1

An Introduction to Identification

Abstract: In this chapter a brief, intuitive introduction to 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. Starting from this example, the advantages of an identification approach for measuring and modeling are shown, and finally a family of estimators is introduced. A comprehensive introduction to identification can be found in, among others, Beck and Arnold (1977), Goodwin and Payne (1977), Norton (1986), Sörenson (1980), and Kendall and Stuart (1979). Basic concepts of statistics such as the expected value, covariance matrix, and probability density function are assumed to be known.

1.1 WHAT IS IDENTIFICATION?

From birth onwards, we interact with our environment. Intuitively, we learn to control our actions by predicting their effect. These predictions are based on an inborn model fitted to reality, using past experiences. Starting from very simple actions (if I push a ball, it rolls), we soon are able to deal with much more complicated challenges (walking, running, biking, playing Ping-Pong). Finally, this process culminates in the design of complicated systems such as radios, airplanes, and mobile phones. We even build models to get a better understanding of our 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 these examples it is seen that we never deal with the whole of nature at once: we focus on the aspects we are interested in and do not 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 today we describe most real-life situations using this "simple" approach. For example, a ball will roll downhill; temperature will rise if the heat has been switched on; it seems it will rain because 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. It soon became clear, however, that it can be very difficult to match a mathematical model to the available observations and experi-

ences. 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, and so on, we always make errors because 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 that makes it impossible to predict exactly their output. Noise in a radio receiver, Brownian motion of small particles, and variation of the wind speed in a thunderstorm are 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 such as "system" and "goodness of fit" will be precisely defined. 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 EXAMPLE

1.2.1 Estimation of the Value of a Resistor

Two groups of students have to measure a resistance. Their measurement setup is shown in Figure 1-1. They pass a constant but unknown current through the resistor. The

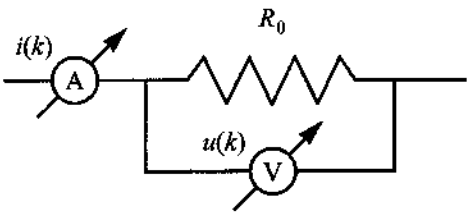

Figure 1-1. Measurement of a resistor using an ammeter (A) and a voltmeter (V).

voltage u_0 across the resistor and the current i_0 through it are 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 $k = 1, 2, ..., N$ is made. The measurement results of each group are shown in Figure 1-2. Because the measurements are very noisy, the groups decide to average their results. Following a lengthy discussion, three estimators for the resistance are proposed:

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

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

$$\hat{R}_{EV}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)}{\frac{1}{N}\sum_{k=1}^{N} i(k)} \tag{1-3}$$

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 process their measurements, and their results are given in Figure 1-3. From this figure a number of interesting observations can be made:

- All estimators have large variations for small values of N and seem to converge to an asymptotic value for large values of N, except $\hat{R}_{SA}(N)$ 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 by 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 two out of the three 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 $\hat{R}_{SA}(N)$ 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 allow us to make a sound selection out of many possibilities and to indicate in advance, before running expensive experiments, whether the selected method is prone to serious shortcomings.

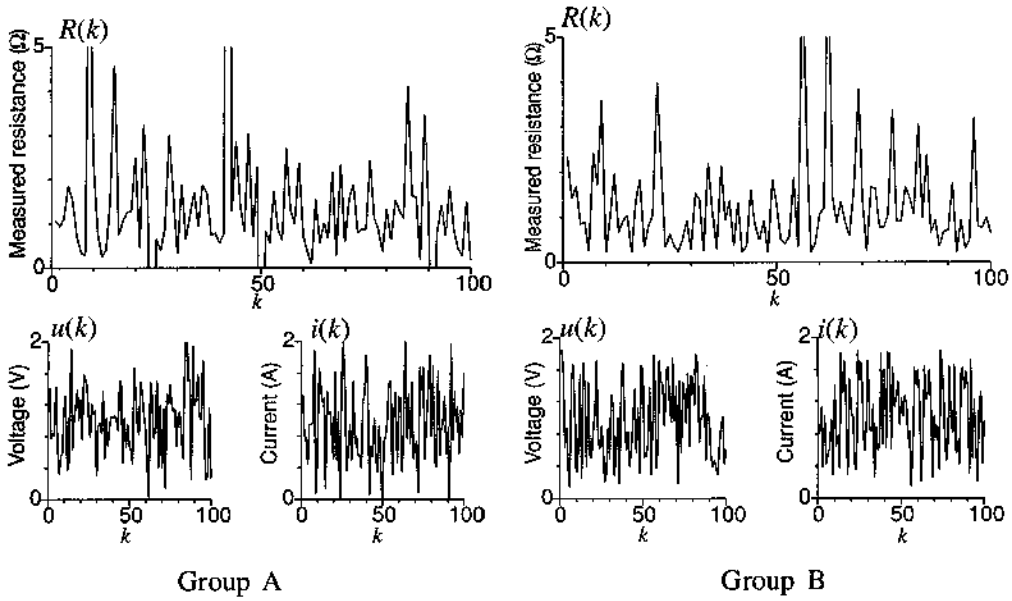

Figure 1-2. Measurement results $u(k)$, $i(k)$ for groups A and B. The plotted value $R(k)$ is obtained by direct division of the voltage by the current: $R(k) = u(k)/i(k)$.

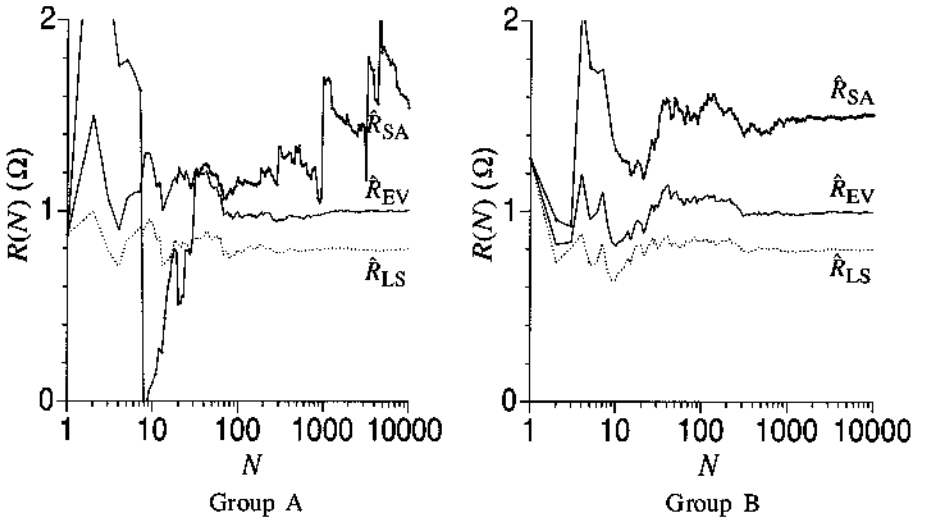

Figure 1-3. Estimated resistance values $\hat{R}(N)$ for both groups as a function of the number of processed data N.

In order to get a better understanding of their results, the students repeat their experiments many times and look 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 Figure 1-4. Again, the students can learn a lot from these figures:

- For small values of N the estimates are widely scattered. As the number of processed measurements increases, the pdf becomes more concentrated.
- The estimates $\hat{R}_{LS}(N)$ are less scattered than $\hat{R}_{EV}(N)$, while for $\hat{R}_{SA}(N)$ 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 still cannot decide which estimator is the best. Moreover, there seems to be a serious problem with the measurements of group A because $\hat{R}_{SA}(N)$ behaves very oddly. First they decide to focus on the scattering of the different estimators, trying to get more insight into the dependence on N. In order to quantify the scattering of the estimates, their standard deviation is calculated and plotted as a function of N in Figure 1-5.

- The standard deviation of $\hat{R}(N)$ decreases monotonically with N except for the pathological case, $\hat{R}_{SA}(N)$, 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 that 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 obtained only for sufficiently large values of N for $\hat{R}_{LS}(N)$ and $\hat{R}_{EV}(N)$.

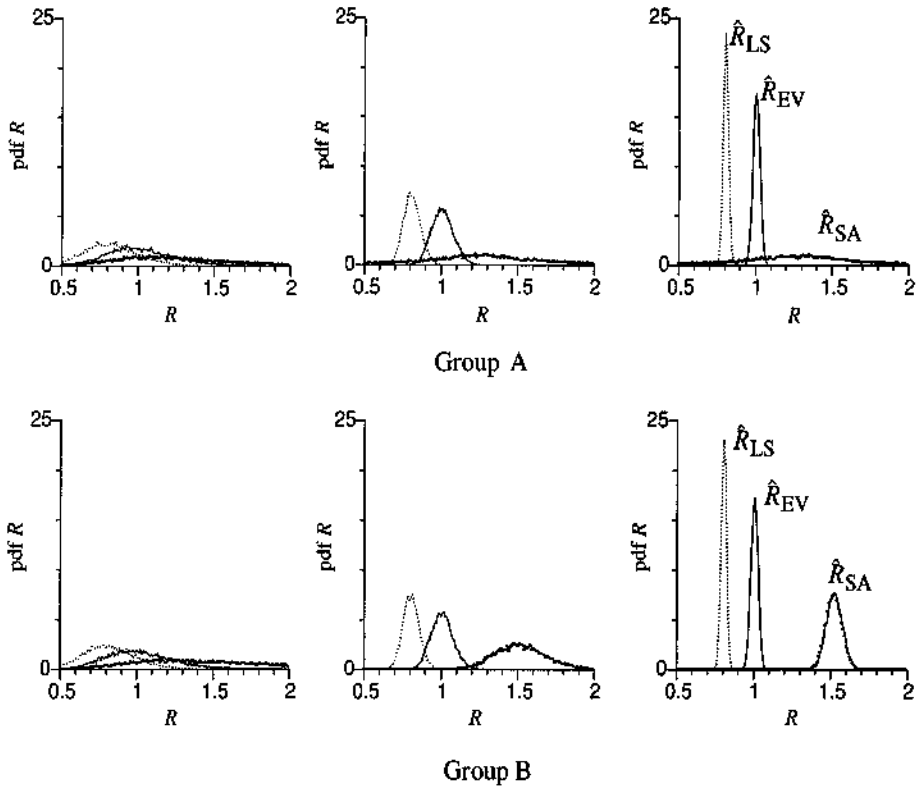

Figure 1-4. Observed pdf of $\hat{R}(N)$ for both groups, from left to right $N = 10$, 100, and 1000.

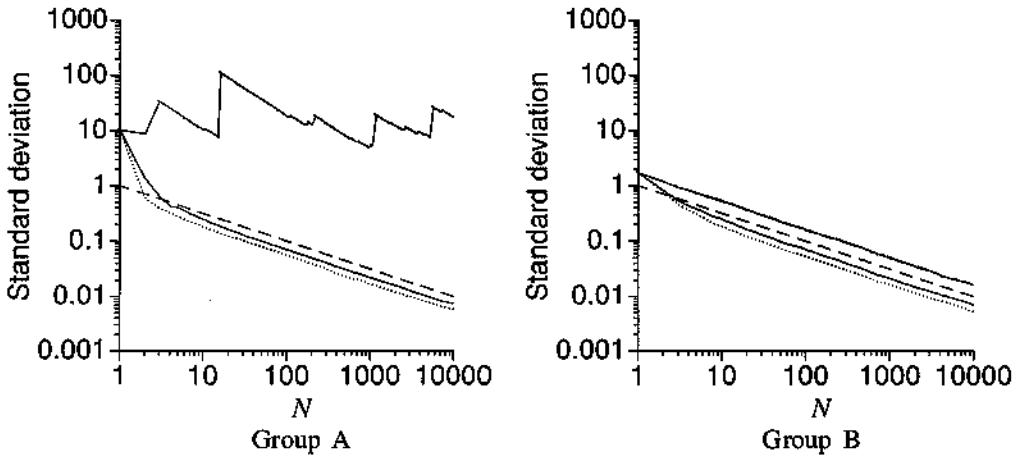

Figure 1-5. Standard deviation of $\hat{R}(N)$ for the different estimators and comparison with $1/\sqrt{N}$; full dotted line: $\hat{R}_{SA}(N)$; dotted line: $\hat{R}_{LS}(N)$, full line: $\hat{R}_{EV}(N)$, dashed line $1/\sqrt{N}$.

Because both groups of students use the same programs to process their measurements, they conclude that the strange behavior of $\hat{R}_{SA}(N)$ in group A should be due to a difference in the raw data. For that reason they take 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 measures some negative values for the current while group B does not. In order to get a better understanding, they make a histogram of the raw current data as shown in Figure 1-6.

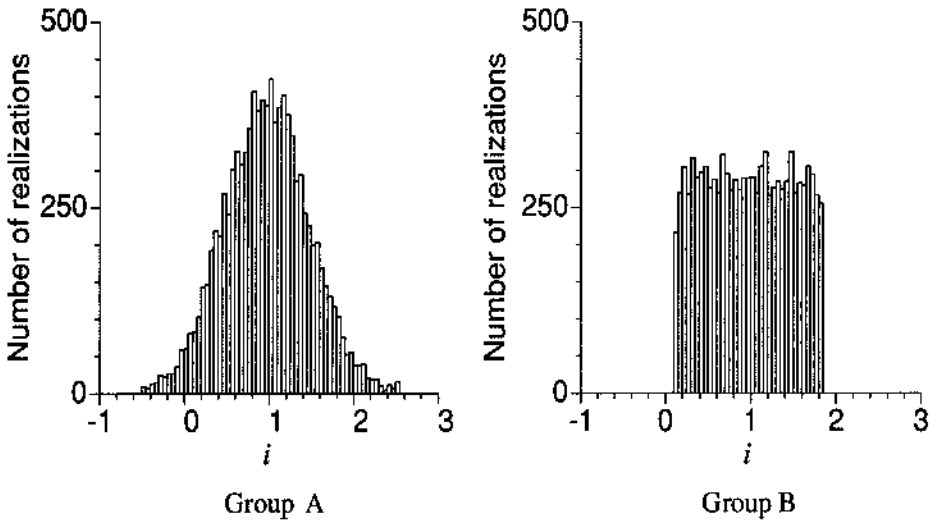

Figure 1-6. Histogram of the current measurements.

These histograms clarify the strange behavior of $\hat{R}_{SA}$ of group A. The noise on the measurements of group A looks completely different from that of group B. Because of the noise on the current measurements, there is a significant risk of getting current values that are very close to zero for group A, whereas this is not so for group B. These small current measurements blow up the estimate $\hat{R}(k) = u(k)/i(k)$ for some k, so that the running average $\hat{R}_{SA}$ cannot converge, or more precisely, the expected value $\mathrm{E}\{u(k)/i(k)\}$ 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 that 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 $\hat{R}_{LS}$ and $\hat{R}_{EV}$ seems to be the same in Figure 1-4 for both groups.

1.2.2 Simplified Analysis of the Estimators

With knowledge obtained from the previous series of experiments, the students eliminate $\hat{R}_{SA}$, but they are still not able to decide whether $\hat{R}_{LS}$ or $\hat{R}_{EV}$ is 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 to Section 16.6. Three observed problems are analyzed in the following:

- Why do the asymptotic values depend on the estimator?
- Can we explain the behavior of the variance?
- Why does the $\hat{R}_{SA}$ 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:

$$i(k) = i_0 + n_i(k) \qquad u(k) = u_0 + n_u(k) \tag{1-4}$$

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 1.1 (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 (see Section 16.9), $\mathbb{E}\{x\}$

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} x(k) = \mathbb{E}\{x\} \tag{1-5}$$

Moreover, if $x(k)$ and $y(k)$ obey Assumption 1.1, then

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} x(k)y(k) = 0 \tag{1-6}$$

Because 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 Section 16.6 for a formal definition).

The first estimator we analyze is $\hat{R}_{LS}(N)$. Taking the limit of (1-2) gives

$$\begin{aligned}\lim_{N\to\infty} \hat{R}_{LS}(N) &= \lim_{N\to\infty} \frac{\sum_{k=1}^{N} u(k)i(k)}{\sum_{k=1}^{N} i^2(k)} \\ &= \frac{\lim_{N\to\infty} \sum_{k=1}^{N} (u_0 + n_u(k))(i_0 + n_i(k))}{\lim_{N\to\infty} \sum_{k=1}^{N} (i_0 + n_i(k))^2}\end{aligned} \tag{1-7}$$

Or, after dividing the numerator and denominator by N,

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N)=\frac{\lim\limits_{N\to\infty}\left[u_0i_0+\frac{u_0}{N}\sum_{k=1}^{N}n_i(k)+\frac{i_0}{N}\sum_{k=1}^{N}n_u(k)+\frac{1}{N}\sum_{k=1}^{N}n_u(k)n_i(k)\right]}{\lim\limits_{N\to\infty}\left[i_0^2+\frac{1}{N}\sum_{k=1}^{N}n_i^2(k)+\frac{2i_0}{N}\sum_{k=1}^{N}n_i(k)\right]}$$

Because 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 converges 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:

$$\lim_{N\to\infty}\hat{R}_{\mathrm{LS}}(N)=\frac{u_0i_0}{i_0^2+\sigma_i^2}=R_0\frac{1}{1+\sigma_i^2/i_0^2} \tag{1-8}$$

This simple analysis gives 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

$$\lim_{N\to\infty}\hat{R}_{\mathrm{EV}}(N)=\lim_{N\to\infty}\frac{\frac{1}{N}\sum_{k=1}^{N}u(k)}{\frac{1}{N}\sum_{k=1}^{N}i(k)}=\frac{\lim\limits_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}(u_0+n_u(k))}{\lim\limits_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}(i_0+n_i(k))} \tag{1-9}$$

or

$$\lim_{N\to\infty}\hat{R}_{\mathrm{EV}}(N)=\frac{u_0+\lim\limits_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_u(k)}{i_0+\lim\limits_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}n_i(k)}=\frac{u_0}{i_0}=R_0 \tag{1-10}$$

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

$$\hat{R}_{SA}(N) = \frac{1}{N}\sum_{k=0}^{N}\frac{u(k)}{i(k)} = \frac{1}{N}\sum_{k=0}^{N}\frac{u_0 + n_u(k)}{i_0 + n_i(k)} \tag{1-11}$$

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

$$\hat{R}_{SA}(N) = \frac{1}{N}\frac{u_0}{i_0}\sum_{k=0}^{N}\frac{1 + n_u(k)/u_0}{1 + n_i(k)/i_0} \tag{1-12}$$

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

$$\frac{1}{1+x} = \sum_{l=0}^{\infty}(-1)^l x^l \text{ for } |x| < 1 \tag{1-13}$$

with $x = n_i(k)/i_0$. Because the terms $n_i^{2l+1}(k)$ and $n_u^l(k)n_i^l(k)$ disappear in the averaging process (the pdfs are symmetrical), the limiting value becomes

$$\lim_{N\to\infty}\hat{R}_{SA}(N) = R_0\left(1 + \frac{1}{N}\sum_{k=1}^{N}(n_i(k)/i_0)^2 + \frac{1}{N}\sum_{k=1}^{N}(n_i(k)/i_0)^4 + \cdots\right) \tag{1-14}$$

with $|n_i(k)/i_0| < 1$. If we neglect all terms of order 4 or more, the final result becomes

$$\lim_{N\to\infty}\hat{R}_{SA}(N) = R_0(1 + \sigma_i^2/i_0^2) \tag{1-15}$$

if $|n_i(k)/i_0| < 1, \ \forall k$.

From this analysis we can draw two important conclusions:

- The asymptotic value exists only if the following condition on the measurements is met: the series expansion must exist otherwise (1-15) is NOT valid. The measurements of group A violate the condition that is given in (1-14), 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 $\mathbb{E}\{\hat{R}_{SA}(N)\}$ exists (see Chapter 17). Because 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, (1-15) 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 $\hat{R}_{EV}(N)$ converges to the exact value, and $\hat{R}_{SA}(N)$ 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 make a variance analysis using first-order Taylor series approximations.

Again they begin with the $\hat{R}_{LS}(N)$. Starting from (1-7) and neglecting all second-order contributions such as $n_u(k)n_i(k)$ or $n_i^2(k)$, it is found that

$$\hat{R}_{LS}(N) \approx R_0\left(1 + \frac{1}{N}\sum_{k=1}^{N}(n_u(k)/u_0 - n_i(k)/i_0)\right) = R_0 + \Delta R \tag{1-16}$$

The approximated variance $\mathrm{var}(\hat{R}_{LS}(N))$ is (using Assumption 1.1)

$$\mathrm{var}(\hat{R}_{LS}(N)) = \mathbb{E}\{(\Delta R)^2\} = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right) \tag{1-17}$$

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

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

$$\mathrm{var}(\hat{R}_{EV}(N)) = \mathrm{var}(\hat{R}_{SA}(N)) = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right) \tag{1-18}$$

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

Again, a number of interesting conclusions can be drawn 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 because 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 variances 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 tune the noise parameters so that they get the same behavior as they observed in their measurements. These values are: $i_0 = 1$ A, $u_0 = 1$ V, $\sigma_i = 1$ A, $\sigma_u = 1$ V. The noise of group A is normally distributed and uniformly distributed for group B. Next they vary the standard deviations and plot the results in Figure 1-7 for $\hat{R}_{EV}(N)$ and $\hat{R}_{LS}(N)$. Here it is clear that for higher SNR the uncertainties coincide, whereas 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); the higher order moments or the pdf of the noise are also required (see Section 16.15).

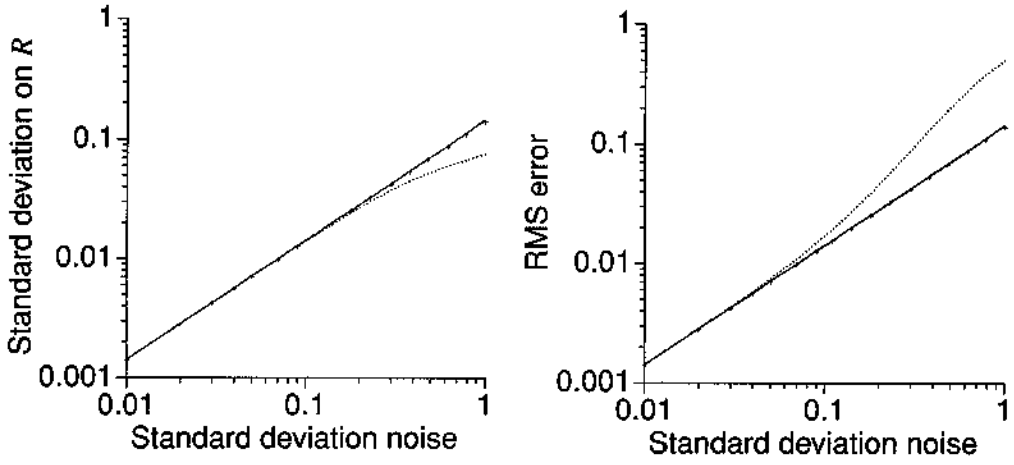

Figure 1-7. Evolution of the standard deviation and the rms error on the estimated resistance value as a function of the standard deviation of the noise ($\sigma_u = \sigma_i$). Solid lines: $\hat{R}_{EV}(N)$, dotted lines : $\hat{R}_{LS}(N)$, and '+' the theoretical value σ_R.

- Although $\hat{R}_{LS}(N)$ has a smaller variance than $\hat{R}_{EV}(N)$ for low SNR, its total root mean square (rms) error (difference with respect to the true value) is significantly larger because of its systematic error. The following is quite a typical observation: many estimators reduce the stochastic error at the cost of systematic errors. For the $\hat{R}_{EV}$ the rms error is completely due to the variability of the estimator because 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 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:

$\hat{R}_{SA}(N)$: 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:

$$\hat{R}_{SA}(N) = \arg\min_R V_{SA}(R, N) \text{ with } V_{SA}(R, N) = \frac{1}{2}\sum_{k=1}^{N}(R(k) - R)^2 \tag{1-19}$$

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

$\hat{R}_{LS}(N)$: 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,

$$\hat{R}_{LS}(N) = \arg\min_R V_{LS}(R, N) \text{ with } V_{LS}(R, N) = \frac{1}{2}\sum_{k=1}^{N} e^2(k, R) \tag{1-20}$$

$\hat{R}_{EV}(N)$: 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), parameterized as (i_p, u_p), keeping in mind the model equation $u_0 = Ri_0$.

$$\begin{aligned} \hat{R}_{EV}(N) &= \arg\min_{R, i_p, u_p} V_{EV}(R, i_p, u_p, N) \text{ subject to } u_p = Ri_p \\ V_{EV}(R, i_p, u_p, N) &= \frac{1}{2}\sum_{k=1}^{N}(u(k) - u_p)^2 + \frac{1}{2}\sum_{k=1}^{N}(i(k) - i_p)^2 \end{aligned} \tag{1-21}$$

This wide variety of possible solutions and motivations illustrates 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 presented here 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 a quadratic cost: first, it is easier to minimize than other functions, and second, 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, a transmission error, etc.), it may 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; for example, 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 BEHAVIOR OF ESTIMATORS

Because 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 in the following. A detailed discussion is given in Chapter 16.

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 such as the median or the mode (Stuart and Ord, 1987) could also be used, but the latter are much more difficult to analyze in most cases. As 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 speak about true parameters only if an exact model exists. It is clear that this is a purely imaginary situation because 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 a 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 1.2 (Unbiasedness): An estimator $\hat{\theta}$ of the parameters θ_0 is unbiased if $\mathbb{E}\{\hat{\theta}\} = \theta_0$ for all true parameters θ_0. Otherwise, it is a biased estimator.

If the expected value equals the true value only 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 (Unbiased and Biased Estimators): At the end of their experiments, the students want 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:

$$\hat{u}(N) = \frac{1}{N}\sum_{k=1}^{N} u(k) \quad \text{and} \quad \hat{\sigma}_u^2(N) = \frac{1}{N}\sum_{k=1}^{N} (u(k) - \hat{u}(N))^2 \tag{1-22}$$

Applying the previous definition, it is readily seen that

$$\mathbb{E}\{\hat{u}(N)\} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\{u(k)\} = \frac{1}{N}\sum_{k=1}^{N} u_0 = u_0 \tag{1-23}$$

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

$$\mathbb{E}\{\hat{\sigma}_u^2(N)\} = \frac{N-1}{N}\sigma_u^2 \tag{1-24}$$

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 $\sum_{k=1}^{N}(u(k)-\hat{u}(N))^2/(N-1)$, 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 is the case for $\hat{R}_{SA}(N)$ of group A. Moreover, unbiased estimators can still have a bad distribution; for example, 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 1.4 (Consistency): An estimator $\hat{\theta}(N)$ of the parameters θ_0 is weakly consistent if it converges in probability to θ_0: $\operatorname{plim}_{N\to\infty}\hat{\theta}(N) = \theta_0$ and strongly consistent if it converges with probability one (almost surely) to θ_0: $\operatorname{a.s.lim}_{N\to\infty}\hat{\theta}(N) = \theta_0$.

The precise explanation of these probability limits is given in Section 16.6. Loosely explained, it means that the pdf of $\hat{\theta}(N)$ contracts around the true value θ_0, or $\lim_{N\to\infty}\operatorname{Prob}(|\hat{\theta}(N)-\theta_0| > \delta > 0) = 0$. 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. A general outline of how to prove consistency is given in Section 17.3. Another nice property of the plim is that it can be interchanged with a continuous function: $\operatorname{plim} f(a) = f(\operatorname{plim}(a))$ if both limits exist (see Section 16.8). In fact, it was this property that we applied during the calculations of the limit values of $\hat{R}_{LS}$ and $\hat{R}_{EV}$, for example,

$$\operatorname*{plim}_{N\to\infty}\hat{R}_{EV}(N) = \operatorname*{plim}_{N\to\infty}\frac{\frac{1}{N}\sum_{k=1}^{N}u(k)}{\frac{1}{N}\sum_{k=1}^{N}i(k)} = \frac{\operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}u(k)}{\operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}i(k)} = \frac{u_0}{i_0} = R_0 \tag{1-25}$$

Consequently, $\hat{R}_{EV}(N)$ is a weakly consistent estimator. Calculating the expected value is much more involved in this case due to the division. Therefore, consistency is better suited 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, that is, to ascertain how much the actual estimator is scattered around its limiting value. Again, this choice, among other possibilities (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, whereas 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, because 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, then 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 parameterized 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 $\int_{z \in \mathbb{R}^N} f(z|\theta_0)dz = 1$. Next we can interpret this relation conversely, namely, 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 now consider a given set of measurements and view the model parameters as the free variables:

$$L(z_m|\theta) = f(z = z_m|\theta) \tag{1-26}$$

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 (1-26) 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 because $\int_\theta L(z|\theta)d\theta \neq 1$. Notice the subtle difference in terminology; that is, probability is replaced by likeliness.

The Cramér-Rao lower bound gives a lower limit on the covariance matrix of parameters. Under quite general conditions, this limit is universal and independent of the selected estimator: no estimator that violates this bound can be found. It is given by (see Section 16.12)

$$\begin{aligned} CR(\theta_0) &= \left(I_{n_\theta} + \frac{\partial b_\theta}{\partial \theta}\right)^T Fi^{-1}(\theta_0)\left(I_{n_\theta} + \frac{\partial b_\theta}{\partial \theta}\right) \\ Fi(\theta_0) &= \mathbb{E}\left\{\left(\frac{\partial l(z|\theta)}{\partial \theta}\right)^T\left(\frac{\partial l(z|\theta)}{\partial \theta}\right)\right\} = -\mathbb{E}\left\{\frac{\partial^2 l(z|\theta)}{\partial \theta^2}\right\} \end{aligned} \tag{1-27}$$

The derivatives are calculated in $\theta = \theta_0$, and $b_\theta = \mathbb{E}\{\hat{\theta}\} - \theta_0$ is the bias on the estimator. Note that for biased estimators ($\partial b_\theta / \partial \theta \neq 0$) the lower bound (1-27) can be zero: $CR(\theta_0) = 0$ (see Example 16.20 on page 590). For unbiased estimators (1-27) reduces to $CR(\theta_0) = Fi^{-1}(\theta_0)$.

$Fi(\theta)$ is called the Fisher information matrix; it is a measure of the information in an experiment: the larger the matrix, the more information there is. In (1-27) it is assumed that the first and second derivatives of the log likelihood function exist with respect to θ.

Example 1.5 (Influence of the Number of Parameters on the Cramér-Rao Lower Bound): A group of students want 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 is not precise and in the end they are not sure about the exact starting time of their experiment. They include it in the model as an additional parameter: $h_0(t) = a(t - t_{\text{start}}) = at + b$, and $\theta = [a, b]^T$. 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:

$$\text{Prob}(h(k), t_k) = \text{Prob}(h(k) - (at_k + b)) = \text{Prob}(n_h(k))$$

where $\text{Prob}(h(k), t_k)$ is the probability of making 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

$$L(h|a,b) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(h(k)-at_k-b)^2} \tag{1-28}$$

and the log likelihood function becomes

$$l(h|a,b) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(h(k)-at_k-b)^2 \tag{1-29}$$

The Fisher information matrix and the Cramér-Rao lower bound are found using (1-27):

$$Fi(a,b) = \frac{N}{\sigma^2}\begin{bmatrix} s^2 & \mu \\ \mu & 1 \end{bmatrix} \rightarrow CR(a,b) = Fi^{-1}(a,b) = \frac{\sigma^2}{N(s^2-\mu^2)}\begin{bmatrix} 1 & -\mu \\ -\mu & s^2 \end{bmatrix} \tag{1-30}$$

with $\mu = \sum_{k=1}^{N} t_k/N$ and $s^2 = \sum_{k=1}^{N} t_k^2/N$. These expressions are very informative. First, 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. This idea will be exploited fully in Chapter 5. Another question we can now answer is what price is paid to include the additional model parameter b to account for the unknown starting time. By comparing $Fi^{-1}(a,b)$ with $Fi^{-1}(a)$ (assuming that b is known), it is found that

$$\sigma_a^2(a,b) = \frac{\sigma^2}{N(s^2-\mu^2)} \geq \frac{\sigma^2}{Ns^2} = \sigma_a^2(a) \tag{1-31}$$

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 Chapter 11. □

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 criterion most used to verify the efficiency of an estimator.

Definition 1.6 (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. For biased estimators, a generalized expression should be used (see Section 16.12).

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/her choice. Clearly, this can result in poor or suboptimal 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 as well as possible the model 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/her 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 depend heavily on the choices that are made. Later, we will thoroughly discuss the selection of the excitation signals.

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, whereas 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 and the motion equations of a piston. 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 nonparametric 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 are more substantial for parametric models than for nonparametric ones. We will concentrate on transfer function models (parametric models), but the problem of frequency response function measurements (nonparametric model) will also 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 related to different scientific fields may be brought into this phase of the identification process. The modeling 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 that allows sufficient description of any observed input and output measurements. This reduces the modeling effort significantly. For example, instead of modeling the loudspeaker using physical laws, an input-output relation, taking the form of a high-order transfer function, can 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 be used only for prediction of the output.

Although, as a rule of thumb, it is advisable to include as much prior knowledge as possible during the modeling process, it is not always easy to do. 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 linearized. This kind of approximation makes it possible to use simple models without jeopardizing properties that 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 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 nonlinear 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 nonlinear in the b_1 and b_2 parameters. Linearity in the parameters is a very important aspect of models because 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. This is illustrated in Section 1.5.1.

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 because 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 also exists a more systematic approach based on stochastic arguments as explained in Section 1.5. Simple tests on the cost function exist (necessary conditions) to check even before deriving the estimator whether it can be consistent (see Chapter 9, Section 9.5).

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 modeled, indicating remaining model errors? In practice, the best model (meaning the one with 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, unmodeled linear dynamics and nonlinear 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 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 to 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 set up 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 users to be aware of the sensitive aspects of their techniques. This will enable them to put all their effort on the most critical decisions. Moreover, they will become aware of the weak points of the final model.

1.5 A STATISTICAL APPROACH TO THE ESTIMATION PROBLEM

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 the-

ory 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, such as 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 are 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 requires only 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.

1.5.1 Least Squares Estimation

One of the simplest estimation techniques is the least squares estimator. In this case, the match between the model and the measurements is quantified by a least squares cost function. As this is an arbitrary choice, initially, it is clear that the result is not necessarily optimal. By choosing other cost functions such as the sum of the least absolute values, it is possible to find other estimators, with different properties, that perform better in specific situations. Some of these 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 the stochastic point of view. Normally distributed noise leads, naturally, to least squares estimation. As seen in the resistance example, even within the class of least squares estimators, there are different possibilities resulting in completely different estimators. A full treatment of the problem is beyond the scope of this book, hence, we focus only on the aspects that are of direct importance to our major goal.

Consider a multiple-input, single-output system modeled by $y_0(k) = g(u_0(k), \theta_0)$ with k the measurement index, $y(k) \in \mathbb{R}$, $u_0(k) \in \mathbb{R}^{1 \times n_u}$, and $\theta_0 \in \mathbb{R}^{n_\theta}$ the true parameter vector. The aim is to estimate the parameters from noisy observations at the output of the system: $y(k) = y_0(k) + n_y(k)$. This is done by minimizing the sum of the squared errors $e(k, \theta) = y(k) - y(k, \theta)$, with $y(k, \theta)$ the modeled output:

$$\hat{\theta}_{\mathrm{NLS}}(N) = \arg\min_{\theta} V_{\mathrm{NLS}}(\theta, N), \text{ with } V_{\mathrm{NLS}}(\theta, N) = \frac{1}{2}\sum_{k=1}^{N} e^2(k, \theta) \tag{1-32}$$

In general, the analytical solution of the nonlinear least squares problem (1-32) is not known, so numerical methods must be used. A number of techniques are described in the literature (Fletcher, 1991), and many are found in commercially available mathematical packages. They vary from very simple techniques such as 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. However, the Gauss-Newton method is very well suited to deal with the least squares minimization problem because it makes explicit use of the structure of the cost function. The second derivatives of the cost function (the Hessian matrix) are approximated in this method by the first-order derivatives of $e(\theta)$. Define the Jacobian matrix $J(\theta) \in \mathbb{R}^{N \times n_\theta}$: $J(\theta) = \partial e(\theta)/\partial \theta$ and consider the Hessian matrix:

$$\frac{\partial^2 V_{\mathrm{NLS}}(\theta, N)}{\partial \theta^2} = J^T(\theta)J(\theta) - \sum_{k=1}^{N} e(k, \theta)\frac{\partial^2 g(u_0(k), \theta)}{\partial \theta^2} \tag{1-33}$$

If the second term in (1-33) is small (for example, $\|e(\theta)\|_2$ is "small") with respect to the first one, then $J^T(\theta)J(\theta)$ will be a good approximation for the second-order derivatives of the cost function. The numerical solution is then found by applying the following iterative process:

$$\theta^{(i+1)} = \theta^{(i)} + \Delta\theta^{(i+1)} \text{ with } J^T(\theta^{(i)})J(\theta^{(i)})\Delta\theta^{(i+1)} = -J^T(\theta^{(i)})e(\theta^{(i)}) \tag{1-34}$$

Equation (1-34) reveals two important advantages. First, only the gradient needs to be calculated, and not the Hessian, thus reducing the calculation time. Moreover, very often, the condition number of the Hessian matrix is the square of that of the Jacobian. This leads us to the second advantage: using, for example, singular value decomposition (SVD) or QR decomposition techniques, (1-34) can be solved without forming the product $J^T(\theta^{(i)})J(\theta^{(i)})$ so that more complex problems can be solved, because the numerical errors are significantly reduced (see Exercise 1.12). If (1-34) converges to the global minimum of (1-32), then $\hat{\theta}_{\mathrm{NLS}}(N) = \theta^{(\infty)}$.

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 Section 19.4. 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 proved. Also, an approximate expression for the covariance matrix $\mathrm{Cov}(\hat{\theta}_{\mathrm{NLS}}(N))$ is available:

$$\mathrm{Cov}(\hat{\theta}_{\mathrm{NLS}}(N)) \approx (J^T(\theta)J(\theta))^{-1}J^T(\theta)\mathrm{Cov}(n_y)J(\theta)(J^T(\theta)J(\theta))^{-1}\Big|_{\theta=\hat{\theta}_{\mathrm{LS}}(N)} \tag{1-35}$$

with $\mathrm{Cov}(n_y) = \mathbb{E}\{n_y n_y^T\}$. Note that this approximation is still a stochastic variable because it depends on $\hat{\theta}_{\mathrm{NLS}}(N)$, while the exact expression should be in θ_0. If the model is linear-in-the-parameters, $y_0 = K(u_0)\theta_0$, and $e(\theta) = y - K(u_0)\theta$, then (1-32) reduces to a linear least squares cost function, and explicit expressions are available for the estimator (note that $K = -\partial e(\theta)/\partial\theta = -J(\theta)$ is parameter independent in this case). In order to keep the expressions compact, we do not include the arguments of K in the following:

$$\hat{\theta}_{\mathrm{LS}}(N) = (K^TK)^{-1}K^Ty \tag{1-36}$$

The covariance matrix still equals (1-35) with $J(\hat{\theta}_{\mathrm{LS}}(N))$ replaced by $-K$, but now it is an exact expression and no longer an approximation. Moreover, it is possible to prove that the estimator is unbiased for zero mean noise:

$$\mathbb{E}\{\hat{\theta}_{\mathrm{LS}}(N)\} = (K^TK)^{-1}K^T\,\mathbb{E}\{y\} = (K^TK)^{-1}K^Ty_0 = (K^TK)^{-1}K^TK\theta_0 = \theta_0 \tag{1-37}$$

This result is valid only if K is not disturbed by noise. If the inputs u are also disturbed by noise, it is no longer possible to bring $(K^TK)^{-1}K^T$ outside the expectation. In this case, additional quadratic noise contributions appear in K^TK so that $\hat{\theta}_{LS}(N)$ underestimates the true values. This was visible in the estimation of the resistance ($K_{[k]} = i(k)$, $y(k) = u(k)$, $\theta = R$) where (1-8) shows the impact of the quadratic contributions of the input noise.

Example 1.7 (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. From a prior analysis, making repeated measurements, it turns out that $n_y(k)$ is zero mean iid with variance σ_y^2. The model becomes $y = K\theta + n_y$ with $K = (1, 1, \ldots, 1)^T$. Using (1-36), the estimate is

$$\hat{\theta}_{LS}(N) = (K^TK)^{-1}K^Ty = \frac{1}{N}\sum_{k=1}^{N} y(k) \tag{1-38}$$

with variance

$$\text{var}(\hat{\theta}_{LS}(N)) = (K^TK)^{-1}K^T(\sigma_y^2 I_N)K(K^TK)^{-1} = \sigma_y^2/N \tag{1-39}$$

This example shows that it is much easier to get the solution when it is possible to formulate the problem under the standard conditions. □

This short analysis shows that the least squares estimator is applicable to a very wide range of problems. No prior information is required to use it, which explains its success. However, its specific properties depend on the actual situation. General statements can be made only if some noise characteristics are known. In that case it is also possible to improve the quality of the estimates by using this knowledge in the estimator. If, for example, the covariance matrix of the noise is known, a weighted least squares can be used.

1.5.2 Weighted Least Squares Estimation

In (1-32) 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.

In general, the weighted nonlinear least squares estimate $\hat{\theta}_{WNLS}(N)$ is

$$\hat{\theta}_{WNLS}(N) = \arg\min_{\theta} V_{WNLS}(\theta, N) \text{ with } V_{WNLS}(\theta, N) = \frac{1}{2}e^T(\theta)We(\theta) \tag{1-40}$$

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 evaluation of this cost function requires $O(N^2)$ operations, which are very time consuming. Consequently, (block) diagonal weighting matri-

ces are preferred in many problems, reducing the number of operations to $O(N)$. All the remarks on the numerical aspects of the least squares estimator are also valid for the weighted least squares. This can be understood easily by applying the following transformation: $\varepsilon(\theta) = Se(\theta)$ with $S^TS = W$ so that $V_{\text{WNLS}}(\theta, N) = \varepsilon^T(\theta)\varepsilon(\theta)/2$, which is a least squares estimator in the transformed variables. This also leads to the following Gauss-Newton algorithm to minimize the cost function

$$\theta^{(i+1)} = \theta^{(i)} + \Delta\theta^{(i+1)} \text{ with } J^T(\theta^{(i)})WJ(\theta^{(i)})\Delta\theta^{(i+1)} = -J^T(\theta^{(i)})We(\theta^{(i)}) \tag{1-41}$$

Equation (1-35) is generalized to (noticing that $W^T = W$)

$$\text{Cov}(\hat{\theta}_{\text{WNLS}}(N)) \approx (J^T(\theta)WJ(\theta))^{-1}J^T(\theta)WC_{n_y}WJ(\theta)(J^T(\theta)WJ(\theta))^{-1}\Big|_{\theta=\hat{\theta}_{\text{WNLS}}(N)} \tag{1-42}$$

with $C_{n_y} = \text{Cov}(n_y)$. By choosing $W = C_{n_y}^{-1}$, the expression simplifies to

$$\text{Cov}(\hat{\theta}_{\text{WNLS}}(N)) \approx [J^T(\hat{\theta}_{\text{WNLS}}(N))C_{n_y}^{-1}J(\hat{\theta}_{\text{WNLS}}(N))]^{-1} \tag{1-43}$$

In Exercise 1.16 it is shown that among all possible positive definite choices for W, the best one is $W = C_{n_y}^{-1}$ because this minimizes the covariance matrix. The results for models that are linear-in-the-parameters are immediately found, analogous to the least squares estimator. Also, in this case, the weighted least squares is unbiased under the same conditions as the least squares estimator.

1.5.3 The Maximum Likelihood Estimator

Using the covariance matrix of the noise as the weighting matrix allows 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, $y_0 = G(u_0, \theta_0)$, that describes the system and the inputs that excite the system. Assuming, again, an additive noise model $y = y_0 + n_y$, with $y, y_0, n_y \in \mathbb{R}^N$, the likelihood function becomes:

$$f(y|\theta_0, u_0) = f_{n_y}(y - G(u_0, \theta_0)) \tag{1-44}$$

The maximum likelihood procedure consists of two steps. First the numerical values y_m of the actual measurements are plugged into (1-44) for the variables y, and 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

$$\hat{\theta}_{\text{ML}}(N) = \arg\max_{\theta} f(y_m|\theta, u_0) \tag{1-45}$$

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 1.8 (Weighing a Loaf of Bread—Continued): Consider Example 1.7 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 an ML estimator:

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\theta)^2}{2\sigma_y^2}} \tag{1-46}$$

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

$$f(y|\theta) = \frac{1}{(2\pi\sigma_y^2)^{N/2}} e^{-\frac{1}{2\sigma_y^2}\sum_{k=1}^{N}(y(k)-\theta)^2} \tag{1-47}$$

Because $(2\pi\sigma_y^2)^{-N/2}$ is parameter independent, the ML estimate is given by the minimizer of $\sum_{k=1}^{N}(y(k)-\theta)^2/(2\sigma_y^2)$ and becomes

$$\hat{\theta}_{\mathrm{ML}}(N) = \frac{1}{N}\sum_{k=1}^{N} y(k) \tag{1-48}$$

This is nothing 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 the case only for normally distributed errors. □

The unbiased behavior may not be generalized because the MLE can also be biased. For example, the sample mean and sample variance are shown to be the ML estimates for the mean and the variance of measurements that are identically independent and normally distributed:

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} y(k), \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N}(y(k)-\hat{\mu}_{\mathrm{ML}})^2.$$

Although the first estimate is unbiased, the second one can be shown to be prone to a bias of σ^2/N that asymptotically disappears in N:

$$\mathbb{E}\{\hat{\sigma}_{\mathrm{ML}}\} = \frac{N-1}{N}\sigma^2.$$

This shows that there is a clear need to understand the properties of the 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 anymore and a dedicated analysis is, again, required. In this introductionary chapter we just make a loose statement of the properties; a very precise description can be found in the literature (Goodwin and Payne, 1977; Caines, 1988).

Properties of the ML Estimator

- *Principle of invariance:* if $\hat{\theta}_{\mathrm{ML}}$ is an ML estimator of $\theta \in \mathbb{R}^{n_\theta}$, then $\hat{\theta}_g = g(\hat{\theta}_{\mathrm{ML}})$ is an ML estimator of $g(\theta)$ where g is a function, $\hat{\theta}_g \in \mathbb{R}^{n_g}$, and $n_g \leq n_\theta$, with n_θ a finite number.
- *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: $\underset{N\to\infty}{\text{a.s.lim}}\, \hat{\theta}_{\mathrm{ML}}(N) = \theta_0$.
  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, and the voltage is calculated from the estimated current and resistance value.
- *Asymptotic normality:* 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 in law to a normal random variable.
  The importance of this property is that it not only allows one to calculate uncertainty bounds on the estimates but also guarantees that most of the probability mass gets more and more unimodally concentrated around its limiting value.
- *Asymptotic efficiency:* if $\hat{\theta}_{\mathrm{ML}}(N)$ is an ML estimator based on N iid random variables, with n_θ independent of N, then $\hat{\theta}_{\mathrm{ML}}(N)$ is asymptotically efficient ($\mathrm{Cov}(\hat{\theta}_{\mathrm{ML}}(N))$ reaches asymptotically the Cramér-Rao lower bound).

1.5.4 The Bayes Estimator

As described before, the Bayes estimator requires the most 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: $f(\theta|u, y)$. 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 $C(\theta|\theta_0)$ that 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 $f(\theta|u, y)$:

$$\hat{\theta}(N) = \arg\min_{\theta_0} \int_{\theta \in \mathbb{D}} C(\theta|\theta_0) f(\theta|u, y) d\theta \tag{1-49}$$

For some specific choices of $C(\theta|\theta_0)$, the solution of (1-49) is well known; for example, $C(\theta|\theta_0) = |\theta - \theta_0|^2$ leads to the mean value, and $C(\theta|\theta_0) = |\theta - \theta_0|$ results in the median, which is less sensitive to outliers because these contribute less to the second criterion than to the first (Eykhoff, 1974).

Another objective criterion is to choose the estimate as

$$\hat{\theta}_{\mathrm{Bayes}}(N) = \arg\max_{\theta} f(\theta|u, y) \tag{1-50}$$

The first and second examples are "minimum risk" estimators, and 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 (1-50) the Bayes rule is applied:

$$f(\theta|u,y) = \frac{f(y|\theta,u)f(\theta)}{f(y)} \tag{1-51}$$

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 $f(y|\theta,u)f(\theta)$. 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 1.9 (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 Prob(observation|elephant) = Prob(observation|horse). However, the probability of seeing an elephant in Belgium is much lower than that of seeing a horse: Prob(elephant in Belgium) « Prob(horse in Belgium) so that from the Bayes principle Joan concludes she was seeing a horse. If she was on safari in Kenya instead of Belgium, the conclusion would be opposite, because Prob(elephant in Kenya) » Prob(horse in Kenya).

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 1.10 (Weighing a Loaf of Bread—Continued): Consider again Example 1.8 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 as a result of 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 $f(y|\theta) = f_v(n_v) = f_v(y-\theta)$, the Bayes estimator is found by maximizing

$$f(y|\theta)f(\theta) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\theta)^2}{2\sigma_y^2}} \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{(\theta-w)^2}{2\sigma_w^2}} \tag{1-52}$$

and the estimated weight becomes

$$\hat{\theta}_{\text{Bayes}} = \frac{y/\sigma_y^2 + w/\sigma_w^2}{1/\sigma_y^2 + 1/\sigma_w^2} \tag{1-53}$$

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

$$\hat{\theta}_{\text{Bayes}}(N) = \frac{\sum_{k=1}^{N} y(k)/\sigma_y^2 + w/\sigma_w^2}{N/\sigma_y^2 + 1/\sigma_w^2} \tag{1-54}$$

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.

1.5.5 Instrumental Variables

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

$$\hat{R}_{\text{LS}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k)}{\frac{1}{N}\sum_{k=1}^{N} i^2(k)}, \text{ with } \lim_{N\to\infty} \hat{R}_{\text{LS}}(N) = R_0 \frac{1}{1 + \sigma_i^2/i_0^2} \tag{1-55}$$

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:

$$\hat{R}_{IV}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k-1)}{\frac{1}{N}\sum_{k=1}^{N} i(k)i(k-1)} \tag{1-56}$$

Making the same analysis as in Section 1.2.2.1, it is seen that all quadratic noise contributions are eliminated by this choice, so that

$$\lim_{N\to\infty} \hat{R}_{IV}(N) = R_0 \tag{1-57}$$

The idea used to generate (1-56) can be generalized as follows. Consider the linear-in-the-parameters model structure $y_0 = K(u_0)\theta_0$ in Section 1.5.1, and replace K^T in (1-36) by G^T, to get

$$\hat{\theta}_{IV}(N) = (G^T K(u))^{-1} G^T y \tag{1-58}$$

The choice of G, a matrix of the same size as $K(u)$, will be defined later. $\hat{\theta}_{IV}(N)$ is the instrumental variables estimate. Consistency is proved by considering the plim for $N \to \infty$ (Norton, 1986). For simplicity, we assume all the plim exists, namely

$$\begin{aligned}\text{plim}\,\hat{\theta}_{IV} &= \text{plim}\,\{(G^T K(u))^{-1} G^T y\} \\ &= (\text{plim}\,\{G^T K(u_0 + n_u)\})^{-1}(\ \text{plim}\{G^T y_0 + G^T n_y\}) \\ &= (\text{plim}\,\{G^T K(u_0 + n_u)/N\})^{-1}(\text{plim}\,\{G^T K(u_0)/N\}\,\theta_0 + \text{plim}\,\{G^T n_y/N\})\end{aligned}$$

If

$$\text{plim}\,\{G^T K(u_0 + n_u)/N\} = \text{plim}\,\{G^T K(u_0)/N\} \quad \text{and} \quad \text{plim}\,\{G^T n_y/N\} = 0 \tag{1-59}$$

then

$$\lim_{N\to\infty} \hat{\theta}_{IV}(N) = \theta_0 \tag{1-60}$$

Equation (1-59) defines 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.

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):

$$\text{Cov}(\hat{\theta}_{IV}(N)) \approx \sigma^2 R_{GK}^{-1} R_{GG} R_{GK}^{-T} \text{ with } R_{GK} = G^T K(u)/N \text{ and } R_{GG} = G^T G/N \tag{1-61}$$

This reveals another condition on the choice of the instrumental variables G: although 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 $\text{Cov}(\hat{\theta}_{IV}(N))$ would become very large. In the case of the resistance estimate, the instrumental variables are the shifted input. Because 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 following exercises the instrumental variables method is applied to the resistance example.

1.6 EXERCISES

1.1. Set up a simulation to measure the value of the resistance using

$$i(k) = i_0 + n_i(k) \qquad u(k) = u_0 + n_u(k) \tag{1-62}$$

Use for n_i and n_u zero mean iid noise with standard deviation σ_i and σ_u. Consider uniformly and normally distributed noise and use $i_0 = 1$ A, $u_0 = 1$ V, $\sigma_i = 0.5(1)$ A, and $\sigma_u = 0.5(1)$ V. Plot $R(k) = u(k)/i(k)$ for $k = 1, \ldots, 100$.

1.2. Apply the estimators $\hat{R}_{\mathrm{LS}}$, $\hat{R}_{\mathrm{EV}}$, $\hat{R}_{\mathrm{SA}}$ from (1-1) to (1-3) to the results of the simulator in Exercise 1.1 and plot the results as a number of the processed measurements N.

1.3. Measure the histogram for the three estimators of Exercise 1.2 for $N = 10, 100, 1000$ and plot the approximated pdf.

1.4. Use the simulator of Exercise 1.1 to estimate the variance of the three estimators of Exercise 1.2 as a function of N and plot the results on a log-log scale. Check the $1/\sqrt{N}$ rule of thumb. Vary N between 1 and 10^4.

1.5. Derive the variance expressions $\mathrm{var}(\hat{R}_{\mathrm{LS}}(N))$, $\mathrm{var}(\hat{R}_{\mathrm{EV}}(N))$, and $\mathrm{var}(\hat{R}_{\mathrm{SA}}(N))$ under Assumption 1.1 using linear approximations as illustrated in (1-16) and (1-17).

1.6. Use the simulator of Exercise 1.1 to estimate the variance of the three estimators of Exercise 1.2 for $N = 100$ as a function of the SNR of the current and the voltage measurements. Compare the results with the theoretical level (see (1-17) and (1-18)) and discuss the results.

1.7. The bias compensated least squares solution of the resistor problem is given by

$$\hat{R}_{\mathrm{BC}}(N) = \frac{\frac{1}{N}\sum_{k=1}^{N} u(k)i(k)}{\frac{1}{N}\sum_{k=1}^{N} i^2(k) - \sigma_i^2} \tag{1-63}$$

Use the simulator of Exercise 1.1 to estimate the variance and the mean square error of $\hat{R}_{\mathrm{LS}}$, $\hat{R}_{\mathrm{EV}}$, and $\hat{R}_{\mathrm{BC}}$ from (1-2), (1-3), and (1-63) with $N = 100$. Vary the noise-to-signal ratio between 0.01 and 1, and plot the results on a log-log scale. Compare the mean square error with the variance. What do you conclude?

1.8. Derive the estimators $\hat{R}_{\mathrm{LS}}(N)$, $\hat{R}_{\mathrm{EV}}(N)$, and $\hat{R}_{\mathrm{SA}}(N)$ by minimizing the cost functions (1-19), (1-20), and (1-21).

1.9. Reformulate the cost functions (1-19), (1-20), and (1-21) for the case that the current is varying from measurement to measurement (the current is no longer a DC source), and derive the new expressions of the estimators. Show that the errors-in-variables estimator $\hat{R}_{\mathrm{EV}}(N)$ minimizes the following cost function w.r.t. R

$$\sum_{k=1}^{N} \frac{(u(k) - Ri(k))^2}{\sigma_u^2(k) + R^2\sigma_i^2(k)} \tag{1-64}$$

with $\sigma_i^2(k)$ and $\sigma_u^2(k)$ the variances of, respectively, the current and voltage measurements (hint: eliminate $i_p(k)$ and $u_p(k) = Ri_p(k)$ in (1-21) via $\partial V_{\mathrm{EV}}/\partial i_p(k) = 0$, $k = 1, 2, \ldots, N$).

1.10. Consider a signal

$$y_0(k) = \sin(2\pi f k T_s + \varphi) \tag{1-65}$$

and its measurement

$$y(k) = y_0(k) + n_y(k), \text{ for } k = 1, \ldots, 1024 \tag{1-66}$$

where $n_y(k)$ is iid normally distributed noise with zero mean and variance σ_y^2. Calculate the Cramér-Rao for the estimates (f, φ). What is the best choice for T_s if we want to estimate the frequency with minimum variance?

1.11. Consider a polynomial model:

$$y_0(k) = \sum_{p=1}^{P} a_p u^p(k) \tag{1-67}$$

that is identified from a set of measurements $y(k) = y_0(k) + n_y(k)$, with $u(k) = [-N{:}N]/N$ and $n_y(k)$ zero mean iid distributed noise with variance σ_y^2. Set up the least squares estimator for this problem, and observe the condition number for growing values of P (put $N = 1000$). What is the maximum order that can be reliably identified?

1.12. Consider the least squares solution $\hat{\theta}_{LS}(N) = (J^T J)^{-1} J^T y$ of the overdetermined set $J\theta = y$ (as they appear in (1-36)). Show that this solution can be calculated using the SVD method of Section 15.5 on matrix algebra without forming the product $J^T J$ as $\hat{\theta}_{LS}(N) = J^+ y$, with $J^+ = V\Sigma^+ U^T$.

1.13. Apply the method of Exercise 1.12. to the polynomial problem of Exercise 1.11, and find the maximum order that can be identified reliably.

1.14. The polynomial identification problem is an ill-posed problem because of the poor numerical conditioning of the normal equations. Using the SVD method, it is already possible to solve higher order problems, but even then the numerical conditioning decreases quickly. A much better solution is to change the model representation and to use orthogonal polynomials $T_p(u)$ such that

$$y_0(k) = \sum_{p=1}^{P} a_p u^p(k) = \sum_{p=1}^{P} t_p T_p(u(k)) \tag{1-68}$$

where $T_p(u) = \sum_{k=1}^{p} a_{pk} u^k$ is a polynomial of degree p. The coefficients a_{pk} are set s.t.

$$\sum_{k=1}^{N} T_r(u(k)) T_s(u(k)) = \delta(r - s)$$

Note that the actual form of $T_p(u)$ (the choice of a_{pk}) depends on the set of input values $u(k)$ that appears in the problem. Reformulate the polynomial identification problem using the orthogonal basis and discuss the condition number of the new estimator.

Remarks:

For the given set of input values, the orthogonal polynomials $T_p(u)$ are given by the following recurrence relation (Ralston and Rabinowitz, 1984):

$$\begin{aligned} &\frac{1}{\alpha_{j+1}} T_{j+1}(u) = \frac{u}{\alpha_j} T_j(u) - \frac{\beta_j}{\alpha_{j-1}} T_{j-1}(u) \\ &\beta_j = \frac{j^2[(2N+1)^2 - j^2]}{4(4j^2 - 1)}, \qquad \alpha_j = \frac{(2j)!}{(j!)^2 (2N)^j} \end{aligned} \tag{1-69}$$

with $T_0(u) = 1$ and $T_{-1}(u) = 0$ for $j = 0, 1, \ldots$. When using orthogonal polynomials the reader should take care not to use the explicit polynomial expressions, but only the values of the orthogonal polynomials. Otherwise the numerical stability is not

guaranteed. As a result, it is also not possible to calculate the coefficients a_p of the original solution; only the value of the solution can be calculated (see Ralston and Rabinowitz, 1984).

1.15. Prove expression (1-42) for the covariance matrix of a weighted least squares for models that are linear-in-the-parameters.

1.16. Show that the covariance matrix of the weighted least squares estimator becomes minimal for $W = C_{n_y}^{-1}$ (hint: use the Schwarz inequality $B^TB \geq (B^TA)(A^TA)^{-1}(A^TB)$, see Eykhoff, 1974, p. 525, and put $C_{n_y}^{-1} = C^TC$, $B = CJ$, and $A = C^{-T}WJ$).

1.17. Consider the linear-in-the-parameters model $y_0 = K(u_0)\theta_0$ and calculate the variance of the modeled output $\hat{y} = K(u_0)\hat{\theta}$ starting from the covariance matrix C_θ given in (1-43).

1.18. Show that the variance on the output of the polynomial model in Exercise 1.11. is independent of the model representation $y_0(k) = \sum_{p=1}^{P} a_p u^p(k)$ or $y_0(k) = \sum_{p=1}^{P} t_p T_p(u(k))$. Check this by a simulation using the estimators of Exercises 1.11. and 1.14. for a polynomial of degree 5 (so that the numerical conditioning of the problem remains acceptable for the direct estimation).

1.19. Consider the system $y_0 = au$. Construct the least squares and the weighted least squares estimator for a starting from the measurements $y(k) = au(k) + n_y(k)$ with $\mathbb{E}\{n_y(k)\} = 0$ and $\sigma_{n_y}^2(k) = u(k)$. Compare the bias and the variance of both estimators for $u(k) = 1, 2, \ldots, 10$. Verify your results by means of a simulation.

1.20. Construct $\hat{R}_{IV}(N)$ for the resistance example of Section 1.2.1 using (1-58). Use the time-shifted current as an instrumental variable. Study the behavior of the estimator (mean value and variance) as a function of the shift by means of a simulation.

1.21. Study the behavior of $\hat{R}_{IV}(N)$ (mean value and variance) of the previous exercise for the situation where $i_0(k)$ is generated as low-pass filtered noise (bandwidth of the filter at $f_s/50$) as a function of the applied delay by means of a simulation.