N. Hence it also decays to zero.

5.3 Recursive least squares with (time)-varying parameters

5.3.1 Introduction

In many problems, the system to be modelled is slowly varying in time. Typical examples are: systems with dynamics that depend on the temperature (chemical plants), linear approximations to nonlinear systems around a changing working point, etc. Such a problem needs special care. A first approach could be to model the variations explicetily, but this increases the modelling effort significantly. An alternative is to built a model with varying parameters. The parameters should track the slow variations. Such an algorithm will be presented in this section. In principal the problem could be solved with the previous developped recursive least squares. By fixing the K to a small value (avoiding convergence towards zero), the new measurements will continue to update the parameters. However, it is possible to make a more systematic approach to the problem. Basically, we should avoid that the data collected long ago are still influencing the actual parameters, since these old measurements do no longer obey the same system equations as the fresh measurements (the system changed in time!). This should be expressed in the cost function. Here, an exponential forgetting factor is proposed, changing the original least squares cost function

$$V_{\mathrm{LS}}(\theta, N) = \frac{1}{2}\sum_{k=1}^{N} e_N^2(k,\theta), \text{ with } e_N(\theta) = y - K_N(u_0)\theta \tag{5.28}$$

to:

$$V_{\mathrm{EF}}(\theta, N) = \frac{1}{2}\sum_{k=1}^{N} g^{N-k} e_N^2(k,\theta), \text{ with } 0 \leq g \leq 1. \tag{5.29}$$

Because g is smaller than 1, measurements from a far past will be exponentially forgotten. Note that this can be interpreted as a first order filter with impuls response $h(l) = g^{\frac{l}{2}}$ acting on $e_N(k, \theta)$. This cost function can also be written using matrix notations:

$$V_{\mathrm{EF}}(\theta, N) = e_N^{\mathrm{T}} R e_N, \tag{5.30}$$

with

$$R_N = \begin{pmatrix} g^{N-1} & 0 & \dots & 0 & 0 \\ 0 & g^{N-2} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & g & 0 \\ 0 & 0 & 0\dots & 0 & 1 \end{pmatrix}. \tag{5.31}$$

5.3.2 The recursive solution

The solution of this weighted least squares problem is

$$\hat{\theta}_{\mathrm{EF}}(N) = \left(K_N^{\mathrm{T}} R_N K_N\right)^{-1} K_N^{\mathrm{T}} R_N y_N \tag{5.32}$$

Define

$$P_N = \left(K_N^{\mathrm{T}} R_N K_N\right)^{-1} \in \mathbb{R}^{n_\theta \times n_\theta}, \tag{5.33}$$

then we get that

$$P_N^{-1} = g P_{N-1}^{-1} + K^{\mathrm{T}}(u_0, N) K(u_0, N). \tag{5.34}$$

From here, the same steps can be made as in the previous section, applying the matrix inversion lemma with $B = \frac{1}{g} P_{N-1}$, and $R = 1$, results in

$$\begin{aligned} \hat{\theta}_{\mathrm{LS}}(N) &= \hat{\theta}_{\mathrm{LS}}(N-1) - K_N \left(K(u_0, N)\hat{\theta}_{\mathrm{LS}}(N-1) - y(N)\right) \\ K_N &= \frac{P_{N-1} K^{\mathrm{T}}(u_0, N)}{g + K(u_0, N) P_{N-1} K^{\mathrm{T}}(u_0, N)} \\ P_N &= \frac{1}{g}\left(P_{N-1} - \frac{P_{N-1} K^{\mathrm{T}}(u_0, N) K(u_0, N) P_{N-1}}{g + K(u_0, N) P_{N-1} K^{\mathrm{T}}(u_0, N)}\right) \end{aligned} \tag{5.35}$$

5.3.3 Discussion

Note that these equations are completely similar to those of the recursive least squares by putting $g = 1$.

The price to be paid for the additional flexibility of the model is an increased uncertainty on the estimates. Due to the forgetting factor, the information matrix P_N^{-1} is not growing anymore towards infinity. Each time a new measurement is added, a fraction of the old information is lost. The smaller g, the more information is lost, but the faster the algorithm can track parameter variations.

Chapter 6

Kalman Filtering

Chapter 6

Kalman Filtering

6.1 Introduction

In practice one frequently encounters situations where a signal disturbed by noise is to be measured. We want to reconstruct the original signal corresponding to the measured disturbed signal. One possibility to do so is by filtering the measured signal. This begs the question how one choses a suited filter. During World War II, this problem was studied by Wiener. His solution is known as 'Wiener filtering'. His proposed solution, however, was hard to implement in practice. During the early nineteen-sixties Kalman and Bucy deduced a more general solution. The problem they regarded can be posed as follows. Consider a system whose model is known. This system is excited by the known input signal $u(t)$ and it is disturbed by the noise source $v(t)$. We wish to study the system by its output quantities $y(t)$, but these observations are disturbed by a noise source $n(t)$. We wish to estimate the state $x(t)$ of the system from the measurements $y(t)$. Depending on what time instance we consider, this is called:

- $x(t-\tau)$: an interpolation problem,
- $x(t)$: a filtering problem,
- $x(t+\tau)$: an extrapolation or prediction problem.

6.2 Construction of the Kalman filter

In the general case, we can consider this case for nonlinear systems. For simplicity's sake, we only consider the linear problem. This can be described as follows.

Continuous state equations:

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t) + v(t) \\ y(t) &= Cx(t) + n(t) \end{cases} \tag{6.1}$$

In these equations x, u, v, y, n are vectors and A, B, C are matrices. A similar formulation can be used for discrete systems:

$$\begin{cases} x(k+1) &= Ax(k) + Bu(k) + v(k) \\ y(k) &= Cx(k) + n(k) \end{cases}$$

The corresponding block diagram is shown in Figure 6.1.

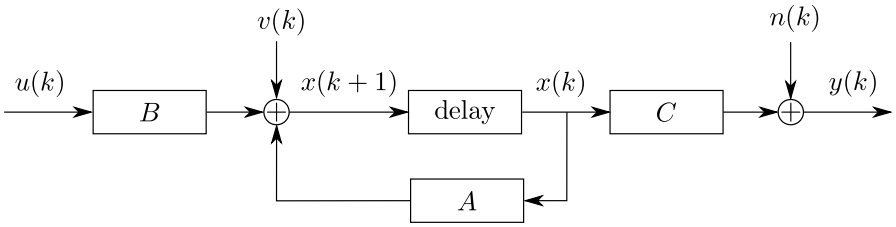


Figure 6.1: Block diagram of the state equation

The information that is available before solving this problem:

- A, B, C are known
- $\mathrm{E}\{v(k)\} = 0 \qquad \mathrm{Cov}\left\{v(k)\, v(j)^{\mathrm{T}}\right\} = R_v \delta_{kj}$
- $\mathrm{E}\{n(k)\} = 0 \qquad \mathrm{Cov}\left\{n(k)\, n(j)^{\mathrm{T}}\right\} = R_n \delta_{kj}$
- v and n are independent variables

In the following derivation, B is kept 0 for simplicity.

Let $Y(k) = \{y(1), y(2), \ldots, y(k)\}$

If we denote $\mathrm{E}\{x(k)|Y(k)\} = X(k)$

We can prove the following relation immediately:

$$\mathrm{E}\{x(k+1)|Y(k)\} = AX(k) \tag{6.2}$$

as

$$\begin{aligned}
\mathrm{E}\{x(k+1)|Y(k)\} &= \mathrm{E}\{Ax(k) + v(k)|Y(k)\} && (6.3)\\
&= A\,\mathrm{E}\{x(k)|Y(k)\} + \mathrm{E}\{v(k)|Y(k)\} && (6.4)\\
&= AX(k) && (6.5)
\end{aligned}$$

Denote $\mathrm{Cov}\{x(k+1)|Y(k)\} = \mathrm{E}\left\{(x(k) - X(k))(x(k) - X(k))^{\mathrm{T}}\right\} = P(k)$

Then we can prove that

$$\mathrm{Cov}\{x(k+1)|Y(k)\} = A \cdot P(k) \cdot A^{\mathrm{T}} + R_v = Q(k+1) \tag{6.6}$$

since

$$\operatorname{Cov}\{x(k+1)|\,Y(k)\} = \operatorname{Cov}\{Ax(k)+v(k)|\,Y(k)\} \tag{6.7}$$

$$= \mathrm{E}\left\{ (Ax(k)+v(k)-AX(k))\,(Ax(k)+v(k)-AX(k))^{\mathrm{T}} \middle| Y(k)\right\} \tag{6.8}$$

$$\begin{aligned} &= \mathrm{E}\left\{ A\cdot(x(k)-X(k))\,(x(k)-X(k))^{\mathrm{T}}\,A^{\mathrm{T}} \middle| Y(k)\right\} \\ &\quad + \mathrm{E}\left\{ v(k)\,x(k)^{\mathrm{T}} \middle| Y(k)\right\} A^{\mathrm{T}} + A\cdot \mathrm{E}\left\{ x(k)\,v(k)^{\mathrm{T}} \middle| Y(k)\right\} \\ &\quad - \mathrm{E}\left\{ v(k)\,X(k)^{\mathrm{T}} \middle| Y(k)\right\} A^{\mathrm{T}} - A\cdot \mathrm{E}\left\{ X(k)\,v(k)^{\mathrm{T}} \middle| Y(k)\right\} \\ &\quad + \mathrm{E}\left\{ v(k)\,v(k)^{\mathrm{T}} \middle| Y(k)\right\} \end{aligned} \tag{6.9}$$

$$= A\cdot \mathrm{E}\left\{ (x(k)-X(k))\,(x(k)-X(k))^{\mathrm{T}} \middle| Y(k)\right\} A^{\mathrm{T}} + R_v \tag{6.10}$$

$$= A\cdot P(k)\cdot A^{\mathrm{T}} + R_v = Q(k+1) \tag{6.11}$$

using the information we deduced above, it is possible to construct the probability density function of $x(k+1)$. In doing so an extension of Bayes' rule is needed.

$$\mathrm{P}\,(a,b,c) = \mathrm{P}\,(a|\,b,c)\,\mathrm{P}\,(b,c) = \mathrm{P}\,(a|\,b,c)\,\mathrm{P}\,(b|\,c)\,\mathrm{P}\,(c) \tag{6.12}$$

$$\mathrm{P}\,(a,b,c) = \mathrm{P}\,(a,b|\,c)\,\mathrm{P}\,(c) \tag{6.13}$$

From equations (6.12) and (6.13) one can immediately derive:

$$\mathrm{P}\,(a|\,b,c) = \frac{\mathrm{P}\,(a,b|\,c)}{\mathrm{P}\,(b|\,c)} \tag{6.14}$$

Substituting $a = x(k+1)$, $b = y(k+1)$, $c = Y(k)$into (6.14) yields:

$$\mathrm{P}\left(x(k+1)|\, y(k+1), Y(k)\right) = \frac{\mathrm{P}\left(x(k+1), y(k+1)|\, Y(k)\right)}{\mathrm{P}\left(y(k+1)|\, Y(k)\right)} \tag{6.15}$$

$$= \frac{\mathrm{P}\left(y(k+1)|\, x(k+1), Y(k)\right)\mathrm{P}\left(x(k+1)|\, Y(k)\right)}{\mathrm{P}\left(y(k+1)|\, Y(k)\right)} \tag{6.16}$$

The numerator of this expression can be expanded even further:

$$\mathrm{P}\left(x(k+1), y(k+1)|\, Y(k)\right) = \mathrm{P}_n\left(y(k+1) - Cx(k+1)\right) \cdot \mathrm{P}\left(x(k+1)|\, Y(k)\right) \tag{6.17}$$

wherein P_n is the probability density function of the noise on the measurements. After substitution of this result into (6.15), we find:

$$\overbrace{\mathrm{P}\left(x(k+1)|\, y(k+1), Y(k)\right)}^{\text{posterior}} = \frac{\mathrm{P}_n\left(y(k+1) - Cx(k+1)\right) \cdot \overbrace{\mathrm{P}\left(x(k+1)|\, Y(k)\right)}^{\text{prior}}}{\mathrm{P}\left(y(k+1)|\, Y(k)\right)} \tag{6.18}$$

This expression is very informative. At the left hand side, we find the so-called 'posterior' (pdf) of $x(k+1)$, which includes the knowledge obtained from the measurement $y(k+1)$. The posterior is calculated from the 'prior' (pdf) by taking the latest measurement $y(k+1)$ into account.

In the following part, we are going to determine $x(k+1)$ such that the probability of realizing $x(k+1)$ after the measurement $y(k+1)$ is maximal. We impose the limitation that the probability density functions of the noise (P_n and P_v) are normal distributions. Since the covariance matrix $\mathrm{Cov}\left\{x(k+1)|\, Y(k)\right\}$ is known from (6.6) and R_n and R_v are given, these probability densitity functions are determined completely. The denominator at the right hand side of (6.18) is independent of $x(k+1)$ and can therefore be considered as a constant when finding the maximum.

$$\mathrm{P}\left(x(k+1)|\, y(k+1), Y(k)\right) = \mathrm{C}^{\underline{\mathrm{te}}} \cdot \exp\left(-\frac{1}{2}\left(x(k+1) - AX(k)\right)^{\mathrm{T}} Q(k+1)^{-1}\left(x(k+1) - AX(k)\right)\right) \times \exp\left(-\frac{1}{2}\left(y(k+1) - CX(k)\right)^{\mathrm{T}} R_n^{-1}\left(y(k+1) - CX(k)\right)\right)$$

One could rearrange this expression to an expression of the form

$$\exp\left(-\frac{1}{2}\left(x(k+1) - \ldots\right)^{\mathrm{T}}\left(Q^{-1}(k+1) + C^{\mathrm{T}} R_n^{-1} C\right)\left(x(k+1) - \ldots\right)\right)$$

This also means that

$$\mathrm{Cov}\left\{x(k+1)|\, Y(k+1)\right\} = P(k+1) = \left(Q^{-1}(k+1) + C^{\mathrm{T}} R_n^{-1} C\right)^{-1} \quad (6.19)$$

This result will be useful further on.

The probability density function $\mathrm{P}\left(x(k+1)|\, y(k+1), Y(k)\right)$ is to be maximized with respect to $x(k+1)$. This corresponds to minimizing the exponent. We consider the condition for stationarity:

$$Q^{-1}(k+1)\left(x(k+1) - AX(k)\right) - C^{\mathrm{T}} R_n^{-1}\left(y(k+1) - Cx(k+1)\right) = 0$$

The value of $X(k+1) = x(k+1)$ that statisfies this equation, is the estimator of $x(k+1)$.

$$\left(Q^{-1}(k+1) + C^{\mathrm{T}} R_n^{-1} C\right) X(k+1) = Q^{-1}(k+1)\, AX(k) + C^{\mathrm{T}} R_n^{-1} y(k+1)$$

At the left hand side, we can identify the value of $P(k+1)$ which was found in 6.19.

Using the matrix inverse lemma

$$P = \left(Q^{-1} + C^{\mathrm{T}} R_n^{-1} C\right)^{-1} = Q - QC^{\mathrm{T}} \left(CQC^{\mathrm{T}} + R_n\right)^{-1} CQ$$

and the following relation

$$\left(Q^{-1} + C^{\mathrm{T}} R_n^{-1} C\right)^{-1} C^{\mathrm{T}} R_n^{-1} = QC^{\mathrm{T}} \left(CQC^{\mathrm{T}} + R_n\right)^{-1}$$

one can derive the solution:

$$X(k+1) = AX(k) + Q(k+1)\, C^{\mathrm{T}} \left(CQ(k+1)\, C^{\mathrm{T}} + R_n\right)^{-1} (y(k+1) - CAX(k))$$

Recursive algorithm The preceding results can be combined into the following recursive algorithm:

$$Q(k+1) = AP(k)\, A^{\mathrm{T}} + R_v \tag{6.20}$$

$$K(k+1) = Q(k+1)\, C^{\mathrm{T}} \left(CQ(k+1)\, C^{\mathrm{T}} + R_n\right)^{-1} \tag{6.21}$$

$$P(k+1) = (\mathrm{I} - K(k+1)\, C)\, Q(k+1) \tag{6.22}$$

$$X(k+1) = AX(k) + K(k+1)\, (y(k+1) - CAX(k)) \tag{6.23}$$

Remarks:

- $Q(k+1) = P(k+1|k)$ is the prior covariance matrix of $X(k+1)$ derived using k measurements,
- $P(k+1)$ is the posterior covariance matrix of $X(k+1)$ derived using $k+1$ measurements,
- $AX(k)$ is the extrapolated state variable given k measurements,
- $CAX(k)$ is the value of the measurement given the extrapolated state,

- The matrices Q, P and K are independent of the measurements. Therefore they can be calculated beforehand.
- This method remains useable when the noise is not normally distributed. However, in that case the solution found will no longer be an optimal solution.

If the input matrix B is not equal to 0, one can derive following recursive equations for Kalman filtering:

$$Q(k+1) = AP(k)\,A^{\mathrm{T}} + R_v \tag{6.24}$$

$$K(k+1) = Q(k+1)\,C^{\mathrm{T}}\left(CQ(k+1)\,C^{\mathrm{T}} + R_n\right)^{-1} \tag{6.25}$$

$$P(k+1) = (\mathrm{I} - K(k+1)\,C)\,Q(k+1) \tag{6.26}$$

$$X(k+1) = AX(k) + Bu(k) + K(k+1)\,(y(k+1) - CAX(k) - CBu(k)) \tag{6.27}$$

In that case, $AX(k)+Bu(k)$ is the extrapolated state and $C\,(AX(k) + Bu(k))$is the measurement corresponding to the extrapolated state.

6.3 Example

Let's examine a system that is described by the following state equations for which we will construct a Kalman filter:

$$x(k+1) = ax(k) + v(k) \qquad \text{met } a = \sqrt{0.5} \tag{6.28}$$

$$y(k) = x(k) + n(k) \tag{6.29}$$

The prior information available is:

- v and n are normally distributed

- $\mathrm{E}\{v(k)\} = 0 \qquad \mathrm{E}\left\{v(k)\, v(j)^{\mathrm{T}}\right\} = \sigma_v^2 \delta_{kj}$
- $\mathrm{E}\{n(k)\} = 0 \qquad \mathrm{E}\left\{n(k)\, n(j)^{\mathrm{T}}\right\} = \sigma_n^2 \delta_{kj}$
- v and n are independent

The solution follows immediately from the equations derived above.

$$Q(k+1) = a^2 P(k) + \sigma_v^2 \tag{6.30}$$

$$K(k+1) = \frac{Q(k+1)}{Q(k+1) + \sigma_n^2} \tag{6.31}$$

$$P(k+1) = (1 - K(k+1)\, C)\, Q(k+1) \tag{6.32}$$

$$X(k+1) = aX(k) + K(k+1)\,(y(k+1) - aX(k)) \tag{6.33}$$

The expressions for $K(k+1)$ and $P(k+1)$ can be simplified even further:

$$K(k+1) = \frac{\sigma_n^2 Q(k+1)}{Q(k+1) + \sigma_n^2} = \frac{\sigma_n^2 \left(a^2 P(k) + \sigma_v^2\right)}{a^2 P(k) + \sigma_v^2 + \sigma_n^2} \tag{6.34}$$

$$P(k+1) = \frac{\sigma_n^2 \left(a^2 P(k) + \sigma_v^2\right)}{\sigma_n^2 + a^2 P(k) + \sigma_v^2} \tag{6.35}$$

In Table 6.1 the gain $K(k)$, $Q(k)$ and $P(k)$ of the corresponding Kalman filter are calculated explicitly. The initial value $x(0)$ was chosen to be 0. As we have no confidence in this initial value, the corresponding covariance matrix $P(0)$ was put at infinity. For simplicity, both σ_n and σ_v were equated to σ.

Table 6.1: Gain $K(k)$, $Q(k)$ and $P(k)$ of the Kalman filter

k	$Q(k)$	$K(k)$	$P(k)$
0			∞
1	∞	1	σ^2
2	$1.5\,\sigma^2$	0.6	$0.6\,\sigma^2$
3	$1.3\,\sigma^2$	0.565	$0.565\,\sigma^2$
4	$1.283\,\sigma^2$	0.56	$0.56\,\sigma^2$
5	$1.28\,\sigma^2$	0.56	$0.56\,\sigma^2$

Table 6.1 allows us to gain some insight into Kalman filtering.

One notices that the gain sequence $K(k)$ is a decreasing sequence that doesn't converge to 0. This means that at first more weight is put in the measurements than in the predictions. After a certain amount of iterations, an equilibrium is reached between the importance of a measurement and a prediction. In contrast to the recursive least squares approach, measurements keep influencing the predictions in Kalman filtering when k grows large. This is due to the fact that the system is internally disturbed by noise (the source $v(k)$ in particular).

The value of $P(k)$ also converges to a nonzero limit value. The Kalman filter is stationary when $P(k+1) = P(k)$. This limiting value can be calculated easily by equating $P(k+1)$ and $P(k)$ in (6.26). By doing so for this example, one obtains:

$$P = P(\infty) = \frac{\sigma_n^2 \left(a^2 P + \sigma_v^2\right)}{\sigma_n^2 + a^2 P + \sigma_v^2}$$

A quadratic equation in P ensues. Its solution for $a = \sqrt{5}$ and $\sigma_n = \sigma_v = \sigma$ is $P \approx 0.56\sigma^2$. This is the value we already encountered in Table 6.1.

Study of the effect of the noise $v(k)$: When the noise $v(k) = 0$, it is possible to simplify expression (6.35):

$$P(k+1) = \frac{\sigma_n^2 \left(a^2 P(k)\right)}{\sigma_n^2 + a^2 P(k)} \tag{6.36}$$

$$\frac{P(k+1)}{P(k)} = \frac{1}{\frac{1}{a^2} + \frac{P(k)}{\sigma_n^2}} = \frac{a^2}{1 + a^2 \frac{P(k)}{\sigma_n^2}} \tag{6.37}$$

For a stable system $|a| < 1$ and the second term in the denominator in (6.37) is positive. Therefore we can conclude that

$$\frac{P(k+1)}{P(k)} \leq a < 1.$$

This means that $P(k)$ converges to 0 for large k. A similar result is valid

for $K(k)$. If one examines the constituent relations of the Kalman filter, one concludes that the Kalman filter will eventually cease to take measurements into account in its predictions when $v(k)$ is assumed to be 0. It will predict the system behavior using only the supplied model and the estimated state. In theory such a situation is plausible, but in a practical setting this is a dangerous set-up. It is impossible to know a given system perfectly, there will always be (small?) model errors. Such model errors will cause the divergence of the estimated state with respect to the measured value, given that the gain factor $K(\infty) = 0$. The filter will correct for these errors no longer. In a nutshell: one always has to include the noise term $v(k)$ in the state equations, even when there is no direct motive to do so. By doing so, $P(k)$ will not converge to 0 (due to a large incertainty) but the Kalman filter will be able to handle model errors more gracefully.

In this chapter the Kalman filter was constructed for linear state equations. It is possible however to construct generalized versions that are applicable for nonlinear equations. Using these techniques, one can estimate both the model parameters and the state of the system at the same time. (Why is this last set-up a nonlinear problem?)

Chapter 7

Exercises

Chapter 7 Exercises

Johan Schoukens and Rik Pintelon. Draft version, 27 April 2007. Vrije Universiteit Brussel, departement ELEC, Pleinlaan 2, B1050 Brussels. email: johan.schoukens@vub.ac.be

What you will learn: The aim of this chapter is to illustrate basic aspects of system identification:
- least squares, weighted least squares and maximum likelihood estimation
- uncertainties and distribution of the estimates
- impact of disturbing noise on the input and the output

7.1 INTRODUCTION

The aim of system identification is to extract a mathematical model $M(\theta)$ from a set of measurements Z. Measurement data are disturbed by measurement errors and process noise, described as disturbing noise n_z on the data:

$$Z = Z_0 + n_z . \tag{1}$$

Since the selected model class M does in general not include the true system S_0, model errors appear:

$$S_0 \in M_0 \text{ and } M_0 = M + M_\varepsilon , \tag{2}$$

with M_ε the model errors. The goal of the identification process is to select M, and to tune the model parameters θ such that the 'distance' between the model and the data becomes as small as possible. This distance is measured by the cost function that is minimized. The selection of these three items (data, model, cost function) sets the whole picture, all the rest are technicalities that do not affect the quality of the estimates. Of course this is an over simplification. The numerical methods used to minimize the cost function, numerical conditioning problems, model parameterizations, ... are all examples of very important choices that should be properly addressed in order to get reliable parameter estimates. Failing to make a proper selection can even drive the whole identification process to useless results. A good understanding of each of these steps is necessary to find out where a specific identification run is failing: is it due to numerical problems, convergence problems, identifiability problems, or a poor design of the experiment?

In this chapter we will study the following issues:

- What is the impact of noise on the estimates (stochastic and systematic errors)?
- What are the important characteristics of the estimates?
- How to select the cost function?
- How does the choice of the cost function affect the results?
- How to select the complexity of the model? What is the impact on the estimates?

7.2 ILLUSTRATION OF SOME IMPORTANT ASPECTS OF SYSTEM IDENTIFICATION

In this section, we illustrates on a simple example some important aspects of system identification. The impact of the noise on the final estimates is illustrated. It will be shown that zero mean measurement noise can result in systematic errors on the estimates (the mean of the parameter errors is not equal to zero!). Also the uncertainty it studied. Depending on the choice of the cost function, a larger or smaller noise sensitivity will be observed. All these aspects are studied on a very simple example: the measurement of the value of a resistance starting from a series of voltage and current measurements.

7.2.1 Least squares estimation: a basic approach to system identification

Exercise 3.a (Least squares estimation of the value of a resistance) Goal: estimate the resistance value starting from a series of repeated current and voltage measurements:

$$u_0(t) = R_0 i_0(t), \; t = 1, 2, \ldots, N \tag{7-1}$$

with u_0, i_0 the exact values of the voltage and the current.

Generate an experiment with $N = 10, 100, 1000,$ and 10000 measurements. The current i_0 is uniformly distributed in $[-i_{\max}, i_{\max}]$ with $i_{\max} = 0.01$ A (use the Matlab™ routine `rand(N,1)`), $R_0 = 1000$. The current is measured without errors, the voltage is disturbed by independent, zero mean, normally distributed noise n_u with: $N(0, \sigma_u^2=1)$.

$$\begin{aligned} i(t) &= i_0(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned}, \; t = 1, 2, \ldots, N \tag{7-2}$$

To measure the distance between the model and the data, we select in this exercise a least squares cost function: $V(R) = \frac{1}{N}\sum_{t=1}^{N} (u(t) - Ri(t))^2$. Notice that many other possible choices can be made.

The least squares estimate $\hat{R}$ is defined as the minimizer of the following cost function:

$$\hat{R} = \arg\min_R V(R) \tag{7-3}$$

- Show that the minimizer of (7-3) is given by:

$$\hat{R} = \sum_{t=1}^{N} u(k)i(k) / \sum_{t=1}^{N} i(k)^2 . \tag{7-4}$$

- Generate 100 data sets with a length $N = 10, 100, 1000, 10000$, and calculate for each of these the estimated value $\hat{R}$.
- Plot the 100 estimates, together with the exact value for each N, and compare the results.

Observations - (see Figure 7-1) From the figure it is seen that the estimates are scattered around the exact value. The scattering decreases for an increasing number N. It can be

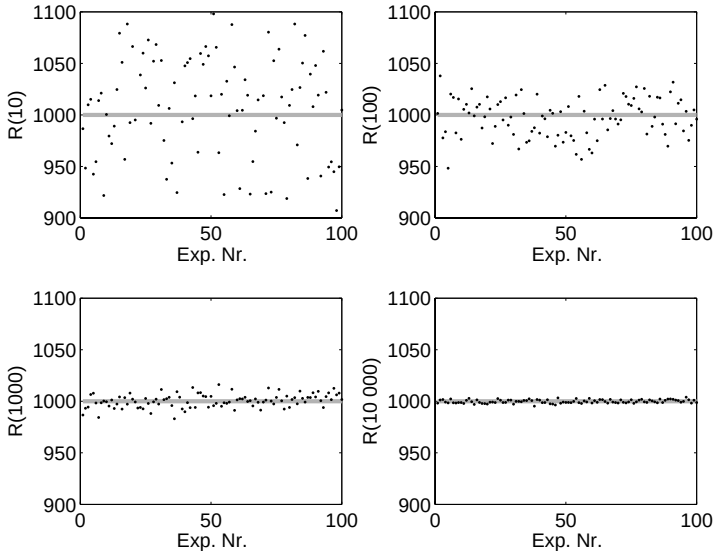


Figure 7-1 Estimated resistance values $\hat{R}(N)$ for N = 10, 100, 1000, 10000 for 100 repeated experiments. Gray line: exact value, dots: estimated value.

shown that under very general conditions, the standard deviation of the estimates decreases as $1/\sqrt{N}$. This is further elaborated in the next exercise.

Exercise 3.b (Analysis of the standard deviation In this exercise, it is verified how the standard deviation varies as a function of N. Consider the resistance

$$u_0(t) = R_0 i_0(t), \; t = 1, 2, \ldots, N. \tag{7-5}$$

with a constant current $i_0 = 0.01$ A, and $R_0 = 1000\Omega$. Generate 1000 experiments with $N = 10, 100, 1000$, and 10000 measurements. The current is measured without errors, the voltage is disturbed by independent, zero mean Gaussian distributed noise n_u in $N(0, \sigma_u^2=1)$:

$$\begin{aligned} i(t) &= i_0(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned}, \; t = 1, 2, \ldots, N \tag{7-6}$$

- Calculate for the four values of N the standard deviation of $\hat{R}$. Make a loglog-plot.
- Compare it with the theoretical value of the standard deviation that is given in this simplified case (constant current) by:

$$\sigma_R = \frac{1}{\sqrt{N}} \frac{\sigma_u}{i_0}. \tag{7-7}$$

Observations - (see Figure 7-2) From the figure it is seen that the standard deviation decreases as $1/\sqrt{N}$. Collecting more data allows to reduce the uncertainty. To get a reduction

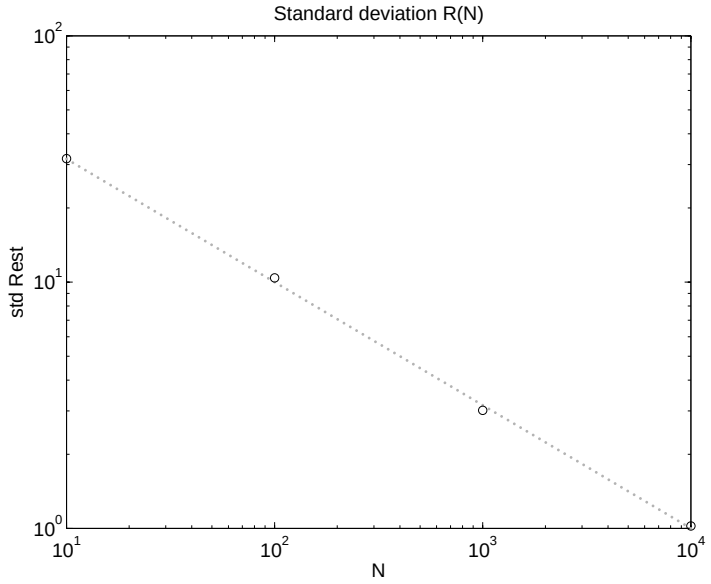


Figure 7-2 Experimental (black bullets) and theoretical (gray dots) standard deviation on $\hat{R}$ as a function of N. The error drops with $\sqrt{10}$ if the number of data N grows with a factor 10.

with a factor 10 in uncertainty, an increase of the measurement time with a factor 100 is needed. This shows that it still pays off to spend enough time on a carefully setup of the experiment in order to reduce the level of the disturbing noise σ_u on the raw data.

Remark: for the general situation with a varying current, the expression for the standard deviation σ_R for a given current sequence $i_0(t)$ is:

$$\sigma_R = \frac{\sigma_u}{\sqrt{\sum_{t=1}^{N} i_0^2(t)}} \tag{7-8}$$

Exercise 3.c (Study of the asymptotic distribution of an estimate) The goal of this exercise is to show that the distribution of an estimate is asymptotically for $N \to \infty$ normally distributed, and this almost independent of the distribution of the disturbing noise (within some regularity conditions, like finite variance, and a restricted 'correlation' length of the noise).

Consider the previous exercise for $N = 1, 2, 4, 8$, and 10^5 repetitions. Use a constant current $i_0 = 0.01$ A, measured without errors. For the voltage we consider two situations. In the first experiment, the voltage is disturbed by independent, zero mean Gaussian distributed noise $N(0, \sigma_u^2=1)$. In the second experiment the voltage noise is uniformly distributed in $[-\sqrt{3}, \sqrt{3}]$.

- Verify that the standard deviation of the uniformly distributed noise source also equals 1.

- Calculate the least squares solution (see eq. 7-4) for $N = 1, 2, 4, 8$ and repeat this 10^5 times for both noise distributions. Plot the estimated pdf for the eight different situations.
- Calculate the mean value and the standard deviation over all realizations (repetitions) for each situation, and compare the results.

h

Observations - (see Figure 7-3) From the figure it is seen the distribution of the esti-

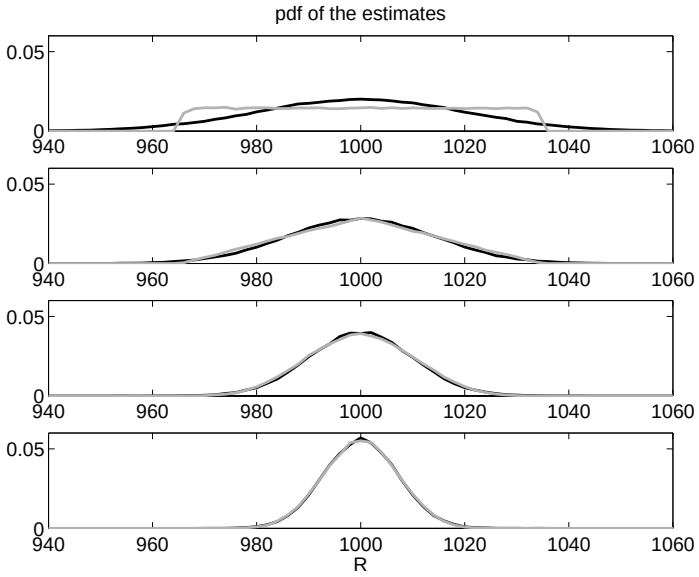


Figure 7-3 Evolution of the pdf of $\hat{R}$ as a function of N, for $N = 1, 2, 4, 8$. Black: Gaussian disturbing noise; Gray: uniform disturbing noise.

mates depends on the distribution of the noise. For example, for $N = 1$, the pdf for the Gaussian disturbed noise case is significantly different from that corresponding to the uniformly disturbed experiment. These differences disappear for a growing number of data per experiment (N increases), and for $N = 8$ it is already hard to see by a simple visual inspection a different shape. The uniform distribution converges to the Gaussian distribution for growing values of N. This is a general valid result.
In this case, the mean value and the variance is the same for all values of N. This is again a general result for models that are linear in the measurements (e.g. $y_0 = au_0$ is linear in u_0, while $y_0 = au_0^3$ is nonlinear in the measurements). The covariance matrix of the estimates depends only on the second order properties of the disturbing noise. This conclusion can not be generalized to models that are nonlinear in the measurements. In the latter case, the estimates will still be Gaussian distributed, but the mean value and variance will also depend on the distribution of the disturbing noise.

7.2.2 Systematic errors in least squares estimation

In the previous section it was shown that disturbing noise on the voltage resulted in noisy estimates of the resistor value, the estimated value of the resistor varies from one ex-

periment to the other. We characterized this behavior by studying the standard deviation of the estimator. The mean value of these disturbances was zero: the estimator converged to the exact value for a growing number of experiments. The goal of this exercise is to show that this behavior of an estimator is not for granted. Compared with the previous Exercises 3.a - 3.c, we add in the next two exercises also disturbing noise on the current. The impact of the current noise will be completely different from that of the voltage noise, besides the variations from one experiment to the other, also a systematic error will become visible.

Exercise 3.d (Impact of disturbing noise on the regressor or input measurements)

Consider the previous exercise for $N = 100$, and 10^5 repetitions. The current i_0 is uniformly distributed between $[-0.01, 0.01]$ A. It is measured this time with white disturbing noise added to it: $i(t) = i_0 + n_i(t)$, with a normal distribution $N(0, \sigma_i^2)$. The voltage measurement is also disturbed with normally distributed noise: $N(0, \sigma_u^2=1)$.

- Repeat the simulations of the previous exercise once without and once with noise on the current. Vary the current noise standard deviation in 3 successive simulations: $\sigma_i = 0, 0.01, 0.02$ A.
- Calculate the least squares solution (see eq. 7-4) for $N = 100$ and repeat this 10^5 times for all situations and plot the pdf for each of them.
- Calculate the mean value and the standard deviation over all realizations (repetitions) for each situation, and compare the results.

h

Observations - (see Figure 7-4) From the figure it is seen that the distribution of the estimates depends strongly on the presence of the noise on the current measurement. Not only

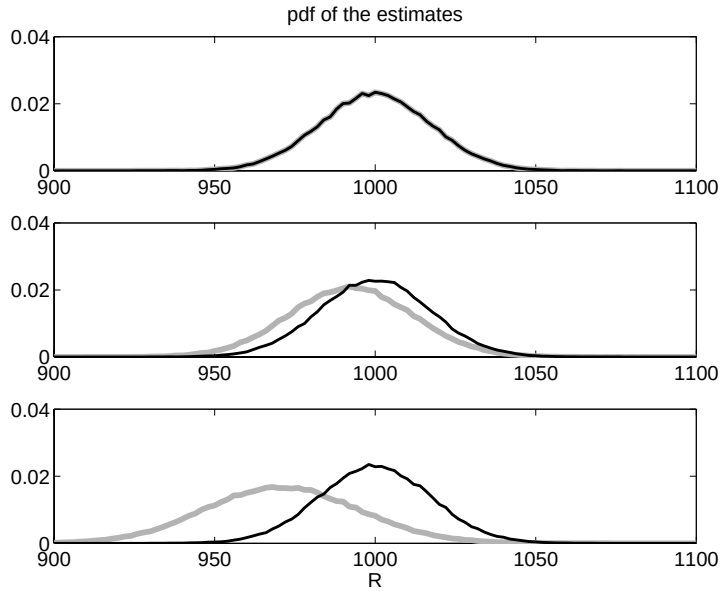


Figure 7-4 Evolution of the pdf of $\hat{R}$ as a function of the noise level at the current. Black: Only noise on the voltage $\sigma_u = 1$ V; Gray: noise on the voltage $\sigma_u = 1$ V and the current. $\sigma_i = 0, 0.1, 0.2$ A (top, middle, bottom).

the standard deviation is affected, also a bias becomes visible that grows with the variance of the current noise. This result is closely linked to the fact that the current is used as regressor or independent variable that explains the voltage as a dependent variable: we used a model where the current is the input, and the voltage is the output. Whenever the measurement of the input variable is disturbed by noise, bias problems will appear unless special designed methods are used. These will be studied in Section 7.6.

Exercise 3.e (The importance of the choice of the independent variable or input)

In the previous Exercise 3.d it became clear that noise on the input or independent variable creates a bias. The importance of this choice is explicitly illustrated by repeating Exercise 3.c. where the disturbing noise is only added to the voltage. In this exercise the same data are processed two times:

- Process the data using the current as independent variable, corresponding to the function $u(t) = Ri(t)$ and an estimate of R:

$$\hat{R} = \sum_{t=1}^{N} u(t)i(t) / \sum_{t=1}^{N} i(t)^2 \tag{7-9}$$

- Process the data using the voltage as independent variable, corresponding to $i(t) = Gu(t)$, with G the conductance:

$$\hat{G} = \sum_{t=1}^{N} u(t)i(k) / \sum_{t=1}^{N} u(t)^2 \text{ and } \hat{R} = 1/\hat{G}. \tag{7-10}$$

- Repeat each experiment 10^5 times, and calculate the pdf of the estimated resistance

h

Discussion - (see Figure 7-5) Whenever the measurement of the variable that appears squared in the denominator of (7-9) or (7-10) is disturbed by noise, a bias will become visible. This shows that the signal with the highest SNR should be used as independent variable or input in order to reduce the systematic errors. The bias will be proportional to the inverse SNR (noise power/signal power).

7.2.3 Weighted least squares: optimal combination of measurements of different quality

The goal of this section is to combine measurements with different quality. A first possibility would be to throw away the poorest data, but even these poor data contain information. Better is to make an optimal combination of all measurements taking into account their individual quality. This will result in better estimates with a lower standard deviation. The price to be paid for this improvement is the need for additional knowledge about the behavior of the disturbing noise. While the least squares (LS) estimator does require no information at all about the disturbing noise distribution, we have to know the standard deviation (or in general the covariance matrix) of the disturbing noise in order to be able to use the improved weighted least squares (WLS) estimator. That is illustrated in this exercise.

Exercise 4.a (combining measurements with a varying SNR: Weighted Least

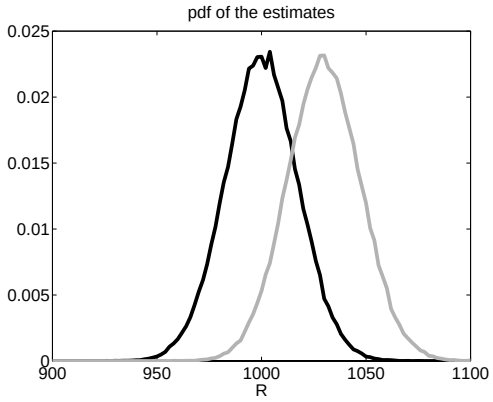


Figure 7-5 Study of the impact of the selection of the independent variable for the estimation of the resistance. Only the voltage is disturbed with noise. The pdf of the estimated resistance is shown for the independent variable being the current (black) or the voltage (gray).

squares estimation) Estimate the resistance value starting from a series of repeated current and voltage measurements:

$$u_0(t) = R_0 i_0(t),\ t = 1, 2, \ldots, N \tag{7-11}$$

with u_0, i_0 the exact values of the voltage and the current. Two different voltmeters are used, resulting in two data sets, the first one with a low noise level, the second one with a high noise level.

- Generate an experiment with N measurements, i_0 uniformly distributed in $[-0.01, 0.01]$ A, $R_0 = 1000\Omega$. The current is measured without errors, the voltage measured by the 2 voltmeters is disturbed by independent, zero mean, normally distributed noise n_u with: $N(0, \sigma_u^2{=}1)$ for the first good voltmeter, and $N(0, \sigma_u^2{=}16)$ for the second bad one.

$$\begin{aligned} i(t) &= i_0(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned},\ t = 1, 2, \ldots, N \tag{7-12}$$

- Calculate the weighted least squares solution, given below:

$$\hat{R} = \frac{\sum_{t=1}^{2N} \frac{u(t)i(t)}{w(t)}}{\sum_{t=1}^{2N} \frac{i(t)^2}{w(t)}}, \tag{7-13}$$

with $w(t)$ the weighting of the t^{th} measurement: $w(t) = \sigma_{u1}^2$ for the measurements of the first voltmeter, and $w(t) = \sigma_{u1}^2$ for the measurements of the second one.

- Repeat this exercise 10000 times for $N = 100$. Estimate the resistance also with the least squares method of Exercise 3.a. Make an histogram of both results.

Observations - (see Figure 7-6) From the figure it is seen that the.estimates are scat-

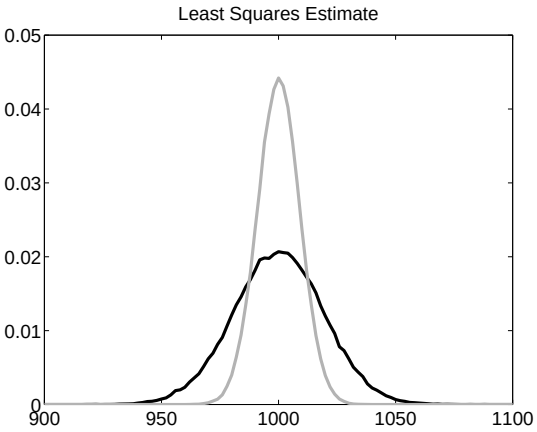


Figure 7-6 Estimated resistance values for $N = 100$, combining measurements of a good and a bad voltmeter. Black: pdf of the least squares; gray: pdf of the weighted least squares estimates.

tered around the exact value. However, the standard deviation of the weighted least squares is smaller than that of the least squares estimate. It can be shown that the inverse of the covariance matrix of the measurements is the optimal weighting for least squares methods.

Exercise 4.b (Weighted least squares estimation: Study of the variance) In this exercise we verify by simulations the theoretical expressions that can be used to calculate the variance of a least squares and a weighted least estimator. It is assumed that there is only noise on the voltage. The exactly measured current is used as regressor (input). The theoretical variance of the least squares estimator (no weighting applied) for the resistance estimate is given by

$$\hat{\sigma}^2_{\mathrm{LS}} = \frac{1}{\sum_{t=1}^{2N} \frac{i(t)^2}{\sigma_u^2(t)}}, \tag{7-14}$$

and the variance of the weighted least squares estimator using the variance on the output (the voltage) as weighting is

$$\hat{\sigma}^2_{\mathrm{WLS}} = \frac{\sum_{t=1}^{2N} \sigma_u^2(t) i^2(t)}{\left(\sum_{t=1}^{2N} i(t)^2\right)^2}. \tag{7-15}$$

- Consider Exercise 4.a, calculate the theoretical value for the standard deviation, and compare this to the results obtained from the simulations.

Observations - A typical result of this exercise is:
theoretical standard deviation LS: 19.4, experimental standard deviation: 19.3

theoretical standard deviation WLS: 9.1, experimental standard deviation: 9.2

Remark: the expressions (7-14) and (7-15) for the theoretical values of the variance are valid for a given input sequence. If the averaged behavior over all (random) inputs is needed, an additional expectation with respect to the input current should be calculated.

7.2.4 Models that are linear-in-the-parameters

The least squares estimates of the resistor that are studied till now were based on the minimization of the weighted cost function

$$V(R) = \frac{1}{N}\sum_{t=1}^{N} \frac{(u(t) - Ri(t))^2}{w(t)}, \tag{7-16}$$

with u, i the measured voltage (output) and current (input) respectively.

In general, the difference between a measured output $y(t)$ and a modelled output $\hat{y}(t) = g(t, u_0, \hat{\theta})$ is minimized for a given input signal $u_0(t)$. All model parameters are grouped in $\theta \in \mathrm{R}^{n_\theta}$. This can be formulized under a matrix notation. Define the signal vectors $\hat{y}, u_0, g \in \mathrm{R}^N$, for example:

$$y^T = \{y(1), \ldots, y(N)\}, \tag{7-17}$$

and a positive weighting matrix $W \in \mathrm{R}^{NxN}$. Then the weighted least squares cost function becomes:

$$V_{\mathrm{WLS}} = (y - g(u_0, \theta))W^{-1}(y - g(u_0, \theta))^T. \tag{7-18}$$

For a diagonal matrix $W_{tt} = w(t)$, and $W_{ij} = 0$ elsewhere, equation (7-18) reduces to

$$V_{\mathrm{WLS}} = \frac{1}{N}\sum_{t=1}^{N} \frac{(y(t) - g(t, u_0, \theta))^2}{w(t)}. \tag{7-19}$$

The estimate $\hat{\theta}$ is found as the minimizer of this cost function:

$$\hat{\theta} = \arg\min_{\theta} V_{\mathrm{WLS}}(\theta). \tag{7-20}$$

In general it will be impossible to solve this minimization problem analytically. However, if the model is linear-in-the-parameters, then it is possible to formulate the solution explicitly, and it is also possible to calculate it in a stable numerical way with one instruction in Matlab™™. A model is called linear-in-the-parameters if the output is a linear combination of the model parameters:

$$y = K(u_0)\theta \text{ with } K \in \mathrm{R}^{Nxn_\theta}. \tag{7-21}$$

Note that K can be a nonlinear function of the input. The explicit solution of the (weighted) least squares problem becomes:

$$\hat{\theta}_{WLS} = (K^T W K)^{-1} W K y \text{ and } \hat{\theta}_{LS} = (K^T K)^{-1} K y . \tag{7-22}$$

Numerical stable solutions to calculate this expression avoid the explicit calculation of the product $K^T W^{-1} K$ to improve the numerical conditioning. The Matlab™™ solution is given by:

$$\hat{\theta}_{WLS} = (W^{1/2}K)\backslash(W^{1/2}y) \text{ with } W = W^{1/2}W^{1/2}$$
$$\hat{\theta}_{LS} = K\backslash y . \tag{7-23}$$

Exercise 5 (Least squares estimation of models that are linear in the parameters) Consider the model `y0=tan(u0*0.9*pi/2)`, evaluated for the inputs `u0=linspace(0,1,N)`.

- Generate a data set $y = y_0$. Put $N = 100$, and vary $n = 1$ to 20. Use the Matlab™ instruction
- Calculate the least squares solution ($W = I^{NxN}$) for the different values of n, using the stable Matlab™ solution (7-23) and the direct implementation (7-22).
- Compare the solutions, and calculate the condition number of K and $K^T T$. This can be done with Matlab™ instruction `cond( )`
- Compare the modeled output with the exact output and calculate the rms-value of the error.

h

Observations - (see Figure 7-7) From this figure, it can be seen that the condition num-

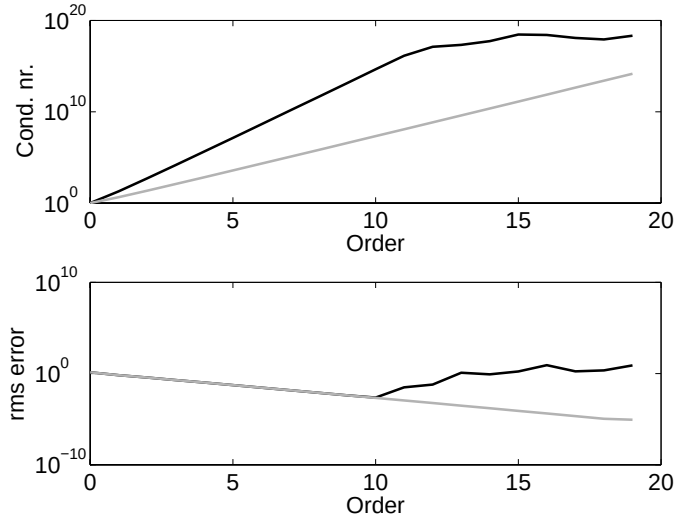


Figure 7-7 Identification of a polynomial model that is linear-in-the-parameters using a method that is numerical stable $\hat{\theta}_{WLS} = K\backslash y$ (gray lines) or numerical unstable $\hat{\theta}_{WLS} = (K^T K)^{-1} K y$ (black lines). Top: condition number as a function of the model order; Bottom: the rms error as a function of the model order.

ber of the numerical unstable method (7-22) grows two times faster on a logarithmic scale

than that of the stable method (7-23). The number of digits required to make the equations is given by the exponent of the condition number. From order 10 or larger more than 15 digits are needed which is more than the calculation precision of Matlab™. As a result, the obtained models are no longer reliable, even if there was no disturbing noise in the experiment. This shows that during system identification procedures, it is always necessary to verify the numerical conditions of the calculations. The condition number of the stable numerical implementation grows less fasts, allowing to solve higher order polynomial approximations.

Remark: if very high order polynomial approximations are needed, other more robust polynomial representations can be used using orthogonal polynomials. The nature of these polynomials will depend upon the applied input signal.

7.2.5 Interpretation of the covariance matrix & Impact experiment design

In the previous Section 7.5.1, a one and a two parameter model was considered. In this section it is shown that: 1) The variance of a set of parameters is not enough to make conclusions on the model uncertainty, the full covariance matrix is needed. 2) The covariance matrix (and the correlation between the parameters) is strongly influenced by the experiment design.

Exercise 6 (Characterizing a 2-dimensional parameter estimate)

Generate a set of measurements:

$$y(t) = au_0(t) + n_t . \tag{7-24}$$

In the first experiment $u_0(t)$ is generated by `linspace(-3,3,N)`, distributing N points equally between -3 and 3. In the second experiment $u_0(t)$ is generated by `linspace(2,5,N)`.

- Choose $a = 1$, $N = 1000$, and $n_k \quad N(0, \sigma_n^2)$ with $\sigma_n^2 = 1$.
- Use as a model $y = au_0 + b$, and estimate the parameters (a, b) using the method of Exercise 5.
- Repeat this experiment 10^5 times.
- Estimate the LS-parameters for both experiments, calculate the covariance matrix, and plot $\hat{a}(i)$ as a function of $\hat{b}(i)$.
- Plot also the estimated lines for the first 50 experiment

Observations - (Figure 7-8) In Figure 7-8, top the parameters are plotted against each other. For the second experiment ($u \sim$ uniform in [2,5]), the parameters are strongly correlated, as can be seen from the linear relation between the estimated values $\hat{a}(i)$ and $\hat{b}(i)$. This is not so for the first experiment ($u = [-3, 3]$), the black cloud has is main axis parallel to the horizontal and vertical axis which is the typical behavior of an uncorrelated variable. This can also be seen in the covariance matrices:

$$C_{\text{exp1}} = \begin{bmatrix} 3.2\times10^{-3} & 0.85\times10^{-4} \\ 0.85\times10^{-4} & 10.5\times10^{-3} \end{bmatrix}, \text{ and } C_{\text{exp2}} = \begin{bmatrix} 1.31\times10^{-2} & -4.6\times10^{-2} \\ -4.6\times10^{-2} & 16.9\times10^{-2} \end{bmatrix}, \tag{7-25}$$

or even better from the correlation matrices

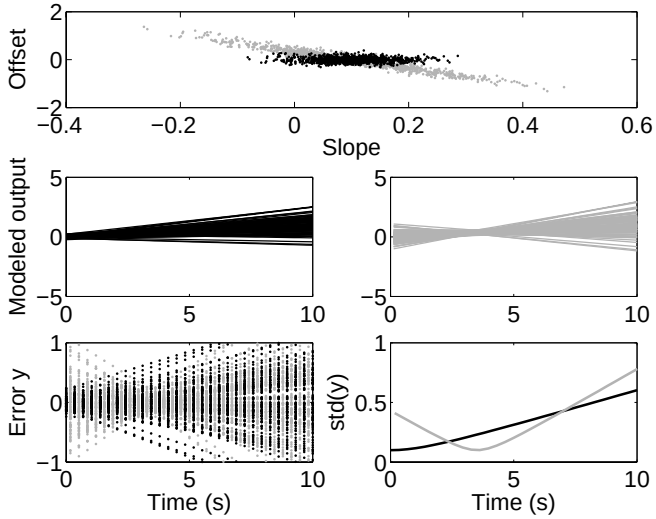


Figure 7-8 Black: experiment in time interval $[-(3, 3])]$; Gray: experiment in time interval $[2, 5]$. Top: Scatter plot (slope, offset). Middle: modeled output. Bottom: Error on modeled output (left) and its standard deviation (right).

$$R_{\text{exp1}} = \begin{bmatrix} 1 & 0.02 \\ 0.02 & 1 \end{bmatrix}, \text{ and } R_{\text{exp2}} = \begin{bmatrix} 1 & -0.97 \\ -0.97 & 1 \end{bmatrix}. \tag{7-26}$$

The correlation in the first matrix is almost zero, while for the second experiment it is almost one, indicating that a strong linear relation between the offset and slope estimate exists. This means that both variables vary considerable (a large standard deviation), but they vary together (a large correlation) so that the effect on the modelled output is small in the input range that the experiment was made (see Figure 7-8, middle and bottom). In that range, the variations of $\hat{a}$ or mostly cancelled by those of $\hat{b}$. Outside this range, the standard deviation of the modelled output will be larger compared to that obtained with the first experiment because there the offset-slope compensation is no longer valid. This shows that the covariances play an important role in the model uncertainty.

7.3 MAXIMUM LIKELIHOOD ESTIMATION FOR GAUSSIAN AND LAPLACE DISTRIBUTED NOISE

In Section 7.2 and 7.3, Gaussian distributed noise was added as disturbances to the measurements. It is shown in theory that least squares estimators, where the cost function is a quadratic function of the errors, perform optimal under these conditions. The smallest uncertainty on the estimators is found if a proper weighting is selected. This picture changes completely if the disturbing noise has no Gaussian distribution. In the identification theory it is shown that for each noise distribution, there corresponds an optimal choice of the cost function. A systematic approach to find these estimators is the maximum likelihood theory. Discussing this theory is out of the scope of this book, but some of its results will be illustrated on the resistance example. The disturbances will be selected once to have a normal dis-

tribution, and once to have a Laplace distribution. The optimal cost functions corresponding to these distributions are a least squares and a least absolute value cost function.

Exercise 7.a (Dependence of the optimal cost function on the distribution of the disturbing noise) Consider a set of repeated measurements:

$$u_0(t) = R_0 i_0(t) \ , \ t = 1, 2, \ldots, N \tag{7-27}$$

with u_0, i_0 the exact values of the voltage and the current. Two different voltmeters are used, resulting in two data sets, the first one disturbed by Gaussian (normal) distributed noise, the second one disturbed with Laplace noise.

Generate an experiment with N measurements, i_0 uniformly distributed in $[0, i_{\max}=0.01\text{A}]$, and $R_0 = 1000\Omega$. The current is measured without errors. The voltage measured with the first voltmeter is disturbed by independent, zero mean, normally distributed noise $n_u \sim N(0, \sigma_u^2=1)$, the second voltmeter is disturbed by Laplace distributed noise with zero mean, and $\sigma_u^2=1$.

$$\begin{aligned} i(t) &= i_0(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned}, \ t = 1, 2, \ldots, N \tag{7-28}$$

For the Gaussian noise, the maximum likelihood solution reduces to a least squares (LS) estimate as in (7-4), for the Laplace distribution the maximum likelihood estimator is found as the minimizer of

$$V_{\text{LAV}}(R) = \frac{1}{N}\sum_{t=1}^{N} |u(t) - Ri(t)| \text{, and } \hat{R}_{\text{LAV}} = \arg\min_R V_{\text{LAV}}(R) \text{,} \tag{7-29}$$

called the least absolute values (LAV) estimate.

- Repeat this exercise 10000 times for $N = 100$.
- Apply both estimators also to the other data set.
- Calculate the mean value, the standard deviation, and plot for each situation the histogram.

Help 1: Laplace distributed noise with zero mean and standard deviation *stdu* can be generated from uniformly distributed noise $[0, 1]$ using the following Matlab™ implementation:

```
x=rand(NData,1);     % generate uniform distributed noise
nLap=zeros(size(x)); % vector used to store the Laplace noise
nLap(x<=0.5) = log(2*x(x<=0.5))/sqrt(2)*stdU;
nLap(x>0.5) =  - log(2*(1-x(x>0.5)))/sqrt(2)*stdU;
```

Help 2: to minimize $V_{\text{LAV}}(R)$, a simple scan can be made over R belonging to [800:0.1:1200]

Observations - (see Figure 7-9) From Figure 7-9, it is seen that the estimates are scattered around the exact value. For the Gaussian case, the LS squares estimate is less scattered than the LAV estimate. For the Laplace case the situation is reversed. The mean and standard deviations are given in TABLE 7-10. This shows that the maximum likelihood estimator is optimal for the distribution that it is designed for. If the noise distribution is not prior known, but the user can guarantee that the variance of the noise is finite, than it can be shown that the least squares estimate is optimal in the sense that it minimizes the worse possible situation among all noise distributions with a finite variance.

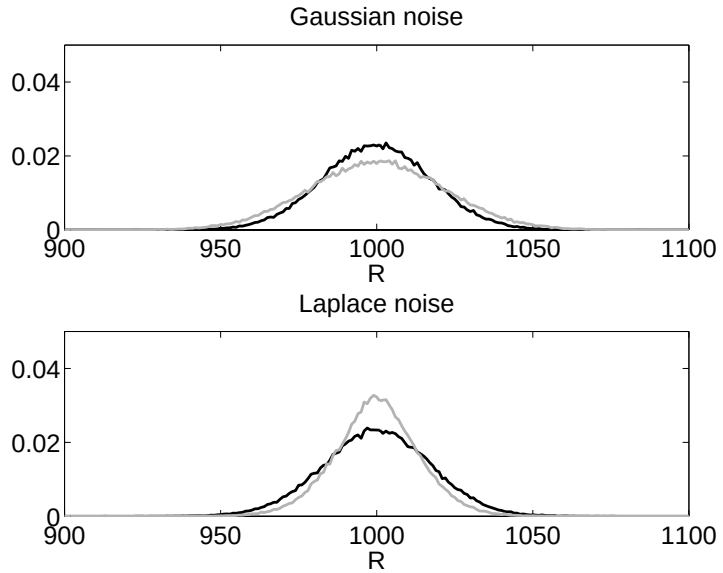


Figure 7-9 PDF of the Gaussian $\hat{R}_{LS}$ and Laplace $\hat{R}_{LAV}$ Maximum Likelihood estimators, applied to a Gaussian and Laplace noise disturbance. Black line: $\hat{R}_{LS}$, gray line: $\hat{R}_{LAV}$.

	$\hat{\mu}_{LS}$	$\hat{\sigma}_{LS}$	$\hat{\mu}_{LAV}$	$\hat{\sigma}_{LAV}$
Gaussian noise	1000.040	17.5]	999.94	22.0
Laplace noise	1000.002	17.3	999.97	13.7

TABLE 7-10 Mean and standard deviation of the Gaussian and Laplace Maximum Likelihood estimators, applied to a Gaussian and Laplace noise disturbance.

7.4 IDENTIFICATION FOR SKEW DISTRIBUTIONS WITH OUTLIERS

In Section 7.3, it was shown that the optimal choice of the cost function depends on the distribution of the disturbing noise. The maximum likelihood theory offers a theoretic framework for the generation of the optimal cost function. In practice a simple rule of thumb can help to select a good cost function. Verify if the disturbing noise has large outliers: large errors appear to be more likely than expected from a Gaussian noise distribution.

In this exercise the LS- and the LAV-estimate are applied to a χ^2-distribution with 1 degree of freedom: this is nothing than a squared Gaussian distributed variable. Compared to the corresponding Gaussian distribution, the extreme large values appear to frequent (due to the square value). Neither of both estimates (LS, LAV) is the MLE for this situation. But from the rule of thumb we expect that the LAV will perform better than the LS estimator. It will turn out that a necessary condition to get good results is to apply a proper calibration procedure for each method, otherwise a bias will become visible.

Exercise 8 (Identification in the presence of outliers) Consider a set of repeated measurements:

$$u_0(t) = R_0 i_0(t) \ , \ t = 1, 2, \ldots, N \tag{7-30}$$

with u_0, i_0 the exact values of the voltage and the current. The voltage measurement is disturbed by noise, generated from a χ^2-distribution with 1 degree of freedom (= squared Gaussian noise).

Generate an experiment with N measurements, i_0 uniformly distributed in $[0, i_{max}=0.01\text{A}]$ (use the Matlab™ routine `rand`), $R_0 = 1000\Omega$. The current is measured without errors. The measured voltage $u(t)$ is disturbed by χ^2-distribution distributed noise n_u with:

$n_u = n^2$, with n generated as $N(0, \sigma_u^2=1)$.

Note that the mean value of $\text{E}\{n_u\} = 1$, and $\text{median}(n_u) = 0.455$.

$$\begin{aligned} i(t) &= i_0(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned} \ , \ t = 1, 2, \ldots, N \tag{7-31}$$

- In order to reduce the systematic errors, calibrate the data first. To do so, the mean value or the median of the noise should be extracted from the measurements. Make first a measurement with zero current, so that $u(t) = n_u(t)$.
- Repeat the exercise 10000 times and estimate each time the LS- and the LAV-estimate for both data sets.
- Estimate the pdf of the estimates, and calculate their mean value and standard deviation.

Observations - (see Figure 7-11) From the figure it is seen that the estimates are no longer scattered around the exact value $R = 1000\Omega$. Only the combination (LS-estimate, mean value calibration) and the combination (LAV-estimate, median value calibration) works well. The other combinations show a significant bias.

The mean and standard deviations are given in Table 7-12. Observe that the standard deviation of the LAV-estimate is smaller than that of the LS-estimate. LAV-estimates are less sensitive to outliers! Note that the mean value of the LAV-estimator, combined with the median calibration has still a small systematic error of 1.85, which is larger than the uncertainty on the mean value: 18.62/sqrt(10000)=0.18. If instead of using the mean, the median value is selected to average the 10000 estimates, the bias disappears completely.

	$\hat{\mu}_{LS}$	$\hat{\sigma}_{LS}$	$\hat{\mu}_{LAV}$	$\hat{\sigma}_{LAV}$
Calibr.: mean value	999.84	24.30	924.29	16.26
Calibr.: median	1081.86	24.43	1001.85	18.62

TABLE 7-12Mean and standard deviation of the Gaussian and Laplace Maximum Likelihood estimators, applied to a Gaussian and Laplace noise disturbance.

Conclusion: The LS-estimate should be combined with a calibration based on the mean, and the mean should be used to average the results. It is sensitive to outliers.

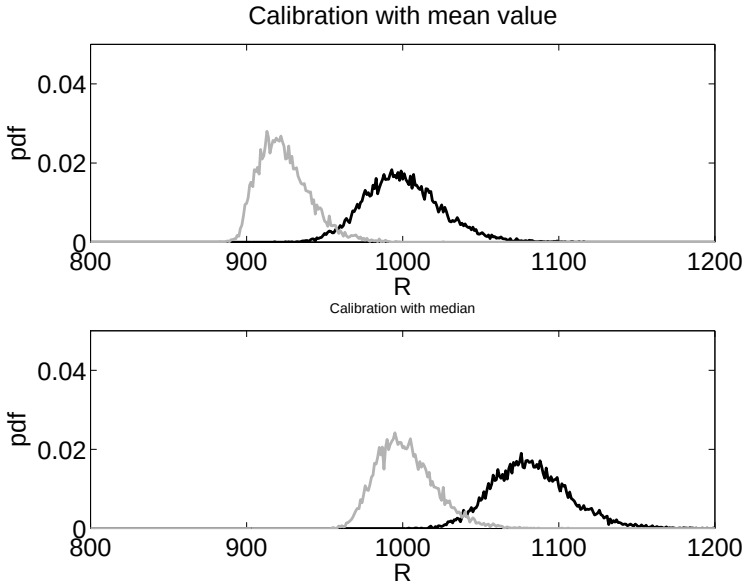


Figure 7-11 PDF of the Gaussian $\hat{R}_{\mathrm{LS}}$ and Laplace $\hat{R}_{\mathrm{LAV}}$ applied to χ^2 disturbed data. Top: calibration with the mean value, bottom: calibration with the median value. Black line: $\hat{R}_{\mathrm{LS}}$, gray line: $\hat{R}_{\mathrm{LAV}}$.

The LAV-estimate should be combined with a calibration based on the median, the median should be used to average the results, and it is less sensitive to the presence of outliers.

7.5 SELECTION OF THE MODEL COMPLEXITY

7.5.1 Influence of the number of parameters on the uncertainty of the estimates

In this exercise it will be shown that, once the model includes all important contributions, the uncertainty grows if the number of model parameters is still increased.

Exercise 9 (Influence of the number of parameters on the model uncertainty) In order to measure the flow of a tap, the height $y(t)$ of the water level in a measuring jug is recorded as a function of time t. However, the starting point of the measurements is uncertain. Hence two models are compared:

$$y(t) = at \text{ and } y(t) = at + b. \tag{7-32}$$

The first model estimates only the flow assuming that the experiment started at time zero, while the second one also estimates the start of the experiment.

Generate a set of measurements:

$$y(t) = at + n(t), \text{ with } t = [0{:}N]/N. \tag{7-33}$$

- Choose $a = 1$, $N = 1000$, and $n_k \sim N(0, \sigma_n^2)$ with $\sigma_n^2 = 1$.
- Repeat this experiment 10^4 times.
- Estimate the LS-parameters of both models, and compare $\hat{a}$ for the one- and two-parameter model by estimating the pdf of $\hat{a}$.
- Calculate the mean value and the standard deviation of the slope.
- Plot also the estimated lines for the first 50 experiments.

h

Observations - The results are shown below in TABLE 7-13 and Figure 7-14. From the

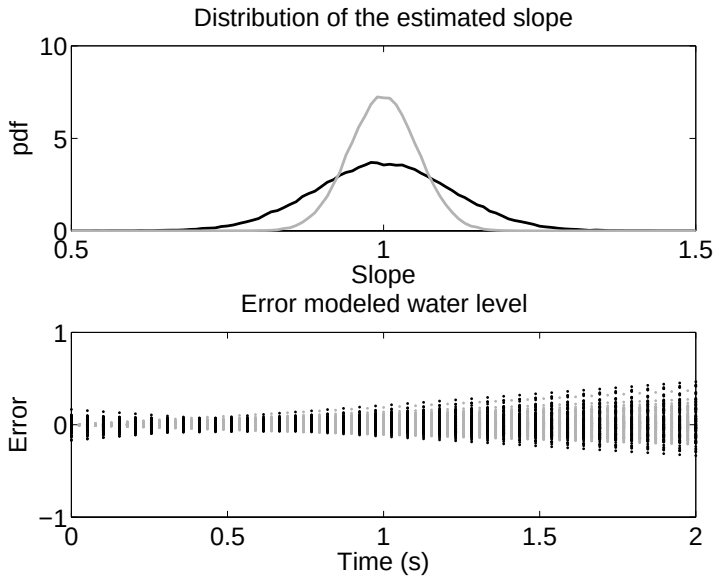


Figure 7-14 Impact of the number of the number of parameters on the uncertainty of the slope estimate and the variability of the model. Black: one-parameter model $y = au$; gray: two-parameter model $y = au + b$. Top: the pdf of the estimated slope; Bottom: the error on the modelled output as a function of time.

	1-parameter model	2-parameter model
mean	0.996	0.987
std. dev.	0.57	1.13

TABLE 7-13Mean and standard deviation of $\hat{a}$ in the one- and two-parameter model.

table it is seen that the uncertainty of the 1-parameter estimate is significantly smaller than that of the 2-parameter model. The mean value of both estimates are unbiased, the error equals the exact value within the uncertainty after averaging 100 experiments. Also in Figure 7-14 the same observations can be made. Notice, that due to the prior knowledge of the one-parameter model (at time zero the height is zero), a significantly smaller uncertainty

on $\hat{a}$ is found for small values of t. The one-parameter model is less scattered than the two-parameter model. If this prior knowledge would be wrong, systematic errors would be made on the flow estimate; if it is correct better estimates are found. An analysis of the residuals can guide the user to find out in which of both cases is faced.

7.5.2 Model selection

The goal of this section is to show how to select an optimal model for a given data set. Too simple models will fail to capture all important aspects of the output, and this will result in too large errors in most cases. Too complex models use too many parameters. As was illustrated in the previous section such models also result in a poor behavior of the modeled output because the model becomes too sensitive to the noise. Hence we need a tool that helps us to select the optimal complexity that balances the model errors against the sensitivity to the noise disturbances. It is clear that this choice will depend on the quality of the data. All these aspects are illustrated in the next exercise where we propose the Akaike information criterion as a tool for model selection.

Consider a single input single output linear dynamic system, excited with an input $u_0(t)$ and output $y_0(t) = g_0(t)*u(t)$. The system has an impulse response $g_0(t)$ that is infinitely long (infinite impulse response or IIR-system). For a stable system $g_{0(t)}$ decays exponentially to zero, so that the IIR system can be approximated by a system with a finite length impulse response $g(t)$, $t = 0, 1, \ldots, I$ (finite impulse response or FIR-system). For $t > I$, the remaining contribution can considered to be negligible. The choice of I will depend not only on $g(t)$, but also on the SNR of the measurements.

$$\hat{y}(t) = \sum_{k=0}^{I} \hat{g}(k)u_0(t-k), \text{ with } u_0(k) = 0 \text{ for } k < 0. \tag{7-34}$$

In (7-34) it is assumed that the system is initially in rest. If this is not the case, transient errors will appear, but these disappear in this model for $t > I$ (why?).

The model parameters θ are in this case the values of the impulse response. θ is estimated from the measured data $u_0(t), y(t)$, $t = 0, 1, \ldots, N$, with $y(t)$ the output measurement that is disturbed with i.i.d. noise with zero mean and variance σ_v^2:

$$y(t) = y_0(t) + v(t). \tag{7-35}$$

The estimates $\hat{\theta}$ are estimated by minimizing the least squares cost function:

$$V_N(\theta, Z^N) = \frac{1}{2N}\sum_{t=0}^{N} |y(t) - \hat{y}(t)|^2, \text{ with } \hat{y}(t) = \hat{g}(t)*u_0(t) \tag{7-36}$$

Note that this model is linear-in-the-parameters, and solution (7-24) can be used.

In order to find the ‘best’ model, a balance is made between the model errors and the noise errors using a modified cost function that accounts for the complexity of the model. Here we propose to use amongst others the AIC criterion:

$$V_{AIC} = V_N(\theta)\left(1 + 2\frac{\dim\theta}{N}\right). \tag{7-37}$$

Exercise 10 (Model selection using the AIC criterion) Consider the discrete time system $g_0(t)$ given by its transfer function:

$$G_0(z) = \sum_{k=0}^{n_b} b_k z^{-k} / \sum_{k=0}^{n_a} a_k z^{-k}, \tag{7-38}$$

Generate the filter coefficients a_k, b_k using the Matlab™ instruction

```
[b,a]=cheby1(3,0.5,[2*0.15  2*0.3])
```
(7-39)

This is a band pass system with a ripple of 0.5 dB in the pass band. Generate two data sets D_{est} and D_{val}, the former with length N_e being used to identify the model, the latter with length N_v to validate the estimated model. Note that the initial conditions for both sets are zero! Use the Matlab™ instruction

```
y0=filter(b,a,u0), y=y0+ny
```
(7-40)

with u_0 zero mean normally distributed noise with $\sigma_{u_0} = 1$, and v zero mean white Gaussian noise with σ_v equal to 0.5 for a first experiment, and 0.05 for a second experiment. Put $N_e = 1000$, and $N_{\text{val}} = 10000$ in both experiments.

- Use the linear least squares procedure (7-24) to estimate the model parameters, and this for varying orders from 0 to 100.
- Calculate for each of the models the simulated output $\hat{y} = \text{filter}(\hat{g}, 1, u_0)$, and calculate the cost function (7-36) on D_{est} and on D_{val}.
- Calculate V_{AIC}.
- Calculate $V_0 = \frac{1}{2N}\sum_{t=0}^{N} |y_0(t) - \hat{y}(t)|^2$ on the undisturbed output of the validation set.
- Plot V_{est}, V_{AIC}, V_{val} as a function of the model order. Normalize the value of the cost function by σ_v^2 to make an easier comparison of the behavior for different noise levels.
- Plot $\sqrt{V_0/\sigma_v^2}$ as a function of the model order.

h

Observations - The results are shown in Figure 7-15, the following observations can be made:

i) Increasing the model order results in a monotonic decreasing cost function V_{est}. This result was to be expected because a simpler model is always included by the more complex model, and the linear LS always retrieve the absolute minimum of the cost function, no local minima exist. Hence increasing the complexity of the model should reduce the value of the cost function.

ii) On the validation data we observe first a decrease and next an increase of V_{val}. In the beginning, the additional model complexity is mainly used to reduce the model errors, a steep descent of the cost function is observed. From a given order on, the reduction of the model errors is smaller than the increased noise sensitivity due to the larger number of parameters, re-

sulting in a deterioration of the capability of the model to simulate the validation output. As a result the validation cost function V_{val} starts to increase.

iii) V_{AIC} gives a good indication, starting from the estimation data only, where V_{val} will be minimum. This reduces the need for long validation records, and it allows to use as much data as possible for the estimation step.

iv) The optimal model order increases for a decreasing disturbing noise variance. Since the plant is an IIR system with an infinite long impulse response, it is clear that in the absence of disturbing noise $\sigma_n = 0$, the optimal order would become infinite. In practice this value is never reached due to the presence of calculation errors that act also as a disturbance.

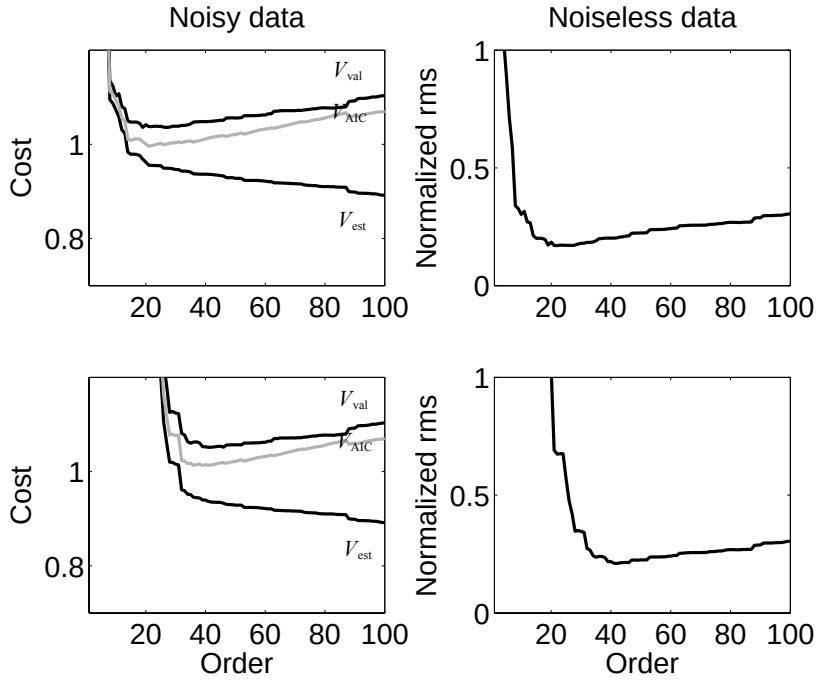


Figure 7-15 : Comparison of the normalized Cost function V_{est}, the AIC-criterion V_{AIC}, and the validation V_{val} for $\sigma_n = 0.5$ (top) and $\sigma_n = 0.05$ (bottom).

v) A fair idea about the quality of the models is given by V_0. The normalized rms-value $\sqrt{V_0/\sigma_v^2}$ is plotted on the right side of Figure 7-15. This figure shows that a wrong selection of the model can result in much larger simulation errors. The good news is that the selection of the best model order is not so critical, the minimum is quite flat and all model orders in the neighborhood of the minimum result in good estimates.

7.6 NOISE ON INPUT AND OUTPUT MEASUREMENTS: THE IV-METHOD

In Section 7.2.2 it was shown that the presence of disturbing noise on the input measurements creates a systematic error. In this set of exercises more advanced identification methods are illustrated that can deal with this situation. Two methods are studied, the first is called the instrumental variables method (IV), the second is the errors-in-variables (EIV) method. The major advantage of the IV-methods is its simplicity. No additional information is required

from the user. The disadvantage is that this method does not always perform well. Both situations are illustrated in the exercises. The EIV performs well in many cases, but in general additional information of the user is required. The covariance matrix of the input-output noise should be known. All methods are illustrated again on the resistance example with measured current and voltage $i(t), u(t)$, $t = 1, 2, \ldots, N$. Both measurements are disturbed by mutually uncorrelated Gaussian noise:

$$\begin{aligned} i(t) &= i_0(t) + n_i(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned} \tag{7-41}$$

The least squares estimate is given by:

$$\hat{R}_{\mathrm{LS}} = \frac{\sum_{t=1}^{N} u(t)i(t)}{\sum_{t=1}^{N} i(t)^2}, \tag{7-42}$$

the instrumental variables estimator (IV) is:

$$\hat{R}_{\mathrm{IV}} = \frac{\sum_{t=1}^{N} u(t)i(t+s)}{\sum_{t=1}^{N} i(t)i(t+s)}, \tag{7-43}$$

with s a user selectable shift parameter. Note that the IV-estimator equals the LS-estimator for $s = 0$.

The EIV estimator is given by

$$\hat{R}_{\mathrm{EIV}} = \frac{\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2} + \sqrt{\left(\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2}\right)^2 + 4\frac{\left(\sum u(t)i(t)\right)^2}{\sigma_u^2\sigma_i^2}}}{\frac{\sum u(t)i(t)}{\sigma_u^2}}, \tag{7-44}$$

with σ_u^2, σ_i^2 the variance of the voltage and current noise, the covariance is assumed to be zero in this expression: $\sigma_{ui}^2 = 0$.

Exercise 11.a (Noise on input and output: the instrumental variables method)

Generate the current $i_0(k)$ from a Gaussian white noise source, filtered by a first order Butterworth filter with cut-off frequency f_{Gen}:

$$i_0 = \mathrm{filter}(b_{\mathrm{Gen}}, a_{Gen}, e_1), \tag{7-45}$$

with `[`b_{Gen}`,` a_{Gen}`]=butter(1,2*`f_{Gen}`)`.

Generate the measured current and voltage (7-41), where $n_u(k)$ is white Gaussian noise: $N(0, \sigma_{n_u}^2)$. The current noise $n_i(k)$ is obtained from a Gaussian white noise source filtered by a second order Butterworth filter with cut-off frequency f_{Noise}:

$$i_0 = \texttt{filter}(\, b_{\text{Noise}}, a_{\text{Noise}}, e_2 \,), \tag{7-46}$$

with $[\, b_{\text{Noise}}, a_{\text{Noise}} \,] = \texttt{butter}(2, 2 * f_{\text{Noise}})$, and e_2 white Gaussian noise. Its variance is scaled to $\sigma_{n_u}^2$.

- Experiment 1: Generate three sets of 1000 experiments with $N = 5000$ measurements each, and the following parameter settings:

$$f_{\text{Gen}} = 0.1\,,\ f_{\text{Noise}} = [0.999, 0.95, 0.6]\,,$$
$$\sigma_{i_0} = 0.1\,,\ \sigma_{n_i} = 0.1\,,\ \sigma_{n_u} = 1\,. \tag{7-47}$$

- Process these measurements with the LS-estimator, and with the IV-estimator with the shift parameter $s = 1$.
- Experiment 2: Generate 1000 experiments with $N = 5000$ measurements each, and the following parameter settings:

$$f_{\text{Gen}} = 0.1\,,\ f_{\text{Noise}} = 0.6\,,\ \sigma_{i_0} = 0.1\,,\ \sigma_{n_i} = 0.1\,,\ \sigma_{n_u} = 1\,. \tag{7-48}$$

- Process these measurements with the LS-estimator, and with the IV-estimator with the shift parameter $s = 1, 2, 5$.

Plot for both experiments:

- the pdf of $\hat{R}_{\text{LS}}$ and $\hat{R}_{\text{IV}}$,
- the auto-correlation function of i_0 and n_i (hint: use the Matlab™ instruction `xcorr`)
- the FRF of the generator and the noise filter.

h

Observations - The results are shown below Figure 7-16 and Figure 7-17. In the first Figure 7-16, the results are shown for a fixed generator filter and a varying noise filter. The shift parameter for the IV is kept constant to 1. From this figure it is clearly seen that the LS are strongly biased. This is due to the noise on the input, the relative bias is in the order of $\sigma_{n_i}^2 / \sigma_{i_0}^2$. For the IV-results, the situation is more complicated. For the white noise situation, no bias is visible. However, once the output noise is filtered, a bias becomes visible. The relative bias is proportional to the ratio of the auto correlation functions of the noise and the current $R_{n_i n_i}(s) / R_{u_0 u_0}(s)$.
The same observations can also be made in Figure 7-17. In this figure, the shift parameter is changed while the filters are kept constant. It can be seen that the bias becomes smaller with increasing shift s, because $R_{n_i n_i}(s) / R_{i_0 i_0}(s)$ is getting smaller. At the same time the dispersion is growing, mainly because $R_{i_0 i_0}(s)$ is getting smaller. Observe also that the sign of the bias depends on the sign of $R_{n_i n_i}(s)$. The IV-method works well if the bandwidth of the generator signal is much smaller than that of the noise disturbances.

Exercise 11.b (Noise on input and output: the errors-in-variables method) : In this exercise the EIV-method is used as an alternative for IV-method to reduce/eliminate the

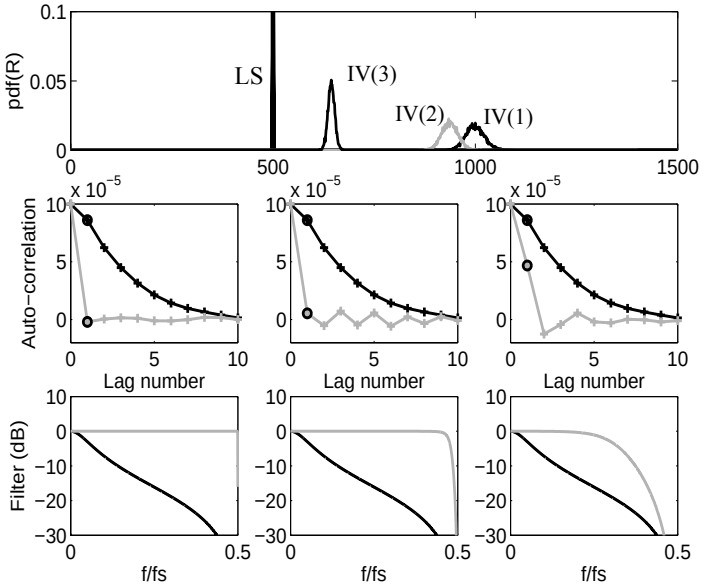


Figure 7-16 Study of the LS- and IV-estimate for a varying noise filter bandwidth and fixed shift $s = 1$. Top: the LS (black line) and IV estimate (black or gray line). IV(1), IV(2), and IV(3) correspond to the first second, and third filter. Middle: the auto correlation of i_0 (black) and n_i (gray) for the different noise filters. Bottom: the filter characteristics of i_0 (black) and the noise n_i (gray).

bias of the least squares estimate. This time no constraint is put on the power spectra (bandwidth) of the excitation and the disturbing noise, but instead the variance of the input and output disturbing noise should be priorly given. This is illustrated again on the resistance example with measured current and voltage $i(t), u(t)$, $t = 1, 2, \ldots, N$.

The least squares estimate is given by

$$\hat{R}_{\mathrm{LS}} = \frac{\sum_{k=1}^{N} u(k)i(k)}{\sum_{k=1}^{N} i(k)^2}, \tag{7-49}$$

the EIV-estimator is

$$\hat{R}_{\mathrm{EIV}} = \frac{\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2} + \sqrt{\left(\frac{\sum u(t)^2}{\sigma_u^2} - \frac{\sum i(t)^2}{\sigma_i^2}\right)^2 + 4\frac{\left(\sum u(t)i(t)\right)^2}{\sigma_u^2\sigma_i^2}}}{\frac{\sum u(t)i(t)}{\sigma_u^2}}, \tag{7-50}$$

where the sum runs over $t = 1, \ldots, N$. It is shown to be the minimizer of the following cost function:

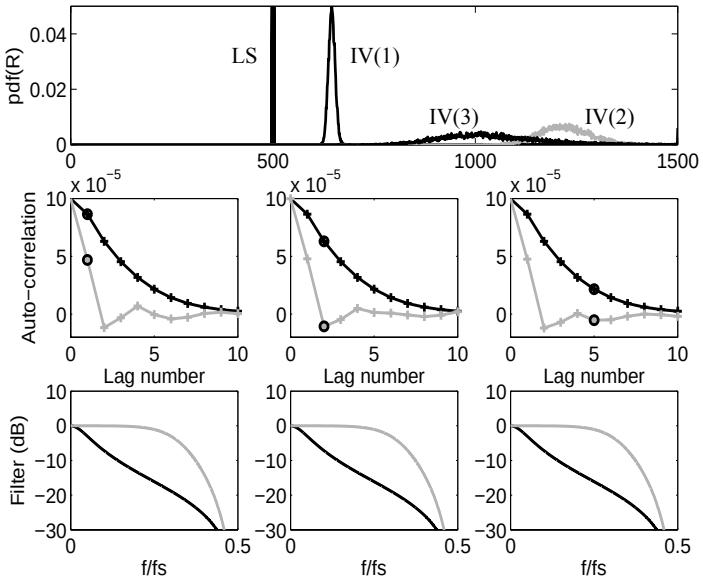


Figure 7-17 Study of the LS- and IV-estimate for a fixed noise filter bandwidth and a varying shift s =1, 2, 5. Top: the LS (black) and IV (black and gray) estimate. IV(1), IV(2), and IV(3) correspond to a shift of 1,2, and 5 tabs. Middle: the auto correlation of i_0 (black) and n_i (gray). Bottom: the filter characteristics of i_0 (black) and the noise n_i (gray)

$$V_{\text{EIV}} = \frac{1}{N}\sum_{t=1}^{N}\left\{\frac{(u(t)-u_0(t))^2}{\sigma_u^2}+\frac{(i(t)-i_0(t))^2}{\sigma_i^2}\right\}, \tag{7-51}$$

with respect to u_0, i_0, R_0 under the constraint $u_0(t) = R_0 i_0(t)$.

- Setup: Generate the current $i_0(t)$ from a white zero mean Gaussian noise source $N(0, \sigma_{i_0}^2)$.

  Generate the measured current and voltage as:

$$\begin{aligned} i(t) &= i_0(t) + n_i(t) \\ u(t) &= u_0(t) + n_u(t) \end{aligned}, \tag{7-52}$$

  $n_u(t)$ and $n_i(t)$ are white Gaussian noise sources with zero mean and variance $\sigma_{n_u}^2$ and $\sigma_{n_i}^2$ respectively.

- Generate a set of 1000 experiments with $N = 5000$ measurements each, and the following parameter settings:

$$R_0 = 1000\,,\ \sigma_{i_0} = 0.01\,,\ \sigma_{n_i} = 0.001\,,\ \sigma_{n_u} = 1\,. \tag{7-53}$$

Calculate the LS- and EIV-estimate. Plot the histogram with $\hat{R}_{\text{LS}}$ and $\hat{R}_{\text{EIV}}$.

Observations - The results are shown below in Figure 7-18., From this figure it is

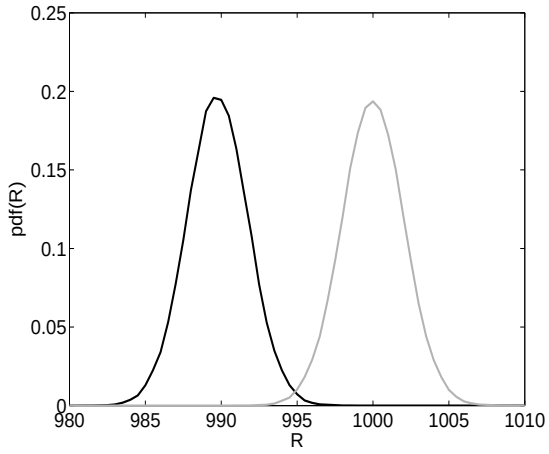


Figure 7-18 Comparison of the pdf of the LS- (black) and the EIV-estimate (gray), calculated with prior known variances.

clearly seen that the LS are strongly biased (mean value is 990.15). This is due to the noise on the input, the relative bias is in the order of $\sigma_{n_i}^2 / \sigma_{i_0}^2$. No systematic error can be observed in the EIV-results (mean value is 999.96). The IV-estimate would fail completely in this situation (why?).

Chapter 8

Further Reading

Chapter 8

Further reading

In this chapter we give a list of books dealing with identification. Most of them deal with the identification of linear dynamic systems using a discrete time representation. Only a few books are focussed on the general identification problem.

- *System Identification*: A frequency domain approach, by R. Pintelon and J. Schoukens, IEEE Press and John Wiley & Sons, Inc.
  - The first and second chapter of this course are based on this book. It presents a general approach to the identification of linear dynamic systems from noisy data. The book gives a general approach to this problem, with both practical examples and theoretical discussions that give the reader a sound understanding of the subject and the pitfalls that might occur on the road from raw data to validated model.
- *System Identification*, theory for the user, by L. Ljung, Prentice-Hall, 1987
  - This is a quite exceptional book that is focused on the identification of linear dynamic systems. It provides the user with working algorithms starting from a basic introduction to the identification theory. A commercial software package (toolbox) MATLAB™is available. Each chapter ends with a summary and exercises.
- *Perturbation Signals for System Identification*, edited by K. Godfrey, Prentice Hall, 1993.

  - An edited book that is focused on the design of excitation signals that can be used to identify a system.

- *System Identification*, T. Söderström and P. Stoica, Prentice-Hall, 1989

  - The book addresses the needs of both newcomers and experts in the field. It introduces the reader to basic concepts and results. There are also advanced results given at the end of the chapters. Each chapter ends with a summary and exercises. The theory is illustrated on a lot of simulation examples, however no real measurement results are shown.

- *Theory of optimal Experiments*, V. Fedorov, Academic Press, 1972

  - This book gives a mathematical treatment of the problem of the optimal design of experiments.This book gives a mathematical treatment of the problem of the optimal design of experiments.

- *System Identification*, Parameter and State Estimation, P. Eykhoff, John Wiley, 1974.

  - This is one of the first books providing a general and coherent introduction to the identification problem of linear dynamic systems.

- *System Identification:* Advances and case studies., edited by R. Mehra and D. Lainiotis, Academic Press, 1976

  - The setup of this book is very similar to the book edited by P. Eykhoff: System Identification.

- *Dynamic Identification:* Experiment Design and Data Analysis, G. Goodwin, R. Payne, 1977, Academic Press

  - This book gives a general introduction to the theory of mathematical model building, using experimental data.

- *Parameter Estimation in Engineering and Science*, J. Beck, K.Arnold, John Wiley & Sons, 1977

  - The objectives of this book are to provide (1) methods for estimating constants (i.e., parameters) appearing in mathematical models, (2) estimates of the accuracy of the estimated parameters, and (3) tools and insights for developing improved mathematical models. The book presents methods that can use all the available statistical information.

- *Optimal Experiment Design for Dynamic System Identification*, M. Zarrop, Springer-Verlag, 1979
  - This book is concerned with the problem of experiment design for the efficient identification of a linear single input, single output dynamic system from input-output data in the presence of disturbances.
- *Parameter Estimation*, principles and problems, H. Sorenson, Marcel Dekker, inc., 1980
  - The purpose of this book is to present the fundamental concepts and major results of parameter estimation theory (not focused on the identification of linear dynamic systems) in a manner that will make the material accessible to as large and audience as possible.
- *Trends and Progress in System Identification*, edited by P. Eykhoff, Pergamon Press, 1981.
  - This book provides a profound introduction to system identification. It is divided into different parts, covering the whole bench of identification problems (modelling, estimating, optimal experimentation) written by specialists on each specific topic.
- *Spectral Analysis and Time Series*, M. Priestly, Academic Press, 1981.
  - This book gives a profound tutorial review of the spectral analysis and time series analysis problem.
- *Adaptive Filtering, Prediction and Control*, G. Goodwin, K. Sang Sin, Prentice-Hall, 1984
  - This book is designed to be a unified treatment of the theory of adaptive filtering, prediction and control. It is largely confined to linear discrete-time systems and explores the natural extensions to nonlinear systems.
- *Adaptive Signal Processing*, B. Widrow, S. Stearns, Prentice Hall, Inc., 1985.
  - The purpose of this book is to present certain basic principles of adoption; to explain the design, operating characteristics, and applications of the simple forms of adaptive systems; and to describe means for their physical realization. The types of systems discussed include those designed primarily for the purposes of adaptive control and adaptive signal processing.

- *An Introduction to Identification*, J.P. Norton, Academic Press, 1986.
  - The aim of this book is to provide a general introduction to identification on the undergraduate and introductory graduate level.
- *Identification of continuous systems*, H. Unbehauen, G. Rao, North-Holland, 1987
  - This book deals with certain recent trends in continuous model identification in view of several advantages in retaining the models of actually time-continuous dynamical systems in continuous time-domain without resorting to discretization for identification.
- *Theory and Practice of Recursive Identification*, L. Ljung and T. Söderström, MIT Press, 1987
  - The book unfolds a systematic framework for developing, describing, and analysing the algorithms that may be used in a wide spectrum of on-line adaptive systems, and will serve as a guide to the large number of algorithms now in use.
- *System Modeling and Identification*, R. Johansson, Prentice Hall, 1993
  - The book provides a general introduction to system identification at the undergraduate level.
- *Identification of Continuous-Time Systems*, edited by N.K. Sinha and G.P. Rao
  - It is an edited book consisting of a series of specialized chapters dealing with the identification of continuous time systems starting from discrete time measurements, written by specialists in the field.
- *Modeling of dynamic systems*, L. Ljung and T. Glad, Prentice Hall, 1994.
  - This book is focused on building models starting from measurements and physical insight.
- *Identification of Parametric Models from Experimental Data*, Éric Walter and Luc Pronzato, Springer, 1997.
  - The emphasis of the book is put on the practical aspects of the general identification problem (numerical aspects of optimization, experiment design, uncertainty calculation). Also attention is paid on other criteria than least squares.

Course on System Identification

Transparencies chapter 1
An introduction to identification

Johan Schoukens

Vrije Universiteit Brussel

Johan.Schoukens@vub.ac.be

Goal of the course

From data to model

Basic steps

- Choice of an experiment: collect data
- choice of a model
- match data and model: choice of an estimator
- model validation/selection

Apply it to the identification of linear dynamic systems

A general introduction to identification

A motivating example: why do *you* need system identification?!

Describing the stochastic behaviour of estimates

Basic steps of the identification process

A statistic approach to system identification

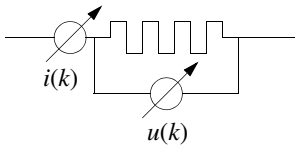
Why do you need identification methods?

A simple experiment

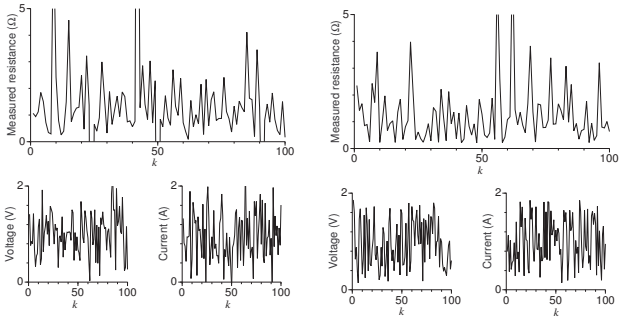
Multiple measurements lead to conflicting results.

How to combine all this information?

Why do you need identification methods
Measurement of a resistance



2 sets of measurements



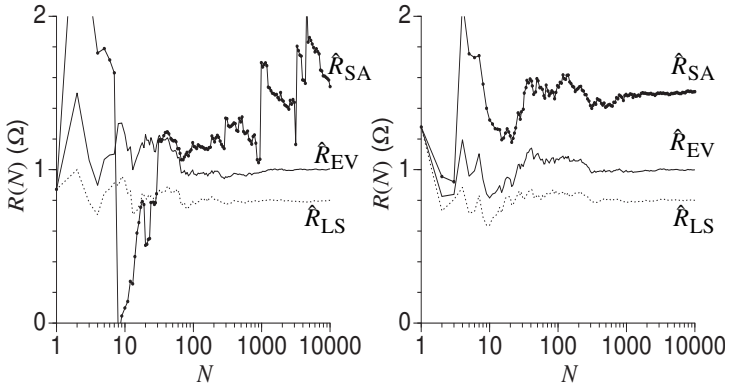
3 different estimators

$$\hat{R}_{\mathrm{SA}}(N) = \frac{1}{N}\sum_{k=1}^{N}\frac{u(k)}{i(k)}$$

$$\hat{R}_{\mathrm{LS}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k)}{\frac{1}{N}\sum_{k=1}^{N} i(k)^2}$$

$$\hat{R}_{\mathrm{EV}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)}{\frac{1}{N}\sum_{k=1}^{N} i(k)}$$

and their results

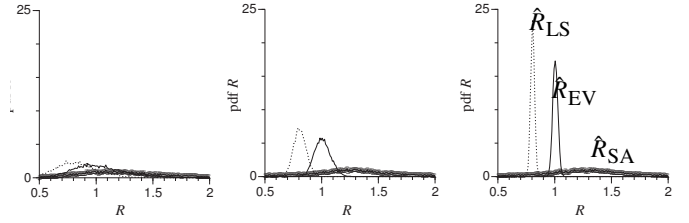


Group A Group B

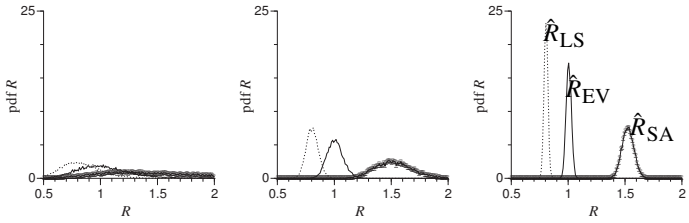
Remarks

- variations decrease as function of N, except for $\hat{R}_{\mathrm{LS}}$
- the asymptotic values are different
- $\hat{R}_{\mathrm{SA}}$ behaves 'strange'

Repeating the experiments.



Group A

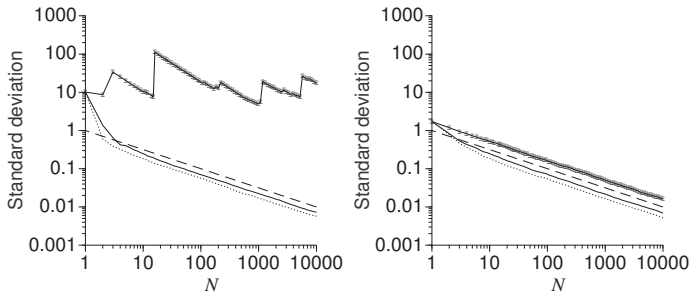


Group B

Observed pdf of $\hat{R}(N)$ for both groups, from the left tot the right $N = 10$, 100, and 1000

- the distributions become more concentrated around their limit value
- $\hat{R}_{SA}$ behaves 'strange' for group A

Repeating the experiments

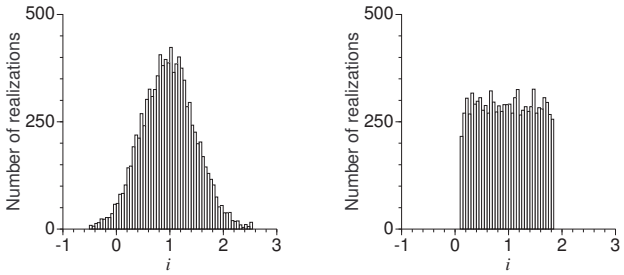


Group Group B

Standard deviation of $\hat{R}(N)$ for the different estimators, and comparison with $1/\sqrt{N}$: full dotted line: $\hat{R}_{SA}$, dotted line: $\hat{R}_{LS}$, full line: $\hat{R}_{EV}$, dashed line $1/\sqrt{N}$.

- the standard deviation decrease in $\sqrt{N}$
- the uncertainty also depends on the estimator

Strange behaviour of $\hat{R}_{\mathrm{LS}}$ for group A



Group A Group B

Histogram of the current measurements.

- The current takes negative values for group A

- the estimators tend to a normal distribution although the noise behaviour is completely different

Simplified analysis

Why do the asymptotic values depend on the estimator?

Can we explain the behaviour of the variance?

Why does the $\hat{R}_{\mathrm{SA}}$ estimator behave strange for group A?

More information is needed to answer these questions

- **noise model of the measurements**

$$i(k) = i_0 + n_i(k) \qquad u(k) = u_0 + n_u(k)$$

- **Assumption**: $n_i(k)$ and $n_u(k)$ are mutually independent zero mean iid (independent and identically distributed) random variables with a symmetric distribution and with variance σ_u^2 and σ_i^2.

Statistical tools

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} x(k) = 0$$

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} x(k)^2 = \sigma_x^2$$

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} x(k)y(k) = 0$$

Asymptotic value of $\hat{R}_{\mathrm{LS}}$

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N) = \lim_{N\to\infty}\left(\sum_{k=1}^{N} u(k)i(k)\right)\Big/\left(\sum_{k=1}^{N} i^2(k)\right)$$

$$= \lim_{N\to\infty}\frac{\frac{1}{N}\sum_{k=1}^{N}(u_0+n_u(k))(i_0+n_i(k))}{\frac{1}{N}\sum_{k=1}^{N}(i_0+n_i(k))^2}$$

Or

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N) =$$

$$\lim_{N\to\infty}\frac{u_0i_0+\frac{u_0}{N}\sum_{k=1}^{N}n_i(k)+\frac{i_0}{N}\sum_{k=1}^{N}n_u(k)+\frac{1}{N}\sum_{k=1}^{N}n_u(k)n_i(k)}{i_0^2+\frac{1}{N}\sum_{k=1}^{N}n_i^2(k)+\frac{2i_0}{N}\sum_{k=1}^{N}n_i(k)}$$

And finally

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N) = \frac{u_0i_0}{i_0^2+\sigma_i^2} = R_0\frac{1}{1+\sigma_i^2/i_0^2}$$

It converges to the wrong value!!!

Asymptotic value of $\hat{R}_{\mathrm{EV}}$

$$\begin{aligned}\lim_{N\to\infty}\hat{R}_{\mathrm{EV}}(N) &= \lim_{N\to\infty}\left(\sum_{k=1}^{N}u(k)\right)\Big/\left(\sum_{k=1}^{N}i(k)\right)\\ &= \lim_{N\to\infty}\frac{\frac{1}{N}\sum_{k=1}^{N}(u_0+n_u(k))}{\frac{1}{N}\sum_{k=1}^{N}(i_0+n_i(k))}\\ &= \left(\lim_{N\to\infty}\frac{u_0+\frac{1}{N}\sum_{k=1}^{N}n_u(k)}{i_0+\frac{1}{N}\sum_{k=1}^{N}n_i(k)}\right)\\ &= R_0\end{aligned}$$

It converges to the exact value!!!

Asymptotic value of $\hat{R}_{\mathrm{LS}}$

$$\hat{R}_{\mathrm{SA}}(N) = \frac{1}{N}\sum_{k=0}^{N}\frac{u(k)}{i(k)} = \frac{1}{N}\sum_{k=0}^{N}\frac{u_0+n_u(k)}{i_0+n_i(k)} = \frac{1}{N}\frac{u_0}{i_0}\sum_{k=0}^{N}\frac{1+n_u(k)/u_0}{1+n_i(k)/i_0}$$

The series expansion exist only for small noise distortions

$$\frac{1}{1+x} = \sum_{l=0}^{\infty}(-1)^l x^l \text{ for } |x|<1$$

Group A: A detailed analysis shows that the expected value does not exist for the data of group A.

The estimator does not converge.

Group B: For group B the series converges and

$$\lim_{N\to\infty}\hat{R}_{\mathrm{SA}}(N) = R_0\left(1+\frac{\sigma_i^2}{i_0^2}\right)$$

The estimator converges to the wrong value!!

Variance expressions

First order approximation

$$\sigma^2_{\hat{R}_{\mathrm{LS}}}(N) = \sigma^2_{\hat{R}_{\mathrm{EV}}}(N) = \sigma^2_{\hat{R}_{\mathrm{SA}}}(N) = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right)$$

- variance decreases in $1/N$

- variance increases with the noise

- for low noise levels, all estimators have the same uncertainty

---> **Experiment design**

Cost function interpretation

The previous estimates match the model $u = Ri$ as good as possible on the data.

A criterion to express the goodness of the fit is needed ----> Cost function interpretation.

$\hat{R}_{\mathrm{SA}}(N)$

$$V_{\mathrm{SA}}(R) = \frac{1}{N}\sum_{k=1}^{N}(R(k) - R)^2.$$

$\hat{R}_{\mathrm{LS}}(N)$

$$V_{\mathrm{LS}}(R) = \frac{1}{N}\sum_{k=1}^{N}(u(k) - Ri(k))^2$$

$\hat{R}_{\mathrm{EV}}(N)$

$$V_{\mathrm{EV}}(R, i_0, u_0) = \frac{1}{N}\left(\sum_{k=1}^{N}(u(k) - u_0)^2 + \sum_{k=1}^{N}(i(k) - i_0)^2\right) \text{ subject to } u_0 = Ri_0$$

Conclusion

- A simple problem

- Many solutions

- How to select a good estimator?

- Can we know the properties in advance?

----> **need for a general framework !!**

Basic steps in identification

1) collect the information: experiment setup

2) select a model

parametric >< nonparametric models

white>< black box models

linear><nonlinear models

linear -in-the-parameters><nonlinear-in-the-parameters

$$\varepsilon = y - (a_1 u + a_2 u^2),\ \varepsilon(\omega) = Y(\omega) - \frac{a_0 + a_1 j\omega}{b_0 + b_1 j\omega} U(\omega)$$

3) match the model to the data

select a cost function

--> LS, WLS, MLE, Bayes estimation

4) validation

does the model explain the data?

can it deal with new data?

Remark: this scheme is not only valid for the classical identification theory. It also applies to neural nets, fuzzy logic, ...

Characterizing estimators

Location properties: are the parameters concentrated around the ‘exact value’ ?

Dispersion properties: is the uncertainty small or large?

Location properties

unbiased and consistent estimators

Unbiased estimates: the mean value equals the exact value

Definition

An estimator $\hat{\theta}$ of the parameters θ_0 is unbiased if $E\{\hat{\theta}\} = \theta_0$, for all true parameters θ_0. Otherwise it is a biased estimator.

Asymptotic unbiased estimates: unbiased for $N \to \infty$

Example

The sample mean

$$\hat{u}(N) = \frac{1}{N}\sum_{k=1}^{N} u(k)$$

Unbiased?

$$E\{\hat{u}(N)\} = \frac{1}{N}\sum_{k=1}^{N} E\{u(k)\} = \frac{1}{N}\sum_{k=1}^{N} u_0 = u_0$$

The sample variance

$$\hat{\sigma}_u^2(N) = \frac{1}{N}\sum_{k=1}^{N} (u(k) - \hat{u}(N))^2$$

Unbiased?

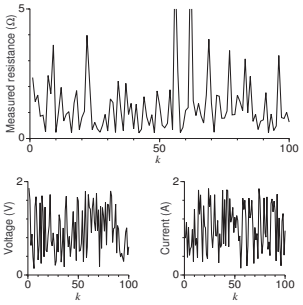
$$E\{\hat{\sigma}_u^2(N)\} = \frac{N-1}{N}\sigma_u^2$$

Alternative expression

$$\frac{1}{N-1}\sum_{k=1}^{N} (u(k) - \hat{u}(N))^2$$

Consistent estimates: the probability mass gets concentrated around the exact value

$$\lim_{N \to \infty} \text{Prob}(\left|\hat{\theta}(N) - \theta_0\right| > \delta > 0) = 0$$



Example

$$\begin{aligned}\operatorname{plim}_{N\to\infty}\hat{R}_{\mathrm{EV}}(N) &= \operatorname{plim}_{N\to\infty}\frac{\frac{1}{N}\sum_{k=1}^{N}u(k)}{\frac{1}{N}\sum_{k=1}^{N}i(k)}\\ &= \frac{\operatorname{plim}_{N\to\infty}\left(\frac{1}{N}\sum_{k=1}^{N}u(k)\right)}{\operatorname{plim}_{N\to\infty}\left(\frac{1}{N}\sum_{k=1}^{N}i(k)\right)}\\ &= \frac{u_0}{i_0}\\ &= R_0\end{aligned}$$

Dispersion properties

efficient estimators

- Mostly the covariance matrix is used, however alternatives like percentiles exist.

- For a given data set, there exists a minimum bound on the covariance matrix: the **Cramér-Rao lower bound**.

$$CR(\theta) = Fi^{-1}(\theta_0)$$

with

$$Fi(\theta_0) = E\left\{\left(\frac{\partial}{\partial\theta}l(Z|\theta)\right)^T\left(\frac{\partial}{\partial\theta}l(Z|\theta)\right)\right\} = -E\left\{-\frac{\partial^2}{\partial\theta^2}l(Z|\theta)\right\}.$$

The derivatives are calculated in $\theta = \theta_0$

The likelihood function

1) Consider the measurements $Z \in \mathrm{R}^N$

2) Z is generated by a hypothetical, exact model with parameters θ_0

3) Z is disturbed by noise --> stochastic variables

4) Consider the probability density function $f(Z|\theta_0)$ with

$$\int_{z \in \mathrm{R}^N} f(Z|\theta_0) dZ = 1 .$$

5) Interpret this relation conversely, viz:

how likely is it that a specific set of measurements $Z = Z_m$ are generated by a system with parameters θ?

In other words, we consider now a given set of measurements and view the model parameters as the free variables:

$$L(Z_m|\theta) = f(Z = Z_m|\theta),$$

with θ the free variables.

$L(Z_m|\theta)$ is called the likelihood function.

Example

Determine the flow of tap water by measuring the height $h_0(t)$ of the water in a measuring jug as a function of time t

Model

$$h_0(t) = a(t - t_{\text{start}}) = at + b \text{ with } \theta = [a, b]$$

Measurements

$$h(k) = at_k + b + n_h(k)$$

Noise model

$$n_h(k): \text{ iid zero mean normally distributed } N(0, \sigma^2)$$

Likelihood function

for the set of measurements $h = \{(h(1), t_1), ..., (h(N), t_N)\}$:

$$L(h|a, b) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{N} (h(k) - at_k - b)^2}$$

Example Continued

Log likelihood function

$$l(h|a,b) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(h(k) - at_k - b)^2$$

Information matrix

$$Fi(\theta_0) = E\left\{\left(\frac{\partial}{\partial\theta}l(Z|\theta)\right)^T\left(\frac{\partial}{\partial\theta}l(Z|\theta)\right)\right\} = -E\left\{-\frac{\partial^2}{\partial\theta^2}l(Z|\theta)\right\}$$

$$Fi(a,b) = \frac{1}{\sigma^2}\begin{bmatrix} Ns^2 & N\mu \\ N\mu & N \end{bmatrix},$$

Cramér-Rao lower bound

$$CR(a,b) = \frac{\sigma^2}{N(s^2-\mu^2)}\begin{bmatrix} 1 & -\mu \\ -\mu & s^2 \end{bmatrix}$$

$$\text{with } \mu = \frac{1}{N}\sum_{k=1}^{N} t_k \text{ and } s^2 = \frac{1}{N}\sum_{k=1}^{N} t_k^2.$$

Example continued

Case 1: a and b unknown: consider $Fi^{-1}(a,b)$

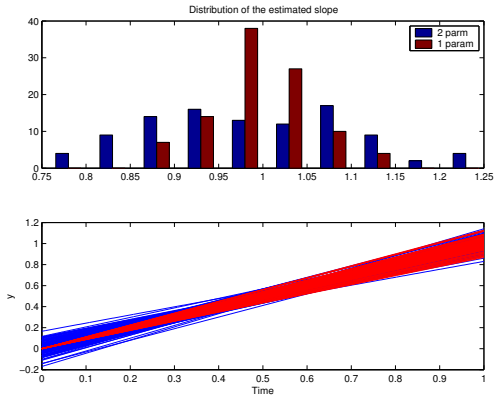
$$\sigma_a^2(a,b) = \frac{\sigma^2}{N(s^2-\mu^2)}$$

Case 2: a unknown: consider $Fi^{-1}(a)$

$$\sigma_a^2(a) = \frac{\sigma^2}{Ns^2}$$

Discussion points

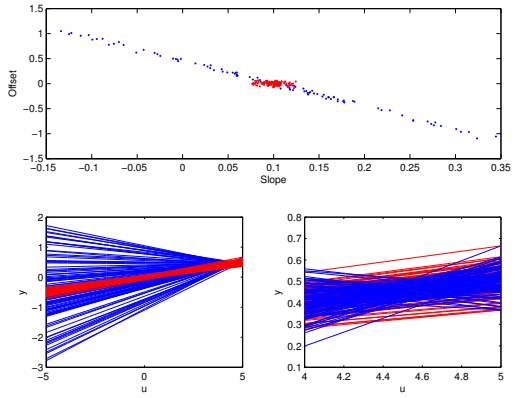
- impact of the number of measurements
- impact of the number of parameters
- the analysis is done without selecting an estimator



$y(t_k) = a_0 t_k + n_k$, with $t_k = [0{:}N]/N$ and $n_k \sim N(0, \sigma^2) = 1$, $a_0 = 1$

model 1: $y = at + b$ (two parameters)

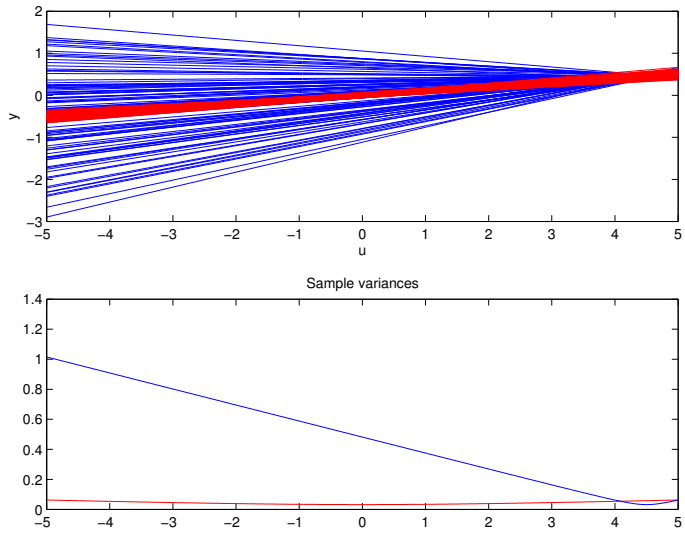
model 2: $y = at$ (one parameter)



IImpact experimtent design: red [-5:5], blue [4 5]

$$C_{\text{exp1}} = \begin{bmatrix} 1.5\times10^{-4} & 0.48\times10^{-4} \\ 0.48\times10^{-4} & 9.1\times10^{-4} \end{bmatrix} \quad \text{, and} \quad C_{\text{exp2}} = \begin{bmatrix} 1.5\times10^{-2} & -6.7\times10^{-2} \\ -6.7\times10^{-2} & 30.1\times10^{-2} \end{bmatrix}$$

$$R_{\text{exp1}} = \begin{bmatrix} 1 & 0.13 \\ 0.13 & 1 \end{bmatrix} \quad \text{, and} \quad R_{\text{exp2}} = \begin{bmatrix} 1 & -0.9985 \\ -0.9985 & 1 \end{bmatrix}$$



Interpretation of the covariance matrix, and the impact of the experiment design on the model uncertainty.

A statistical framework: choice of the cost functions

$$y_0 = G(u, \theta_0), \; y = y_0 + n_y, \; e = y - G(u, \theta_0)$$

Least squares estimation

$$V_{\mathrm{LS}}(\theta) = \frac{1}{N}\sum_{k=1}^{N} e^2(k, \theta)$$

Weighted least squares estimation

$$V_{\mathrm{WLS}}(\theta) = \frac{1}{N} e(\theta)^T W e(\theta)$$

Maximum likelihood estimation

$$f(y|\theta_0) = f_{n_y}(y - G(u, \theta_0))$$

$$\theta_{\mathrm{ML}} = \underset{\theta}{\mathrm{argmax}} f(y_m|\theta)$$

Transparencies
A statistical framework

$$y_0 = G(u, \theta_0), y = y_0 + n_y, e = y - G(u, \theta_0)$$

Least squares estimation

$$V_{\mathrm{LS}}(\theta) = \frac{1}{N}\sum_{k=1}^{N} e^2(k, \theta)$$

Weighted least squares estimation

$$V_{\mathrm{WLS}}(\theta) = \frac{1}{N} e(\theta)^T W e(\theta)$$

Maximum likelihood estimation

$$f(y|\theta_0) = f_{n_y}(y - G(u, \theta_0))$$

$$\theta_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}} f(y_m|\theta)$$

Bayes estimators

Transparencies Chapter 2

A statistical framework

Least squares: principle

Model

$$y_0(k) = g(u_0(k), \theta)$$

with k the measurement index, and

$$y(k) \in \mathrm{R},\ u(k) \in \mathrm{R}^{1xM},\ \theta \in \mathrm{R}^{n_\theta x 1}$$

Measurements

$$y(k) = y_0(k) + n_y(k)$$

Match model and measurements

Choose:

$$e(k, \theta) = y(k) - y(k, \theta),$$

with $y(k, \theta)$ the modelled output.

Then

$$\hat{\theta}_{\mathrm{LS}} = \arg\min_{\theta} V_{\mathrm{LS}}(\theta),$$

with

$$V_{\mathrm{LS}}(\theta) = \frac{1}{N}\sum_{k=1}^{N} e^2(k, \theta)$$

Least squares: minimization of the cost

Example: Gauss-Newton method

1) Jacobian matrix $J \in \mathrm{R}^{N \times n_\theta}$:

$$J(\theta) = \frac{\partial}{\partial \theta} e(\theta)$$

2) Consider the Hessian matrix:

$$\frac{\partial^2}{\partial \theta^2} V_{\mathrm{LS}}(\theta) = J(\theta)^T J(\theta) - \frac{1}{N}\sum_{k=1}^{N} e(k, \theta) \frac{\partial^2}{\partial \theta^2} g(u_0(k), \theta)\ .$$

3) Approximation

$$\frac{\partial^2}{\partial \theta^2} V_{\mathrm{LS}}(\theta) \approx J(\theta)^T J(\theta)$$

4) Algorithm

$$\theta_{l+1} = \theta_l + \delta_l(\theta_l)$$

with

$$J(\theta_l)^T J(\theta_l) \delta_l = -J(\theta_l)^T e(\theta_l)$$

Least squares: special case
model that is linear-in-the-parameters

$$y_0 = K(u_0)\,\theta_0$$

$$e(\theta) = y - K(u)\theta, K = -\frac{\partial e}{\partial \theta} = -J$$

$$\hat{\theta}_{\mathrm{LS}} = (K^T K)^{-1} K^T y$$

Example

weight of a loaf of bread

model:

$$y_0 = \theta_0$$

measurements:

$$y(k) = y_0 + n_y(k)$$

estimator:

$$e(k) = y(k) - \theta$$

Standard formulation

$$y = K\theta + n_y \text{ with } K = (1, 1, \ldots, 1)^T$$

Solution

$$\hat{\theta}_{\mathrm{LS}} = (K^T K)^{-1} K^T y = \frac{1}{N}\sum_{k=1}^{N} y(k)$$

Weighted least squares

Goal: bring your confidence in the measurements into the problem

Model

$$y_0(k) = g(u_0(k), \theta)$$

with k the measurement index, and

$$y(k) \in \mathrm{R},\ u(k) \in \mathrm{R}^{1xM},\ \theta \in \mathrm{R}^{n_\theta x 1}$$

Measurements

$$y(k) = y_0(k) + n_y(k)$$

confidence in measurement k: $w(k)$

Match model and measurements

Choose:

$$e(k, \theta) = y(k) - y(k, \theta),$$

Then

$$\hat{\theta}_{\mathrm{LS}} = \arg\min_{\theta} V_{\mathrm{LS}}(\theta),$$

with

$$V_{\mathrm{WLS}}(\theta) = \frac{1}{N}\sum_{k=1}^{N} \frac{e^2(k, \theta)}{w(k)}$$

Weighted least squares (continued)

Generalization: use a full matrix to weight the measurements

define: $e(\theta) = (e(1, \theta) \ldots e(N, \theta))$

consider a positive definite matrix W

Then

$$V_{\mathrm{WLS}}(\theta) = \frac{1}{N} e(\theta)^T W^{-1} e(\theta)$$

Special choice:

$$W = E\{n_y n_y^T\} \in R^{NxN}$$

This choice minimizes C_θ

Maximum likelihood estimation

Model

$$y_0(k) = g(u_0(k), \theta)$$

with k the measurement index, and

$$y(k) \in \mathrm{R},\ u(k) \in \mathrm{R}^{1xM},\ \theta \in \mathrm{R}^{n_\theta x 1}$$

Measurements

$$y(k) = y_0(k) + n_y(k)$$

with f_{n_y} the pdf of the noise n_y

Match model and measurements

Choose the experiments such that the model becomes most likely:

Then

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} f(y_m | \theta, u)$$

with

$$f(y | \theta, u) = f_{n_y}(y - G(u, \theta))$$

Maximum likelihood: example

weight of a loaf of bread

Model:

$$y_0 = \theta_0$$

Measurements:

$$y(k) = y_0 + n_y(k)$$

Additional information

The distribution f_y of n_y is normal with zero mean and standard deviation σ_y

Likelihood function:

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\theta)^2}{2\sigma_y^2}} = \frac{1}{\sqrt{2\pi\sigma_y^2}^N} e^{-\frac{1}{2\sigma_y^2}\sum_{k=1}^{N}(y(k)-\theta)^2}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} y(k)$$

Bayes estimator: principle

Choose the parameters that have the highest probability:

$$\hat{\theta} = \arg\max_{\theta} f(\theta|u, y)$$

Problem: prior distribution of the parameters is required

$$f(\theta|u, y) = \frac{f(y|\theta, u)f(\theta)}{f(y)}$$

Properties of the Maximum likelihood estimator

principle of invariance: if $\hat{\theta}_{\mathrm{ML}}$ is a MLE of $\theta \in \mathrm{R}^{n_\theta}$, then $\hat{\theta}_g = g(\hat{\theta}_{\mathrm{ML}})$ is a MLE of $g(\theta)$ where g is a function, $\hat{\theta}_g \in \mathrm{R}^{n_g}$ and $n_g \leq n_\theta$, with n_θ a finite number.

***consistency*:** if $\hat{\theta}_{\mathrm{ML}}(N)$ is an MLE based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}$ converges to θ_0 almost surely: $\underset{N\to\infty}{\mathrm{a.s.lim}}\,\hat{\theta}_{\mathrm{ML}}(N) = \theta_0$.

***asymptotic normality*:** if $\hat{\theta}_{\mathrm{ML}}(N)$ is a MLE based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ converges in law to a normal random variable with the Cramér-Rao lower bound as covariance matrix.

Bayes estimator: example 1

Use of Bayes estimators in our daily life.

Bayes estimator: example 2

weight of a loaf of bread

Model:

$$y_0 = \theta_0$$

Measurements:

$$y(k) = y_0 + n_y(k)$$

Additional information 1: disturbing noise

The distribution f_y of n_y is normal with zero mean and standard deviation σ_y

Additional information 2: prior distribution of the parameters

The bread is normally distributed: $N(800\,\text{gr}, \sigma_w)$

Bayes estimator:

$$f(y|\theta)f(\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\theta)^2}{2\sigma_y^2}} \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{(\theta-w)^2}{2\sigma_w^2}}$$

$$\hat{\theta} = \frac{z/\sigma_y^2 + w/\sigma_w^2}{1/\sigma_y^2 + 1/\sigma_w^2}$$

Example continued

After making several independent measurements $y(1), \ldots, y(N)$

$$f(y|\theta)f(\theta) = \frac{1}{\left(\sqrt{2\pi\sigma_y^2}\right)^N} e^{-\sum_{k=1}^{N} \frac{(y(k)-\theta)^2}{2\sigma_y^2}} \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{(\theta - w)^2}{2\sigma_w^2}}$$

the Bayes estimator becomes

$$\hat{\theta} = \frac{\sum_{k=1}^{N} y(k)/\sigma_y^2 + w/\sigma_w^2}{N/\sigma_y^2 + 1/\sigma_w^2}$$

For a large number of measurements:

$$\hat{\theta} = \frac{\sum_{k=1}^{N} y(k)/\sigma_y^2}{N/\sigma_y^2} = \frac{1}{N}\sum_{k=1}^{N} y(k)$$