# 11

# Bayes and Maximum Likelihood Methods

While the parameter estimation methods presented so far assumed that the parameters $\boldsymbol{\theta}$ and the observations of the output $\boldsymbol{y}$ are deterministic values, the parameters themselves and/or the output will now be seen in a stochastic view as a series of random variables. In Bayes estimation, the parameter vector has the probability density function $p(\boldsymbol{\theta})$ and the output can be described by the conditional probability density function $p(\boldsymbol{y}|\boldsymbol{\theta})$. One can then derive a solution to the parameter estimation problem based on this statistical information. As especially information about the probability density function of the parameters, $p(\boldsymbol{\theta})$, is seldom available in practical applications, the maximum likelihood estimator will be derived subsequently. It is based on the probability density function of the observed output $p(\boldsymbol{y}|\boldsymbol{\theta})$.

## 11.1 Bayes Method

For a given set of measurements $\boldsymbol{y}$, one can infer the parameters from the conditional probability density function $p(\boldsymbol{\theta}|\boldsymbol{y})$. This conditional probability density function can only be determined once the experiment has been conducted since it obviously depends on the measurements. Hence, it is an a posteriori probability density function. Based on this a posteriori probability density function, one is now interested in finding "best" parameter estimates $\hat{\boldsymbol{\theta}}$. To judge the optimality, once again an optimality criterion has to be introduced as $W(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$, for which a cost function must then be minimized,

$$\min_{\hat{\boldsymbol{\theta}}} \int_m W(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}^m \boldsymbol{\theta} \ , \tag{11.1.1}$$

and to find the minimum

$$\frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \int_m W(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}^m \boldsymbol{\theta} = \mathbf{0} \ , \tag{11.1.2}$$

where $\int_m$ is the $m$-dimensional integral over all components $\mathrm{d}\theta_1, \mathrm{d}\theta_2, \ldots, \mathrm{d}\theta_m$ of $\boldsymbol{\theta}$. The optimality criterion can for example be a quadratic function such as

R. Isermann, M. Münchhof, *Identification of Dynamic Systems*,
DOI 10.1007/978-3-540-78879-9_11, © Springer-Verlag Berlin Heidelberg 2011

$$W = \big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)^{\mathrm{T}}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \; . \tag{11.1.3}$$

In the one dimensional case, one can then write (11.1.1) as

$$\min_{\hat{\theta}} \int (\hat{\theta} - \theta)^2 p(\theta|\boldsymbol{y})\mathrm{d}\theta \; . \tag{11.1.4}$$

Taking the first derivative to determine the optimal value $\hat{\theta}$ leads to

$$\hat{\theta} = \int \theta p(\theta|\boldsymbol{y})\mathrm{d}\theta \; . \tag{11.1.5}$$

which is just the expected value of the parameter $\theta$ for the given probability density function $p(\theta|\boldsymbol{y})$.

A different approach is to choose the most likely value as indicated by the probability density function i.e. to choose the maximum of the probability density function as the estimate,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} p(\boldsymbol{\theta}|\boldsymbol{y}) \; . \tag{11.1.6}$$

In this setting, the PDF is termed *likelihood function*.

The key issue for both approaches is the determination of the conditional probability density function $p(\boldsymbol{\theta}|\boldsymbol{y})$, which can be determined by Bayes rule (Papoulis, 1962) as

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{y}) \; . \tag{11.1.7}$$

Here, $p(\boldsymbol{\theta}, \boldsymbol{y})$ is the joint PDF and $p(\boldsymbol{y})$ is the a posteriori PDF which follows from the measurements conducted during the experiment. Furthermore

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \; . \tag{11.1.8}$$

Hence, it follows from (11.1.7) that

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} \; , \tag{11.1.9}$$

and with (11.1.8) that

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \; , \tag{11.1.10}$$

where the PDF of $\boldsymbol{\theta}$ must be known a priori. If this is the case, then one could for example solve (11.1.2) directly.

Similarly, (11.1.6) can be written using the above results as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \; . \tag{11.1.11}$$

If no assumption can be made about $\boldsymbol{\theta}$ and hence it is assumed to be distributed uniformly over the parameter space, then

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} p(\boldsymbol{y}|\boldsymbol{\theta}) \tag{11.1.12}$$

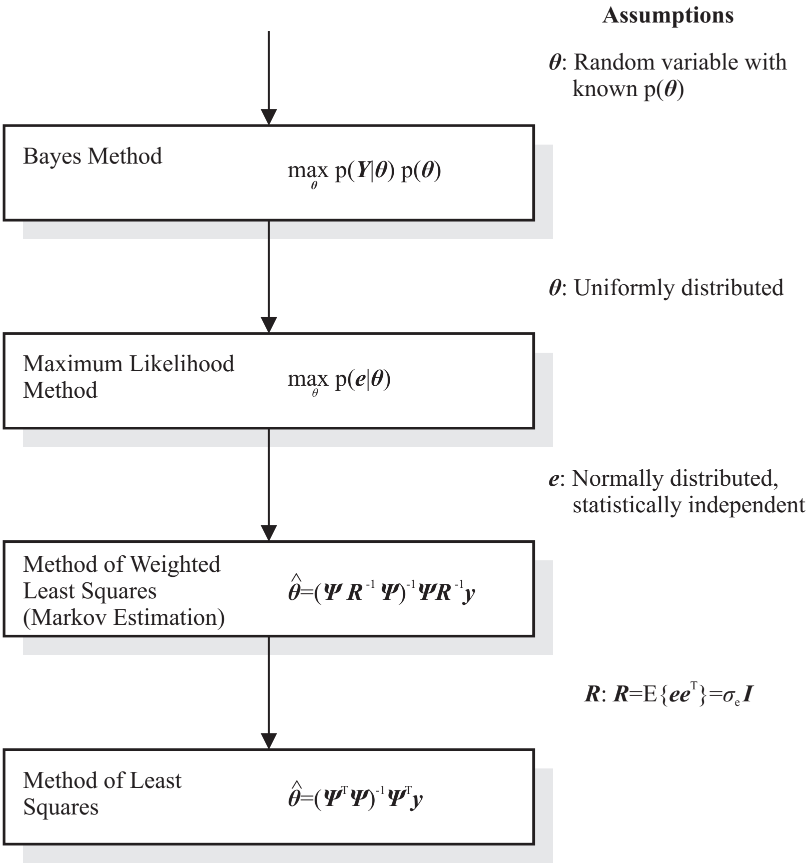
Assumptions
$\boldsymbol{\theta}$: Random variable with known p($\boldsymbol{\theta}$)
Bayes Method
$\max_{\theta} \mathrm{p}(\boldsymbol{Y}|\boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta})$
$\boldsymbol{\theta}$: Uniformly distributed
Maximum Likelihood Method
$\max_{\theta} \mathrm{p}(\boldsymbol{e}|\boldsymbol{\theta})$
$\boldsymbol{e}$: Normally distributed, statistically independent
Method of Weighted Least Squares (Markov Estimation)
$\hat{\boldsymbol{\theta}}=(\boldsymbol{\Psi}\,\boldsymbol{R}^{-1}\,\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}\boldsymbol{R}^{-1}\boldsymbol{y}$
$\boldsymbol{R}$: $\boldsymbol{R}=\mathrm{E}\{\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\}=\sigma_{\mathrm{e}}\boldsymbol{I}$
Method of Least Squares
$\hat{\boldsymbol{\theta}}=(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{y}$

**Fig. 11.1.** Derivation of different parameter estimation methods from the Bayes method through specializing assumptions

results, which is the *maximum likelihood* estimate, that has been introduced in Sect. 8.5. In those cases, where the prior PDF has negligible influence on the estimation results, the maximum a posteriori estimation is also close to the maximum likelihood estimation (Ljung, 1999).

The main drawback is the fact that the Bayes estimation necessitates knowledge of the probability density function of the parameters $\boldsymbol{\theta}$ and that the conditional probability density function can only be established under a high mathematical burden. Hence, the Bayes estimation has little relevance for practical applications in the area of system identification. It can however be seen as the most comprehensive parameter estimation technique and it serves as a starting point for the development of many other algorithms, such as e.g. the maximum likelihood estimate, which is covered in the following section and can be seen as a specialization of the Bayes estimation.

The maximum likelihood method in turn can be brought into liaison with the least squares parameter estimation under certain assumptions about the noise (Isermann, 1992), see Fig. 11.1. For further reading, one can consult e.g. (Lee, 1964; Nahi, 1969; Eykhoff, 1974; Peterka, 1981; Ljung, 1999). The Bayes rule is also often applied in classification problems (e.g. Isermann, 2006) .

There number of publications about the application of Bayes method for parameter estimation is relatively sparse. This can mainly be attributed to the problems in the computational problems in determining the conditional probability density functions and the fact that the probability density functions of the parameters are typically unknown. Therefore, the Bayes estimation is mainly of theoretical value. It can be regarded as the most general and most comprehensive estimation method. Other fundamental estimation methods can be derived from this starting point by making certain assumptions or specializations.

This is depicted in Fig. 11.1. Upon the assumption of uniformly distributed parameters, i.e. $p(\boldsymbol{\theta}_0) = \text{const}$, then the Bayes estimation (11.1.11) becomes a maximum likelihood estimation (11.1.12). As will be shown in the following derivation of the maximum likelihood estimation for dynamic systems, one uses the equation error $\boldsymbol{e}$ instead of the measured signals $\boldsymbol{y}$ due to the easier treatment, i.e.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} p(\boldsymbol{e}|\boldsymbol{\theta}) \ . \tag{11.1.13}$$

If one furthermore assumes that the error $\boldsymbol{e}$ is statistically independent, Gaussian distributed with $\mathrm{E}\{\boldsymbol{e}\} = 0$, and has the error covariance matrix $\boldsymbol{R} = \mathrm{E}\{\boldsymbol{e}\,\boldsymbol{e}^{\mathrm{T}}\}$, then

$$p(\boldsymbol{e}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}(\det \boldsymbol{R})^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{e}\right) \tag{11.1.14}$$

results, which will be derived in Sect. 11.2 in a more detailed way. From there follows

$$\ln p(\boldsymbol{e}|\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{e} + \text{const} \tag{11.1.15}$$

and

$$\frac{\partial}{\partial \boldsymbol{\theta}} = -\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{e} = \boldsymbol{0} \ . \tag{11.1.16}$$

Hence, one has to minimize the quadratic cost function (11.1.15), where the errors are weighted with the inverse of their covariance matrix. According to (9.5.6) and (9.5.4) this is the method of weighted least squares with minimal variance of the parameter estimates,

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y} \ , \tag{11.1.17}$$

and therefore is a Markov estimator. For uncorrelated errors, one obtains

$$\boldsymbol{R} = \sigma_{\boldsymbol{e}}^2 \boldsymbol{I} \ , \tag{11.1.18}$$

so that (11.1.17) results in the estimation equation for the method of least squares as

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{y} \ . \tag{11.1.19}$$

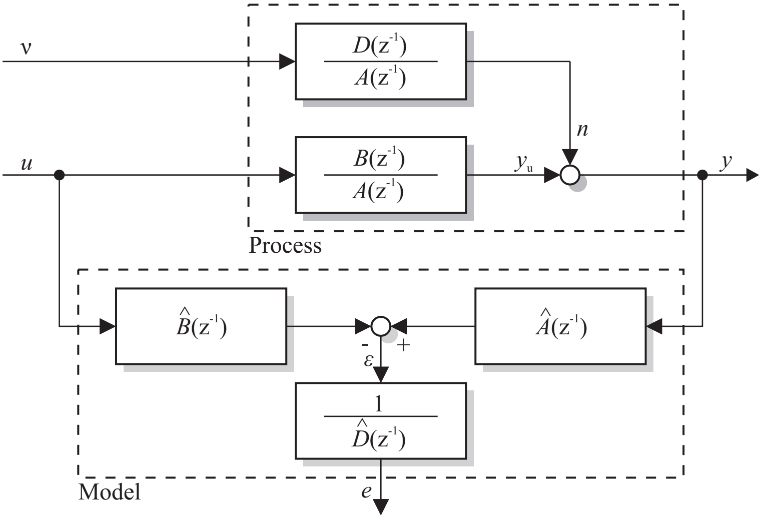

**Fig. 11.2.** Schematic diagram of the arrangement for the maximum likelihood method for dynamic systems

The maximum-likelihood method, which is described in Sect. 11.2, can be regarded as a method of least squares as it is assumed that $\boldsymbol{e}$ is uniformly distributed and statistically independent. The estimation equation must be solved iteratively due to the non-linear relation between $\boldsymbol{e}$ and the coefficients of the noise form filter polynomial $D(z^{-1})$.

## 11.2 Maximum Likelihood Method (ML)

In the following, the maximum likelihood estimator for linear dynamic discrete-time systems will first be formulated in a non-recursive formulation. It is then shown that under certain simplifying assumptions, it can also be formulated in a recursive fashion.

### 11.2.1 Non-Recursive Maximum Likelihood Method

While the maximum likelihood method has been introduced in Sect. 8.5 for static systems, it shall now be applied to linear dynamic systems in discrete-time. The process shall be governed by the model

$$A(z^{-1})y(z) - B(z^{-1})u(z) = D(z^{-1})e(z) \tag{11.2.1}$$

with

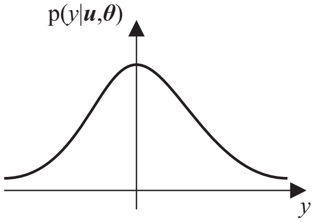

**Fig. 11.3.** Conditional probability density function of the observed signal $y(k)$

$$A(z^{-1}) = 1 + a_1 z^{-1} + \ldots + a_m z^{-m} \tag{11.2.2}$$

$$B(z^{-1}) = b_1 z^{-1} + \ldots + b_m z^{-m} \tag{11.2.3}$$

$$D(z^{-1}) = 1 + d_1 z^{-1} + \ldots + d_m z^{-m} \ , \tag{11.2.4}$$

where $e(k)$ shall be a Gaussian distributed, statistically independent signal with $(0, \sigma_{\mathrm{e}})$ and all roots of $D(z^{-1})$ shall lie within the unit circle. Compared to the method of least squares as introduced in Sect. 9.1, the model in (11.2.1) filters the equation error $\varepsilon(k)$ with the filter $1/\hat{D}(z^{-1})$, i.e.

$$\varepsilon(z) = \hat{D}(z^{-1}) e(z) \ \Leftrightarrow \ e(z) = \frac{1}{\hat{D}(z^{-1})} \varepsilon(z) \ . \tag{11.2.5}$$

The equation error $\varepsilon(k)$ is hence assumed to be a correlated signal, which by means of the filter is converted into an uncorrelated error $e(k)$, see Fig. 11.2.

In order to derive the maximum likelihood estimator (see also the development in Sect. 8.5), the probability density function of the measured, disturbed output has to be considered. In the following, it shall be assumed that the measured output $y(k)$ follows a Gaussian distribution, which allows to analytically treat the resulting equations.

The conditional probability density function of the observed signal samples $\{y(k)\}$ for a given input signal $\{u(k)\}$ and for given process parameters

$$\boldsymbol{\theta} = \big(a_1 \ldots a_m \big| b_1 \ldots b_m \big| d_1 \ldots d_m\big) \tag{11.2.6}$$

shall be denoted as

$$p(\{y(k)\}|\{u(k)\}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta}) \ . \tag{11.2.7}$$

and shall be known, see Fig. 11.3. One can now insert the measured values $y_p(k)$ and $u_p(k)$ into the above equation. Then, one obtains the *likelihood function*

$$p(\boldsymbol{y}_{\mathrm{P}}|\boldsymbol{u}_{\mathrm{P}}, \boldsymbol{\theta}) \ , \tag{11.2.8}$$

which is analyzed in dependence of the unknown parameters $\theta_i$, see Fig. 11.4.

As the parameters $\theta_i$ are constants and hence no stochastic variables, the likelihood function is not a probability density function of the parameters. The underlying principle of the maximum likelihood estimation is the idea that the best estimates for the unknown parameters $\theta_i$ are those values that attribute maximum possibility (or

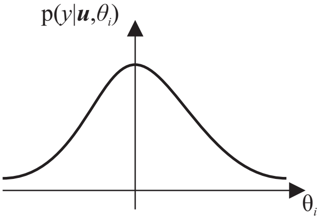

**Fig. 11.4.** Likelihood function for a single parameter $\theta_i$

*likelihood*) to the observed results. Mathematically speaking, one is looking for those values of $\theta_i$ that maximize the likelihood function. Hence, the parameters $\boldsymbol{\theta}$ can be determined by locating the maximum of the likelihood function or correspondingly by taking the first derivative and equating it to zero

$$\frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \boldsymbol{0} \,. \tag{11.2.9}$$

Since the individual measurements $y(k)$ are not statistically independent, the probability density function is difficult to calculate. Hence, the following derivation will be based on the error $e(k)$ which is assumed to be Gaussian distributed and statistically independent. In this case, one can consider the likelihood function

$$p(\boldsymbol{e}|\boldsymbol{u},\boldsymbol{\theta}) \tag{11.2.10}$$

and can determine the estimates by

$$\frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{e}|\boldsymbol{u},\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \boldsymbol{0} \,. \tag{11.2.11}$$

Because $e(k)$ is assumed to be statistically independent, one can now write the probability density function $p(\boldsymbol{e}|\boldsymbol{u},\boldsymbol{\theta})$ as

$$p(\boldsymbol{e}|\boldsymbol{u},\boldsymbol{\theta}) = \prod_{k=1}^{N} p(e(k)|\boldsymbol{u},\boldsymbol{\theta}) \,. \tag{11.2.12}$$

As the individual errors $e(k)$ are assumed to be Gaussian distributed, it is beneficial to take the logarithm of the likelihood function as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \ln\left(\left(\frac{1}{\sigma_{\mathrm{e}}\sqrt{2\pi}}\right)^{N} \prod_{k=1}^{N} \mathrm{e}^{-\frac{1}{2}\frac{e^2(k)}{\sigma_{\mathrm{e}}^2}}\right) \\ &= -\frac{1}{2\sigma_{\mathrm{e}}^2}\sum_{k=1}^{N} e^2(k) - N\ln\sigma_{\mathrm{e}} - \frac{N}{2}\ln 2\pi \,. \end{aligned} \tag{11.2.13}$$

One can see that maximizing the log-likelihood function (which should more precisely be called ln-likelihood function) with respect to the parameters $\boldsymbol{\theta}$ is the same as minimizing the sum of squared errors,

$$V=\sum_{k=1}^{N} e^2(k)\ . \tag{11.2.14}$$

Hence, for a Gaussian distributed error $e(k)$, the maximum likelihood and the least squares estimator yield identical results for the system structure as shown in Fig. 11.2.

The solution can only be determined iteratively, since the cost function is linear in the parameters of $A(z^{-1})$ and $B(z^{-1})$, but non-linear in the parameters $D(z^{-1})$. Åström and Bohlin (1965) employed a Newton-Raphson algorithm to solve this optimization problem. The first and second order derivatives will be denoted as

$$\boldsymbol{V}_{\boldsymbol{\theta}}^{\mathrm{T}}(\boldsymbol{\theta})=\left(\frac{\partial V}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}=\left(\frac{\partial V}{\partial\theta_1}\ \frac{\partial V}{\partial\theta_2}\ \cdots\ \frac{\partial V}{\partial\theta_p}\right) \tag{11.2.15}$$

with the Hesse matrix being

$$\boldsymbol{V}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta})=\frac{\partial^2 V}{\partial\boldsymbol{\theta}^{\mathrm{T}}\partial\boldsymbol{\theta}}=\begin{pmatrix}\dfrac{\partial^2 V}{\partial\theta_1\partial\theta_1} & \cdots & \dfrac{\partial^2 V}{\partial\theta_p\partial\theta_1}\\ \vdots & & \vdots\\ \dfrac{\partial^2 V}{\partial\theta_1\partial\theta_p} & \cdots & \dfrac{\partial^2 V}{\partial\theta_p\partial\theta_p}\end{pmatrix}. \tag{11.2.16}$$

The corresponding partial derivatives can be summarized as

$$\frac{\partial V}{\partial\theta_i}=\sum_{k=1}^{N} e(k)\frac{\partial e(k)}{\partial\theta_i} \tag{11.2.17}$$

$$\frac{\partial^2 V}{\partial\theta_i\partial\theta_j}=\sum_{k=1}^{N}\frac{\partial e(k)}{\partial\theta_i}\frac{\partial e(k)}{\partial\theta_j}+\sum_{k=1}^{N} e(k)\frac{\partial^2 e(k)}{\partial\theta_i\partial\theta_j}\ . \tag{11.2.18}$$

One therefore needs the partial derivatives of the error $e(k)$ with respect to the individual parameters, which can be provided as follows

$$D(q^{-1})\frac{\partial e(k)}{\partial a_i}=y(k)q^{-i} \tag{11.2.19}$$

$$D(q^{-1})\frac{\partial e(k)}{\partial b_i}=-u(k)q^{-i} \tag{11.2.20}$$

$$D(q^{-1})\frac{\partial e(k)}{\partial d_i}=-e(k)q^{-i} \tag{11.2.21}$$

$$D(q^{-1})\frac{\partial^2 e(k)}{\partial a_i\partial d_j}=-q^{-j}\frac{\partial e(k)}{\partial a_i}=-q^{-i-j+1}\frac{\partial e(k)}{\partial a_1} \tag{11.2.22}$$

$$D(q^{-1})\frac{\partial^2 e(k)}{\partial b_i\partial d_j}=-q^{-j}\frac{\partial e(k)}{\partial b_i}=-q^{-i-j+1}\frac{\partial e(k)}{\partial b_1} \tag{11.2.23}$$

$$D(q^{-1})\frac{\partial^2 e(k)}{\partial d_i\partial d_j}=-2q^{-j}\frac{\partial e(k)}{\partial d_i}=-2q^{-i-j+1}\frac{\partial e(k)}{\partial d_1}\ , \tag{11.2.24}$$

where the time shift operator $q$ has been introduced and defined as

$$y(k)q^{-l} = y(k-l) \tag{11.2.25}$$

Furthermore

$$D(q^{-1})\frac{\partial^2 e(k)}{\partial a_i \partial a_j} = 0 \tag{11.2.26}$$

$$D(q^{-1})\frac{\partial^2 e(k)}{\partial a_i \partial b_j} = 0 \tag{11.2.27}$$

$$D(q^{-1})\frac{\partial^2 e(k)}{\partial b_i \partial b_j} = 0\,. \tag{11.2.28}$$

Since the update equation for the optimization algorithm is given as

$$\begin{aligned}\boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) - \left(\frac{\partial^2 V}{\partial \boldsymbol{\theta}^{\mathrm{T}} \partial \boldsymbol{\theta}}\right)^{-1}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)} \left(\frac{\partial V}{\partial \boldsymbol{\theta}}\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}(k)} \\ &= \boldsymbol{\theta}(k) - \boldsymbol{V}_{\boldsymbol{\theta\theta}}\big(\boldsymbol{\theta}(k)\big)^{-1} \boldsymbol{V}_{\boldsymbol{\theta}}(\boldsymbol{\theta}(k))\,,\end{aligned} \tag{11.2.29}$$

the term $D(q^{-1})$ in this case cancels out.

A prerequisite for the convergence of the maximum likelihood estimate are appropriate initial values. It is suggested to set $D(z^{-1}) = 1$, i.e. $d_i = 0$, in the first iteration, which leads to the normal method of least squares and allows to obtain (biased) initial values by the direct solution of the least squares problem.

**Theorem 11.1 (Convergence of the Maximum Likelihood Estimate).**

*The Maximum Likelihood estimator delivers for an ARMAX process as depicted in Fig. 11.2 consistent asymptotically efficient parameter estimates, that fulfill the Cramér-Rao bound (Åström and Bohlin, 1965; van der Waerden, 1969; Deutsch, 1965), if the following conditions are met:*

- $u(k) = U(k) - U_{00}$ *is exactly known*
- $Y_{00}$ *is exactly known and belongs to* $U_{00}$
- *The elements of* $e(k)$ *are statistically independent and Gaussian distributed*
- *The roots of* $D(z) = 0$ *lie within the unit circle*
- *Appropriate initial values* $\hat{\boldsymbol{\theta}}(0)$ *are known*

□

The described method should also converge for many other noise distributions, but will in most cases not be asymptotically efficient any longer.

The maximum likelihood estimation for dynamic systems has also been outlined in (Raol et al, 2004). Here, the maximum likelihood estimation is applied to the output error model, partial derivatives of the cost function with respect to the parameters have been determined by finite differencing and a corresponding perturbation of the parameters. The monograph by van den Bos (2007) also discusses the maximum likelihood estimation in combination with non-linear optimization algorithms. The maximum likelihood estimation can according to Ljung (1999) also be interpreted as a maximum entropy or minimum information distance estimate.

### 11.2.2 Recursive Maximum Likelihood Method (RML)

The recursive Maximum Likelihood method can be derived by an approximation of the partial derivatives of the non-recursive method (Söderström, 1973; Fuhrt and Carapic, 1975). For the derivation, the process model in (11.2.1) is first expressed as

$$y(k) = \boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{\theta} + \nu(k) \tag{11.2.30}$$

with

$$\begin{aligned}\boldsymbol{\psi}^{\mathrm{T}}(k) = \big(&-y(k-1)\,\ldots\,-y(k-m)\big|u(k-d-1)\,\ldots\,u(k-d-m)\big|\\ &\nu(k-1)\,\ldots\,\nu(k-m)\big)\end{aligned} \tag{11.2.31}$$

and

$$\boldsymbol{\theta}^{\mathrm{T}} = \big(a_1 \,\ldots\, a_m\big|b_1 \,\ldots\, b_m\big|d_1 \,\ldots\, d_m\big)\,. \tag{11.2.32}$$

The cost function is given as

$$V(k+1,\hat{\boldsymbol{\theta}}) = V(k,\hat{\boldsymbol{\theta}}) + \frac{1}{2}e^2(k+1,\hat{\boldsymbol{\theta}})\,. \tag{11.2.33}$$

The first and second partial derivative are then given as

$$\boldsymbol{V}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}},k+1) = \underbrace{\boldsymbol{V}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}},k)}_{\approx\boldsymbol{0}} + e(\hat{\boldsymbol{\theta}},k+1)\frac{\partial e(\boldsymbol{\theta},k+1)}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{11.2.34}$$

and

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}},k+1) = \boldsymbol{V}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}},k) &+ \Big(\frac{\partial e(\boldsymbol{\theta},k+1)}{\partial\boldsymbol{\theta}}\Big)^{\mathrm{T}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\Big(\frac{\partial e(\boldsymbol{\theta},k+1)}{\partial\boldsymbol{\theta}}\Big)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\\ &\underbrace{+\,e(\hat{\boldsymbol{\theta}},k+1)\Big(\frac{\partial^2 e(\boldsymbol{\theta},k+1)}{\partial\boldsymbol{\theta}^2}\Big)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}_{\approx\boldsymbol{0}}\,,\end{aligned} \tag{11.2.35}$$

where the indicated terms have been approximated by zero (Söderström, 1973). These equations now allow to formulate the estimation algorithm as

$$\hat{\boldsymbol{\theta}}(k+1) = \hat{\boldsymbol{\theta}}(k) + \boldsymbol{\gamma}(k)e(k+1) \tag{11.2.36}$$

with

$$\boldsymbol{\gamma}(k) = \boldsymbol{P}(k+1)\boldsymbol{\varphi}(k+1) = \frac{\boldsymbol{P}(k)\boldsymbol{\varphi}(k+1)}{1+\boldsymbol{\varphi}^{\mathrm{T}}(k+1)\boldsymbol{P}(k)\boldsymbol{\varphi}(k+1)} \tag{11.2.37}$$

$$\boldsymbol{P}(k) = \boldsymbol{V}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}(k-1),k) \tag{11.2.38}$$

$$\boldsymbol{P}(k+1) = \big(\boldsymbol{I} - \boldsymbol{\gamma}(k)\boldsymbol{\varphi}^{\mathrm{T}}(k+1)\big)\boldsymbol{P}(k) \tag{11.2.39}$$

$$\boldsymbol{\varphi}(k+1) = -\frac{\partial e(\boldsymbol{\theta}(k),k+1)}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{11.2.40}$$

$$e(k+1) = y(k+1) - \hat{\boldsymbol{\psi}}^{\mathrm{T}}(k+1)\hat{\boldsymbol{\theta}}(k) \tag{11.2.41}$$

$$\hat{\nu}(k+1) = \hat{e}(k+1) \tag{11.2.42}$$

and hence one approximates $\boldsymbol{\psi}^{\mathrm{T}}$ from (11.2.31) by

$$\begin{aligned}\hat{\boldsymbol{\psi}}^{\mathrm{T}}(k+1) = \big(&-y(k-1)\,\ldots\,-y(k-m)\big|u(k-d-1)\,\ldots\,u(k-d-m)\big|\\ &e(k-1)\,\ldots\,e(k-m)\big)\,.\end{aligned} \tag{11.2.43}$$

The elements of the vector $\boldsymbol{\varphi}^{\mathrm{T}}(k+1)$ can now be determined as

$$\begin{aligned}\boldsymbol{\varphi}^{\mathrm{T}}(k+1) = -\bigg(&\frac{\partial e(k+1)}{\partial a_1}\ldots\frac{\partial e(k+1)}{\partial a_m}\,\frac{\partial e(k+1)}{\partial b_1}\ldots\frac{\partial e(k+1)}{\partial b_m}\\ &\frac{\partial e(k+1)}{\partial d_1}\ldots\frac{\partial e(k+1)}{\partial d_m}\bigg)\end{aligned} \tag{11.2.44}$$

with $e(k)=\hat{\nu}(k)$ and (11.2.1) are given as

$$z\frac{\partial e(z)}{\partial a_i} = \frac{1}{\hat{D}(z^{-1})}y(z)z^{-(i-1)} = y'(z)z^{-(i-1)} \tag{11.2.45}$$

$$z\frac{\partial e(z)}{\partial b_i} = -\frac{1}{\hat{D}(z^{-1})}u(z)z^{-(i-1)}z^{-d} = -u'(z)z^{-(i-1)}z^{-d} \tag{11.2.46}$$

$$z\frac{\partial e(z)}{\partial d_i} = -\frac{1}{\hat{D}(z^{-1})}e(z)z^{-(i-1)} = -e'(z)z^{-(i-1)} \tag{11.2.47}$$

for $i=1,\ldots,m$. These entries can be understood as filtered signals

$$\begin{aligned}\hat{\boldsymbol{\varphi}}^{\mathrm{T}}(k+1) = \big(&-y'(k-1)\,\ldots\,-y'(k-m)\big|u'(k-d-1)\,\ldots\\ &u'(k-d-m)\big|e'(k-1)\,\ldots\,e'(k-m)\big)\end{aligned} \tag{11.2.48}$$

which can be generated by the difference equation

$$y'(k) = y(k)-\hat{d}_1 y'(k-1)-\ldots-\hat{d}_m y'(k-m) \tag{11.2.49}$$

$$u'(k-d) = u(k-d)-\hat{d}_1 u'(k-d-1)-\ldots-\hat{d}_m u'(k-d-m) \tag{11.2.50}$$

$$e'(k) = e(k)-\hat{d}_1 e'(k-1)-\ldots-\hat{d}_m e'(k-m)\,. \tag{11.2.51}$$

For the $\hat{d}_i$, one can use the current estimates $\hat{d}_i(k)$. Due to the simplifying approximations at the beginning of the derivation, one will only obtain an approximation of the solution of the non-recursive maximum likelihood method.

As initial values, one can use

$$\hat{\boldsymbol{\theta}}(0)=\mathbf{0},\;\; \boldsymbol{P}(0)=\alpha\boldsymbol{I},\;\; \boldsymbol{\varphi}(0)=\mathbf{0}\,. \tag{11.2.52}$$

The convergence criteria are identical to those of the non-recursive maximum likelihood estimation. In particular, the roots of $D(z)=0$ must be within the unit circle, so that (11.2.49), (11.2.50), and (11.2.51) are stable.

### 11.2.3 Cramér-Rao Bound and Maximum Precision

The Cramér-Rao bound (Eykhoff, 1974), see (8.5.14), can also be evaluated for the maximum likelihood estimation of linear dynamic systems. In the case of multiple parameters, the Cramér-Rao bound is given as

$$\operatorname{cov}\Delta\hat{\boldsymbol{\theta}} = \mathrm{E}\big\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\big\} \geq \boldsymbol{J}^{-1} \tag{11.2.53}$$

with the *information matrix*

$$\boldsymbol{J} = \mathrm{E}\bigg\{\bigg(\frac{\partial L}{\partial\boldsymbol{\theta}_0}\bigg)\bigg(\frac{\partial L}{\partial\boldsymbol{\theta}_0}\bigg)^{\mathrm{T}}\bigg\} = -\mathrm{E}\bigg\{\frac{\partial^2 L}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\bigg\}\,. \tag{11.2.54}$$

Here, $\boldsymbol{\theta}_0$ denotes the true parameters. For a Gaussian distributed error $e(k)$, one can equate

$$\frac{\partial L}{\partial\boldsymbol{\theta}_0} = -\frac{1}{\sigma_{\mathrm{e}}^2}\frac{\partial V}{\partial\boldsymbol{\theta}_0} \tag{11.2.55}$$

and hence

$$\boldsymbol{J} = \frac{1}{\sigma_{\mathrm{e}}^4}\mathrm{E}\bigg\{\bigg(\frac{\partial V}{\partial\boldsymbol{\theta}_0}\bigg)\bigg(\frac{\partial V}{\partial\boldsymbol{\theta}_0}\bigg)^{\mathrm{T}}\bigg\} = \frac{1}{\sigma_{\mathrm{e}}^2}\mathrm{E}\bigg\{\frac{\partial^2 L}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\bigg\}\,. \tag{11.2.56}$$

From this follows for the covariance of the parameter estimates

$$\operatorname{cov}\Delta\hat{\boldsymbol{\theta}} \geq \frac{2V}{N}\mathrm{E}\big\{\boldsymbol{V}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\big\}\,. \tag{11.2.57}$$

This result shows that under the given assumptions, there is no other unbiased estimator that delivers estimates with a smaller variance than the maximum likelihood estimator. The maximum likelihood estimate is hence *asymptotically efficient*.

If the Cramér-Rao bound is applied to the fundamental equation of the least squares parameter estimation, (9.1.12),

$$\boldsymbol{y} = \boldsymbol{\Psi}\hat{\boldsymbol{\theta}} + \boldsymbol{e}\ , \tag{11.2.58}$$

then the ln-likelihood function is given as

$$L(\boldsymbol{\theta}) = -\frac{1}{2\sigma_{\mathrm{e}}^2}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} + \mathrm{const} \tag{11.2.59}$$

and the information matrix is given as

$$J = \frac{1}{\sigma_{\mathrm{e}}^2}\mathrm{E}\big\{\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\big\}\ , \tag{11.2.60}$$

and hence, compare (9.1.24),

$$\operatorname{cov}\Delta\hat{\boldsymbol{\theta}} \geq \sigma_{\mathrm{e}}^2\mathrm{E}\Big\{\big(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\big)^{-1}\Big\}\,. \tag{11.2.61}$$

The lower bound is thus identical with (9.1.69). A further comparison with (9.5.7) shows that for the case of a non-correlated error signal and a model according to (9.1.12) or (11.2.58) respectively, the estimation by means of the method of least squares, by the Markov estimation and by the maximum likelihood method all yield parameter estimates with the smallest possible variance. A comparison of the Cramér-Rao bound with simulation results by van den Boom (1982) shows a good match for the best parameter estimation methods. Ninness (2009) discussed the error quantification also for finite (and especially short length) data sequences.

## 11.3 Summary

This chapter has presented the Bayes estimator and the maximum likelihood estimator, which were now specifically tailored to the identification of linear dynamic systems in discrete-time. The Bayes estimator treats the parameters as random variables and incorporates information about their probability density functions into the solution of the parameter estimation problem. As this information however is seldom available in practical applications, the Bayes estimator is limited in its applicability for parameter estimation from experimental data. Still, it can be shown that the maximum likelihood estimator and the least squares estimator  can both be derived from the Bayes estimator, see Fig. 11.1.

The maximum likelihood estimation is based on a stochastic treatment of the measured signals. The parameter estimates are determined based on the probability density function of the observed measurements. For an ARMAX model structure and a normally distributed, statistically independent error signal, a maximum likelihood estimation technique has been derived for linear dynamic discrete-time systems, which can be solved by a non-linear optimization algorithm. After certain simplifying approximations, also a recursive maximum likelihood estimator could be formulated. While the computational burden for the maximum likelihood estimator is high, it can on the other hand be shown that the estimator is asymptotically efficient, i.e. that it reaches the Cramér-Rao bound and yields estimates with the smallest possible variance for specified conditions. Maximum likelihood estimators can also be formulated for many other settings, e.g. for frequency domain identification (McKelvey, 2000).

## Problems

**11.1. Bayes Estimation**

How can the Bayes rule be used to determine the conditional probability density function of the parameters for a given set of observed measurements $p(\boldsymbol{\theta}\,|\,\boldsymbol{y})$?

**11.2. Bayes Estimation, Maximum Likelihood, and Least Squares**

How do these parameter estimation methods relate to each other. Which assumptions lead from one estimator to the other?

**11.3. Cramér-Rao Lower Bound**
Derive the Cramér-Rao inequality for one parameter $\theta_0$. (Solution can be found in (, p. 14 Isermann, 1992)

## References

Åström KJ, Bohlin T (1965) Numerical identification of linear dynamic systems from normal operating records. In: Proceedings of the IFAC Symposium Theory of Self-Adaptive Control Systems, Teddington

van den Boom AJW (1982) System identification - on the variety and coherence in parameter- and order erstimation methods. Ph. D. thesis. TH Eindhoven, Eindhoven

van den Bos A (2007) Parameter estimation for scientists and engineers. Wiley-Interscience, Hoboken, NJ

Deutsch R (1965) Estimation theory. Prentice-Hall, Englewood Cliffs, NJ

Eykhoff P (1974) System identification: Parameter and state estimation. Wiley-Interscience, London

Fuhrt BP, Carapic M (1975) On-line maximum likelihood algorithm for the identification of dynamic systems. In: 4th IFAC-Symposium on Identification, Tbilisi, USSR

Isermann R (1992) Identifikation dynamischer Systeme: Besondere Methoden, Anwendungen (Vol 2). Springer, Berlin

Isermann R (2006) Fault-diagnosis systems: An introduction from fault detection to fault tolerance. Springer, Berlin

Lee KI (1964) Optimal estimation, identification, and control, Massachusetts Institute of Technology research monographs, vol 28. MIT Press, Cambridge, MA

Ljung L (1999) System identification: Theory for the user, 2nd edn. Prentice Hall Information and System Sciences Series, Prentice Hall PTR, Upper Saddle River, NJ

McKelvey T (2000) Frequency domain identification. In: Proccedings of the 12th IFAC Symposium on System Identification, Santa Barbara, CA, USA

Nahi NE (1969) Estimation theory and applications. J. Wiley, New York, NY

Ninness B (2009) Some system identification challenges and approaches. In: Proceedings of the 15th IFAC Symposium on System Identification, Saint-Malo, France

Papoulis A (1962) The Fourier integral and its applications. McGraw Hill, New York

Peterka V (1981) Bayesian approach to system identification. In: Trends and progress in system identification, Pergamon Press, Oxford

Raol JR, Girija G, Singh J (2004) Modelling and parameter estimation of dynamic systems, IEE control engineering series, vol 65. Institution of Electrical Engineers, London

Söderström T (1973) An on-line algorithm for approximate maximum likelihood identification of linear dynamic systems. Report 7308. Dept. of Automatic Control, Lund Inst of Technology, Lund van der Waerden BL (1969) Mathematical statistics. Springer, Berlin