FUNDAMENTALS OF PARAMETER ESTIMATION

H. KRONMÜLLER

Institut für Prozeßmeßtechnik und Prozeßleittechnik, Universität Karlsruhe, German Federal Republic

Summary

Since the time that man first became involved with measurements, he has had to deal with the problems caused by discrepancies in different measurements of the same object. The attempt for reducing the effects of these discrepancies had led to the development of estimation theory.

The discrepancies or errors are generally regarded as being unknowable or random. To reduce their effect with respect to the quantity of interest one is led to the problem of defining an estimator.

The problem of estimating parameters from observational data can be traced from antiquity. From about 300 B.C. Babylonian astronomers dealt with this problem. Up to present times astronomical studies have provided a major stimulus for the development of estimation theory. In the 18. and 19. century we find essential contributions by Bernoulli, Euler, Legendre, Gauss and Bayes. In these days we recognize wide applications in the space technology, control and measurement theory.

Definitions

Sample. The outcome of an experiment, performed under a given set of conditions, is called a sample. Generally a sample is a set of measured variables y_i, $i=1 \ldots n$. We write a sample as a vector:

$$y_1 \ldots y_n \rightarrow \mathbf{y}^T.$$

Estimator. An estimator $\hat{\mathbf{b}}=\hat{\mathbf{b}}(\mathbf{y})$ is an algorithm using the sample to calculate approximate values of a set of unknown parameters $b_1 \ldots b_k=\mathbf{b}^T$, $k \leq n$.

Obviously, for one experiment a large number of estimators may be invented. For example, take k "good measurements" out of the sample $\mathbf{y}$ and calculate the k unknown parameters $\mathbf{b}$. But this procedure drops off the information of the $(n-k)$ events of the experiment here considered to be "bad".

Estimation theory deals with all kinds of estimators and gives criteria for optimal or at least "good" estimators. The performance of estimators is judged with the following qualities:

Bias of estimators

An estimator is called unbiased if

$$E\{\hat{\mathbf{b}}\}=E\{\mathbf{b}\}. \tag{1}$$

An estimator has a bias $\mathbf{r(b)}$ given by

$$\mathbf{r} = E\{\hat{\mathbf{b}}\} - E\{\mathbf{b}\}. \tag{2}$$

$E\{\ \}$ is the expectance operator:

$$E\{\ldots\} = \int \ldots \int \ldots p(\mathbf{b}, \mathbf{y}) \mathrm{d}\mathbf{y} \mathrm{d}\mathbf{b};$$

$$\mathrm{d}\mathbf{y} = \mathrm{d}y_1 \ldots \mathrm{d}y_n;$$

$$\mathrm{d}\mathbf{b} = \mathrm{d}b_1 \ldots \mathrm{d}b_k.$$

The feature "unbiased" guarantees that in the case of a large number of tests $j = 1 \ldots m$ under identical conditions the mean of the estimates $\hat{\mathbf{b}}_j$ tends to $\mathbf{b}$ with $m \to \infty$.

Another appreciated quality is that the deviations of repeated estimations $\hat{\mathbf{b}}_j$ are to be small. This challenge leads to the definition of efficient estimators.

Efficient estimators

For the sake of simplicity $\mathbf{b}$ is assumed to be unbiased. A measure of performance with respect to the deviations is the covariance matrix of an estimated vector $\hat{\mathbf{b}}$

$$E\{(\mathbf{b} - \hat{\mathbf{b}})(\mathbf{b} - \hat{\mathbf{b}})^T\} = \mathbf{V}_{\hat{\mathbf{b}}}$$

An estimator $\hat{\mathbf{b}}_0$ is called to be efficient, if for all other estimators $\hat{\mathbf{b}}_i$

$$\mathbf{V}_{\hat{\mathbf{b}}_0} \leq \mathbf{V}_{\hat{\mathbf{b}}_i} \tag{3}$$

(The covariance matrix $\mathbf{V}_{\hat{\mathbf{b}}}$ is a positive definite matrix. A positive definite matrix $\mathbf{A}$ is greater than a positive definite matrix $\mathbf{B}$, $\mathbf{A} > \mathbf{B}$, if for any vector $\mathbf{x} \neq 0$ $\mathbf{x}^T\mathbf{A}\mathbf{x} > \mathbf{x}^T\mathbf{B}\mathbf{x}$.)

Unbiased and efficient estimators are in some sense ideal estimators, but these qualities are missed in general. Very often consistent estimators are reliable and satisfying.

Consistent estimators

An estimator is called to be consistent, if with increasing sample size n the algorithm $\hat{\mathbf{b}}_n = \hat{\mathbf{b}}_n(y_1 \ldots y_n)$ converges in probability to $\mathbf{b}$:

$$\lim_{n \to \infty} \hat{\mathbf{b}}_n = \mathbf{b} \tag{4}$$

(lim = limes in the mean, i.e. $\lim_{n\to\infty} E\{\mathbf{b}_n\} = \mathbf{b}$, no bias for $n\to\infty$, and $\lim_{n\to\infty} \mathbf{V}_{\hat{\mathbf{b}}_n} = 0$, all covariances of the parameters vanish for $n\to\infty$).

All definitions above were introduced by Fisher [1]. Additionally, he gave the definition of a sufficient estimator. The basic objection of estimation theory is to determine a function of the measured data, the sample, which approximates the unknown parameter. Such a function $\mathbf{\Phi}(\mathbf{y})$ is called a statistic. Obviously, the statistic should use as much "information", whatever this is, about $\mathbf{b}$ as is contained in the sample itself. A first step in the systematic development of estimators using all information of the sample, is the class of sufficient statistics.

Sufficient estimators

A statistic $\mathbf{\Phi}(\mathbf{y})$ or $\hat{\mathbf{b}}(\mathbf{y})$ is said to be sufficient, if the "a posteriori density" of $\mathbf{b}$, given $\mathbf{\Phi}(\mathbf{y})$ is equivalent to "a posteriori density" of $\mathbf{b}$, given sample $\mathbf{y}$:

$$p(\mathbf{b}|\mathbf{y}) = p(\mathbf{b}|\mathbf{\Phi}(\mathbf{y}))\,. \tag{5}$$

Another equivalent formulation is (the proof is straightforward), using the Bayes' rule:

$$p(\mathbf{y}|\mathbf{\Phi}, \mathbf{b}) = p(\mathbf{y}|\mathbf{\Phi})\,. \tag{6}$$

From Eq. (6) we recognize, a statistic $\mathbf{\Phi}(\mathbf{y})$ is sufficient, if the sample $\mathbf{y}$ or any other statistic $\mathbf{\Phi}$ does not provide an additional information about the distribution of the variables $\mathbf{b}$.

Examples: Estimation of the voltage of a battery (Fig. 1).

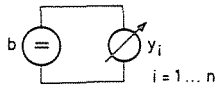

Fig. 1. Voltage of a battery

We use the a priori information that the voltage b will be constant but unknown during the experiment. The noise n_i is additive in the signal: $y_i = b + n_i$. The mean of the noise shall be zero, the variance constant and the measurements uncorrelated:

$$E\{n_i\} = 0,\quad E\{n_i n_j\} = \sigma^2 \delta_{ij};$$

$$\delta_{ij} = \begin{cases} 1 & i = j; \\ 0 & i \neq j. \end{cases}$$

In the example we choose a linear estimator, i.e.

$$\hat{b} = \sum_{i=1}^{n} \alpha_i y_i \,. \tag{7}$$

The bias of the estimator:

$$E\{\hat{b}\} = E\{\Sigma \alpha_i y_i\} = b\Sigma\alpha_i + \Sigma\alpha_i E\{n_i\} = b\Sigma\alpha_i \,. \tag{8}$$

The estimator is unbiased, if $\sum_{i=1}^{n} \alpha_i = 1$.

The variance of the estimator:

$$V_{\hat{b}} = E\{(b-\hat{b})^2\} = E\{(b-\Sigma\alpha_i(b+n_i))^2\} = \sigma^2 \Sigma\alpha_i^2 \,.$$

The best unbiased estimator in the class of linear estimators is, remember the constraint $\Sigma\alpha_i = 1$:

$$\frac{\partial V_{\hat{b}}}{\partial \alpha_i} = 2\alpha_i\sigma^2 + \lambda = 0 \mid i = 1 \ldots n \,;$$

$$\curvearrowright \alpha_i = \frac{1}{n} \,; \tag{9}$$

$$\hat{V}_{b_{\text{best}}} = \frac{\sigma^2}{n} \,.$$

Efficiency: This question will be answered later on in section 2.
Consistency: The unbiased estimator (7) is consistent because

$$\lim_{n\to\infty} E\{\hat{b}_n\} = b \quad \text{and} \quad \lim_{n\to\infty} \frac{\sigma^2}{n} = 0 \,. \tag{10}$$

Sufficiency: The estimator for b is following (7) and (9):

$$\hat{b} = \frac{1}{n} \Sigma y_i \,;$$

for the variance σ^2 we choose an estimator $\hat{\sigma}^2 = \frac{1}{n}\Sigma(y_i - \hat{b})^2$. The noise $n_i = y_i - b$ is assumed to have normal distribution:

$$p(\mathbf{y} \mid b, \sigma^2) = a \exp\left\{ -\frac{1}{2\sigma^2} \Sigma(y_i - b)^2 \right\} . \tag{11}$$

It is true that

$$\Sigma(y_i - b)^2 = \Sigma(y_i - \hat{b} + \hat{b} - b)^2 = \Sigma(y_i - \hat{b})^2 + n(\hat{b} - b)^2 + 2\Sigma(y_i - \hat{b}) \cdot (\hat{b} - b) = \tag{12}$$

$$= n\hat{\sigma}^2 + n(\hat{b} - b)^2 \,.$$

Using Bayes' rule:

$$p(b, \sigma^2|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|b, \sigma^2)p(b, \sigma^2)$$

one yields

$$p(b, \sigma^2|\mathbf{y}) = \frac{a\, e^{-\frac{\hat{\sigma}^2 n}{2\sigma^2} - n\frac{(b-\hat{b})^2}{2\sigma^2}} \cdot p(b, \sigma^2)}{\int\int p(b, \sigma^2)p(\mathbf{y}|b, \sigma^2)\mathrm{d}b\mathrm{d}\sigma^2}\, p(b, \sigma^2|\hat{b}, \hat{\sigma}^2)\,. \tag{13}$$

The last equation is established with Eq. (11) and (12). $p(y|b, \sigma^2)$ is expressed in terms of $\hat{\sigma}^2$ and $\hat{b}$ (Eq. (12)).

Eq. (13) corresponds to Eq. (5) so the estimators for $\hat{b}$ and σ^2 are sufficient.

Optimal estimators, lower bound for covariance

In probability theory all about the process is known, if the distribution or the probability density function — p.d.f. — is known. In estimation theory we have an optimum of information about the process, if the class of the p.d.f. is known. The p.d.f. depends on the random variables $\mathbf{y}$ of the sample and further on the unknown parameter $\mathbf{b}$. The problem in this field is, to find out which parameter vector $\hat{\mathbf{b}}$ is valid for our experiment. We start with $p(\mathbf{y}|\mathbf{b})$ and adjust the parameter vector $\mathbf{b} \rightarrow \hat{\mathbf{b}}$ in such a way that $\hat{\mathbf{b}}$ fits best our experimental results.

A measure how good a parameter vector $\mathbf{b}_1$ matches the random process given by $p(\mathbf{y}, \mathbf{b}_0)$ is the H-function introduced by Boltzmann in statistical mechanics:

$$H(\mathbf{b}_0, \mathbf{b}_1) = \int \{\log p(\mathbf{y}|\mathbf{b}_1)\}\, p(\mathbf{y}|\mathbf{b}_0)\mathrm{d}\mathbf{y}$$

The H-function is related tightly to the entropy, introduced by Shannon in communication theory in 1948.

The H-function is maximized by varying the parameter vector $\mathbf{b}_1$.

To become familiar with this definition, let us take one parameter b, the mean of the population $y_1 \ldots y_n$, and let us start independent tests.

The idea is to adjust the p.d.f. in such a way, that probability of all possible tests is a maximum. Instead of working with the densities $p(\mathbf{y}|\mathbf{b})$, we may maximize the function $\log p$, because log is a monotonic function. In our example of independent tests, we have

$$p(\mathbf{y}|b) = \prod_{i=1}^{n} p(y_i|b) \quad \text{and} \quad \log p(\mathbf{y}|b) = \Sigma \log p(y_i|b)$$

Figure 2 shows, how $p(y_i, b)$ is adjusted in the case of 3 tests to get a maximum.

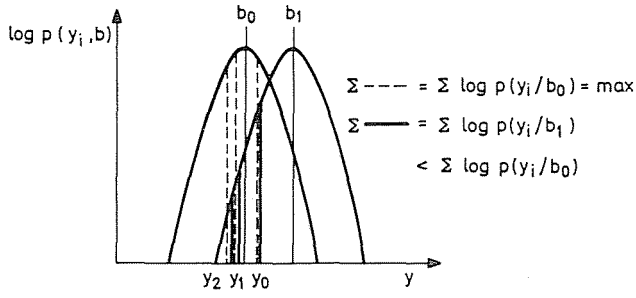

Fig. 2. Adjusting parameter b to b_0 to get a maximum of probability

If we start a large number of tests, it is comfortable to calculate with relative frequencies $h(y)\Delta y$. $h(y)\Delta y$ is the relative frequency of all events having a result of y between y and $y+\Delta y$.

The H-function is calculated with respect to the classification of the above results by:

$$H = \Sigma \log p(\mathbf{y}|b)h(\mathbf{y})\Delta \mathbf{y} .$$

For a very large number of tests $h(\mathbf{y})\Delta\mathbf{y}$ may be replaced by the density $p(\mathbf{y})$ $h(\mathbf{y})\Delta\mathbf{y} \to p(\mathbf{y}|b_0)\mathrm{d}\mathbf{y}$. The function to be maximized is

$$H = \int \log \{p(\mathbf{y}, b_1)\} p(\mathbf{y}, b_0)\mathrm{d}\mathbf{y} .$$

It is easy to show mathematically that the H-function has a relative maximum for $b_1 = b_0$:

$$\frac{\partial H}{\partial b_1} = \int \frac{\frac{\partial p}{\partial b_1}}{p(\mathbf{y}, b_1)} p(\mathbf{y}, b_0)\mathrm{d}\mathbf{y} ,$$

$$\frac{\partial H}{\partial b_1}\Big|_{b_1 = b_0} = 0 \quad \text{as} \quad \int \frac{\partial p}{\partial b_1}\mathrm{d}\mathbf{y} = \frac{\partial}{\partial b_1}\int p\mathrm{d}\mathbf{y} = 0 .$$

For a maximum, $\dfrac{\partial^2 H}{\partial b_1^2} < 0$, we obtain

$$\frac{\partial^2 H}{\partial b_1^2} = \int \frac{\frac{\partial^2 p}{\partial b_1^2}}{p(\mathbf{y}, b_1)} p(\mathbf{y}, b_0)\mathrm{d}\mathbf{y} - \int \left(\frac{\partial}{\partial b_1} \log p(\mathbf{y}, b_1)\right)^2 p(\mathbf{y}, b_0)\mathrm{d}\mathbf{y} ;$$

$$\frac{\partial^2 H}{\partial b_1^2}\Big|_{b_1=b_0} = -\int \left\{\frac{\partial}{\partial b_1}\log p(\mathbf{y}, b_0)\right\}^2 p(\mathbf{y}, b_0)\mathrm{d}\mathbf{y} \leq 0\,. \tag{14}$$

It is really appealing to formulate a problem as a principle of extremum. In natural science this way has an old tradition, remember for example mechanics with the Hamilton—Jacobian equations, thermodynamics with the maximum of entropy or optics with the principle of Fermat. But with the problem of designing a practical estimator this principle does not help, because the wanted quantity $\mathbf{b}_0$ is hidden in $p(\mathbf{y}, \mathbf{b}_0)\mathrm{d}\mathbf{y}$. But it is extremely useful to elaborate the lower bound of variances. Another application is the use in Kullback's information integral [2]:

$$I(\mathbf{b}_0, \mathbf{b}_1) = \int \log \frac{p(\mathbf{y}|\mathbf{b}_0)}{p(\mathbf{y}|\mathbf{b}_1)}\, p(\mathbf{y}|\mathbf{b}_0)\mathrm{d}\mathbf{y}\,. \tag{15}$$

By (15) the "information distance" between a density with $\mathbf{b}_0$ and $\mathbf{b}_1$ may be calculated.

For the sake of simplicity again, only one parameter b and an unbiased estimator is assumed, then

$$\int (\hat{b} - b) p(\mathbf{y}|b)\mathrm{d}\mathbf{y} = 0\,.$$

Differentiation with respect to b gives:

$$\int \underbrace{(\hat{b} - b)}_{x} \underbrace{\frac{\partial}{\partial b}\log\left(p(\mathbf{y}|b)\right)}_{y} \cdot p(\mathbf{y}|b)\mathrm{d}\mathbf{y} = 1\,.$$

Applying the Schwartz-inequality:

$$E\{x^2\}E\{y^2\} \geq E\{xy\}^2\,, \tag{16}$$

thus yielding

$$1 = \left[\int (\hat{b} - b)\frac{\partial}{\partial b}\{\log p\} p\, \mathrm{d}\mathbf{y}\right]^2 \leq \int (\hat{b} - b)^2 p\mathrm{d}\mathbf{y} \int \left(\frac{\partial \log p}{\partial b}\right)^2 p\mathrm{d}\mathbf{y}\,. \tag{17}$$

For the lower bound σ_{eff}^2 of the variance, using Eqs (14) and (17) we get

$$\sigma_{\mathrm{eff}}^2 \geq \frac{1}{-\dfrac{\partial^2 H(b, \hat{b})}{\partial \hat{b}^2}\Big|_{\hat{b}=b}}\,. \tag{18}$$

The Schwartz-inequality has the equation sign only, if $x = cy$ or in our case

$$c(\hat{b}(\mathbf{y}) - b) = \frac{\partial}{\partial b}\log p(\mathbf{y}|b)\,. \tag{19}$$

Example: The former example, measuring the battery voltage, is used again. We take the best linear estimator for the mean $\hat{b} = \frac{1}{n} \Sigma y_i$, the noise $n_i = y_i - b$ normally distributed again, so yielding:

$$p(\mathbf{y}|b) \sim \exp\left\{-\frac{1}{2\sigma^2} \Sigma (y_i - b)^2\right\};$$

$$\left(\frac{\partial}{\partial b} \log p(\mathbf{y}|b)\right)^2 = \left\{\frac{\Sigma (y_i - b)}{\sigma^2}\right\}^2;$$

$$E\left\{\frac{\Sigma (y_i - b)}{\sigma^2}\right\}^2 = \frac{n}{\sigma^2};$$

and with Eqs (17) and (18):

$$\sigma^2_{\text{eff}}\, \hat{b} = \frac{\sigma^2}{n}. \tag{20}$$

Comparing with Eq. (9) we shall find the estimator $\hat{b} = \frac{1}{n} \Sigma y_i$ to be efficient. Or using the condition (19) for the efficient estimator

$$c(\hat{b} - b) = \frac{\partial}{\partial b} \log p(y|b) = \frac{1}{\sigma^2} \Sigma (y_i - b),$$

the condition is fulfilled with $c = \frac{n}{\sigma^2}$ and $\hat{b} = \frac{1}{n} \Sigma y_i$.

For a parameter vector $\mathbf{b}$, the lower bound of the covariance matrix $\mathbf{V}_{\hat{\mathbf{b}}}$ was given by Cramér and Rao [3]:

$$\mathbf{V}_{\hat{\mathbf{b}}} = E\{(\hat{\mathbf{b}} - \mathbf{b})(\hat{\mathbf{b}} - \mathbf{b})^T\} \geq \mathbf{J}^{-1}. \tag{21}$$

The matrix $\mathbf{J}$ is the Fisher information matrix:

$$\mathbf{J} = E\left\{\left[\frac{\partial}{\partial \mathbf{b}} \log p(\mathbf{y}, \mathbf{b})\right]\left[\frac{\partial}{\partial \mathbf{b}} \log p(\mathbf{y}, \mathbf{b})\right]^T\right\}.$$

Formula (21) shows identity, if and only if

$$\frac{\partial}{\partial b} \ln p(\mathbf{y}|\mathbf{b}) = c(\mathbf{b})\,[\hat{\mathbf{b}} - \mathbf{b}]. \tag{22}$$

Practical estimators

The most fruitful approach to get practical estimators is the loss-function $c(\hat{\mathbf{b}}-\mathbf{b})$ introduced by Bayes. The expectation of the loss function is called risk. The risk is minimized by the choice of $\hat{\mathbf{b}}$,

$$r(\hat{\mathbf{b}}-\mathbf{b})=E\{c(\hat{\mathbf{b}}-\mathbf{b})\}\stackrel{!}{=}\min. \tag{23}$$

The cost function will be chosen freely. The function c usually has a minimum for $\hat{\mathbf{b}}=\mathbf{b}$.
Remember the Bayes' rule, $p(\mathbf{b},\mathbf{y})=p(\mathbf{b}|\mathbf{y})p(\mathbf{y})$, and

$$\hat{\mathbf{b}}=\hat{\mathbf{b}}(\mathbf{y}),$$

thus yielding

$$\frac{\partial}{\partial\hat{\mathbf{b}}}E\{E_{\mathbf{y}}\{c(\hat{\mathbf{b}}-\mathbf{b})|\mathbf{y}\}\}=0, \tag{24}$$

or

$$E_{\mathbf{y}}\left\{\frac{\partial}{\partial\hat{\mathbf{b}}}c(\hat{\mathbf{b}}-\mathbf{b})|\mathbf{y}\right\}=0. \tag{25}$$

Interpreting equation (25), one gets the result: the best estimate in the sense of Bayes is a conditional expectation.
Very common is the quadratic loss function ($\mathbf{G}$ is the weighting matrix):

$$c(\hat{\mathbf{b}}-\mathbf{b})=(\hat{\mathbf{b}}-\mathbf{b})^T\mathbf{G}(\hat{\mathbf{b}}-\mathbf{b}). \tag{26}$$

Applying this loss function, one yields the *minimum mean square estimator*

$$\hat{\mathbf{b}}_{\mathrm{MS}}=E\{\mathbf{b}|\mathbf{y}\}. \tag{27}$$

The MS-estimator is unbiased because

$$E\{\hat{\mathbf{b}}_{\mathrm{MS}}\}=E\{E\{\mathbf{b}|\mathbf{y}\}\}=E\{\mathbf{b}\}.$$

The MS-estimator is independent of the weighting matrix $\mathbf{G}$ in Eq. (26).

This quality is very important in practice, for example, if we are only interested in one parameter b_j of the vector $\mathbf{b}$. The minimum mean square estimator in this case is again

$$\hat{b}_j=E\{b_j|\mathbf{y}\}.$$

The minimum mean square estimator is optimal not only for a quadratic

function c, but for 2 classes L_1 and L_2 of loss functions (Sherman [4], Mereditch [5]).

$$L_1 \begin{cases} c(\hat{\mathbf{b}}-\mathbf{b}) \text{ is symmetric}; \\ c(\hat{\mathbf{b}}-\mathbf{b}) \text{ is convex}; \end{cases}$$

$$L_2 \begin{cases} c(\hat{\mathbf{b}}-\mathbf{b}) \text{ is symmetric}; \\ c \text{ is nondecreasing}, \end{cases}$$

$$\text{i.e.} \quad c(\hat{\mathbf{b}}_1-\mathbf{b}_1) \leq c(\hat{\mathbf{b}}_2-\mathbf{b}_2)$$

$$\text{if} \quad (\hat{\mathbf{b}}_1-\mathbf{b}_1)^T(\hat{\mathbf{b}}_1-\mathbf{b}_1) \leq (\hat{\mathbf{b}}_2-\mathbf{b}_2)^T(\hat{\mathbf{b}}_2-\mathbf{b}_2).$$

For this 2 classes $\hat{\mathbf{b}}_{MS}$ is the optimal solution if

$$\lim_{|\mathbf{b}|\to\infty} c(\hat{\mathbf{b}}-\mathbf{b})p(\mathbf{b}|\mathbf{y})=0$$

An important representative of class L_2 is the *uniform loss function.*

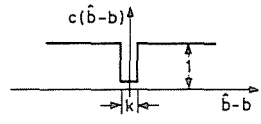

Fig. 3. Uniform loss function

The risk r is:

$$r = \int_{-\infty}^{\infty} p(\mathbf{b}|\mathbf{y})d\mathbf{b} - kp(\mathbf{b}|\mathbf{y}),$$

$$\frac{\partial r}{\partial \mathbf{b}}\Big|_{\mathbf{b}=\hat{\mathbf{b}}} = -k\frac{\partial}{\partial \mathbf{b}} p(\mathbf{b}|\mathbf{y})|_{\mathbf{b}=\hat{\mathbf{b}}} = 0. \tag{28}$$

This estimator is called "*maximum a posteriori*" or *MAP-estimator.*

Example: The mean b of a normal distribution $N(b, \sigma_y^2)$ has to be estimated, the sample of uncorrelated measurements is $\mathbf{y}^T=(y_1 \ldots y_n)$. The a priori distribution of b is known: normal distribution $N(0, \sigma_b^2)$. The variances σ_y^2 and σ_b^2 are known.

For the MAP estimator the a-posteriori-density $p(b|\mathbf{y})$ is wanted. With the Bayes'rule we get:

$$p(b|\mathbf{y}) = p(\mathbf{y}|b) \cdot p(b)/p(\mathbf{y}),$$

$$\frac{\partial p(b|\mathbf{y})}{\partial b} = \frac{\partial}{\partial b} p(\mathbf{y}|b) \cdot p(b) =$$

$$= \frac{\partial}{\partial b} \{(2\pi\sigma_y^2)^{-\frac{n}{2}} e^{-\frac{\Sigma(y_i-b)^2}{2\sigma_y^2}} (2\pi\sigma_b^2)^{-\frac{1}{2}} e^{-\frac{b^2}{2\sigma_b^2}}\}|_{b=\hat{b}} = 0,$$

or

$$\frac{\Sigma(y_i - \hat{b})}{\sigma_y^2} - \frac{\hat{b}}{\sigma_b^2} = 0;$$

$$\hat{b}_{\mathrm{MAP}} = \frac{\sigma_b^2}{\sigma_y^2 + n\sigma_b^2} \Sigma y_i. \tag{29}$$

In the case of large samples, $n \to \infty$ the estimator tends to $\hat{b} = \frac{1}{n} \Sigma y_i$. It is the same estimator we had in section 1 Eqs (7) and (9). The influence of the a priori information $(0, \sigma_b^2)$ diminishes with the increasing sample size.

Maximum likelihood estimator (ML-estimator). This estimator is very well known, some favourable features like asymptotic efficiency for large samples have to be mentioned.

The estimator is defined by the likelihood equation

$$\frac{\partial}{\partial \mathbf{b}} \log p(\mathbf{y}|\mathbf{b}) = 0. \tag{30}$$

Similar to the assignments of H-function or information integral by Kullback Eq. (15), the parameter vector $\mathbf{b}$ is adjusted so that the density has a maximum for the given sample $\mathbf{y}$. Maximum-likelihood estimators belong to the class of estimators with the uniform loss function, but no a priori information $p(\mathbf{b})$ is used or available. This can be easily shown with Bayes' rule:

$$\log p(\mathbf{b}|\mathbf{y}) = \log p(\mathbf{y}|\mathbf{b}) + \log p(\mathbf{b}) - \log p(\mathbf{y}).$$

The MAP estimator is, following Eq. (28):

$$\frac{\partial}{\partial \mathbf{b}} \log p(\mathbf{b}|\mathbf{y}) = 0 = \frac{\partial}{\partial \mathbf{b}} \log p(\mathbf{y}|\mathbf{b}) + \frac{\partial}{\partial \mathbf{b}} \log p(\mathbf{b}) = 0.$$

If there is no a priori information, the p.d.f. $p(\mathbf{b})$ is very flat and wide and so $\frac{\partial p(\mathbf{b})}{\partial \mathbf{b}} \to 0$.

The ML-estimator is one of the most developed estimators with excellent performance. For large samples, $n \to \infty$, the ML-estimator is asymptotically unbiased and efficient.

Fig 4. A flat and wide probability density function $p(b)$

The proof cannot be given here, but remember the Cramér—Rao condition (19):

$$\frac{\partial \ln p(\mathbf{y}|\mathbf{b})}{\partial \mathbf{b}} = c(\mathbf{b})\,(\mathbf{b} - \hat{\mathbf{b}}).$$

If the estimator is consistent, $\hat{\mathbf{b}} - \mathbf{b}$ vanishes asymptotically and the condition turns over to likelihood equation.

Examples for a ML-estimator:

A linear process-model is assumed: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{n}$. The noise n_i is normally distributed, the measurements are uncorrelated, the variances are known:

$$E\{n_i n_j\} = \sigma^2 \delta_{ij},$$

$$\log p(\mathbf{y}|\mathbf{b}) = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})}{2\sigma^2} + c.$$

The function to be minimized is, besides some factors, nothing else as the sum of least squares. Minimizing $\log p(\mathbf{y}|\mathbf{b})$ leads straightforward to the famous least square estimator introduced by Gauss (Eq. 42).

Linear estimators

Some properties of linear vector spaces

The following operations are admitted:

$$\begin{gathered} \text{addition } x' + y' = y' + x' = z' \\ \text{norm, length of a vector } (x', x') = \|x'\|^2 \\ \text{scalar product of vectors } (x', y') = (y', x') = \|x\|\,\|y\| \cos \Theta \\ |\cos \Theta| \leq 1 \end{gathered} \tag{31}$$

(Vectors in this section are written x' to make differences with the vectors $\mathbf{y}$, used in matrix calculus before!)

Vector space of dimension n:

n linear independent vectors x_i' form a linear vector space V_n of dimension n. Any vector z' of this space can be written by $z'=\Sigma c_i x_i'$. The independent vectors x_i', $i=1 \ldots n$ are said to form a base of V_n.

A set of n vectors x_i' is said to be linear dependent only if $\Sigma c_i x_i'=0$ for any $c_i \neq 0$.

Orthonormal base. A special base in a vector-space V_n is an orthonormal base e_i' with the definition

$$(e_i', e_j')=\delta_{ij}, \quad \delta_{ij}=\begin{cases}1 \text{ for } i=j \\ 0 \text{ for } i\neq j\end{cases}. \tag{32}$$

Principle of orthogonality or projection theorem:

Let be $x_1' \in V_n$ and $x_1'=\Sigma c_i e_i'$ and x' another vector not necessary an element of V_n, $x' \in V_m$, $m \geq n$.

The problem considered here is how to approximate the vector x' by an optimal vector $x_{10}' \in V_n$ so that the norm of the error vector $x_e'=x'-x_1'$ is a minimum.

We get the norm of the error-vector:

$$\begin{aligned}\|x_e'\|^2 &= \|x'-x_1'\|^2=((x'-\Sigma c_i e_i'), (x'-\Sigma c_i e_i')) \\ &= \|x'\|^2-2\Sigma c_i(e_i', x')+\Sigma c_i^2 \\ &= \|x'\|^2+\Sigma((x', e_i')-c_i)^2-\Sigma(x', e_i')^2 .\end{aligned}$$

$\|x_e'\|$ is independent of the choice of base e_i'.

The absolute minimum for the norm of x_e' is obtained, if $c_i=(x', e_i')$ and following this, the optimal vector x_{10}' gets

$$x_{10}'=\Sigma(x', e_i')e_i' \tag{33}$$

Straightforward, it follows:

$$\begin{aligned}&\text{Bessel inequality: } \|x_e'\| \geq \Sigma c_i^2 ; \\ &\text{Principle of orthogonality: } ((x'-x_{10}'), e_i')=0 .\end{aligned} \tag{34}$$

In the case of the optimal vector x_{10}' the error vector $(x'-x_{10}')$ is orthogonal to any vector of space V_n. The optimal vector x_{10}' is the projection of vector x' onto the vector space V_n.

The same axioms, operations and definitions, we used in the linear vector space, are valid to stochastic variables.

For example: A stochastic variable X corresponds to a vector x';

the norm is given by $E\{x^2\} \circ\!\!-\!\!\circ \|x'\|^2$;

addition: $E\{x\}+E\{y\}=E\{y\}+E\{x\} \circ\!\!-\!\!\circ x'+y'=y'+x'$;

scalar product: $E\{xy\} \circ\!\!-\!\!\circ (x'y')$;

Schwartz inequality:

$$E\{xy\}^2 \leq E\{x^2\}E\{y^2\} \circ\!\!-\!\!\circ (x', y')^2 \leq \|x'\|^2 \|y'\|^2 .$$

For stochastic variables the principle of orthogonality is given by (b stochastic variable, $\hat{b}=\hat{b}(\mathbf{y})$ stochastic variable):

$$E\{(\hat{b}-b)\hat{b}\}=0 \text{ or } E\{(\hat{b}-b)\Phi(\mathbf{y})\}=0 \tag{35}$$

with $\Phi(\mathbf{y})$ any function of the sample vector $\mathbf{y}$.

Example: In the class of linear estimators, an estimator for the mean b shall be designed: $\hat{b}=\Sigma\alpha_i y_i$.

$$\text{Process-model: } y_i=b+n_i .$$

The following properties are known: $E\{n_i\}=0$; $E\{n_i n_j\}=\sigma^2\delta_{ij}$;

$$E\{b^2\}=K_b,\ E\{n_i b\}=0.$$

Using the principle of orthogonality (35):

$$E\{(\hat{b}-b)\mathbf{y}^T\}=0$$

$$E\{\sum_i \alpha_i(b+n_i)(b+n_j)\}=E\{b(b+n_j)\}, \qquad j=1 \ldots n$$

$$K_b\Sigma\alpha_i+\alpha_j\sigma^2=K_b$$

$$\curvearrowright \mathrm{d}_i=\alpha_i=\alpha_j=\frac{K_b}{\sigma^2}(1-\Sigma\alpha_i)$$

or

$$\hat{b}=\frac{K_b}{\sigma^2+nK_b}\Sigma y_i . \tag{36}$$

The result is the same as in the example for the MAP-estimator section 2, Eq. (29), if $K_b=\sigma_b^2$.

Gauss—Markoff-estimators

The general formulation of a linear estimator is

$$\hat{\mathbf{b}} = \mathbf{Q}\mathbf{y} . \tag{37}$$

Applying the principle of orthogonality (35) one yields:

$$E\{\mathbf{b}\,\hat{\mathbf{b}}^T\} = E\{\mathbf{b}\,\mathbf{y}^T\}\,\mathbf{Q}^T = E\{\hat{\mathbf{b}}\,\hat{\mathbf{b}}^T\} = \mathbf{Q}E\{\mathbf{y}\,\mathbf{y}^T\}\,\mathbf{Q}^T$$

and

$$\mathbf{Q} = \mathbf{K}_{by}\,\mathbf{K}_{yy}^{-1} \quad \text{with} \quad \mathbf{K}_{by} = E\{\mathbf{b}\,\mathbf{y}^T\} \quad \text{and} \quad \mathbf{K}_{yy} = E\{\mathbf{y}\,\mathbf{y}^T\} .$$

With the assumptions $E\{\mathbf{n}\} = 0$; $E\{\mathbf{n}\mathbf{n}^T\} = \mathbf{V}_n$, $E\{\mathbf{n}\,\mathbf{b}^T\} = 0$ and $E\{\mathbf{b}\,\mathbf{b}^T\} = \mathbf{K}_b$, and the linear process model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{n}$, the Gauss—Markoff-estimator can be written:

$$\hat{b} = \mathbf{K}_b\,\mathbf{X}^T\,\{\mathbf{X}\mathbf{K}_b\,\mathbf{X}^T + \mathbf{V}_n\}^{-1}\,\mathbf{y} . \tag{38}$$

The inverse of the matrix $\{\ \}$ is cumbersome to calculate because of the large matrix size $n \times n$.

Here a matrix inversion lemma is helpful:

$$\mathbf{A}\mathbf{B}^{-1} = \mathbf{C}^{-1}\mathbf{D} \text{ is valid if } \mathbf{C}\mathbf{A} = \mathbf{D}\mathbf{B} .$$

This can be easily verified by multiplying with $\mathbf{C}$ from the left and with $\mathbf{B}$ from the right. With $\mathbf{A} = \mathbf{K}_b\,\mathbf{X}^T$ and $\mathbf{B} = (\mathbf{X}\mathbf{K}_b\,\mathbf{X}^T + \mathbf{V}_n)$ we get $\mathbf{C}\mathbf{K}_b\mathbf{X}^T = \mathbf{D}\mathbf{X}\mathbf{K}_b\,\mathbf{X}^T + \mathbf{D}\mathbf{V}_n$.

Let be $\mathbf{C} = \mathbf{C}_1 + \mathbf{C}_2$, one yields

$$\mathbf{D} = \mathbf{C}_2\mathbf{K}_b\mathbf{X}^T\,\mathbf{V}_n^{-1} ,$$

$$\mathbf{C}_1 = \mathbf{D}\mathbf{X} = \mathbf{C}_2\mathbf{K}_b\mathbf{X}^T\,\mathbf{V}_n^{-1}\,\mathbf{X} .$$

Choose $\mathbf{C}_2 = \mathbf{K}_b^{-1}$ and get $\mathbf{C}_1 = \mathbf{X}^T\,\mathbf{V}_e^{-1}\,\mathbf{X}$; $\mathbf{D} = \mathbf{X}^T\,\mathbf{V}_e^{-1}$.

The equivalent, more handsome formulation for the Gauss—Markoff-estimator is then

$$\hat{\mathbf{b}}_{\mathrm{GM}} = (\mathbf{K}_b^{-1} + \mathbf{X}^T\,\mathbf{V}_n\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{V}_n\,\mathbf{y} . \tag{39}$$

If there is no a priori information, i.e. $E\{b_i^2\} \to \infty$ and $E\{b_i b_j\} = 0\ i \neq j$, then $\mathbf{K}_b^{-1}$ vanishes:

$$\mathbf{K}_b^{-1} = 0 . \tag{40}$$

The result is the so-called minimum variance estimator

$$\hat{\mathbf{b}}_{MV} = (\mathbf{X}^T \mathbf{V}_n \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}_n \mathbf{y} \tag{41}$$

with the covariance matrix $\mathbf{V}_{\hat{b}} = (\mathbf{X}^T \mathbf{V}_n^{-1} \mathbf{X})^{-1} = \mathbf{J}^{-1}$.

In the case of normal distribution of the noise, the estimator is efficient and $\mathbf{V}_{\hat{b}}$ is proportional to the inverse of the Fisher information matrix $\mathbf{J}$.

If the noise $\mathbf{n}$ is uncorrelated, i.e. $\mathbf{V}_n = \sigma^2 \mathbf{I}$, the famous least squares estimator is received:

$$\hat{\mathbf{b}}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{42}$$

with the covariance matrix $\mathbf{V}_{\hat{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 = \sigma^2 \mathbf{J}^{-1}$.

$\mathbf{V}_{\hat{b}}$ is again the inverse of Fisher's information matrix $\mathbf{J}$ and the estimator is efficient in the case of normal distribution of the noise. All examples before with uncorrelated noise belong to the linear Gauss—Markoff or least-squares estimators.

Some intercorrelations

The paper deals with the basic ideas of the estimation theory. Only a few ideas are really fundamental. But the number and the names of the different estimators may be confusing for beginners. So some intercorrelations between the most important estimators will be given without any pretension for completeness.

If there is a linear process model and a normal distribution of noise $\mathbf{n}$ and parameter-vector $\mathbf{b}$, the Gauss—Markoff-estimator is efficient and so the best. The Gauss—Markoff-estimator, the MAP- and the Bayes MS-estimator are identical.

In the case of other distributions, the Gauss—Markoff-estimator is the best linear estimator.

ML-estimators, — no a priori-information is needed —, supply an asymptotically unbiased and efficient estimate.

For normal distributions the ML-estimator is identical with the minimum-variance-estimator. The minimum variance-estimator is the best linear estimator at all, if no a priori information is available.

If there is no knowledge of the density function of the parameter vector $\mathbf{b}$ and as well no knowledge of the covariance matrix $\mathbf{V}_n$ of the noise, the best linear estimator is the LS-estimator. If the noise is uncorrelated and normally distributed, the LS-estimator is unbiased and efficient and identical for the Bayes MS-estimator and the ML-estimator.

References

1. FISHER, R. A.: On the mathematical foundations of theoretical statistics, Phil. Trans. Royal Soc. London *222*, 309 (1922).
   FISHER, R. A.: Theory of statistical estimation, Proc. Cambridge Phil. Soc. *22*, 700 (1925).
2. KULLBACK, S.: Information Theory and Statistics, John Wiley, New York, 1959.
3. CRAMER, H.: Mathematical methods of statistics, Princeton University Press, 1946.
4. SHERMAN, S.: Non mean square error criteria, IRE Trans. Inform. Theory IT—4: 125 (1958).
5. MEDITCH, J.: Stochastic optimal linear estimation and control, McGraw-Hill, New York, 1967.
6. NAHI, NASSER: Estimation Theory and its Applications, John Wiley & Sons, 1969.
7. DEUTSCH, R.: Estimation Theory, Prentice-Hall, Inc., 1965.
8. WILKS, S. S.: Mathematical Statistics, John Wiley & Sons, 1962.
9. SAGE, A. P.—MELSA, J. L.: Estimation Theory with Applications to Communications and Control, McGraw-Hill, Inc., 1971.
10. SORENSON, H. W.: Parameter Estimation, Marcel Dekker, Inc., 1980.

Prof. Dr. Heinz KRONMÜLLER Institut für Prozeßmeßtechnik und Prozeßleittechnik, Universität Karlsruhe, Hertzstr. 16. D-7500 Karlsruhe 21, German Federal Republic