Process parameter and state estimation

Eykhoff, Pieter

Published: 01/01/1967

Document Version
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 02. Dec. 2017
Group Measurement and Control

PROCESS PARAMETER AND STATE ESTIMATION

P. Eykhoff.

Jan. 1967

Typ.: Miss M.J. van Schaik

The paper presents a coherent picture of the parameter-estimation problem. Starting from the theory of minimum risk- or Bayes estimation the paper shows how other statistical estimation techniques can be interpreted as special cases (viz. maximum likelihood, Markov-, and least squares estimation). The most important properties of these estimates are summarized.

The engineering approaches based on these statistical techniques can be divided into two classes, viz. "using explicit mathematical relations" and "using adjustment of a model". Each of these classes is discussed briefly. The majority of parameter estimation techniques can be embodied in this framework.

A very brief discussion is given on the problem of process state estimation which is related to parameter estimation. A few examples are used to illustrate the notions presented and to indicate some engineering considerations.

Summary

The paper presents a coherent picture of the parameter-estimation problem. Starting from the theory of minimum risk- or Bayes estimation the paper shows how other statistical estimation techniques can be interpreted as special cases (viz. maximum likelihood, Markov-, and least squares estimation). The most important properties of these estimates are summarized.

The engineering approaches based on these statistical techniques can be divided into two classes, viz. "using explicit mathematical relations" and "using adjustment of a model". Each of these classes is discussed briefly. The majority of parameter estimation techniques can be embodied in this framework.

A very brief discussion is given on the problem of process state estimation which is related to parameter estimation. A few examples are used to illustrate the notions presented and to indicate some engineering considerations.

Contents

1 Introduction
2 Elements of statistical estimation theory
   minimum risk estimate
   unconditional maximum likelihood
   conditional minimum risk
   Markov
   least squares
3 The use of explicit mathematical relations
4 The use of model-adjustment techniques
5 Process state estimation
6 Some examples
7 Concluding remarks
References

1. Introduction

For the engineering scientist there is a real challenge in the attempt to analyse the common factor behind the immense number of applications of automatic control systems. In the first place there is the wide spectrum of man-made control systems in engineering, ranging from the simplest forms of on-off control to complicated computer control in process industries. Besides this there are also the control mechanisms in society (e.g. the enforcement of the law), and the abundant number of control loops in biology.

For such a wide variety of "systems" only a rather loose notion can be expected to indicate the common factor behind the applications of control. Such a notion may be "uncertainty".

Since the celebrated application of Watt's regulator it has become apparent in innumerable cases that the introduction of some feedback mechanism can be used as an effective expedient for combatting uncertainties. These uncertainties may be the result of unpredictable influences (disturbances) of the environment on a system, or they may originate inside the system being considered (e.g. wear, aging, catalyst poisoning, etc.)

Simple feedback can diminish the influence of uncertainty, or at least shift the adverse effect of uncertainty from an important system quantity to an unimportant quantity. Of course there are limitations to the application of this remedy. Such limitations are found in those cases where very large parameter variations occur; such a situation may lead to the use of adaptive control principles. Another limitation may be found in the requirements for optimizing some economic criterion, leading to optimal or self-optimizing systems. In some of these cases there is a definite need for accurate knowledge on the system, both with respect to parameters and state variables.

In ordinary control applications the knowledge required on the process behaviour is limited to the data needed for stability considerations. For systems with large parameter variations (leading to adaptive control) and systems with strict "economic" criteria (leading to optimal or self-optimizing control) the need for process identification, -modelling and -parameter estimation is apparent.

In most engineering situations the black box approach is not a very realistic one. The experimenter, in many cases, has derived some a priori knowledge from physical insight into the process under consideration. This may give information on the topology of a conceptual model for that process, and perhaps even an approximate knowledge on the values of the coefficients (parameters) in that model.

For a survey of some examples of model building and the use of state description cf. ref. [1]. On account of the wide variety of different "processes" for which models have to be build (the range includes high performance aircrafts as well as chemical production plants) model building is quite strongly object-oriented. For this reason this survey paper is mainly devoted to a part of the identification procedure that permits the presentation of a more or less coherent picture, viz. process parameter and -state estimation.

Among the basic considerations on process-parameter estimation are the problems of time and cost. Generally speaking the time interval needed for the estimation has to be as short as possible in view of the need for timely informa-
tion on the (slowly changing) parameters. On the other hand the minimal time interval is bounded by the statistical effects (inherent variance of the measurements, influence of additive noise). In a sense the time interval needed for the estimation can be related to the instrumentation costs; cf. fig. 1.

Assume that only one parameter \( b \) has to be determined. One could then use \( n \) models, each with a different parameter value \( \beta_i \) (\( i=1, \ldots, n \)). If the output of the process is compared with that of each of the models (the difference or error being weighted according to some criterion) the value of \( b \) is known after one "measuring time-interval" \( T \); \( b = \beta_i \) if \( E_i < E_j \) for each \( j \neq i \).

Under stationary conditions the same knowledge can be obtained using one model and setting in the \( j \) th measuring time-interval \( \beta_j \). Apparently in this case the desired information is obtained only after an interval of length \( nT \); the instrumentation costs, however, are much less than in the previous case. For this reason one has to be careful in comparing different parameter-estimating schemes.

For the purpose of this paper the parameter estimation problem can be represented by the block diagram fig. 2 where \( P \) is the process with parameters \( b \) and \( M \) is a (conceptual) model with parameters \( \beta \);
The first class of methods uses a mathematical expression that explicitly provides the numerical values of the parameter estimates in terms of known a priori knowledge and measured variables. This is discussed in section 3: the use of explicit mathematical relations.

The second class of methods uses some kind of "model" of the process; estimates for the parameters are determined by successive adjustments of this model (either continuously or intermittently). These adjustments are made using some quality criterion with respect to process and model correspondence. This is outlined in section 4: the use of model-adjustment techniques.

2. Elements of statistical estimation theory.

The parameter estimation cases to be considered can all be illustrated by fig. 2. We want to derive an estimate, i.e. a fundamental relationship:

\[ \hat{\beta}_1 = \beta_1(z(0), \ldots, z(k-\tau)) = \hat{\beta}_1(z) \]  

so that a numerical value can be assigned for the process parameter \( \beta_1 \) from the sequence of observations on the output signal \( z \). When a number of parameters \( \{\beta_1, \ldots, \beta_n\} = \beta \) have to be estimated the relationship is indicated by

\[ \hat{\beta} = \hat{\beta}(z) \]

In statistical literature, e.g. ref. [2] some properties are defined for such functional relationships:

- **Unbiased estimate**: if \( \varepsilon[\beta] = b \)
- **Efficient estimate**: \( \beta(z) \) is an efficient estimate if

\[ \varepsilon[(\beta - b)^2] \leq \varepsilon[(\gamma - b)^2] \]

for all \( \gamma = \gamma(z) \)

- **Consistent estimate**: \( \lim_{k \to \infty} p((\hat{\beta} - b) = 0) = 1 \)

\( \varepsilon[ \cdot ] \) denotes the mathematical expectation. The first two properties may also hold only for \( k = \infty \); in that case they are called asymptotic unbiasedness and efficiency.

As a starting point for deriving estimates we will choose a situation where much a priori knowledge is available [3], viz.:

a) the probability density function of the noise \( n \). From this function the probability density of the measurements \( z \) follows; this function, being dependent on the process parameters \( b \), is denoted as \( p(z|b) \).

b) the probability density functions of the parameter values \( b \). This function is written as \( q(b) \).

c) the cost of choosing the value \( \hat{\beta} \) for the estimate if the true value of the process parameters is \( b \). This cost or loss function \( C(\hat{\beta}, b) \) has a minimum for \( \hat{\beta} = b \).

After considering the use of all this information we will indicate the effect of dropping the assumptions c), b) and a) successively.

The conditional risk of choosing \( \hat{\beta}(z) \) if the true process parameter value is \( b \) can be written as the expectation of the cost function with respect to the probability of the observations \( z \):

\[ s^2[C(\hat{\beta}, b)] = \int \int C(\hat{\beta}, b) p(z|b) d^{k+1}z \]  

where the following notation is used:

\[ s^2[C(\hat{\beta}, b)] \]

is the expectation with respect to \( z \)

\[ \int_{k+1} \]

indicates an \( k+1 \) fold integral

\[ d^{k+1}z = dz(0) \ dz(\tau) \ldots \ dz(k \tau) \]

The average risk for this estimating situation is the expectation with respect to the probability of the values of the process parameter \( b \)

\[ R(\hat{\beta}) = e_b \int \int \int C(\hat{\beta}, b) \ p(z|b) \ q(b) \ d^{k+1}z \ d^mb \]

The estimate that minimizes this expression is called the minimum risk-, minimum cost- or Bayes estimate.

On account of the well-known relationships:

\[ q(b) \]

the average risk can be written as:

\[ R(\hat{\beta}) = \int \int d^{k+1}z \ p(z|b) \ C(\hat{\beta}, b) \ q(b) \ d^mb \]

As \( p(z) > 0 \) the average risk \( R(\hat{\beta}) \) can be minimized by making the second integral as small as possible for each \( z \):

\[ \min_\beta \int \int C(\hat{\beta}, b) \ p(b|z) \ d^mb \]

A necessary condition for such a minimum is simply:

\[ \frac{\partial}{\partial \beta} \int \int C(\hat{\beta}, b) \ p(b|z) \ d^mb = 0 \]

As \( C(\hat{\beta}, b) \) has a minimum for \( \hat{\beta} = b \), as it presumably has small values for values of \( \hat{\beta} \) in the vicinity of \( b \), and as

\[ \int p(b|z) \ d^mb = 1 \]

it will be clear that eq. (6) is satisfied if \( \hat{\beta} \) is chosen in the neighbourhood of that \( b \) where the conditional probability \( p(b|z) \) is max.

Now we will drop the assumption c) i.e. the knowledge about an adequate cost or loss function \( C(\hat{\beta}, b) \). In that case it is reasonable to choose \( \hat{\beta} \) at that value of \( b \) for which \( p(b|z) \) is a maximum.
As according to eq. (4):
\[
p(b | z) = \frac{p(z | b) q(b)}{p(z)}
\]
with
\[
p(z) = \int p(z | b) q(b) \, db
\]
both assumptions a) and b) are still being used.

Next we will consider the consequences of dropping both assumptions b) and c); the a priori probability \( q(b) \) of the process parameters \( b \) is also unknown. This ignorance can be expressed by assuming a uniform distribution \( q(b) = A \) over the interval under consideration. In that case for any \( z \):
\[
\max_b p(b | z) = \frac{\lambda}{p(z)} \max_b p(z | b)
\]
As \( b \) is no longer a random variable but an unknown constant parameter the following necessary conditions for finding the maximum can be given:
\[
\frac{\partial}{\partial b} p(z | b) = 0
\]
or as the logarithmic function is monotonic
\[
\frac{\partial}{\partial b} \ln p(z | b) = 0
\]
Picking that root of this set of equations which yields the largest value for \( p(z | b) \) we have obtained the celebrated maximum likelihood estimate (M.L.E.) [4].

The M.L.E. has been extensively discussed in literature. Some of its interesting properties are the following:
- asympt. normality, i.e. \( p(\hat{b} | b) \) approaches a normal distribution for \( k \to \infty \), for a meaning of \( k \) cf. eq. (1)
- asympt. unbiasedness, i.e. \( \mathbb{E}[\hat{b}] = b \) for \( k = \infty \)
- asympt. efficiency, i.e. approaching the best accuracy or minimum variance as given by the Cramer-Rao (in)equality [5].

Now let us refer again to fig. 2 and assume that \( n \) has a \( k+1 \) variate gaussian (normal) distribution i.e.:
\[
p(n) = \frac{1}{(2\pi)^{k+1/2} |\Sigma|^{1/2}} \exp{-\frac{1}{2}(n - \mu)^\top \Sigma^{-1} (n - \mu)} \tag{9}
\]

with
\[
\mu = 0 \quad \text{and} \quad \Sigma^{-1} = \Sigma
\]
Then we can write for the logarithm of the probability density function of \( n = z - U \hat{b} \):
\[
\ln p(z - U \hat{b}) = \frac{1}{2} \ln (2\pi)^{k+1} |\Sigma| + \frac{1}{2} (z - U \hat{b})^\top \Sigma^{-1} (z - U \hat{b})
\]
Maximizing this function leads to:
\[
\frac{\partial}{\partial \hat{b}} \left\{ (z - U \hat{b})^\top \Sigma^{-1} (z - U \hat{b}) \right\} = 0 \tag{10}
\]
or
\[
U' \Sigma^{-1} U \hat{b} = U' \Sigma^{-1} z = 0 \tag{11}
\]
If \( U' \Sigma^{-1} U \) has an inverse then eq. (11) can be written:
\[
\hat{b} = (U' \Sigma^{-1} U)^{-1} U' \Sigma^{-1} z
\]
This is the expression for the Markov estimate. It has the following properties:
- linearity i.e. \( \hat{b} = 0 \) \( z \)
- unbiasedness i.e. \( \mathbb{E}[\hat{b}] = b \)
- minimum variance of all linear unbiased estimates. This variance follows from:
\[
\text{cov. } \hat{b} = \mathbb{E}[(\hat{b} - b)(\hat{b} - b)']
\]
If knowledge on the covariance matrix of the noise is also lacking \( \Sigma \) is best to choose \( \Sigma = I \), the identity matrix, assuming that the noise is white. Consequently
\[
\hat{b} = (U' \Sigma^{-1} U)^{-1} U' \Sigma z
\]
This is the expression for the least squares estimate.

The Markov and least squares estimate have been "derived" from the maximum likelihood estimate under the assumption of gaussian noise. This has only been done to indicate the type of relationship that exists between the different estimates. These estimates, however, can be derived irrespective of the type of probability distribution of the noise by minimizing respectively the conceptual errors, cf. eq. (10):
\[
E = e' \Sigma^{-1} e \tag{14}
\]
and
\[
E = e' \Sigma^{-1} e = e' e \tag{15}
\]
with
\[
e = z - U \hat{b}
\]
Up to now the discussion has been restricted to sampled signals. By increasing the number of samples and decreasing the sampling interval the corresponding expressions can be derived for continuous signals.

Some of the relations between the different types of estimates are summarized in table 1. One may note that in this table there are only two expressions which give \( \hat{b} \), the estimate of the parameter vector, explicitly. This leads to the distinction between the two classes of parameter estimation techniques that compare in a very general way as follows:

\( \text{cov. } \hat{b} = \mathbb{E}[(\hat{b} - b)(\hat{b} - b)'] \)
Bayes unconditional estimation

<table>
<thead>
<tr>
<th>Bayes estimation</th>
<th>unconditional maximum likelihood estimation</th>
<th>conditional maximum likelihood estimation</th>
<th>Markov estimation</th>
<th>least squares estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>a priori knowledge</td>
<td>p(z</td>
<td>h)</td>
<td>p(z</td>
<td>h)</td>
</tr>
<tr>
<td>conditions</td>
<td>min (f(C(b)\cdot p(h</td>
<td>z)dE))</td>
<td>max (p(z</td>
<td>h)) or (max \frac{p(z</td>
</tr>
<tr>
<td>properties</td>
<td>min. risk or cost</td>
<td>asympt. normality unbiasedness</td>
<td>linearity unbiasedness (min.variance)</td>
<td>linearity unbiasedness</td>
</tr>
<tr>
<td>relations</td>
<td></td>
<td>if gaussian noise</td>
<td>if white noise</td>
<td></td>
</tr>
</tbody>
</table>

Table 1

class I

using explicit mathematical relations or explicit methods or open loop methods or direct methods,

where the solution:
- is available after a finite number of elementary operations
- requires considerable memory
- is not available in an approximate form as an intermediate result

class II

using model-adjustment techniques or implicit methods or closed loop methods or iterative methods,

where the solution:
- is available after (in principle) an infinite number of elementary operations
- requires less memory
- is available in an approximate form as an intermediate result
- is found by a self-correcting procedure.

3. The use of explicit mathematical relations.

In the previous section two explicit expressions \(\hat{b} = \hat{b}(z)\) were given in eq. (12) and (13). In this section we will discuss some aspects of the instrumentation of these expressions.

We will start with the simplest case, the least squares estimation, following from

\[
\min E \left( \hat{b} \right) \\
\text{with eq. (15)} \\
E = e'e = \sum_{i=0}^{k} e_i^2(\tau_i) \\
\text{resulting in eq. (13)} \\
\hat{b} = (U'U)^{-1}U'z \\
\text{The instrumentation of this equation leads to correlation techniques. As an example let us assume that only one parameter } \beta \text{ has to be determined. Equation (13) then reduces to:} \\
\hat{b}_j = (u_i'u_i)^{-1}u_i'z = \sum_{i=0}^{k} u_i^2(\tau_i) z(\tau_i) \\
\text{The numerator and the denominator can be recognized as the time averaged products that approximate the respective cross and auto correlations for } k+\infty; \text{ cf. fig. 3.} \\
\text{In the majority of cases } u_i \text{ is chosen as a delayed version of the input signal } u. \text{ In that case the parameter } \hat{b}_j \text{ is an approximation of one point of the process impulse response.}
For a number of parameters this expression can be written as:

\[
H = \begin{bmatrix}
1 & u_1 & u_2 & \cdots & u_n \\
0 & 1 & u_1 & \cdots & u_{n-1} \\
0 & 0 & 1 & \cdots & u_{n-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix}^{-1}
\begin{bmatrix}
u_2 \\
u_3 \\
\vdots \\
u_n \\
u_1
\end{bmatrix}
\]

(16)

A number of remarks are in order here:
- The matrix \( H \) is symmetric around its main diagonal as \( u^T u = u u^T \).
- If \( u^T u = n \); the Kronecker delta, \( M = I \), the identity matrix. Such conditions of orthornormality can be fulfilled by a suitable choice of the input signal \( u \) and appropriate transfer functions \( G_i \), cf. fig. 4.
- Instead of operating on signals with a continuous amplitude scale one may work with quantized signals that can have only a limited number of predetermined amplitude values. The limiting case of quantization gives a binary signal.
- If one may use a test signal then \( u \) or \( u(t) \) can be chosen with full emphasis on its properties. Well known choices are the multi-frequency and the maximum-length-binary sequences. Both types of signals can be given interesting properties with respect to orthogonality and generation.
- In the derivation of the estimate no assumption was made that the process \( P \) has to be linear. The operations \( G_i \) in fig. 4 may be nonlinear. The only requirements, imposed by the condition \( \mathbf{w} = \mathbf{u} \) as a model for \( x = \mathbf{u} b \) (cf. section 1 and fig. 2), is the linearity of \( \mathbf{w} \) with respect to the parameters to be determined. Consequently classes of nonlinear systems may be amenable for correlation techniques; cf. Wiener's characterization of nonlinear systems, Volterra expansions, Pugashev's systems-reducible-to-linear.

The next degree of complication is found in the instrumentation of the Markov estimate, following from

\[
\min \mathbf{\hat{e}}
\]

with eq. (14)

\[
\mathbf{\hat{e}} = \mathbf{e} \mathbf{N}^{-1} \mathbf{e}
\]

resulting in eq. (12)

\[
\mathbf{\hat{e}} = (\mathbf{u}^T \mathbf{N}^{-1} \mathbf{u})^{-1} \mathbf{u}^T \mathbf{N}^{-1} \mathbf{e}
\]

The matrix \( \mathbf{N}^{-1} \) can be separated into a lower triangular matrix \( D \) and its transpose:

\[
\mathbf{N}^{-1} = D^T D
\]

Using this notation eq. (14) and (12) can be written as

\[
\mathbf{\hat{e}} = (D \mathbf{e})^T D \mathbf{e}
\]

\[
\mathbf{\hat{e}} = (D \mathbf{u})^T D \mathbf{u}
\]

(17)

The matrix \( D \) represents a "noise-whitening" filter; given \( u \) as an input sequence, the output of that filter is white noise. Fig. 5 indicates how these filters may be introduced into the instrumentation of the correlation technique. As was mentioned before this provides the minimum variance estimate (of all linear unbiased estimates that can be found). In spite of this interesting property and the relative simplicity of the instrumentation it seems that it has found little application.

The same remarks that have been made regarding \( \mathbf{N}^{-1} \) in the least squares case, may be made as regards the matrix \( (D \mathbf{u})^T D \mathbf{u}^{-1} \). For the non-sampled or "continuous" case the error is defined analogously to eq. (14) by

\[
E = \int_0^T \int_0^T v(\tau, \eta) e(\tau) e(\eta) \, d\tau \, d\eta
\]

where \( \mathbf{v} \) can be determined from the autocorrelation function \( \mathbf{v}_{nn}(\tau) \) of the noise.
4. The use of model-adjusting techniques.

In section 1 a (conceptual) model was introduced. In the discussion on the use of explicit mathematical relations no use was made of a physical realisation of such a model. There may be an advantage in the actual application of a model. This can be instrumented using analog or digital means.

Referring again to fig. 2 we find that the goal of parameter estimation is now formulated as: adjusting the model parameters $\theta$ in such a way that the actual error $e$ is minimal in some predefined sense.

If the representation chosen is adequate for describing the process behaviour then

$$\begin{align*}
\dot{y} &= U \theta \\
\dot{z} &= U \theta + n
\end{align*}$$

The model equation is

$$\begin{align*}
\dot{y} &= U \theta \\
\dot{z} &= U \theta + n
\end{align*}$$

and consequently the error is found to be

$$e = z - w = U(b - \theta) + n$$

Apparently a quadratic form

$$e' R e$$

will be minimized if $b = \bar{\theta}$. Here the relation of model adjustment techniques with the use of explicit mathematical techniques is clear by comparing eq. (18) with eq. (14) and (15); cf. ref. [1]. An analogous formulation of the problem holds for non-sampled or "continuous" signals.

Now the task is to find an instrumentation that, through the use of the error $e$, gives an automatic adjustment of the parameter $\theta$. Such an instrumentation can be derived along the following lines:

Define

$$\dot{\theta} = \frac{1}{2} \dot{\varphi}(e' R e)$$

where $\dot{\varphi}$ determines the gradient of the error with respect to $\theta$. Starting with this knowledge one has to use such a control policy for $\dot{\theta}$ that the error will be minimized. The adjustment can be continuous or intermittent.

- continuous adjustment scheme. A favoured policy is obtained by choosing

$$\dot{\theta} = -\gamma e = -\frac{1}{2} \varphi (e' R e)$$

Such a choice leads to the gradient method. Strictly speaking $\theta$ has to be time-invariant while the gradient is determined. This is not the case in continuous adjustment schemes. For this reason eq. (20) offers an approximate description of the adjustment dynamics only if that adjustment is comparatively slow.

- intermittent adjustment scheme. The problem of the gradient determination just mentioned does not occur if one uses an intermittent scheme: measuring the gradient while $\theta$ is kept constant, adjusting $\theta$, measuring again etc. This problem is studied in the theory of stochastic approximation [6], which may lead to an adjustment algorithm as e.g.:

$$\theta^{(i+1)} = \theta^{(i)} + \gamma(i) \varphi \{ e^{(i)} - e^{(i)} \}$$

with $\theta^{(i)}$ the parameter vector of the model after the $i$-th adjustment

$e^{(i)}$ the $i$-th sequence of error samples

$\gamma(i)$ a factor governing the speed of convergence.

A simple example may indicate the connection between model adjustment techniques and the use of explicit mathematical relations. Assume that the adjustment criterion is:

$$\min \ e' e$$

After the $i$-th model adjustment we know from eq. (13) that the optimal parameter is given by:

$$\bar{\theta} = (U'U)^{-1} U'(w + \xi) = (U'U)^{-1} U'(\bar{\theta}) + (U'U)^{-1} U'e$$

This is the best least squares estimation, given the "a priori" model adjustment and the particular series of error samples. For consecutive model adjustments one finds:

$$\theta^{(i+1)} = \theta^{(i)} + \gamma(i) \varphi \{ e^{(i)} - e^{(i)} \}$$

This is the best least squares estimation, given the "a priori" model adjustment and the particular series of error samples. For consecutive model adjustments one finds:

$$\theta^{(i+1)} = \theta^{(i)} + \gamma(i) \varphi \{ e^{(i)} - e^{(i)} \}$$

Grov. $\theta^{(i+1)}$ can also be determined. In the same way

$$\min \ e' N^{-1} e$$

can be used as the model adjustment criterion.

The crucial problem connected with equation (19) is the determination of this gradient. Several approaches are available [1] - the use of parameter influence coefficients or parameter sensitivity functions [7] - the use of two models with parameters $\theta$ and $\theta + \Delta \theta$ - the use of one model with measurements taken before and after making a step-change $\Delta \theta$ - the use of one model with (periodically) varying parameters.
- the use of a (generalized) model providing the parameter sensitivity functions simultaneously for all parameters to be estimated.


For our purpose it suffices to recall that the state of a process is defined as a variable which, at any instant of time together with the subsequent input to the process, completely determines its subsequent behaviour [1].

In quite a number of applications this notion of state plays an important part; this is particularly true for optimal control problems. As the state variables need not be measurable directly and as the quantities that are measurable will be subject to noise the process state can only be estimated [8]. This estimation problem is adequately solved for linear systems for which the dynamic properties are known completely (including the numerical values of the process parameters) [9].

The problem becomes more complicated if the parameter values of the process are not known beforehand but have to be estimated together with the state variables. This type of problem can be formulated in the following way [10].

We assume that the process behaviour is described by

\[ \dot{x}(t) = f(x, u, b) \]  

(21)

with:

\[ x'(t) = \{x_1, \ldots, x_n\} \]  

the state of the process

\[ u'(t) = \{u_1, \ldots, u_l\} \]  

the control vector

\[ b'(t) = \{b_1, \ldots, b_m\} \]  

the parameter vector

The input \( u \) to the process can be observed as well as the output \( y \), where \( y \) may be a subset of \( x \). The observations on \( y \) are given by

\[ y(t_j) = c_j \]  

(22)

Although \( b \) is considered to be constant during the observation interval it will be treated as a function of time by the equation

\[ b = 0 \]  

(23)

Expressions (21) and (23) provide \((n+m)\) equations that have to satisfy a sufficient number of boundary conditions given by (22). The computational solution using the method of quasi linearization [11] uses as the \((i+1)\)-th approximation of the time function \( x(t) \) over a certain interval of time:

\[ \dot{x}^{(i+1)} = f(x^{(i)}, u, b^{(i)}) + F_1^{(i)}(x^{(i)}) \Delta x^{(i)} + F_2^{(i)}(x^{(i)}, u^{(i)}) \]  

(24)

with

\[ x^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\} \]

\[ y^{(i)} = \{y_1^{(i)}, \ldots, y_m^{(i)}\} \]

\[ F_1^{(i)} = \{F_1^{(i)}(x^{(i)})\} \]

\[ F_2^{(i)} = \{F_2^{(i)}(x^{(i)}, u^{(i)})\} \]

\[ u = \{u_1, \ldots, u_l\} \]

\[ b = \{b_1, \ldots, b_m\} \]

\[ \Delta x^{(i)} = x^{(i+1)} - x^{(i)} \]

\[ \Delta b^{(i)} = b^{(i+1)} - b^{(i)} \]

\[ c_j = \{c_1, \ldots, c_m\} \]

This procedure estimates are found for both the parameters \( b \) and those of the states \( x \) that are not measurable.

6. Some examples.

In order to illustrate the ideas presented in the previous sections a few simple examples will be discussed.

The explicit mathematical relations discussed in section 3 result in the celebrated correlation techniques. If one wants to estimate a point \( h(t) \) of the impulse response of a "process" \( P \), then the instrumentation shown in diagram 6 can be used. If one wants to estimate a number of such points, e.g. \( h(t_1), \ldots, h(t_n) \) the same measurement is done a number of times, each time taking another value \( t = t_i \). Apart from the change of \( t \) all measurements are made in an identical way. This is remarkable since, while performing the \( j \)-th measurement, the estimates of \( h(t_1), \ldots, h(t_{j-1}) \) can be considered to be a priori knowledge on the process.

In the following we would like to stress that if this a priori knowledge is not used measurement time is being wasted. After this statement the most important question is how the a priori knowledge from the \((j-1)\)measurements can be incorporated into the \( j \)-th measurement. This can easily be done by imbedding the correlation technique in the model-adjustment procedure.

Before studying this in some detail we would like to recall that the correlation technique provides estimation of the unknown parameters in the sense of a mean squared error, cf. section 3. For continuous signals this can also easily be shown. Cf. fig. 6 and assume the process to be linear.

\[ \text{fig. 6} \]
One wants to estimate a point $h(\tau)$ of the impulse response of $P$. As a result of the linearity:

$$z(t) = y(t) + n(t) = \int_0^\infty h(\theta) u(t-\theta) d\theta + n(t).$$

**Correlation technique.** Multiplying both sides with $u(t-\tau)$ and taking the mathematical expectation we find:

$$E[u(t-\tau)z(t)] = \int_0^\infty h(\theta) E[u(t-\tau)u(t-\theta)] d\theta + E[u(t-\tau)n(t)].$$

or

$$\varphi_{uz}(\tau) = \int_0^\infty h(\theta) \varphi_{uu}(\theta) d\theta + 0$$

if $n(t)$ and $u(t)$ are independent. If in addition $u(t)$ can be assumed to be white on account of its large bandwidth compared to the bandwidth of the process then

$$\varphi_{uu}(\tau-\delta) = \varphi_{uu}(0) \delta(\tau-\delta)$$

and

$$\varphi_{uz}(\tau) = h(\tau) \varphi_{uu}(0)$$

(25).

**Minimization of the average squared error.** This requires a minimization of:

$$E = \lim_{T \to \infty} \frac{1}{T} \int_0^T e(t)^2 dt$$

with

$$e(t) = z(t) - u(t-\tau)$$

The necessary condition $\partial E/\partial \theta$ leads to:

$$\frac{\partial E}{\partial \theta} = -\lim_{T \to \infty} \frac{1}{T} \int_0^T u(t-\tau) dt = 0$$

or

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T z(t)u(t-\tau) dt = \lim_{T \to \infty} \frac{1}{T} \int_0^T u(t-\tau) dt = 0$$

Under the assumption of ergodicity this can be written as

$$\varphi_{uz}(\tau) = \beta \varphi_{uu}(0)$$

(26).

The second case can be instrumented using a model adjustment technique. A comparison of equations (25) and (26) indicates that the model-parameters in fig. 7 can be adjusted according to the result of the correlation measurement.

With respect to the instrumentation one can distinguish the following cases:

a) estimation of the parameters one by one using correlation techniques,

b) model adjustment of the parameters one by one in a sequential way.

c) model adjustment of all parameters simultaneously.

The requirements as regards instrumentation increase in this order.

ad a) estimation of the parameter one by one using correlation techniques.

The system to be discussed first is represented in fig. 6. The essential part is shown in fig. 8, where $p$ and $q$ are ergodic (stationary) stochastic signals. The switch $S$ is closed at $t=0$, the initial condition $r(0) = 0$.

![fig. 7](image)

![fig. 8](image)
The contribution due to $R_2$ is the uncertainty due to the additive noise.

ad b) estimation of the parameters one by one in a sequential way.

Up to now only the determination of one parameter (viz. one point of the impulse response) has been discussed. A new element enters the discussion if a number of parameters have to be determined. After one quantity has been estimated this estimate can be considered as being a priori knowledge for the next estimation cycle, etc. Intuitively it is clear that the use of a priori knowledge may result in a more efficient estimation procedure. A crucial question, however, is in what way such knowledge can be incorporated into the instrumentation. This can be done by using a model of the process that is well adapted to the parameter-description of the process, e.g. if the attention is focussed on the impulse response then a time-delay circuit with taps and potentiometers is adequate.

Fig. 9 gives an illustration of this situation. In order to obtain some insight into the effect of the model the relations (27) through (30) will be applied to an arbitrary simple example, viz. a process with impulse response:

$$h(t) = \frac{6}{\tau} e^{-6\tau}$$

for $\tau > 0$

$$h(t) = 0$$

for $\tau < 0$

We will consider the determination of one point of this function, viz. for $\tau = 0,225$ (fig. 10a) using a white noise input signal.

If the switch $S$ in fig. 9 is open there is no difference with the correlation diagram of fig. 6. In that case we find for the standard deviation

$$\sigma_x(t)$$

with respect to the expected value

$$\mu_x = \psi_{x0}(t)$$

the relation given in fig. 11, curve a. From this diagram the correlation interval needed for a certain accuracy (standard deviation) of the estimated parameter can be found.

Now we will assume that in fig. 9 the parameters $\beta_1$ through $\beta_6$ have already been estimated. The determined values are used for adjusting those potentiometers, the other potentiometers are set equal to zero and switch $S$ is closed.

The adjusted part of the model, with the hold-circuit included, has the impulse response $m(t)$ shown in fig. 10b. Consequently the error signal $e$ is determined by the difference between $h(t)$ and $m(t)$ as shown in fig. 10c; yet in this impulse response $h(t) - m(t)$ the parameter to be estimated still has its original value.

Now turning our attention to equation (30) we notice that in this case the index $y$ has to be replaced by the index $e$. This implies that the corresponding terms are smaller than the original ones, which results in a smaller variance or a reduction of the correlation interval that is needed for obtaining the same variance.
For the example given we find the reduction of the correlation interval given by fig. 11, curve b; now the same standard deviation is obtained in about one tenth of the original time interval.

Fig. 12a and b give a number of actual recordings of the integrator output r(t) for the cases indicated above; S open resp. closed. The influence of additive noise has been neglected; it can be taken into account using the equation (31).

ad c) model adjustment of all parameters simultaneously.

The choice of the model is now of crucial importance.

In the previous paragraphs the timedelay model for the (linear) process and the error were chosen respectively as (cf. fig. 13a):

\[
\begin{align*}
    e(t) &= z(t) - \sum_{i=1}^{n} a_i u_i(t - \tau_i) \\
    e(t) &= z(t) - \sum_{i=1}^{n} u_i(t - \tau_i) 
\end{align*}
\]  
\[ (32) \]

A more general case of a linear or nonlinear model and the corresponding error may be chosen as (cf. fig. 13b):

\[
\begin{align*}
    e(t) &= z(t) - \sum_{i=1}^{n} a_i u_i(t) \\
    e(t) &= z(t) - \sum_{i=1}^{n} a_i u_i(t) 
\end{align*}
\]  
\[ (33) \]

where \( u_i(t) \) represents the outputs of the linear or nonlinear filters operating on the process input \( u(t) \), e.g. a set of orthogonal filters.

A generalized model operates on both the process input \( u \) and the process output \( z \) to build an error given by (cf. fig. 13c)

\[
\begin{align*}
    e(t) &= z(t) - \sum_{i=1}^{n} a_i u_i(t) - \sum_{i=1}^{n} \beta_i v_i(t) 
\end{align*}
\]  
\[ (34) \]

If the differential or the difference equations of the process are known it may seem to be appropriate to instrument the model using the same equation.
In that case the error can be chosen as:
\[ e(t) = z(t) - \mathcal{G}[u(t), \Gamma] \]  (35)
where \( \mathcal{G}[\cdot] \) is the dynamic operator provided by the model and \( \Gamma \) is the parameter vector with components \( \alpha_i \) and \( \beta_j \).

In the first three cases there is a linear relation between the error and the parameters to be determined. This does not hold in the last case which implies a number of special problems in the instrumentation of the model adjustment (stability, additional hardware).

These models offer different possibilities of using a priori knowledge:

<table>
<thead>
<tr>
<th>type of model</th>
<th>used or usable a priori knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>using timedelay elements</td>
<td>- process is (approx.) linear</td>
</tr>
<tr>
<td>using orthogonal filters</td>
<td>- approx. length of the process impulse response</td>
</tr>
<tr>
<td>using a topologically identical model</td>
<td>- order and form of the process differential equation</td>
</tr>
<tr>
<td>using a generalized model</td>
<td>- order of magnitude of the coefficients to be determined</td>
</tr>
<tr>
<td></td>
<td>- values of the already known coefficients</td>
</tr>
</tbody>
</table>

Analogous considerations hold for nonlinear models.

The model-adjustment criterion may be chosen as:
\[ \min \langle e, e \rangle \]
where \( \langle \cdot, \cdot \rangle \) may stand for \( \langle e(t), e(t) \rangle \). Consequently a number of error criteria can be represented by these bracket symbols, e.g.
\[ c[i \int e^2 dt + \int e|dt| + \int \mathcal{E}^2(e) k(t-e)da] \]
If the generalized model and the error equation (34) are used together with the mean square error criterion, the gradient with respect to the unknown parameters \( \Theta = 1 \sum_{i} \langle e, e \rangle \)
has the components
\[ \epsilon_i \text{ def } \frac{3}{3} \langle e, e \rangle = \langle e, u_i \rangle \]
\[ \eta_j \text{ def } \frac{3}{3} \langle e, e \rangle = \langle e, v_j \rangle \]  (36)
This provides us with the instrumentation that is needed to find the gradient. This can be used, e.g., in instrumenting a stochastic-approximation algorithm.

The dynamics of the gradient measurement can be studied using the matrix equation:
\[ \Theta = \Theta - M \Gamma \]  (37)

with
\[ M = \begin{pmatrix}
\langle e, u_1 \rangle \\
\vdots \\
\langle e, v_n \rangle 
\end{pmatrix} \]
\[ \Gamma = \begin{pmatrix}
\alpha_1 \\
\vdots \\
\beta_n 
\end{pmatrix} \]

Strictly speaking this implies the assumption that the parameter vector \( \Gamma \) is constant as otherwise \( \langle e, u, v, \rangle \neq 0 \), etc. On the other hand it will be clear that this equation holds not only for the generalized model but also for the other situations where the error is linear in the parameters.

We will make a sketchy comparison of the use of different types of models. A simple linear process will be taken as an example:
\[ b_2 y + b_1 y + y = a_0 u + a_0 u \]
with unknown parameters \( b_2, b_1, a_1, a_0 \).

If a timedelay model (with a wideband stationary stochastic input signal \( u(t) \)) or a set of orthogonal filters is used the instrumentation for finding the gradient is:
\[ \int \langle e, e \rangle = \int \langle e, u_i \rangle \]
The matrix \( M \) is a diagonal matrix with elements \( \langle u_i, u_j \rangle \neq 0 \). From the definition it follows that \( \langle \cdot, \cdot \rangle \) denotes the expectation; in practice only one realization of the ensemble is available. This implies that in the actual instrumentation the off-diagonal elements of \( M \) are not identically zero and some interaction between the adjustment of parameters will occur. Summarizing the properties of the timedelay model or the orthogonal model adjustment we find:
- there is no easy way of introducing the a priori knowledge provided by the differential equation.
- there is a little interaction between parameter adjustments due to the off-diagonal terms of \( M \) in spite of the zero expectation.
- extra noise causes some extra variance but no bias.
If a generalized model is used the gradient instrumentation is given by eq. (36) and the behavior is described by eq. (37). In general the expectations of the off-diagonal terms of $M$ are not equal to zero and this leads to interaction. Summarizing some considerations we find:
- the a priori knowledge is well used; one obtains a direct estimation of the coefficients of the differential equation.
- there is an interaction between parameter adjustments due to the off-diagonal terms of $M$, which may have an expectation $\neq 0$.
- additive noise causes an extra variance and bias.

If a topologically identical model is used
\[
\begin{align*}
z &= y + n \\
y &= a_1 \hat{u} + a_0 u - b_2 \hat{y} - b_1 y \quad \text{process} \\
w &= a_1 \hat{u} + a_0 u - b_2 \hat{w} - b_1 \hat{w} \quad \text{model}
\end{align*}
\]

As $w = w(a_1, a_0, b_2, b_1)$ the error $e = z - w$ is not linearly related to the parameters to be determined. This implies that the determination of the gradient with respect to the unknown parameters does not lead to the simple expression given before; additional models are needed [7]. Partial differentiation with respect to the unknown parameters yields:
\[
\begin{align*}
\frac{\partial w}{\partial a_1} &= \hat{u} \\
\frac{\partial w}{\partial a_0} &= u \\
\frac{\partial w}{\partial b_2} &= b_2 \hat{y} + b_1 \hat{w} + v_2 = -\hat{w} \quad \text{with} \quad v_2 = \frac{\partial w}{\partial b_2} \\
\frac{\partial w}{\partial b_1} &= b_2 \hat{u} + b_1 \hat{w} + v_1 = -\hat{w} \quad \text{with} \quad v_1 = \frac{\partial w}{\partial b_1}
\end{align*}
\]

and $v_2 = \hat{y}$. This results in a block diagram using two models.

Further considerations lead to the following aspects:
- the a priori knowledge is well used.
- as a result of the nonlinear relation between the parameter and error there is an interaction between parameter adjustments which is stronger than in the case of the generalized model. The instrumentation for determining the gradient (parameter influence coefficients) is also more complicated.
- additive noise causes an extra variance but no bias.

Using these considerations as a basis we chose the generalized model for a further discussion on the simple example indicated above.
\[
b_2 \hat{z} + b_1 \hat{x} + z - a_1 \hat{u} - a_0 u = 0 \quad (38)
\]
provided $n(t) \equiv 0$. The use of a generalized model gives:
\[
b_2 \hat{z} + b_1 \hat{x} + z - a_1 \hat{u} - a_0 u = e(t) \quad (39)
\]
and by subtraction of these equations:
\[
(b_2 - b_2) \hat{z} + (b_1 - b_1) \hat{x} - (a_1 - a_2) \hat{u} - (a_0 - a_0) u = e
\]

In engineering terms this type of model is quite impractical on account of the use that is made of differentiators (accentuation of noise; stability). There is, however, no objection to applying the same linear dynamic operator $G([\cdot])$ to each term of equation (40). This results in
\[
(b_2 - b_2) v_2 + (b_1 - b_1) v_1 - (a_1 - a_2) u_1 - (a_0 - a_0) u_0 = e
\]

By a suitable choice of $G$, in this case a second order filter, the necessary derivative signals ($v_2 = \hat{y}$, $u_1 = \hat{u}$) can be generated. Moreover the operator can be chosen taking into account the statistical properties of the noise $n(t)$.

The gradient with respect to the unknown parameters is found as:
\[
\begin{align*}
\frac{\partial e}{\partial e_1} &= \langle v_1, v_2 \rangle \\
\frac{\partial e}{\partial e_2} &= \langle v_2, v_2 \rangle \\
\frac{\partial e}{\partial e_1} &= \langle v_1, v_2 \rangle \\
\frac{\partial e}{\partial e_2} &= \langle v_1, v_2 \rangle \\
\end{align*}
\]

Due to the orthogonality of ($v_2 = \hat{y}$ and $v_1$) and ($u_1 = \hat{u}$ and $u_0$) four elements of $M$ are zero. If a steepest descent instrumentation is chosen:
\[
\Gamma = \text{constant}
\]

Then for a slow model adjustment the following equation approximately holds
\[
\dot{\Gamma} = -2k \Gamma \langle e_1, e_1 \rangle
\]

where $\Gamma_0$ is the parameter vector at $t = 0$. According to the theory of matrix exponentials $\Gamma = 0$ for $t = \infty$ if all the characteristic roots of $M$ have real positive parts. As $M$ is a real symmetric matrix the characteristic roots are real. Fig. 14 gives an indication of the "parameter tracking" capabilities of such a continuously adjusting model.

The restriction $\Gamma = \text{constant}$ is fulfilled if the adjustment policy is an intermittent one: measuring the gradient with $\Gamma$ constant, adjusting $\Gamma$, measuring again etc. In terms of expectation one may indicate the convergence properties as follows. Put
\[
\Delta = \Gamma (i+1) - \Gamma (i) = -k \Gamma (i)
\]
or
\[
\Gamma (i+1) = (\Gamma - k M) \Gamma (i)
\]

In terms of actual observations, i.e. one element of the ensemble, the theory of stochastic approximation is directly applicable.
fig. 14

\[ y + 0.5y + 0.10 + 0.5u \]

Model:

- Change in parameter \( b \)
- Change in parameter \( a \)
- Adjustment of the four parameters simultaneously

---

Graphs showing time series data with various parameters: \( u, y, a, \alpha_1, \alpha_2, \beta_1, \beta_2 \) over a period of 10 seconds.
7. Concluding Remarks.

There is little need to stress the point that this paper is incomplete in many respects. The list of additional topics that ought to be discussed includes the following:
- the types of description of process dynamics for linear, vary-linear and nonlinear cases
- the description of signal properties; the construction of (test) signals with good estimation properties
- convergence rates for different estimation procedures
- stability properties for model-adjustment techniques
- the approximation of a process by a model of a lower order or reduced complexity
- the relation between financial aspects of the instrumentation and the optimum properties of the estimation schemes
- the connection between the problems discussed and related theories and techniques in mathematics and engineering
- the application of these theories to problems in process industry, power generation, aerospace vehicles, automation of measurement and decisions, biology and medicine and other fields outside the realm of engineering.

In fact this list of topics can also be used as an inventory of problems that still need to be tackled in order to convert the art of process parameter estimation into a science. On each of the topics cited our knowledge is partial and in many cases not too well adapted to the estimation problem.

In other fields of engineering science there are theorems that indicate the ultimate limits of action and observation that can be reached (e.g. thermodynamics; communication/information theory; uncertainty principle). Limiting theorems of that kind are much needed in the realm of parameter and state estimation, answering such questions as: "what is the amount of knowledge that can be derived during this time interval for that particular situation?" Little work has been done along this line, some work is under way, but probably much more effort is needed.

Beside the knowledge of the ultimate limits (the theoretical optimum) one also has to have an insight into the "economic" aspects of a particular situation as e.g. the hardware and software needed.

A more comprehensive knowledge would even include the relation between the "economic" and the "theoretic optimality" of a range of solutions for a particular estimation problem. As an example we can think of the different estimation methods indicated in table 1 applied to the same problem. Starting from the simple least squares method one may ask whether the increasing complexity of other approaches is worthwhile in a certain situation.

These problems of a general nature obtain their importance in the engineering sense through the practical applications. Each application, however, has its own specific salient aspects that warrant many case studies over a wide spectrum: "diagnostic" measurements, control engineering, communication engineering, automatic (industrial) measurements, automatic decisions, automatic adjustments, to mention just a few.

References

The literature on parameter and state estimation is quite extensive. The reader may consult [1] for a partial list of approximately 60 references to publications on estimation.


Acknowledgement.

The author gratefully acknowledges fruitful discussions with and contributions from his co-workers, particularly Mr. A.A. van Rede, and (former) students, a.o. Mr. J.P.M.Driessen. For experimental work he is indebted to Messrs. R.E. Langers and E. Sies.