# BAYESIAN APPROACH TO SYSTEM IDENTIFICATION

**V. Peterka**

## Contents

# 1 Introduction

In this chapter the identification problems are approached via *Bayesian statistics*. In Bayesian view the concept of probability is not interpreted in terms of limits of relative frequencies but more generally as a *subjective measure of belief* of a rationally and consistently reasoning person (here called the statistician) which is used to describe quantitatively the uncertain relationship between the statistician and the external world. Originally, the concept of subjective probability was not introduced in the anticipation of radical changes in statistical practice. Saying with [30] "the idea was, rather, that subjective probability would lead to a better justification of statistics as it was then taught and practised, without having any urgent practical consequences. However, it has since become more and more clear that the concept of subjective probability is capable of suggesting and unifying important advances in statistical practice". It is one of the objectives of this chapter to show that the latter applies also to systems identification.

One has to agree with [2] that the field of systems identification, as it has naturally developed, "appears to look more like a bag of tricks than a unified subject" and that "it seems to be highly desirable to achieve some unification". This is a natural reflection of a similar situation in the field of data analysis which was characterized by [19] as "a field in which bright ideas of a few clever men abound, but these ideas are, because of the informality of the subject, difficult, if not impossible, to convey to the average statistical practicioner". It is another objective of this chapter to show that a systematic application of the Bayesian approach is capable to make from systems identification a consistent theory with formal structure. Once this status is reached one can be quite sure what it is we are talking about and the solutions of particular identification problems can be obtained by deduction without any need of developing special methods.

The chapter is organized in the following manner. In Section 2 the main distinguishing features of the Bayesian position are briefly recalled and compared to classical frequency interpretation of probability commonly accepted in the present day statistics. For a more detailed discussion of Bayesian standpoint the reader is referred to [19]. Fuller statement and justification can be found in [10], [18], [29], [6] and [7]. An interesting discussion, including also the opposite opinions, is registered in [31].

In Section 2 also two basic operations on uncertainties are introduced. The understanding of these two basic operations is actually all what is required to be able to solve, at least conceptually, a rather wide spectrum of identification problems in a unique and consistent way.

In Section 3 the notion of a system model (or process model) is revised from Bayesian viewpoint. The characteristic feature of the Bayesian position is that the final purpose of statistical inference is to provide a rational basis for some kind of decision. This final goal has to be kept in mind right from the formulation of any statistical problem. The purpose of system identification, consider throughout the chapter, is its potential use for prediction and digital control of an uncertain process. As a matter of fact, control is nothing else than sequential decision making and the ability of prediction is the most important prerequisite for a rational control. Therefore, the problem of suitable process model is posed as the question: What is required to be able to predict and control an uncertain process?

Once the model structure is chosen or given, the problem of system identification is reduced to the problem of parameter estimation which is the main topic of Section 4. In Bayesian view the "estimate" is the probability distribution conditional on the given data and any point estimate is nothing else than some (more or less suitable) partial description of this distribution. In Baye4sian statistics the unknown parameters are actually not "estimated" but the aposterior probability distribution for them is *calculated*. Therefore, the problems like "biasedness","efficiency", "confidence interval", etc. disappear or are irrelevant. Both one-shot and real-time parameter estimation are considered in Section 4. The Bayesian approach is especially fruitful when the parameter estimation is a part (a sub-problem) of adaptive control and is performed in closed control loop.

The problem of time-varying parameters is addressed in Section 5 where also a general Bayesian view on adaptivity is given. The discussion includes the Kalman filtering, possibly performed in a closed control loop, as a special case. In practical applications the case of "slowly varying" parameters is often handled using the technique sometimes called exponential "*forgetting*", or "age weighting", or "discounting". The Bayesian interpretation of this technique is presented and its possible extension is outlined.

In many practical cases the in internal mechanism or physics of the system is not understood enough

to be able to specify the model structure uniquely. Then, the following question arises: Which one of the possible model structures has to be preferred when a finite set of input-output data is available? As a matter of fact, in most practical situations this question should be answered as one of the first steps towards system identification. Here, for didactic reasons, it is left to the last Section 6, where it is answered again in terms of the aposterior probability distribution on the set of hypotheses. A special case of this kind is the uncertain order of a linear model.

As it has been mentioned above, *the role of Bayesian statistics is to provide a rational basis for some kind of decision.* Similarly, system identification is only a part of a more complex problem for instance of control problem. Being limited by the scope of the monograph only to system identification the exposition of the Bayesian approach inevitably must be able to apply the presented results in a proper way and to complete the story according to his particular need if he understands the basis. Therefore, the emphasis is given to the conceptual side of the exposition.

In the following sections the general Bayesian solutions of the identification problems outlined above are accompanied by two kinds of examples: simple and practical ones. Th simple examples have to help the reader to understand the principles on which the solution is based. In more complicated practical examples the emphasis is given to explanation of how the given particular results can be obtained rather than to technical details of their derivations which are often left to the reader as an exercise.

Not to promise too much, it should be said in advance that it is often not easy to apply the conceptual case. Nevertheless, even when the exact Bayesian solution is practically not feasible, it clearly shows the essence feasible, it clearly shows the essence of the problem and helps to construct reasonable approximations.

## 2 Underlying Philosophy and Basic Relations

In Bayesian view *random* means *uncertain.* Any quantity the true value of which is not known to the statistician, is a random variable. Thus, not only time-varying quantities, like input-output data, but also unknown or uncertain constants, like model parameters, are random. Similarly, a hypothesis about which the statistician, on the level of his knowledge, is not able to decide whether it is true or not, is a random event.

A random variable can take on only one true value. If this true value is not known to the statistician, he has to take into account the whole set of values which the random variable could possibly take. Dealing with such a situation one has to distinguish a general possible value, say $x$, of a random variable from its true but unknown value which will be denoted by $\underline{x}$. The set of all possible values $x$ will be denoted by $\mathcal{S}_x$.If $\mathcal{S}_x$ is an interval on a real axis, or more generally a connected space of vector valued quantities, the random variable is said to be of continuous type. If $\mathcal{S}_x$ countable set of discrete real numbers, or vectors say $\mathcal{S}_x = \{x_1, x_2, x_3, \ldots\}$, the random variable is of discrete type and $x$ a general representant for any $x_i$.

In the sense of higher credence, the statistician may prefer a particular possible value to another possible value when, according to his knowledge or experience, the former is "more likely" than the latter. To describe his system of preferences numerically the Bayesian statistician uses the notion of subjective probability which can be introduced as one unit (i.e. 100%) of his belief distributed over the set $\mathcal{S}_x$ of the values which he considers as possible. In the case of discrete random variable the probability assigned to the event $\underline{x} = x_i$ is

$$\Pr[\underline{x} = x_i] = P(x_i) \tag{1}$$

and $P(x)$ means a function (real and nonnegative) defined on the set $\mathcal{S}_x$.

From the interpretation of subjective probability as distributed probability mass directly follows its additivity property [1],

$$\Pr[\underline{x} = x_i \text{ or } \underline{x} = x_j] = P(x_i) + P(x_j),\ i \neq j$$

[1] For a thorough discussion of the question whether such a description of the system of statistician's preferences is relevant at all the reader is referred to [10].

As on the set $\mathcal{S}_x$ the total statistician's belief (the total unit of the probability mass) is distributed, the following relation must hold.

$$\sum_{x \in \mathcal{S}_x} P(x) = 1 \tag{2}$$

The same notation can be used if $\mathcal{S}_x$ is a set of elementary (mutually exclusive) events of non-numeric character. For instance, if $x$ means a side of a tossed coin, then $\mathcal{S}_x = \{$"head","tail"$\}$ and according to (2)

$$P(\text{"head"}) + P(\text{ "}tail\text{"}) = 1$$

In the case when $\underline{x}$ is a random variable of continuous type, the set $\mathcal{S}_x$ contains uncountably many elements and the probability $\Pr[\underline{x} = x] = P(x)$ is zero in general, even when the event $\underline{x} = x$ is not impossible. In that case it is more suitable to describe the probability distribution by a probability density function $p(x)$ defined by the relation

$$\Pr[x \in \Omega_x] = \int_{x \in \Omega_x} p(x)\, dx \tag{3}$$

where $\Omega_x$ is any subset of $\mathcal{S}_x$, $\Omega_x \subset S_x$. [2] Apparently, the probability density $p(x)$ must fulfill the relation

$$\int_{\mathcal{S}_x} p(x)\, dx = 1 \tag{4}$$

If, for given two subsets $\Omega_1 \subset \mathcal{S}_x$ and $\Omega_2 \subset \mathcal{S}_x$, it holds

$$\int_{x \in \Omega_1} p(x)\, dx > \int_{x \in \Omega_2} p(x)\, dx$$

then it means that the statistician may expect that the true value of the random variable $\underline{x}$ will lie (or lies but it is not known to him) rather in the subset $\Omega_1$ than in the subset $\Omega_2$. For the moment let us leave aside the question how such a probability distribution can be obtained. We shall come to it later on. Notice that $P(\cdot)$ as well as $p(\cdot)$ do not have any meaning if it is not given what random variable they concern. For instance, $p(x) = f(x)$ is a function in general different from $p(y) = g(y)$ and $p(2)$ itself does not say whether $f(2)$ or $g(2)$ is meant. If we leave the arguments to identify the probability distribution we also may use, for the sake of simplicity and generality, the same notation $p(\cdot)$ both for the probability densities and probabilities $P(\cdot)$ letting the arguments indicate also which one of these two possibilities is meant. In this way a significant simplification and unification of all formulas, we shall make use of, can be achieved. One only has to keep in mind that the integration has to be replaced by regular summation whenever the argument is discrete [3]

If we have a reason to consider all or some of the components of a multi-dimensional random variable separately we speak about joint (or simultaneous)probability distribution of two or more random variables. For instance, if $\underline{x} = (\underline{a}, \underline{b})$ and $\mathcal{S}_x$ is the Cartesian product $\mathcal{S}_x = \mathcal{S}_a \times \mathcal{S}_b$, i.e. a set of ordered pairs $(a, b)$ where $a \in \mathcal{S}_a$ and $b \in \mathcal{S}_b$, then $p(x) = p(a, b)$ is the joint probability distribution for two random variables $\underline{a}$ and $\underline{b}$. For illustration let us consider the case when $a$ is continuous defined on the interval $\mathcal{S}_a = (a_1, a_2)$ while $b$ is discrete with $\mathcal{S}_b = (b_1, b_2, b_3)$. Then $p(a, b)$ is a set of three functions $\{p(a, b_i) = f_i(a),\ i = 1, 2, 3\}$, sketched in Fig. 1 which fulfill the relation

$$\sum_{i=1}^{3} \int_{a_1}^{a_2} f_i(a)\, da = 1$$

[2] In the integral (3) $dx$ means an elementary subset of $\mathcal{S}_x$ or, more precisely, $dx = \mu(dx)$, where $\mu(\cdot)$ is a measure defined on $\mathcal{S}_x$

[3] A mathematically educated reader may employ the measure theory and operate in a uniform way with probability densities generalized in Radon-Nikodym sense.The practical effect is the same and therefore it seems to us neither necessary nor very helpful, at least for our purposes.

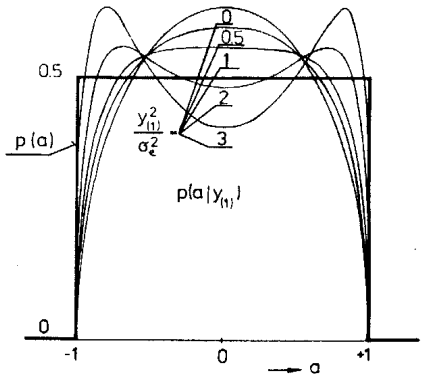



Fig. 1 Joint probability distribution $p(a, b)$ of mixed type: $a$ continuous, $b$ discrete

The concept of subjective probability distribution introduced above would be little practical value, at least in engineering and natural sciences, if it were not given how experimental evidence can be incorporated into it. In Bayesian view the statistical inference can be understood as correction of prior subjective probability distribution by objective data. Put in other words, to provide a rational basis for a decision means to provide the probability distribution conditional on data. This is the task of Bayesian statistics. In performing this task Bayesian statistics rests on the fact that *uncertainty has probability structure*. The meaning of this statement is that the mathematical discipline called probability theory, in which the notion of probability is defined axiomatically without any relation to reality, can be employed to operate with subjective probability distributions. This can be proved on the basis of a few simple and sound principles which nobody of us would wish to violate when acting in the face of uncertainty. As an example the "sure-thing principle" [29]can be given. It says that"if $A$ is preferred to $B$ when $C$ does not obtain, then $A$ is preferred to $B$ when $C$ obtains and also when $C$ does not obtain, then $A$ is preferred to $B$ when one is uncertain about $C$" [19].

It is out of the scope of this presentation and also out of the author's competence to go deeper into these philosophical and logical fundaments. Following our practical objectives we feel it to be more appropriate if we give a pertinent Bayesian interpretation of two basic operations on probability distributions. In fact, the Bayesian solutions of identification problems, we shall deal with in the following sections, are nothing else than systematic applications of these two basic operations and therefore their deeper-rooted understanding is vital.

## 2.1 Two Basic Operations on Uncertainties

The first operation, we need to be able to solve our problems, can be stated as follows. Given the joint probability distribution of two random variables $\underline{a}$ and $\underline{b}$ determine the probability distribution for $\underline{b}$ without taking into account what value the random variable $a$ may take. Expressed mathematically, given $p(a, b)$, $a \in \mathcal{S}_a$, $b \in \mathcal{S}_b$, determine $p(b)$ defined as

$$Pr[\underline{b} \in \Omega_b] = \int_{\Omega_b} p(b)\, db \tag{5}$$

for any $\Omega_b$. The answer to this question directly follows from the additivity property of subjective probability. As $a \in \mathcal{S}_a$ is a certain event it holds

$$Pr[\underline{b} \in \Omega_b] = Pr[\underline{b} \in \Omega_b \; and \; \underline{a} \in S_a] = \tag{6}$$
$$\int_{\Omega_b} \int_{\mathcal{S}_a} p(a, b)\, da\, db$$

and from comparison of (5) and (6) it follows

$$p(b) = \int_{\mathcal{S}_a} p(a,b)\, da \tag{7}$$

According to the convention accepted, the integration in (7) has to be replaced by regular summation if a is discrete. The probability distribution $p(b)$, when related to $p(a,b)$, is sometimes called marginal. If the range if integration in (7) is not given the entire set $\mathcal{S}_a$ will be meant.

The second basic operation on uncertainties cannot be derived from the concept we have already defined, but must be introduced exogenously on the basis of sound reasoning. Consider the situation when the statistician is uncertain about two quantities, $\underline{a}$ and $\underline{b}$, and somehow has determined his subjective probability distribution $p(a,b)$. Now, he obtains the information that the true value of the random variable $\underline{b}$ is $\beta$, $\underline{b} = \beta$. The uncertainty of the quantity $\underline{b}$ disappeared but the uncertainty of the quantity $\underline{a}$ remains. How has the statistician to recalculate his probability distribution to match this new situation? The problem is: given $p(a,b)$ determine the conditional distribution $p(a|b=\beta)$. Apparently, the distribution $p(a,b)$ for $b \neq \beta$ becomes irrelevant, but the statistician has no reason to change his system of preferences in the direction of $a$ for $b = \beta$. Therefore, it is natural to determine $p(a|b=\beta)$ as proportional to $p(a,b)$ for $b = \beta$

$$p(a|\underline{b} = \beta) = \kappa p(a,b)|_{b=\beta} \tag{8}$$

where $\kappa$ is the coefficient of proportionality. Obviously, for all $a$ where $p(a,b)|_{b=\beta} = 0$ also $p(a|b=\beta) = 0$. The coefficient $\kappa$ can be determined from the condition

$$\int p(a|\underline{b} = \beta)\, da = 1$$

Hence

$$\kappa = \frac{1}{\int p(a,b)|_{b=\beta}\, da}$$

and using (7) we have

$$\kappa = \frac{1}{p(b)|_{b=\beta}} \tag{9}$$

As we are interested in general relation for any $\beta \in \mathcal{S}_b$, i.e. in $p(a|\underline{b} = \beta)$ as a function of $\beta$, it does not have much sense to distinguish in notation the variables $\beta$ and $b$. We also may write instead of $p(a|\underline{b} = b)$ more simply $p(a|b)$. With this change in notation the relation described by (8) and (9) can be written as follows.

$$p(a|b) = \frac{p(a,b)}{p(b)} \tag{10}$$

Rewritten as

$$p(a,b) = p(a|b)p(b) \tag{11}$$

the relation can be understood as a rule how to construct joint probability distribution when conditional $p(a|b)$ and marginal $p(b)$ distributions are given. For illustration consider again the simple example pictured in Fig.1, where $\underline{a}$ is a continuous random variable while $\underline{b}$ is discrete with three possible values $b_1$, $b_2$, $b_3$. Suppose that the statistician is given the three probabilities $p(b_1)$, $p(b_2)$ and $p(b_3)$ and he also knows how to distribute his subjective probability if it were $\underline{b} = b_i$, i.e. he knows the functions $p(a|b_i) = g_i(a)$ for all three $i$'s. In order to be consistent with the two basic operations introduced above, he has to determine the joint probability density $p(a,b)$ in such a way that the functions $p(a,b_i) = f_i(a)$, see Fig. 1, are

$$f_i(a) = p(b_i)g_i(a),\ i = 1,2,3.$$

It is clear from the way how the two basic relations (7) and (11) have been introduced that they apply also for conditional distributions. Actually, they determine the logical structure of the system called Bayesian statistics and we shall register them for further references in the following form.

$$p(b|c) = \int p(a,b|c)\, da \tag{12}$$

$$p(a,b|c) = p(a|b,c)p(b|c) \tag{13}$$

It should be recalled once more that the integral in (12) has to be replaced by a regular sum if $a$ is discrete, or by a sum of integrals if $a$ is multivariate and mixed.

## 2.2 Independent Uncertain Quantities

We shall call the uncertain quantity $\underline{a}$ independent of the quantity $\underline{b}$ if the knowledge of the true value of $\underline{b}$ does not bring any information about $\underline{a}$ and therefore

$$p(a|b) = p(a) \tag{14}$$

If the quantity $\underline{b}$ is also uncertain with probability distribution $p(b)$ then from (11) and (14) follows

$$p(a,b) = p(a)p(b) \tag{15}$$

Moreover, as

$$p(a,b) = p(b|a)p(a)$$

it also holds

$$p(b|a) = p(b)$$

It means that if an uncertain quantity does not depend on another uncertain quantity then they are mutually independent.

In probability theory the independence of two random variables is usually defined by the relation (15). We took as primary the relation (14) as it has a clear Bayesian interpretation.

It is useful to define also *conditional independence*. If the true value of an uncertain quantity $\underline{c}$ is known and if the knowledge of the true value of the uncertain quantity $\underline{b}$ does not bring any additional information about the uncertain quantity $\underline{a}$ then the uncertain quantities $\underline{a}$ and $\underline{b}$ are called conditionally independent, under the condition that $\underline{c}$ is known. In this case it holds

$$\begin{aligned} p(a|b,c) &= p(a|c) \quad \text{and consequently} \\ p(b|a,c) &= p(b|c) \end{aligned}$$

Note that conditional independence does not imply unconditional independence and also that $p(a|b,c) = p(a|c)$ in general does not imply $p(a|b,c) = p(a|b)$.

## 2.3 Derived Relations

Solutions of all identification problems we shall deal with can be obtained by an appropriate application of the two basic relations (12) and (13). However, some formulae appear so often that it is worth while to derive them generally in advance and use them as standard.

First, we shall derive the famous Bayes formula which gave the name to the Bayesian statistics the application of which we shall deal with. From (12) and (13) we have

$$p(a|b,c) = \frac{p(a,b|c)}{p(b|c)} = \frac{p(a,b|c)}{\int p(a,b|c)\, a} \tag{16}$$

If we interchange the role of $a$ and $b$ in (13) we also have

$$p(a,b|c) = p(b|a,c)p(a|c) \tag{17}$$

Substitution of (17) into (16) gives the *Bayes formula*

$$p(a|b,c) = \frac{p(b|a,c)p(a|c)}{\int p(b|a,c)p(a|c)\, da} \tag{18}$$

The second standard formula, we shall often make use of, is the so-called chain rule. To derive this rule consider the joint probability distribution of $N$ random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N$ and apply successively (11). After $N$ steps the *chain rule* is obtained

$$p(x_N, x_{N-1}, \ldots, x_1) = \prod_{k=2}^{N} p(x_k | x_{k-1}, \ldots, x_1) \cdot p(x_1) \tag{19}$$

## 2.4 Additional Remarks

We conclude this section with several general remarks.

Often, Bayesian statistics is not distinguished clearly enough from decision theory. Statistical inference is only a part of decision making. Bayesian statistics provides probability distributions, conditional on data as a rational basis for decisions. Decision theory adds the utility (or risk), calculates expectations and performs maximization (or minimization). If one has a reason to choose some single value from the set of possible values value from the set of possible values value from the set of possible values of an uncertain quantity (or accept as true a single hypothesis from the set of mutually exclusive hypotheses none of which is known to be certainly true) one has to solve a decision problem.

As a rule, *system identification* is only a part of a more complex decision problem (forecasting, control, some kind of diagnosis, etc.) for which point estimates of model parameters are, actually, not required, at least not directly as the final objective. It is true that some point estimates often appear as a natural (or reasonable) inter-step in the exact or approximate solution of a given decision problem, but examples also can be given where no point estimate can be chosen as a suitable representant for the unknown parameter (see e.g. Peterka, 1977, par. 7.2). Therefore, when dealing solely with systems identification we shall give the solutions of particular problems in the form of probability distributions. Not given a clearly defined purpose for which the system identification is performed we have to provide this full information.

Frequency interpretation of probability (Von Mises) rests on the idea of repeated experiments performed under "similar" conditions. Outcomes of these trials are interpreted as different realizations of the same random variable (or random event). In Bayesian view each random variable can take just one true value. *"Act of observation changes the status of the quantity from a random variable to a number"* [19]. Repeated trials are just a sequence of random events and all they have in common is that they can be assumed to have the same probability distribution and to be conditional independent under the condition that the common probability distribution is a priori given. However, this is in no contradiction with the intuitive conception of probability as the limit of relative frequencies the stationarity of which may appear to an outer observer (with a given observation ability) as an objective property of the external world. On the contrary, the idea of existence of such limits can be very helpful in constructing probabilistic models but by no means can be taken as a basic for a consistent theory. Bayesian statistics can serve as a means for finding out what these "objective" probabilities are but its applicability is much wider.

One may say that Bayesian statistics is nothing else than probability theory applied to statistical problems.It is true when the above given interpretation of probability, conditioning and statistical independence is added. However, probability theory as such can only transform probability distributions, it cannot create them. Similarly, Bayesian statistics requires the prior probability distribution which the statistician has to assign to unknown quantities or uncertain events before the observed data are incorporated into his knowledge. The prior probability distribution is a model of the statistician's prior uncertainty. Like any other mathematical model, for mathematics it is an input. It is the user, not the theory, who is responsible for all models which make the link between mathematics and the true world. One also cannot expect a reasonable answer to an ill-posed question. Man thinks, theory helps him to think and to maintain his thinking consistent in complex situations. This is the role of any theory.

# 3 System Model, Reexamined from Bayesian Viewpoint

Throughout this chapter the term "system" is understood very generally as a part of the external world the statistician wishes to identify, i.e. to describe mathematically for a given purpose. To perform his task the statistician has the possibility to observe on the system a time-oriented sequence of quantities [4] (a process), say

$$D_{(1)}, D_{(2)}, D_{(3)}, \ldots, D_{(t)}, \ldots$$

The values of these quantities, which are known to the statistician at a given time point, will be called the data. In general, there are two kinds of quantities which can be observed on a system: inputs, which will be denoted by $u_{(t)}$, and outputs denoted by $y_{(t)}$

$$D_{(t)} = (u_{(t)}, y_{(t)}) \tag{20}$$

The inputs are the quantities the values of which are enforced on the system, contingently by the statistician himself, while the outputs can be observed only passively and if they can be influenced by the statistician then only through the preceding inputs. The systems which have no observable inputs, i.e. $D_{(t)} = y_{(t)}$, are sometimes called *autonomous*.

If the model of the system has to be used for control purposes then it is essential to define what data are available when the value of the particular input, say $u_{(t)}$ is decided. We choose the time-indexing in such a way that by the output $y_{(t)}$ we denote a set of quantities the values of which are available when the decision concerning $u_{(t+1)}$ is taken but are not yet known when $u_{(t)}$ is decided [5]. Thus, in our time-indexing the sequence of inputs and outputs ordered in the way how they become to be known to the statistician, who is in the position of an outer observer and an actual or potential decision maker controlling the system, is

$$u_{(1)}, y_{(1)}, u_{(2)}, y_{(2)}, \ldots, u_{(t-1)}, y_{(t-1)}, u_{(t)}, y_{(t)}, \ldots$$

where $(u_{(1)}, y_{(1)}) = D_{(1)}$ is the first input-output pair observed. To shorten the writing when dealing with sets of inputs and outputs the following notation will be used where $x$ stands for either $u$ or $y$ or $D$.

$$x_{(i)}^{(j)} = \{x_{(i)}, x_{(i+1)}, \ldots, x_{(j)}\} \tag{21}$$

For $j < i$ the set (21) is empty. Clearly

$$x_{(i)}^{(j)} = \{x_{(j)}, x_{(i)}^{(j-1)}\}$$

and

$$D_{(i)}^{(j)} = \{y_{(j)}, u_{(j)}, D_{(i)}^{(j-1)}\} \tag{22}$$

If the lower time index $(i)$ is omitted then it means $i = 1$. This is used to denote the set of all data from the beginning of observation, e.g.

$$D^{(t)} = \{D_{(1)}, D_{(2)}, \ldots, D_{(t-1)}, D_{(t)}\} \tag{23}$$

As it has been emphasized already, when approaching any modeling problem the purpose for which the model will be used has to be considered right from the beginning. In this chapter it is assumed that the purpose of modeling and identification of the given system is to provide a rational basis for the control of the future course of the output. Therefore, the problem of a suitable system model is posed as a question: What does the statistician need to know to be able to solve his control problem?

Assume that the input-output data up to and including the time-index $t_0$, i.e. $D^{(t_0)}$ are known to the statistician and his task is to design a control strategy for the next $N$ steps, where $N$ arbitrarily

[4] Saying quantities we may generally mean also events of non-numeric character.

[5] As misunderstandings concerning this point are met in control literature it may be worth noting that the choice of time-indexing is, to a certain extent, a question of convention. The output we denoted by $y_{(t)}$ could be equally well denoted by $y_{(t-1)}$. This is not essential, but essential is to define whether this output is available or not when $u_{(t+1)}$ is determined.

large but finite. If the statistician picked a particular strategy and performed the experiment he would be able to judge the quality of his performance according to the actual values of inputs and outputs in the time-interval considered, i.e. according to the true values of $D_{(t_0+1)}^{(t_0+N)}$. This is the full information the experiment could yield to him as to an outer observer. As he has to choose in advance the control strategy which is optimal in some sense, he must be able to forecast, before the input $u_{(t_0+1)}$ is applied, what the future input-output data would be for any control strategy he might apply. Hence what he needs is the conditional probability distribution

$$p(D_{(t_0+1)}^{(t_0+N)}|D^{(t_0)}) \tag{24}$$

for any admissible control strategy. Applying the chain rule (19) to (34)

$$p(D_{(t_0+1)}^{(t_0+N)}|D^{(t_0)}) = \prod_{t=t_0+1}^{t_0+N} p(D_{(t)}|D^{(t-1)})$$

and making use of the basic relation (13)

$$p(D_{(t)}|D^{(t-1)}) = p(y_{(t)}, u_{(t)}|D^{(t-1)}) = p(y_{(t)}|u_{(t)}, D^{(t-1)})p(u_{(t)}|D^{(t-1)})$$

we obtain

$$p(D_{(t_0+1)}^{(t_0+N)}|D^{(t_0+1)}) = \prod_{t=t_0+1}^{t_0+N} p(y_{(}t|u_{(t)}, D^{(t-1)})p(u_{(t)}|D^{(t-1)}) \tag{25}$$

The factors in (25) have the following interpretation. The conditional probability distribution

$$p(u_{(t)}|D^{(t-1)}) \tag{26}$$

describes the transformation, in general stochastic, by which the input $u_{(t)}$ is determined on the basis of the known past history of the process. The set of functions (26) for $t = t_0 + 1, \ldots, t_0 + N$ is, actually, the control strategy the statistician has to determine when he solves his control problem. If the control strategy is deterministic[6], i.e. $u_{(t)} = f_{(t)}(D^{(t-1)})$ then (26) is

$$p(u_{(t)}|D^{(t-1)}) = \delta(u_{(t)} - f_{(t)}(D^{(t-1)}))$$

where $\delta(\cdot)$ is either the Dirac $\delta$-function when $u_{(t)}$ is of continuous type, or the Kronecker's $\delta$ ($\delta(0) = 1$ and $\delta(x) = 0$ for $x \neq 0$) when $u_{(t)}$ is discrete. If the input is generated in open loop, i.e. independently of the outputs, then

$$p(u_{(t)}|D^{(t-1)}) = p(u_{(t)}|u^{(t-1)}) \tag{27}$$

Hence, the probability distribution (26) is a description of the feedback or of the input generator, not of the system itself.

The remaining factors in (25), i.e. the set of conditional probability distributions

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) \tag{28}$$

describe, for each $t$, the dependence of the output $y_{(t)}$ on the known past history of the input-output process including the last input. The set of conditional probability distributions (28) is the most general description of the system from the viewpoint of an outer observer. It is this set of functions the statistician needs to be able to design a control strategy or to forecast the outputs for a given control strategy.

By a *system model* (or process model) we shall mean any mathematical model which *defines the set of conditional probability distributions (28) for the time period required through a finite set of parameters.* By a parameter we mean here a time-invariant quantity, a constant.

Clearly, all models which define the same set of conditional probability distributions (28) are equivalent from the viewpoint of an outer observer, they cannot be distinguished by him and,for the purpose of forecasting and control of the future outputs, also do not need to be distinguished.

[6]It is possible to prove that under very general conditions optimal strategies are deterministic.

Notice that the conditional probability distribution (28) can be considered as one-step-ahead predictor. If the process model is not given directly in this form it has to be recalculated into this form when it has to be used for the purpose of forecasting and control of the output.

Consider the situation when a finite set of some or all model parameters, say $\theta$,is unknown or uncertain. In such a case the model (its structure) does not fully define the distributions (28) but only distributions conditional, in addition, on $\theta$.

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta) \tag{29}$$

In (29) $\theta$ has to be considered as a variable, in general multi-dimensional, ranging over the set $\mathcal{S}_\theta$ of all possible values of the uncertain quantity (random variable) $\underline{\theta}$.

Not knowing the true value of $\underline{\theta}$, the statistician cannot make a direct use of (29), he has to eliminate the unknown parameters first. This can be done by application of the two basic operations (12) and (13) in the following way.

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = \int p(y_{(t)}, \theta|u_{(t)}, D^{(t-1)})\, d\theta = \tag{30}$$

$$\int p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta) p(\theta|u_{(t)}, D^{(t-1)})\, d\theta$$

The first factor of the integrand in (26) is the distribution (29) defined by the model structure. The second factor

$$p(\theta|u_{(t)}, D^{(t-1)}) \tag{31}$$

is the probability distribution describing the uncertainty of the parameters at the given time point.

Hence, the system identification performed for the purpose of control or forecasting of the outputs can be decomposed into two steps:

**i)** choice of model structure defining the conditional probability distributions (29)

**ii)** estimation of model parameters, i.e. the determination of the conditional probability distribution (31).

The problem of parameter estimation will be considered in detail in the next Section 4 assuming that the model structure is given. The problem the statistician is faced when he is uncertain also about the model structure is called the system classification and will be solved in Section 6.

The following simple example may serve to illustrate the ideas.

Example 3.1 Consider an autonomous system (with no observable inputs) the output process of which is a sequence of random events with just two possible outcomes, say $A$ and $\overline{A}$. Thus, either $\underline{y}_{(t)} \equiv A$ or $\underline{y}_{(t)} \equiv \overline{A}$ but it is a priori not known which one of identities is true. Clearly, the set of all possible outcomes of $\underline{y}_{(t)}$ consists of only two elements, $\mathcal{S}_y = \{A, \overline{A}\}$ and $p(y_{(t)}) = f_{(t)}(y_{(t)})$, which must fulfill the relation $f_{(t)}(A) + f_{(t)}(\overline{A}) = 1$, determined by just one number $\alpha_{(t)}$

$$f_{(t)}(A) = \alpha_{(t)}, \quad f_{(t)}(\overline{A}) = 1 - \alpha_{(t)} \tag{32}$$

To construct a model of the process means to accept some assumptions. Let, in our example, these assumptions be:

**(a)** The statistician who has determined on the basis of the prior information about the system modelled, his probability distribution $p(y_{(t)})$ i.e. the number $\alpha_{(t)}$, assumes, also on the basis of the prior information, that the past history of the process cannot bring any additional information about the expected output $y_{(t)}$. Therefore, he does not change his opinion when this information is given to him

$$p(y_{(t)})|y^{(t-1)}) = p(y_{(t)}),\ t = 1, 2, \ldots \tag{33}$$

**(b)** Considering the physical nature of the system (again prior information) the statistician assumes that the number $\alpha_{(t)}$ is the same for each $t$.

$$\alpha_{(t)} = \alpha, \ t = 1, 2, \ldots$$

Hence, the process model is

$$\begin{aligned} p(y_{(t)})|y^{(t-1)}) &= \alpha \text{ for } y_{(t)} \equiv A \text{ and any } y^{(t-1)} \\ p(y_{(t)})|y^{(t-1)}) &= 1 - \alpha \text{ for } y_{(t)} \equiv \overline{A} \text{ and any } y^{(t-1)} \end{aligned} \tag{34}$$

and is fully defined by a single parameter $\theta = \alpha$.

It should be emphasized that, actually, all probability distributions are conditional.However, it does not make much sense to state explicitly and repeatedly all conditions which do not change during the solution of a given problem. Moreover, some of these "permanent" conditions are often difficult to express in a simple way. For instance, in this example the model obtained is conditional on the prior information which allows the statistician to determine both the simple structure and the single parameter $\alpha$. To be more explicit let us consider that the process modelled is a "fair" tossing of a "fair" coin ($A \equiv$"head" $\overline{A} \equiv$"tail, or reversely). Because of symmetry and for "insufficient reasons" for preferring some of the two possible outcomes as more likely, the statistician can assume, $p(A) = p(\overline{A})$ from which follows $\alpha = \frac{1}{2}$.

However, if the prior information does not allow the statistician to determine the parameter $\alpha$ (for instance, he is sure about the fairness of the tossing but in doubt whether the coin is fair) he should re-formulate the assumption (a) in the following way.

**(a′)** If I knew more about the system and could determine the parameter $\alpha$ then the knowledge of the past history of the process would not bring any additional information about the expected output of the process $\underline{y}_{(t)}$.

This means that the independence (33) has to be replaced by a weaker assumption of conditional independence

$$p(y_{(t)}|y^{(t-1)}, \alpha) = p(y_{(t)}|\alpha) \tag{35}$$

and the unknown parameter has to be considered as a continuous random variable $\underline{\alpha}$ the possible values of which are real numbers between 0 and 1, $\mathcal{S}_\alpha = [0, 1]$. Instead of (34) we now have for any $y^{(t-1)}$ and $\alpha \in \mathcal{S}_\alpha$

$$\begin{aligned} p(y_{(t)}|y^{(t-1)}, \alpha) &= \alpha \text{ for } y_{(t)} = A \\ p(y_{(t)}|y^{(t-1)}, \alpha) &= 1 - \alpha \text{ for } y_{(t)} = \overline{A} \end{aligned} \tag{36}$$

where $\alpha$ is not a constant but a variable.

As the past history of the process carries information about the unknown parameter $\underline{\alpha}$, (35) does not imply (33). To predict the future output $\underline{y}_{(t)}$ the statistician can make use of the formula (30) which, applied to this case, reads

$$p(y_{(t)}|y^{(t-1)}) = \int_0^1 p(y_{(t)}|\alpha) p(\alpha|y^{(t-1)}) \, d\alpha$$

and particularly for $y_{(t)} \equiv A$:

$$p(y_{(t)}|y^{(t-1)}) = \int_0^1 \alpha p(\alpha|y^{(t-1)}) \, d\alpha \tag{37}$$

for $y_{(t)} \equiv \overline{A}$:

$$p(y_{(t)}|y^{(t-1)}) = \int_0^1 (1 - \alpha) p(\alpha|y^{(t-1)}) \, d\alpha \tag{38}$$

The conditional probability distribution $p(\alpha|y^{(t-1)})$ will be determined and the integrals evaluated in Section 4 where we shall deal with parameter estimation.

## 3.1 Discrete White Noise

If the output $\underline{y}_{(t)}$ is a random variable of continuous type it may be useful to introduce a related variable $\underline{e}_{(t)}$ as a difference between $\underline{y}_{(t)}$ and its mean value conditioned on the past history of the input-output process. If the output $\underline{y}_{(t)}$, a set $\nu$ of quantities, is ordered into a column $\nu$ -vector then we define $\underline{e}_{(t)}$ as follows.

$$\underline{e}_{(t)} = \underline{y}_{(t)} - \hat{y}_{(t)}(u_{(t)}, D^{(t-1)}) \tag{39}$$

$$\hat{y}_{(t)}(u_{(t)}, D^{(t-1)}) = E[y_{(t)}|u_{(t)}, D^{(t-1)}] = \int y_{(t)} p(y_{(t)}|u_{(t)}, D^{(t-1)})\, dy_{(t)} \tag{40}$$

Clearly,

$$E[e_{(t)}|u_{(t)}, D^{(t-1)}] = 0 \tag{41}$$

But the sequence $\{\underline{e}_{(t)};\ t = 1, 2, \ldots\}$ has also the following properties.

$$E[e_{(t)}] = 0 \tag{42}$$

$$E[e_{(t)} e^T_{(t-i)}] = 0;\ i \neq 0,\ i < t \tag{43}$$

$$E[e_{(t)} y^T_{(t-i)}] = 0,\ 0 < i < t \tag{44}$$

$$E[e_{(t)} u^T_{(t-i)}] = 0,\ 0 \leq i < t \tag{45}$$

A sequence of random variables with zero unconditional mean (42),which are mutually uncorrelated, (43), is often called a *discrete white noise.* We shall prove only (43), the remaining properties can be proved in a very similar way.

Consider $i > 0$ first. From the definition (39) of $\underline{e}_{(t)}$ it follows

$$E[e_{(t)} e^T_{(t-i)}] = \int [y_{(t)} - \hat{y}_{(t)}(u_{(t)}, D^{(t-1)})][y_{(t-i)} - \hat{y}_{(t-i)}(u_{(t-i)}, D^{(t-1-i)})]^T p(D^{(t)})\, dD^{(t)} \tag{46}$$

Using the basic operation (13) we may write

$$p(D^{(t)}) = p(y_{(t)}|u_{(t)}, D^{(t-1)}) p(u_{(t)}, D^{(t-1)})$$

The substitution into (46) gives

$$\begin{aligned} &E[e_{(t)} e^T_{(t-i)}] = \\ &= \int \left[ \int y_{(t)} p(y_{(t)}|u_{(t)}, D^{(t-1)})\, dy_{(t)} - \hat{y}_{(t)} \right] \\ &\quad [y_{(t-i)} - \hat{y}_{(t-i)}(u_{(t-i)}, D^{(t-1-i)})]^T p(u_{(t)}, D^{(t-1)})\, d(u_{(t)}, D^{(t-1)}) \end{aligned}$$

which is a zero matrix due to the fact that the difference in the first brackets is a zero vector according to the definition of $\hat{y}_{(t)}$ (40). This proves (43) for $i > 0$. For $i < 0$ shift the time index by introducing $\tau = t - i$ instead of $t$ and proceed similarly.

According to (39) the conditional mean value of $\underline{e}_{(t)}$ is equal to zero independently of the past history of the input-output process. A significant simplification can be assumed that not only this mean value but also the entire form of the distribution of $\underline{e}_{(t)}$ is independent of the past input-output data and that this form, say $g(\underline{e}_{(t)})$, is the same for each $t$.

$$p(e_{(t)}|u_{(t)}, D^{(t-1)}) = p(e_{(t)}) = g(e_{(t)}) \tag{47}$$

Clearly, if $g(\cdot)$ is time-invariant then the covariance matrix of $e_{(t)}$ is constant.

$$E[e_{(t)} e^T_{(t)}|u_{(t)}, D^{(t-1)}] = E[e_{(t)} e^T_{(t)}] = R \tag{48}$$

For given $u_{((t))}$ and $D^{(t-1)}$ the random variables $y_{(t)}$ and $e_{(t)}$ are related by the one-to-one transformation (39) the Jacobian of which is equal to one. Therefore

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = g(y_{(t)} - \hat{y}_{(t)}(u_{(t)}, D^{(t-1)})). \tag{49}$$

The relation (39) between random variables $\underline{y}_{(t)}$ and $\underline{e}_{(t)}$ holds, of course, also for any pair of their possible values and the process model can be given the form of a stochastic equation.

$$y_{(t)} = \hat{y}_{(t)}(u_{(t)}, D^{(t-1)}) + e_{(t)}. \tag{50}$$

If, in addition to (47), it can be assumed that $g(\cdot)$ is normal (Gaussian)

$$g(e_{(t)}) = (2\pi)^{-\frac{\nu}{2}} |R|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} e_{(t)}^T R^{-1} e_{(t)}\right] \tag{51}$$

or in shortened notation

$$g(e_{(t)}) \sim \mathcal{N}(0, R)$$

then the process model is fully defined if the covariance matrix $R$ is given and if the conditional mean value $\hat{y}_{(t)}$ is expressed through a finite set of parameters as a (deterministic) function of the past input-output data.

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) \sim \mathcal{N}(\hat{y}_{(t)}, R) \tag{52}$$

In this way the modelling problem is considerably reduced but it should be emphasized that the assumptions (47) and (51) may sometimes be rather restrictive.

The random variable $\underline{e}_{(t)}$ defined by (39) is sometimes called innovation, see [13]. The decomposition of the system description (50) into the deterministic term, the conditional mean of the output (40), and the random term, innovation (39), is used in prediction error methods of systems identification, see [20].

In the following examples three most commonly used linear input-output models are introduced. The purpose of these examples is to make clear the assumptions on which the model is based so that the potential user can judge himself whether the model suits his particular case or whether he has to look for another model more suitable for his need. For a case study on modelling of a nonlinear non-stationary multi-output macro-economic process the reader is referred to [23].

Example 3.2 (linear regression model) Consider a system with $\mu$ inputs and $\nu$ outputs, both continuous in magnitudes, $u_{(t)} \in \mathcal{R}^{\mu}$, $y_{(t)} \in \mathcal{R}^{\nu}$. Suppose that the probability distribution of the output $\underline{y}_{(t)}$ conditional only on $u_{(t)}$ and $n$ previous input-output pairs $D_{(t-n)}^{(t-1)}$ is given or can be determined (including both structure and parameters) on the basis of prior information about the system modelled. If the past history considered in the condition is long enough ($n$ is sufficiently large) it may be reasonable to assume that the older input-output data $D^{(t-n-1)}$ cannot bring any additional information about the expected output $\underline{y}_{(t)}$. Mathematically, it is assumed that

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = p(y_{(t)}|u_{(t)}, D_{(t-n)}^{(t-1)}) \tag{53}$$

$$\hat{y}_{(t)}(u_{(t)}, D^{(t-1)}) = \hat{y}_{(t)} p(u_{(t)}, D_{(t-n)}^{(t-1)}) \tag{54}$$

If the conditional mean (54) is a linear function of its arguments

$$\hat{y}_{(t)} = B_{(0)} u_{(t)} + \sum_{i=1}^{n} \left(A_i y_{(t-i)} + B_i u_{(t-i)}\right) + c \tag{55}$$

and the assumption (47) is added the linear regression model of $n$-th order is obtained. Written in the form (50) this model is

$$y_{(t)} = B_{(0)} u_{(t)} + \sum_{i=1}^{n} \left(A_i y_{(t-i)} + B_i u_{(t-i)}\right) + c + e_{(t)} \tag{56}$$

where $\{\underline{e}_{(t)}\}$ is discrete white noise with constant covariance matrix $R$ (48).

If also the normality (51) is assumed then the process model (52) is, for $t > n$, fully specified by the parameter set

$$\theta = \{A_i\,(i = 1, \ldots, n),\ B_i\ (i = 0, 1, \ldots, n),\ c,\ R\} \tag{57}$$

The constant term $c$ in (55) and (56) can be eliminated by a proper choice of origins of the scales in which inputs and/or outputs are measured. However, if the parameters (57) are not known and the system describable by the regression model has to be identified then, in general, this term has to be considered.

Example 3.3 (incremental regression model). The practical experience shows that often the real processes are non-stationary. The nonstationarity usually follows from the fact that from time to time something happens which causes that the constant $c$ in the regression model (55) is, actually, not constant but varies in a rather unpredictable way. In such situations it may be more appropriate to assume the conditional mean $\hat{y}_{(t)}(u_{(t)}, D^{(t-1)})$ in the following form

$$\hat{y}_{(t)} = y_{(t-1)} + B_0 \Delta u_{(t)} + \sum_{i=1}^{n} \left(A_i \Delta y_{(t-i)} + B_i \Delta u_{(t-i)}\right) \tag{58}$$

where

$$\Delta y_{(\tau)} = y_{(\tau)} - y_{(\tau-1)} \tag{59}$$

The assumed form (58) of the conditional mean of the expected output $\underline{y}_{(t)}$ can be understood as a linear extrapolation of the output process related to the last known output $y_{(t-1)}$. The model can be written also in the following form

$$y_{(t)} = B_0 u_{(t)} + \sum_{i=1}^{n} \left(A_i y_{(t-i)} + B_i u_{(t-i)}\right) + c_{(t)} \tag{60}$$

where $\{c_{(t)}\}$ is a stochastic process with independent increments

$$\underline{c}_{(t)} = \underline{c}_{(t-1)} + \underline{e}_{(t)} \tag{61}$$

i.e. a summed discrete white noise $\{\underline{e}_{(t)}\}$.

Example 3.4 (ARMA model). In the previous two examples the mean value $\hat{y}_{(t)}$ has been assumed to be a function only of the finite number of the foregoing inputs and outputs. However,in general, this mean value can be a deterministic function of the entire past history of the input-output process. To express this function through a finite number of parameters assume that $\hat{y}_{(t)}$ is defined recursively by the following difference equation

$$\hat{y}_{(t)} + \sum_{i=1}^{n} C_i \hat{y}_{(t-i)} = B_0 u_{(t)} + \sum_{i=1}^{n} \left(G_i y_{(t-i)} + B_i u_{(t-i)}\right) + c \tag{62}$$

Of course, such a recursive definition of the conditional mean value $\hat{y}_{(t)}$ makes sense only when the homogenous part of the difference equation (62) is stable, i.e. when all roots $\zeta_i$, $i = 1, \ldots, n\nu$ of the polynomial

$$|I + \sum_{i=1}^{n} C_i \zeta^i| \tag{63}$$

lie outside the unit circle.

If $\hat{y}_{(t)} = y_{(t)} - e_{(t)}$ is substituted into (62) the following popular form of this model is obtained

$$y_{(t)} + \sum_{i=1}^{n} A_i y_{(t-i)} = B_0 u_{(t)} + \sum_{i=1}^{n} B_i u_{(t-i)} + e_{(t)} + \sum_{i=1}^{n} C_i e_{(t-i)} + c \tag{64}$$

where

$$A_i = C_i - G_i \tag{65}$$

Usually, the model is considered without the constant term $c$ which can be eliminated by a proper shift of the scales for $u_{(t)}$ and/or $y_{(t)}$, however, only when the matrix coefficients are known. The model got its name ARMA according to the *autoregressive and moving-average* parts in (64). We prefer the form (62) to (64) as the former is directly related to the set of probability distribution $p(y_{(t)}|u_{(t)}, D^{(t-1)})$ which are required for the purpose of forecasting and control.

If the normality of $e_{(t)}$ (51) is assumed, then the model defines the distributions (28) for $t > n$ through the following set of parameters

$$\theta = \{G_i(i = 1, 2, \ldots, n),\ B_i(i = 0, 1, \ldots, n),$$

$$C_i(i = 1, 2, \ldots, n),\ c,\ R,\ \hat{y}_{(i)}(i = 1, 2, \ldots, n)\}$$

where $\hat{y}_{(i)}(i = 1, 2, \ldots, n)$ are initial conditions for the difference equation (62). If the known past history of the input-output process is long enough (i.e. $t_0 >> n$ in (24) the influence of the initial conditions $\hat{y}_{(i)}$ $(i = 1, 2, \ldots, n)$ may be negligible, they can be set to zero and considered as known. Nevertheless, even then the estimation of parameters $C_i$ $(i = 1, 2, \ldots, n)$ is technically difficult and therefore the ARMA model is less suitable for real-time identification in adaptive control systems except when the parameters $C_i$ are fixed as a priori known.

Example 3.5 (state space model in innovation from). Problem of system modeling, from our point of view, can be understood as parameterization of the family of conditional probability distributions (28) for $t > t_0$. In general, each member of this family is a scalar function defined on a set variables the dimension of which is different for each $t$. If it is required that the entire family be described by a finite set of parameters then it is appropriate to assume that there exists a finite dimensional quantity, say $s_{(t-1)}$, into which the information about the known past history of the process $D^{(t-1)}$ can be reduced. This quantity $s_{(t-1)}$ can be understood as a state of the system or, more precisely, as a sufficient statistic [7] for the output $y_{(t)}$.

Under this condition it holds

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = \psi(y_{(t)}, u_{(t)}, s_{(t-1)}) \tag{66}$$

and the deterministic relations for the updating of the sufficient statistics $s_{(t-1)}$ completes the general form of the model.

$$s_{(t)} = \phi(s_{(t-1)}, u_{(t)}, y_{(t)}) \tag{67}$$

In this way the modeling problem is reduced to the choice of the dimension of $s_{(t)}$ and to the parametrization of one scalar function $\psi(\cdot)$ and one multidimensional function $\phi(\cdot)$, both defined on he same set of variables of fixed dimension. The simplest possible way how to perform this parametrization is to assume linearity and normality as follows. The decomposition (50) in this case reads

$$y_{(t)} = \hat{y}_{(t)}(u_{(t)}, s_{(t-1)}) + e_{(t)} \tag{68}$$

If linearity of both functions $\hat{y}_{(t)}(\cdot)$ and $\phi(\cdot)$ is assumed the relations (68) and (67) get the form

$$y_{(t)} = Cs_{(t-1)} + Du_{(t)} + e_{(t)} \tag{69}$$

$$s_{(t)} = Hs_{(t-1)} + Gy_{(t)} + Fu_{(t)} \tag{70}$$

If, in addition, the assumptions (47), (48) and (51), concerning the stochastic term $e_{(t)}$, are accepted then

$$p(e_{(t)}|u_{(t)}, s_{(t-1)}) = p(e_{(t)}) \sim N(O, R) \tag{71}$$

together with (69) defines the function $\psi(\cdot)$ in (66). Substitution of $y_{(t)}$ from (69) into (70) gives

$$s_{(t)} = A\ s_{(t-1)} + B\ u_{(t)} + H\ e_{(t)} \tag{72}$$

where

$$A = H + GC,\ B = GD + F \tag{73}$$

[7] Sufficient statistics will be discussed in more detail in Section 4

The couple of equations (69) and (72) is sometimes called the *innovation form of the state space model.* In some applications the form (70) may be more suitable because of its determinism. Notice, that the stability of the matrix $H$ is more important then the stability of the matrix $A$, if the model is constructed for the purpose of prediction (opposed to system simulation when the stability $A$ is crucial). Notice also the parameter redundancy due to the invariance of the input-output relation with respect to any regular transformation of $s_{(t)}$.

To show the relations between different forms of models we shall bring the ARMA model from Example 3.4 to the state space form (69 - 70).

If we denote

$$S_{k(t-k)} = \sum_{i=k}^{n} (-C_i\ \hat{y}_{(t-i)} + G_i\ y(t-i) + B_i\ u_{(t-i)}) \tag{74}$$

for $1 \leq k \leq n$, then from (62), where we set $c = 0$ for simplicity, we have

$$y_t = B_0 u_{(t)} + s_{1(t-1)} + e_{(t)} \tag{75}$$

From (74) it follows for $1 \leq k \leq n$

$$\begin{aligned} s_{k(t)} &= +C_k\hat{y}_{(t)} + G_k y_{(t)} + B_k u_{(t)} + s_{k+1(t-1)} = & (76)\\ &= -C_k s_{1(t-1)} + G_k y_{(t)} + (B_k - C_k B_0) u_{(t)} + s_{k+1(t-1)} & \\ s_{n(t)} &= -C_n s_{1(t-1)} + G_n y_{(t)} + (B_n - C_n B_0) u_{(t)} & (77) \end{aligned}$$

The set of equations (76) and (77) written in the matrix form, is the canonical form of (70) with the state defined as

$$s^T_{(t)} = [s^T_{1(t)}, s^T_{2(t)}, \ldots, s^T_{n(t)}]$$

## 3.2 Measurable External Disturbances

When modeling a system the statistician, as a rationally reasoning person, has to use all prior information in order to make the model as certain as possible. Some prior information is always available. No prior information is a fallacy: an ignorant has no problems to solve. If, for instance, the statistician would not know what variables can be manipulated on the system he would not be able to distinguish the inputs from the outputs and neither would he be able to formulate the control problem.

An important prior information which is often available to the statistician trying to identify a given system is that the output $y_{(t)}$, i.e.a set of quantities the statistician manipulate, can be decomposed into two subsets

$$y_{(t)} = (v_{(t)}, y_{s(t)}) \tag{78}$$

where $v_{(t)}$ are *measurable external disturbances* which depend only on their own past history but not on the present and past values of all other quantities observed on the system

$$p(v_{(t)}|y_{s(t)}, u_{(t)}, D^{(t-1)}) = p(v_{(t)}|v^{(t-1)}) \tag{79}$$

External disturbances $v_{(t)}$ can be considered as a measurable output of an the external world sometimes called the *environment.* The subset $y_{s(t)}$ is the output of the system proper, i.e.,of the controlled part of the external world. Every practicioner knows how useful it can be to introduce the measurable external disturbances into the control algorithm.

Considering the decomposition (78) and applying the basic operation (13) we may write

$$\begin{aligned} p(y_{(t)}|u_{(t)}, D^{(t-1)}) &= p(v_{(t)}, y_{s(t)}|u_{(t)}, D^{(t-1)}) = \\ = p(v_{(t)}|y_{s(t)}, u_{(t)}, D^{(t-1)}) & p(y_{s(t)}|u_{(t)}, D^{(t-1)}) \end{aligned}$$

and according to the definition of external disturbances (79)

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = p(v_{(t)}|v^{(t-1)}) p(y_{s(t)}|u_{(t)}, D^{(t-1)}). \tag{80}$$

In this way the process model is decomposed into two models. The first factor in (80) is the model of measurable external disturbances (model of the autonomous and uncontrollable part of the external world) while the second factor is the description of the system proper (controlled part of external world). In the next section it will be shown that these two models can be identified separately, as one could intuitively expect.

Notice, that the probability distribution (26)

$$p(u_{(t)}|D^{(t-1)}) = p(u_{(t)}|u^{(t-1)}, y_s^{(t-1)}, v^{(t-1)})$$

is a general description of the control law including the feed-forward from the measurable disturbances.

# 4 Parameter Estimation and Output Prediction

Suppose that the statistician knows the system model up to a finite set of parameter $\theta$. This means that for a certain time interval, say $\tau = t_0 + 1, t_0 + 2, \ldots, t,$, the conditional probability distributions

$$p(y(\tau)|u_{(\tau)}, D^{(\tau-1)}, \theta) \tag{81}$$

are given. The statistician had the possibility to observe the system up to and including the time index $(t)$, i.e. the data $D^{(t)}$ are known to him. The first question we shall consider is

1. How can the statistician extract the information about the unknown parameters which is contained in the known input-output data? In Bayesian view the question reads: how to calculate the posterior probability distribution

   $$p(\theta|D^{(t)}) \tag{82}$$

   This is what we call Bayesian estimation. We already mentioned that a point estimate of the parameter set $\theta$ is just a partial description of the distribution (82) and that to choose such a point means to solve a decision problem. In a remark to Section 2 we also claimed that the system identification is, as a rule, only an inter-step in the solution of a more complex problem for which the point estimate of unknown parameters is, actually, not required. To demonstrate this fact we shall consider together with the first question also the following related question.

2. How can the statistician predict, for any given input $u_{(t+1)}$, the next output $y_{(t+1)}$ using only the known past history of the input-output process but not the parameter values which are not known to him? In Bayesian view this means to calculate the probability distribution of the output $y_{(t+1)}$ conditional on $u_{(t+1)}$ and $D^{(t)}$ but not $\theta$.

Making use of the two basic operations (12) and (13) we obtain similarly to (30)

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)} = \int p(y_{(t+1)}|u_{(t+1)}, D^{(t)}, \theta) p(\theta|u_{(t+1)}, D^{(t)}) d\theta \tag{83}$$

where the first factor is the conditional probability distribution (81) defined by the model structure. Clearly, the second posed question will be answered when the first one is solved and the relation between the posterior probability distribution (82) and the second factor of the integrated function (83), i.e.

$$p(\theta|u_{(t+1)}, D^{(t)}) \tag{84}$$

is established. It may be useful to distinguish two situations which may occur in practical applications. In the first case the amount of data is fixed and they have to be processed in one shot. In the other case the data are growing and the estimation is required in real time for every $t$ as, for instance, in adaptive control. In real-time estimation the problem is to update the probability distribution for $\underline{\theta}$ with respect to the new input-output pair, i.e. to calculate $p(\theta|D^{(t)})$ when $p(\theta|D^{(t-1)})$ and $D_{(t)}$ are given. To solve

both one-shot and real-time estimation at the same time let us formulate the problem as follows. Given $p(\theta|D^{(t_1)} =$ and the data $D^{(t)}_{(t_1+1)}$, $t_1 < t$, determine $p(\theta|D^{(t)})$. If we succeed to solve this problem then for $t_1 = 0$ the formula for one-shot estimation will be obtained while setting $t_1 = t-1$ we get the recursive relations for real-time estimation. Applying the Bayes formula (18) for $a = \theta$, $b = D^{(t)}_{(t_1+1)}$ and $c = D^{(t_1)}$ we obtain

$$p(\theta|D^{(t)}) = \frac{p(D^{(t)}_{(t_1+1)}|D^{(t_1)},\theta)p(\theta|D^{(t_1)})}{\int p(D^{(t)}_{(t_1+1)}|D^{(t_1)},\theta)p(\theta|D^{(t_1)})d\theta} \tag{85}$$

To be able to use this formula we have to express the conditional probability distribution

$$p(D^{(t)}_{(t_1+1)}|D^{(t_1)},\theta) \tag{86}$$

through probability distributions which are known.

The following four cases can be distinguished with respect to the way how the inputs are generated.

(a) The system is autonomous - has no observable input. In this case $D_{(\tau)} = y_{(\tau)}$ and instead of (81) we have the set of conditional distributions

$$p(y_{(\tau)}|y^{(\tau-1)},\theta) \tag{87}$$

which are given by the model structure.

(b) The input is deterministic, i.e. $u^{(t)}$ is a priori known before the experiment is performed. All $u_{(k)}$, $k = 1, 2, \ldots, \tau$ contained in (81) can be considered for each $\tau$ as known constants (parameters) of these functions and therefore can be omitted in (81) In this way the case is reduced to the case (a)

(c) The sequence of inputs is stochastic, i.e., not a priori known, but it is generated in open loop, i.e. independently of the outputs and of the unknown system parameters $\theta$

$$p(u_{(\tau)}|D^{(\tau-1)},\theta) = p(u_{(\tau)}|u^{(\tau-1)}) \tag{88}$$

(d) The inputs are generated in closed control loop, possibly by an adaptive controller or by the statistician himself during the experiment. They depend on the past outputs and through them also on the unknown parameters $\underline{\theta}$.

As it will appear that, under very general conditions, all cases listed above can be solved in the same way, we shall attack directly the most complex case (d)

## 4.1 Estimation in Closed Control Loop: Natural Conditions of Control

The joint probability distribution (86), which is required in (85), can be expressed, similarly to (25), as follows.

$$p(D^{(t)}_{(t_1+1)}|D^{(t_1)},\theta) = \prod_{\tau=t_1+1}^{t} p(y_{(\tau)}|u_{(\tau)},D^{(\tau-1)},\theta)p(u_{(\tau)}|D^{(\tau-1)},\theta) \tag{89}$$

The first factor in (89) is the conditional probability distribution (81) defined, for $\tau > t_0$, by the model structure, while the second factor, namely

$$p(u_{(\tau)}|D^{(\tau-1)},\theta) = p(u_{(\tau)}|u^{(\tau-1)},y^{(\tau-1)},\theta) \tag{90}$$

is a general description (from the statistician's viewpoint) of the law by which the input is generated.

Before we substitute (89) into (85) we shall show first that, under rather general conditions, a significant simplification can be achieved.

If the control strategy, generally described by (90), does not use more information about the unknown parameters than the information contained in the past input-outputs data $D^{(\tau-1)}$ then $\theta$ in the condition part of (90) is redundant and it holds

$$p(u_{(\tau)}|D^{(\tau-1)},\theta) = p(u_{(\tau)}|D^{(\tau-1)}) \tag{91}$$

Clearly, (91) holds when the statistician (observer) is at the same time also the decision maker (controller) controlling the system as, for instance, in adaptive control. The relation (91) cannot be derived mathematically, it must be introduced externally as a definition of conditions which will be called natural conditions of control. To throw more light on these conditions consider the joint probability distribution $p(u_{(\tau)},\theta|D^{(\tau-1)})$ which can be expressed, using the basic operation (13), in the following two ways.

$$p(u_{(\tau)}|D^{(\tau-1)},\theta)p(\theta|D^{(\tau-1)}) = p(\theta|u_{(\tau)},D^{(\tau-1)})p(u_{(\tau)}|D^{(\tau-1)})$$

From this identity follows that if (91) holds then also the following relation holds (and inversely).

$$p(\theta|u_{(\tau)},D^{(\tau-1)}) = p(\theta|D^{(\tau-1)}) \tag{92}$$

This relation can be used as a definition of natural conditions of control instead of (91). If says that the distribution for $\underline{\theta}$ remains unchanged when the true value of the single $u_{(\tau)}$ is obtained. This is, under the conditions discussed, self-evident as for decision concerning the input $u_{(\tau)}$ only that information about the unknown parameters $\theta$ could be used which could be extracted from the known past history of the process and therefore the result of this decision $\underline{u}_{(\tau)}$ cannot bring any additional information about $\underline{\theta}$. As it can be seen from (91) and (92), under natural conditions of control the random variables $\underline{u}_{(\tau)}$ and $\underline{\theta}$ are conditionally independent when the past input-output data $D^{(\tau-1)}$ are given. From (92) also follows that the formula for one-step-ahead prediction (83) under natural conditions of control reads

$$p(y_{(t+1)}|u_{(t+1)},D^{(t)}) = \int p(y_{(t+1)}|u_{(t+1)},D^{(t)},\theta)p(\theta|D^{(t)})d\theta \tag{93}$$

It should be emphasized that the natural conditions of control defined by (91) or by (92) are not fulfilled in all possible cases. Consider, for instance, the situation when the decision maker and the observer are two different persons. If the decision maker had more information about the parameters $\underline{\theta}$ and the observer knew his strategy then the observer could gain a new piece of information about $\underline{\theta}$ also from the single $\underline{u}_{(\tau)}$. However, this is not the case of our interest. Throughout the rest of this chapter, when not noted explicitly, it will be assumed that natural conditions of control are fulfilled.

Now we can return to our estimation problem. If the relation (89) is substituted into (85) then all functions (91) can be brought in front of the integral in the denominator (they do not depend on $\theta$) and cancelled with the same functions in the numerator.

$$p(\theta|D^{(t)}) = \frac{\prod_{\tau=t_1+1}^{t} p(y_{(\tau)}|u_{(\tau)},D^{(\tau-1)},\theta)p(\theta|D^{(t_1)})}{\int \prod_{\tau=t_1+1}^{t} p(y_{(\tau)}|u_{(\tau)},D^{(\tau-1)},\theta)p(\theta|D^{(t_1)})d\theta} \tag{94}$$

Notice that the formula covers all cases we want to consider. If the input is generated in open loop (case `(c)`) the natural conditions of control can be replaced by the stronger condition (88) which leads to the same result. If the system has no observable inputs or when they are deterministic (cases `(a)` and `(b)` then all $u_{(\tau)}$ either can be omitted or enter the formula (94) as a priori given constants.

If external disturbances $v_{(\tau)}$ can be observed on the system then, according to (80), the overall system model can be decomposed into two models

$$p(y_{(\tau)}|u_{(\tau)},D^{(\tau-1)},\theta) = p(v_{(\tau)}|v^{(\tau-1)},\theta_v)p(y_{s(\tau)}|u_{(\tau)},D^{(\tau-1)},\theta_s) \tag{95}$$

where $\theta_v$ is the set of unknown parameters in the model of external disturbances while $\theta_s$ is the set of unknown parameters in the model of the system proper, $\theta = \{\theta_v,\theta_s\}$. Substitution of (95) into (94) shows that if $p(\theta|D^{(t_1)}) = p(\theta_v|v^{(t_1)})p(\theta_s|D^{(t_1)})$ than also $p(\theta|D^{(t)}) = p(\theta_v|v^{(t)})p(\theta_s|D^{(t)})$ for any $t \geq t_1$. This means that the parameters of the two models can be estimated separately.

For $\theta_s$ we have

$$p(\theta_s|D^{(t)}) = \frac{\prod_{\tau=t_1+1}^{t} p(y_{s(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta_s) p(\theta_s|D^{(t_1)})}{\int \prod_{\tau=t_1+1}^{t} p(y_{s(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta_s) p(\theta_s|D^{(t_1)}) d\theta_s} \tag{96}$$

where

$$D^{(\tau-1)} = \{y_s^{(\tau-1)}, u^{(\tau-1)}, v^{(\tau-1)}\}$$

For the sake of simplicity, we shall omit the index $s$ in the sequel. In other words, we shall operate with the general formula (94) and shall leave the reader to perform the decomposition outlined, if it appears advantageous.

## 4.2 One-Shot Estimation

By setting $t_1 = 0$ into (94) the following formula for one-shot parameter estimation of the set of unknown parameters $\theta$ is obtained

$$p(\theta|D^{(t)}) = \frac{L_{(t)}(\theta, D^{(t)}) p(\theta)}{\int L_{(t)}(\theta, D^{(t)}) p(\theta) d\theta} \tag{97}$$

where

$$L_{(t)}(\theta, D^{(t)}) = \prod_{\tau=1}^{t} p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta) \tag{98}$$

and $p(\theta)$ is the probability distribution which models the statisticians prior uncertainty about the parameters $\underline{\theta}$ before the observed data $D^{(t)}$ are incorporated into his knowledge. The operation described by the formula (97) can be understood as a correction of the prior subjective probability distribution $p(\theta)$ by objective data. The product (98), considered for given data $D^{(t)}$ as a function of possible values of unknown parameters $\theta$, will be called the likelihood function or simply likelihood. The likelihood function reflects all what the experiment can say about the unknown parameters.

Some abuse of language should be noted. Usually, by likelihood function the prior joint probability density for all observed data, considered as a function of unknown parameters $\theta$, is meant. As the factors $p(u_{(\tau)}|D^{(\tau-1)}, \theta)$ are missing in (98), one should, to be precise, call (98) the significant part of the likelihood function for the case when the input-output data $D^{(t)}$ are observed under natural conditions of control.

As an introduction to more complex situation we give a simple example. This example may also help to understand clearly the relation between subjective probability and relative frequencies.

Example 4.1 This example is a continuation of Example 3.1. We consider a sequence of random events $\{\underline{y}_\tau; \tau = 1, 2, \ldots, t\}$ with two possible outcomes, either $\underline{y}_\tau \equiv A$ or $\underline{y}_\tau \equiv \overline{A}$. The model of the process is fully defined by a single parameter $\alpha \in S_\alpha$, $S_\alpha =< 0, 1 >$, such that, according to (36) for all $\tau > 1$, any $y^{(\tau-1)}$ and $\alpha \in S_\alpha$

$$\begin{aligned} p(y_{(\tau)}|y^{(\tau-1)}, \alpha) &= \alpha \quad \text{for} \quad y_{(\tau)} \equiv A \\ p(y_{(\tau)}|y^{(\tau-1)}, \alpha) &= 1 - \alpha \quad \text{for} \quad y_{(\tau)} \equiv \overline{A} \end{aligned} \tag{99}$$

Let $t$ be the total number of observations made and let $n$ be the number of observations the result of which was $y_\tau \equiv A$, $1 \leq \tau \leq t$. Hence $\alpha$ appears $n$ times as a factor in the likelihood function (98) while $(1 - \alpha)$ enters this product $(t - n)$ times.

$$L_{(t)}(\alpha, y^{(t)}) = \alpha^n (1 - \alpha)^{t-n} \tag{100}$$

Suppose that the statistician has no prior information about the parameter $\underline{\alpha}$ and therefore he has to consider, before the result of observation is known to him, all possible values $\alpha \in S_\alpha$ as equally likely. The model reflecting such a situation is

$$p(\alpha) = 1, \quad 0 \leq \alpha \leq 1 \tag{101}$$

Substitution of (100) and (101) into (97) gives

$$p(\alpha|y^{(t)}) = \kappa_{(t)}\alpha^n(1-\alpha)^{t-n} \tag{102}$$

where $\kappa_{(t)}$ is the normalizing factor not depending on $\alpha$

$$\kappa_{(t)} = \frac{1}{\int_0^1 \alpha^n(1-\alpha)^{t-n}d\alpha} = \frac{(t+1)!}{n!(t-n)!} = (t+1)\binom{t}{n} \tag{103}$$

It is easy to find that the maximum of the aposterior probability distribution (102) lies in the point

$$\hat{\alpha}_{(t)} = \frac{n}{t} \tag{104}$$

which is the maximum likelihood (ML) estimate of $\alpha$ well known from non-bayesian statistics. However, the statistician does not need such an estimate when he wants to predict the next output. What he needs is

$$p(y_{(t+1)}|y^{(t)}) = \int_0^1 p(y_{(t+1)}|y^{(t)},\alpha)p(\alpha|y^{(t)})d\alpha \tag{105}$$

Substitution of (99) yields for

$$y_{(t+1)} \equiv A : p(y_{(t+1)}|y^{(t)}) = Pr[\underline{y}_{(t+1)} \equiv A|y^{(t)})] = \int_0^1 \alpha p(\alpha|y^{(t)})d\alpha \tag{106}$$

and using (102) we obtain

$$Pr[\underline{y}_{(t+1)} \equiv A|y^{(t)}] = \kappa_{(t)}\int_0^1 \alpha^{n+1}(1-\alpha)^{t-n}d\alpha = \frac{n+1}{t+2} \tag{107}$$

Similarly

$$Pr[\underline{y}_{(t+1)} \equiv \overline{A}|y^{(t)}] = \frac{t-n+1}{t+2} \tag{108}$$

The conditional probability (107) can also be considered as a point estimate of $\underline{\alpha}$. It can be seen from (106) that in this particular example the point estimate which is optimal for the purpose of prediction is not the maximum of $p(\alpha|y^{(t)})$ but the mean value of this distribution. However, this observation must not be generalized. In other cases other point estimates may be more suitable and, as mentioned earlier, there exist also cases when no point estimate can be chosen as a suitable representant for the unknown parameter.

Notice that for small $t$ (107) behaves much more reasonably than (104). Even for $t = n = 0$, when the ML-estimate (104) is not defined, the prediction (107) gives $\frac{1}{2}$ which is logically correct.

Similarly for $t = 1$ and $n = 0$ (or $n = 1$)the prediction $\frac{1}{3}$ (or $\frac{2}{3}$) obtained from (100) is much more reasonable then the ML-estimate $\hat{\alpha}_{(1)} = 0$ (or 1). However, for large $t$ the difference is insignificant and asymptotically for $t \to \infty$ the "objective" (actually indiscernible) probability is formally obtained in both cases as the limit of relative frequency,

$$\alpha = \lim_{t\to\infty}\frac{n}{t} = \lim_{t\to\infty}\frac{n+1}{t+2}$$

## 4.3 Problem of Initial Data

Often the conditional probability distributions (81) are not defined by the model structure right from the beginning of observation, i.e. from $\tau = 1$ but only for $\tau > t_0 > 0$. For instance, the regression model

$$y_{(\tau)} = bu_{(\tau)} + ay_{(\tau-1)} + e_{(\tau)}$$

does not define (81) for $\tau = 1$ ($y_{(0)}$is missing, $t_0 = 1$). In such cases the formula (97) cannot be directly used as the first $t_0$ factors in the likelihood function (98) are not specified. This difficulty can be overcome in different ways, according to the prior information available.

To make the situation more transparent let us write the formula (94) for $t_1 = t_0$

$$p(\theta|D^{(t)}) = \frac{\tilde{L}_{(t)}(\theta, D^{(t)})p(\theta|D^{(t_0)})}{\int \tilde{L}_{(t)}(\theta, D^{(t)})p(\theta|D^{(t_0)})d\theta} \tag{109}$$

where

$$\tilde{L}_{(t)}(\theta, D^{(t)}) = \prod_{\tau=t_0+1}^{t} p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta) \tag{110}$$

Notice that the only difference between (110) and (98) is that in (110) the first $t_0$ factors are missing. When considered as a function of $\theta$ for given $D^{(t)}$, (110) may be called the conditional likelihood function. The adjective "conditional" is added to the likelihood function (110) because it is obtained from the joint probability distribution of all observed data except the initial data $D^{(t_0)}$ which are left in the condition. Often, see [17] and others, the conditional likelihood is introduced as the joint probability distribution of observed outputs conditioned on initial outputs and all inputs, $p(y^{(t)}_{(t_0+1)}|y^{(t_0)}, u^{(t)}, \theta)$. However, it does not make a clear sense if the inputs are allowed to be functions of previous outputs, not only of $y^{(t_0)}$. Notice that only the systematic application of Bayesian interpretation of probability and introduction of natural conditions of control fully justifies its usage. Moreover, it also shows that, except a small correction we are going to discuss, nothing better can be found, under natural conditions of control, of course.

According to the formula (109) the essence of our problem can be stated as follows: How the piece of information about the unknown parameters $\theta$, which is possibly contained in the initial data $D^{(t_0)}$, can be extracted? From practical point of view the question is not very important if the total amount of data $D^{(t)}$ is large compared to the initial data $D^{(t_0)}$, i.e. if $t >> t_0$ and the conditional likelihood (110) dominates. Then the information contained in $D^{(t_0)}$ can be neglected and the approximation

$$p(\theta|D^{(t_0)}) \approx p(\theta) \tag{111}$$

is well acceptable.

To throw more light on the approximation (111), we shall make use of, let us consider the Bayes formula relating the two probability distributions.

$$p(\theta|D^{(t_0)}) = \frac{p(D^{(t_0)}|\theta)p(\theta)}{\int p(D^{(t_0)}|\theta)p(\theta)d(\theta)} \tag{112}$$

Notice that the approximation (111) is an exact solution if the initial data $D^{(t_0)}$ can be considered as initial conditions of a stochastic difference equation which have nothing to do with the parameters. Then the probability distribution

$$p(D^{(t_0)}|\theta) = p(D^{(t_0)}) \tag{113}$$

can be brought out of the integral and cancelled in (112).

The Bayesian approach makes it possible to handle also the situations when the assumption (113) does not suit the given case. Then the prior information about the foregoing input-output data, $D_{(\tau)}(\tau < 1)$, which are not known but are required by the model for $\tau \leq t_0$, must be employed. We shall not follow this line as we consider it rather academic than of practical importance, at least in engineering and natural sciences. Instead of that we feel more appropriate to give a simple example which clearly shows that not much can be gained even when such a strong prior information, like stationarity of the process observed, is available.

Example 4.2 Consider an autonomous system the output process of which is describable by the auto-regression model of first order

$$y_{(\tau)} = ay_{(\tau-1)} + e_{(\tau)} \tag{114}$$

where the random component $e_{(\tau)}$, defined by (39), is assumed to be normal with constant and known variance $\sigma_e^2$. The parameter $a$ is unknown but it is a priori known that the system is stable and that at the moment when the observation starts, i.e. for $\tau = 1$ the output process has reached its stationarity. Hence, according to this prior information the parameter a must lie within the interval $S_a = (-1, +1)$ and

$$E[y_{(1)}] = E[y_{(0)}] = 0 \tag{115}$$

$$E[y_{(1)}^2|a] = E[y_{(0)}^2|a] = \frac{\sigma_e^2}{1-a^2} \tag{116}$$

As no other prior information about the true value of the unknown parameter is available it is appropriate to chose

$$p(a) = \frac{1}{2} \quad for \quad a \in S_a \tag{117}$$
$$p(a) = 0 \quad for \quad a \notin S_a$$

The model structure (114) defines the conditional likelihood (110) for $t_0 = 1$ and for the estimation formula we should know the probability distribution $p(a|y_{(1)})$. To investigate the relevancy of the approximation (111) we will calculate how the prior distribution (117) is modified by the single observation of $\underline{y}_{(1)}$.

According to (112) it holds

$$p(a|y_{(1)}) = \frac{p(y_{(1)}|a)\ p(a)}{\int p(y_{(1)}|a)\ p(a)da} \tag{118}$$

The mean value (115) and the variance (116) together with the assumed normality define

$$p(y_{(1)}|a) = -\frac{1}{\sqrt{2\pi}} \frac{\sqrt{1-a^2}}{\sigma_e} \exp\{-\frac{1-a^2}{2\sigma_e^2} y_{(1)}^2\} \tag{119}$$

Substitution of (119) and (117) into (118) gives for $a \in S_a$

$$p(a|y_{(1)}) = \frac{\sqrt{1-a^2}\exp\{-\frac{y_{(1)}^2}{\sigma_e^2}\frac{1-a^2}{2}\}}{\int_{-1}^{1}\sqrt{1-a^2}\exp\{-\frac{y_{(1)}^2}{\sigma_e^2}\frac{1-a^2}{2}\}da} \tag{120}$$

The probability density (120) is plotted for different ratios $\frac{y_{(1)}^2}{\sigma_e^2}$ in Fig. 2

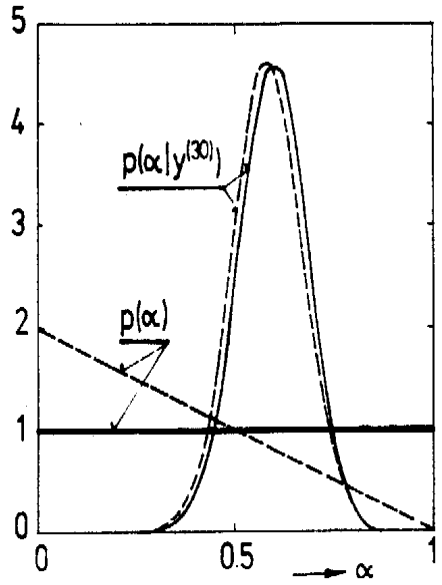



Fig. 2 Probability distribution for the unknown parameter $a$ in the model (114) after single observation $\underline{y}_{(1)} = y_{(1)}$

## 4.4 Non-informative Prior and Principle of Stable Estimation

The statistician, applying the logical system called Bayesian statistics to system identification, has to furnish the theory with three inputs:

(i) the structure of the system or process model defining, up to a finite set of parameters $\theta$, the family of conditional probability distributions

$$\{p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta), t_0 < \tau < t\} \tag{121}$$

(ii) the data and

(iii) the model of prior uncertainty of the unknown parameters $p(\theta)$ or more precisely $p(\theta|D^{(t_0)})$.

The latter has been, and still is, a matter of dispute and a heart of controversy between Bayesians and their opponents. It is true that, rather often then not, it is not easy to specify numerically and uniquely one's own state of mind in terms of prior probability distribution and that a certain degree of arbitrariness is present in any choice of this model. However, what mathematical model of a real world is not arbitrary, at least to some extent?

Engineers and natural scientists have a natural tendency to base their conclusions and decisions rather an objective measurements than on subjective and vague prior guess. This attitude is often expressed by the slogan "Let the data speak for themselves!".The endeavaour to make the Bayesian statistics free of prior and purely subjective probability distributions led to a number of studies on the so-called non-informative prior distributions. However, it turns out that it is impossible to give a satisfactory definition of "knowing nothing" and that a model of an" absolute ignorant", in fact, does not exist. The expression "non-informative"( as well as the concept of information in general) always has only a relative meaning

and all what can be done is to suggest a reasonable mathematical model of the situation when "little is known a priori" relatively to what the data can say and relatively to what they to speak about.

Fortunately, for large or medium length of observation, say for $t$ of order of several tens and more, if the data carry the information about the unknown parameters $\theta$ and the likelihood function is well peaked, then even a rather drastic modification of the prior distribution $p(\theta)$ does not significantly change the aposterior distribution $p(\theta|D^{(t)})$. This favorable fact is sometimes referred to, following [8] and [30], as the principle of "stable estimation" or "precise measurement".

Consider, for instance, the likelihood function (100) from Example 4.1, $L(\alpha, D^{(t)}) = \alpha^n(1-\alpha)^{t-n}$ and two rather different prior distributions $p(\alpha) = 1$ and $p(\alpha) = 2(1-\alpha)$. The resulting aposterior distributions,both for $t = 30$ and $n = 18$, are plotted in Fig. 3 for comparison.

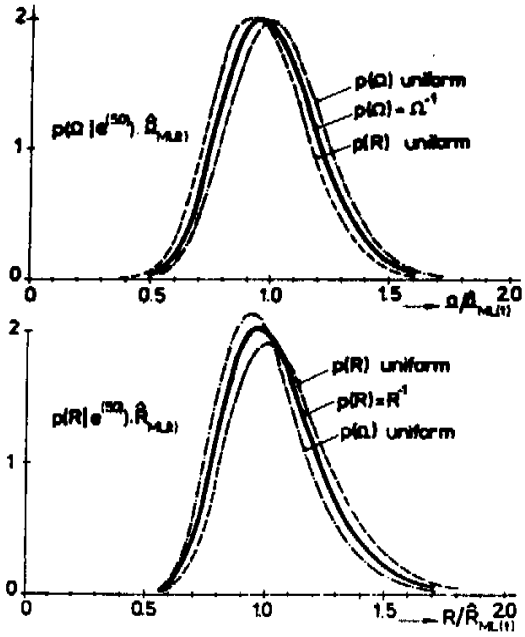



Fig. 3 Demonstration of the principle of stable estimation

The practical implication of the principle of stable estimation is that one does not need to worry much about the choice of the prior distribution and that any prior distribution which is flat relatively to the likelihood function is good enough. From what has been said follows that the uniform distribution could be recommended as a reasonable choice of prior distribution whenever the statistician's prior information about the unknown parameters is negligible relatively to the information which is expected to be provided by the intended experiment. However, such a recommendation must not be applied mechanically. There are two kinds of difficulties associated with uniform prior which have to be considered.

Strictly taken, the uniform probability distribution

$$p(\theta) = k, \quad \theta \in S_\theta \tag{122}$$

$$k = \frac{1}{\int_{S_\theta} d\theta} \tag{123}$$

can be introduced only on sets $S_\theta$ with a finite measure, i.e. when the integral in the denominator in (123) is finite. On the other hand, it is often more convenient when the aposterior distribution (97) is defined by a single formula on the entire Euclidian space $R^\lambda$ (here $\lambda$ is the number of unknown parameters in the set $\theta$) rather than on its subset $S_\theta \subset R^\lambda$. This difficulty is of technical nature and can be easily overcome if the integral of the likelihood function is finite

$$\int_{R^\lambda} L_{(t)}(\theta, D^{(t)}) d\theta < \infty \tag{124}$$

Then, according to (97), it holds

$$p(\theta|D^{(t)}) = \frac{L_{(t)}(\theta, D^{(t)})k}{\int_{S_\theta} L_{(t)}(\theta, D^{(t)})kd\theta} = \tag{125}$$

$$= \frac{L_{(t)}(\theta, D^{(t)})}{\int_{R^\lambda} L_{(t)}(\theta, D^{(t)})d\theta - \int_{R^\lambda - S_\theta} L_{(t)}(\theta, D^{(t)})d\theta}$$

and, as a limit for $S_\theta \to R^\lambda$, the aposterior distribution is obtained in the form of the standardized likelihood

$$p(\theta|D^{(t)}) = \frac{L_{(t)}(\theta, D^{(t)})}{\int L_{(t)}(\theta, D^{(t)})d\theta} \tag{126}$$

[8] investigated the influence of modifications in the prior distribution on the aposterior distribution and established a theorem (see also [7], *par.*10.4 which relates quantitatively the standardized likelihood (126) to the aposterior distribution based on a more carefully chosen prior.

The second difficulty associated with the uniform prior distribution is more substantial. The parameterisation of the model defining the family of conditional probability distributions (121) is often not unique and the same system can be equally well characterized by two different sets of parameters, say $\theta$ and $\tilde{\theta}$, which are related by a regular (one-to-one) transformation

$$\tilde{\theta}_i = \mu_i(\theta), \qquad i = 1, 2, ..., \lambda \tag{127}$$

where $\tilde{\theta}_i$ is the i-th member of the set $\tilde{\theta}$. If the uncertainty of the parameters $\tilde{\theta}$ is described by the probability density

$$p(\tilde{\theta}) = \phi_{\tilde{\theta}}(\tilde{\theta}) \tag{128}$$

then the probability density for the parameter set $\theta$ is determined by the relation (see any course of probability theory)

$$p(\theta) = \phi_{\tilde{\theta}}(\mu(\theta))|J_\mu(\theta)| \tag{129}$$

where $|J_\mu(\theta)|$ is the absolute value of the determinant (Jacobian) of the transformation (127)

$$J_\mu(\theta) = \left|\frac{\mathcal{D}\mu(\theta)}{\mathcal{D}\theta}\right| = \begin{vmatrix} \frac{\partial\mu_1(\theta)}{\partial\theta_1}, & \cdots, & \frac{\partial\mu_1(\theta)}{\partial\theta_\lambda} \\ \vdots & & \vdots \\ \frac{\partial\mu_\lambda(\theta)}{\partial\theta_1}, & \cdots, & \frac{\partial\mu_\lambda(\theta)}{\partial\theta_\lambda} \end{vmatrix} \tag{130}$$

The Jacobian is, in general, a function of the parameters $\theta$ and therefore, as it is seen from the relation (129) the prior probability density which is uniform with respect to $\tilde{\theta}$, i.e. $\phi(\tilde{\theta}) = k$, is not neccessarily uniform with respect to $\theta$ and inversely. The statistician, who wants to model his "knowing nothing" by a uniform prior distribution, has to choose such a parameterisation of his model which corresponds to his lack of prior information.

The case often met in practical applications, in which confusion may occur, is the unknown covariance matrix of the normal distribution (51). Instead of the covariance matrix $R$ it is equally well possible (and usually more convenient) to consider the precision matrix [7]

$$\Omega = R^{-1} \tag{131}$$

In this parameterisation the probability density (51) reads

$$p(e_{(t)}|\Omega) = (2\pi)^{-\frac{\nu}{2}}|\Omega|^{\frac{1}{2}}\exp\{-\frac{1}{2}e_{(t)}^T\Omega e_{(t)}\}$$

where $\nu$ is the dimension of $e_{(t)}$. The symmetric matrices $R$ and $\Omega$ consist of $\lambda = \frac{1}{2}\nu(\nu+1)$ distinct elements and are related by the one-to-one transformation (131) the Jacobian of which is (see e.g. [4], appendix A 8.1)

$$J(\Omega) = |\frac{\mathcal{D}R}{\mathcal{D}\Omega}| = |\Omega|^{-(\nu+1)} \tag{132}$$

Hence, the probability densities $p(R) = \phi_R(R)$ and $p(\Omega) = \phi_\Omega(\Omega)$ are related as follows

$$\phi_\Omega(\Omega) = \phi_R(\Omega^{-1})|\Omega|^{-(\nu+1)}$$

$$\phi_R(R) = \phi_\Omega(R^{-1})|R|^{-(\nu+1)}$$

and one may hesitate which one should be chosen as uniform if one wants to be "objective". The compromise

$$p(\Omega) = |\Omega|^{-\frac{\nu+1}{2}}, \quad p(R) = |R|^{-\frac{\nu+1}{2}} \tag{133}$$

is the most often made choice the justification of which can based on different grounds (see [4]; [16]; [26]. The Example 4.3 appended to this subsection may help the reader to get a quantitative idea how the three different choices; $p(R)$ uniform, $p(\Omega)$ uniform and the compromise (133) may influence the aposterior distribution in the case of medium data size.

Summing up we can see that, in the lack of prior information, the basic formula for one-shot parameter estimation (109) can be applied when formally, but only formally, the prior probability distribution $p(\theta|D^{(t_0)})$ is substituted by

$$p(\theta)|D^{(t_0)} \approx p(\theta) \approx kJ_\mu(\theta) \tag{134}$$

where $k$ is an arbitrary constant and $J_\mu(\theta)$ is the Jacobian of the transformation (127) between the parameter space $S_{\tilde{\theta}}$ on which the probability is distributed uniformly and the parameter space $S_\theta$ considered in the estimation problem.

The probability densities of the type (134) usually do not fulfill the basic property of probability distributions (4), i.e. they do not integrate to one over $R^\lambda$. Because of this deficiency they are sometimes called improper prior distributions. However it should be emphasized that caution must be exercised when dealing with improper prior. They can be employed in estimation problems only when the integral in the denominator of (112) is finite and only in the sense of the limit we applied to obtain the formula (126) from (125)

We will conclude this discussion on the arbitrariness in the choice of the prior distributions by quoting two opinions which seem very reasonable. "In applied (as opposed to pure) mathematics, arbitrariness is in-admissible only in so far as it produces results outside acceptable limits of approximation", [4]. "Any theory that pretends to produce exactness where it is unjustified is a false servant", [31]. Perhaps, it should be recalled that what has been said concerns only one of the inputs which has to be supplied to the theory by the user. The logical structure of the theory itself does not leave any space for arbitrariness, produces sensible results whenever the inputs are sensible and provides insight where common sense fails. Nevertheless, in order to be able to exploit all potentialities of the Bayesian theory we have to learn more how to construct models of our prior uncertainties in particular situations.

Example 4.3 The purpose of this example is to demonstrate how three different prior probability distributions may influence the estimation of an unknown variance.

Consider discrete uni-variate and normal white noise with unknown variance $\sigma^2 = R$. The variance, or equivalently $\Omega = \frac{1}{R}$, has to be estimated from $t = 50$ samples.

If $\Omega$ is considered as the unknown parameter, then the corresponding likelihood function reads

$$L_{(t)}(\Omega, e^{(t)}) = (2\pi)^{-\frac{t}{2}}\Omega^{\frac{t}{2}}\exp\{-\frac{\Omega}{2}\sum_{\tau=1}^{t}e_{(\tau)}^2\}$$

and the maximum-likelihood point estimate is

$$\hat{\Omega}_{ML(t)} = \frac{1}{\hat{R}_{ML(t)}} = \frac{t}{\sum_{\tau=1}^{t} e_{(\tau)}^2}$$

We shall consider the following three different prior distributions

(a) $p(\Omega)$ uniform, $p(R) = \frac{1}{R^2}$

(b) $p(R)$ uniform, $p(\Omega) = \frac{1}{\Omega^2}$

(c) $p(\Omega) = \frac{1}{\Omega}, \quad p(R) = \frac{1}{R}$

Application of the formula (97) gives the aposterior probability distribution for $\Omega$ in the form of gamma-distribution which can be brought into the following form

$$p(\Omega|e^{(t)}) = \hat{\Omega}_{ML(t)}^{-1} \frac{(\frac{t}{2})^{\frac{t}{2}+1-m}}{\Gamma(\frac{t}{2}+1-m)} \xi^{\frac{t}{2}-m} e^{-\xi\frac{t}{2}} \tag{135}$$

where

$$\xi = \frac{\Omega}{\hat{\Omega}_{ML(t)}}$$

and $m = 0$ in the case (a), $m = 2$ in the case (b) and $m = 1$ in the case (c)

Using the relation (129) we also have

$$p(R|e^{(t)}) = \hat{R}_{ML(t)}^{-1} \frac{(\frac{t}{2})^{\frac{t}{2}+1-m}}{\Gamma(\frac{t}{2}+1-m)} \mu^{\frac{t}{2}+2-m} e^{-\mu\frac{t}{2}} \tag{136}$$

where

$$\mu = \frac{\hat{R}_{ML(t)}}{R}.$$

The probability distributions (135) and (136) multiplied by $\hat{\Omega}_{ML(t)}$ and $\hat{R}_{ML(t)}$, respectively, are plotted in Fig. 4 for $t = 50$ and for the three priors considered $m = 0, 1, 2$.

## 4.5 Redundant and Non-identifiable Parameters

It is necessary to emphasize that the principle of stable estimation is not a generally valid principle. It applies only when the data really carry the information about the parameters which are to be estimated. It does not apply in the cases of redundant, non-identifiable or weakly identifiable parameters which will be discussed now.

Consider a regular transformation of parameters

$$\overline{\theta} = H(\theta)$$

If the set of transformed parameters $\overline{\theta}$ can be decomposed into two subsets

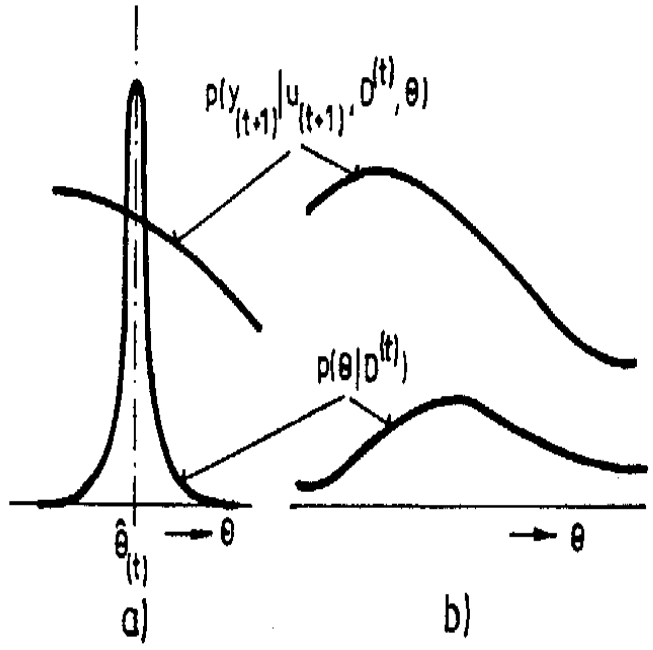



Fig. 4 Estimation of unknown variance $R$ or precision measure $\Omega = R^{-1}$ influence of different priors after 50 observations

$$\begin{aligned} \overline{\theta} &= \{\overline{\theta_a}, \overline{\theta_b}\} \\ \overline{\theta}_a &= H_a(\theta) \\ \overline{\theta}_b &= H_b(\theta) \end{aligned} \tag{137}$$

so that

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta) = p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \overline{\theta}_a), \quad \tau > t_0 \tag{138}$$

i.e. the conditional probability distribution (138) depends, for any $u_{(\tau)}$ and $D^{(\tau-1)}$, only on $\overline{\theta}_a$, but not on $\overline{\theta}_b$, then the subset of $\theta$, determined as the image of $\overline{\theta}_b$ defined by the mapping (137), will be called input-output redundant, or simply redundant.

From (138) it follows that also the conditional likelihood (110) does not depend on the subset of redundant parameters $\overline{\theta}_b$

$$\tilde{L}_{(t)}(\theta, D^{(t)}) = \tilde{L}_{(t)}(\overline{\theta}_a, D^{(t)}) \quad for\ any\ \overline{\theta}_b \tag{139}$$

and according to (109) it holds

$$p(\overline{\theta}|D^{(t)}) = \frac{\tilde{L}_{(t)}(\overline{\theta}_a, D^{(t)}) p(\overline{\theta}_a, \overline{\theta}_b|D^{(t_0)})}{\int \tilde{L}_{(t)}(\overline{\theta}_a, D^{(t)})[\int p(\overline{\theta}_a, \overline{\theta}_b|D^{(t_0)}) d\overline{\theta}_b] d\overline{\theta}_a} =$$

$$= \frac{\tilde{L}_{(t)}(\overline{\theta}_a, D^{(t)}) p(\overline{\theta}_a, |D^{(t_0)})}{\int \tilde{L}_{(t)}(\overline{\theta}_a, D^{(t)}) p(\overline{\theta}_a, |D^{(t_0)}) d\overline{\theta}_a} p(\overline{\theta}_b|\overline{\theta}_a, D^{(t_0)})$$

This shows that only the marginal prior distribution $p(\overline{\theta}_a|D^{(t_0)})$ is corrected by the observed data

$$p(\overline{\theta}_a|D^{(t)}) = \frac{\tilde{L}_{(t)}(\overline{\theta}_a, D^{(t)})p(\overline{\theta}_a, |D^{(t_0)})}{\int \tilde{L}_{(t)}(\overline{\theta}_a, D^{(t)})p(\overline{\theta}_a, |D^{(t_0)})d\overline{\theta}_a}$$

while the conditional prior $p(\overline{\theta}_b|\overline{\theta}_a, D^{(t_0)})$ remains unchanged

$$p(\overline{\theta}_b|\overline{\theta}_a, D^{(t)}) = p(\overline{\theta}_b|\overline{\theta}_a, D^{(t_0)}) \tag{140}$$

Now we shall investigate how the redundancy in estimated parameters may influence the prediction. As a regular transformation of variables does not change the integral, the prediction (93)

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \int p(y_{(t+1)}|u_{(t+1)}, D^{(t)}, \theta)\ p(\theta|D^{(t)})d\theta \tag{141}$$

can be written as follows

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \tag{142}$$

$$= \int p(y_{(t+1)}|u_{(t+1)}, D^{(t)}, \overline{\theta_a})\ p(\overline{\theta_a}|D^{(t)})[\int p(\overline{\theta_b}|\overline{\theta_a}, D^{(t)})d\overline{\theta_b}]d\overline{\theta_a} =$$

$$= \int p(y_{(t+1)}|u_{(t+1)}, D^{(t)}, \overline{\theta_a})\ p(\overline{\theta_a}|D^{(t)})d\overline{\theta_a}$$

The equality between (141) and (142) clearly shows that the redundancy in identified parameters does not effect the Bayesian prediction. Put in other words, the same predictive probability distribution is obtained whether the parameter set is reduced by excluding the redundant parameters after a suitable transformation or when the redundancy is ignored. Practically it means that canonical and non-canonical forms of input-output models are equally good for the purpose of prediction and control of the output.

This conclusion indicates that the similar can be expected when the set of parameters $\theta$ contains some subset on which the probability distribution $p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta)$, and consequently also the conditional likelihood, depends only weakly. In such a case the data carry only little information about this subset of parameters, the subset is difficult to identify from medium data size and any point estimate of $\theta$ may be very unreliable. However, it does not necessarily mean that the prediction of the output is unreliable, too.

Non-identifiability, or weak identifiability, of model parameters may be caused not only by the existence of a redundant, or almost redundant subset of parameters but also by the way how the input of the system is generated during the experiment.

Usually, the concept of identifiability is introduced and treated as the question of consistency of certain point estimates of parameters. As pointed out by [20](Section 4.1) the sense of this concept is to test the identification methods (i.e.different constructions of point estimators) on artificial systems which can be exactly described by the given model. However, the test on consistency of point estimates is relevant only for $t \to \infty$ and does not guarantee that the unknown parameter values can be replaced by their point estimates also for the finite data size available. Bayesian approach does not operate with point estimates and therefore does not rely on such a tool. By using the Bayesian approach that information about the unknown quantities is extracted which is contained in the data - of course, also under the assumption that the assumed model structure is a suitable representation of reality - and this information is presented in the form of the aposterior distribution for further use. If the data do not carry information about some subset of unknown parameters, then it can be recognized from the form of this distribution. Sometimes it may be difficult to investigate the entire form of the aposterior distribution. Nevertheless, if the parameter values are of final interest, it must be recommended to investigate at least the vicinity of the point in which the distribution reaches its maximum (as outlined in Example 4.4 in order to check whether there does not exist some "ridge"in the aposterior distribution along which the system parameters have not been identified or have been only weakly identified.

The formula (141) indicates under what conditions the unknown parameters can be simply replaced by their point estimate. Consider the probability distribution $p(y_{t+1}|u_{t+1}, D^{(t)}, \theta)$ as a function of $\theta$ for

given $y_{(t+1)}$ and any but fixed $\{u_{(t+1)}, D^{(t)}\}$, as shown in Fig. 5a it is evident that at good approximation of the integral (141) can be obtained if the variable in the first factor of the integrated function is simply replaced by some reasonable point estimate $\hat{\theta}_t$

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = p(y_{(t+1)}|u_{(t+1)}, D^{(t)}, \theta)|_{\theta=\hat{\theta}_{(t)}} \tag{143}$$

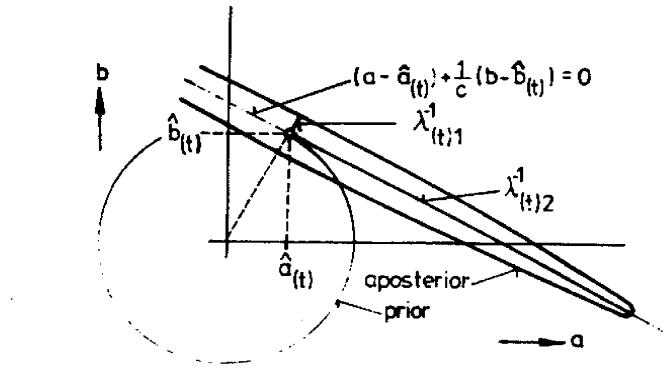



Fig. 5 Two extreme situation in Bayesian prediction - two factors in the integral (141) plotted as functions of $\theta$ for $y_{(t+1)}$, $u_{(t+1)}$ and $D^{(t)}$ fixed

However, if the situation is like Fig. 5b, the approximation (143) does not hold and the integration in (141) has to be performed. Unfortunately, it is usually not easy to recognize what situation occurs without a more detailed investigation. Moreover, due to an insufficient excitation of the system by the input signal, it may well happen that for some $u_{(t+1)}$ the situation is like in Fig. 5a while for an other $u_{(t+1)}$ like in Fig. 5b (see Example 4.4) This is, in fact, the reason why the duality of control actions is required when controlling a system with uncertain parameters.

Example 4.4 The purpose of this Example is to demonstrate on a simple case the influence of a time-invariant feedback on the estimation of system parameters.

Consider a system describable by the normal regression model

$$y_{(\tau)} = ay_{(\tau-1)} + bu_{(\tau)} + e_{(\tau)}$$

or equivalently by the conditional probability density

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta) = (2\pi)^{-\frac{1}{2}}\sigma_e^{-1}\exp\{-\frac{1}{2\sigma_e^2}(y_{(\tau)} - ay_{(\tau-1)} - bu_{(\tau)})^2\} \tag{144}$$

For the sake of simplicity let us assume that the variance of $e_{(\tau)}$, $\sigma_e^2$, is known and that only the regression coefficients $\underline{a}$ and $\underline{b}$ are unknown. Hence, the set of unknown parameters $\underline{\theta} = \{\underline{a}, \underline{b}\}$

Suppose that the input is derived from the previous output by the feedback

$$u_{(\tau)} = cy_{(\tau-1)} + v_{(\tau)} \tag{145}$$

where $\{v_{(\tau)}\}$ is sequence of random variables with zero mean and constant variance $\sigma_v^2$, which are uncorrelated with the foregoing inputs and outputs. The feedback gain $c$ as well as the variance $\sigma_v^2$ can be known or unknown, but it is assumed that these parameters have no relation to the unknown parameters $\underline{b}$ and $\underline{a}$, which could be exploited by the statistician. Hence, then natural conditions of control are satisfied for $\tau > 1$ and the formula (109) can be applied.

Suppose that the statistician identifying the system is very uncertain about the parameters $\underline{a}$ and $\underline{b}$, he has no idea whether their true could be positive or negative, and therefore he chooses the prior

distribution in the normal form with zero mean and very large variance

$$\sigma_a^2 = \sigma_b^2 = \sigma_\theta^2 \tag{146}$$

$$p(\theta|y_{(1)}) = p(a,b|y_{(1)}) = p(a,b) = p(a)p(b) = (2\pi)^{-1}\sigma_\theta^{-2}\exp\{-\frac{1}{2\sigma_\theta^2}(a^2+b^2)\}$$

To obtain the conditional likelihood (110) in a convenient form it is suitable to rewrite the exponent of the probability density (144) as follows

$$(y_{(\tau)} - ay_{(\tau-1)} - bu_{(\tau)})^2 = \begin{bmatrix} -1 \\ a \\ b \end{bmatrix}^T \begin{bmatrix} y_{(\tau)} \\ y_{(\tau-1)} \\ u_{(\tau)} \end{bmatrix} \begin{bmatrix} y_{(\tau)} \\ y_{(\tau-1)} \\ u_{(\tau)} \end{bmatrix}^T \begin{bmatrix} -1 \\ a \\ b \end{bmatrix}$$

Then it is easily seen that

$$\tilde{L}_{(t)}(a,b,D^{(t)})p(a,b|y_{(1)}) = (2\pi)^{-\frac{t+1}{2}}\sigma_e^{-(t-1)}\sigma_\theta^{-2}\times \tag{147}$$

$$\times\exp\{-\frac{t-1}{2\sigma_e^2}\begin{bmatrix} -1 \\ a \\ b \end{bmatrix}^T \begin{bmatrix} V_{o(t)} & V_{a(t)} & V_{b(t)} \\ V_{a(t)} & V_{aa(t)} & V_{ab(t)} \\ V_{b(t)} & V_{ab(t)} & V_{bb(t)} \end{bmatrix} \begin{bmatrix} -1 \\ a \\ b \end{bmatrix}\}$$

where

$$V_{0(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau)}^2 \tag{148}$$

$$V_{a(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau)}y_{(\tau-1)} \tag{149}$$

$$V_{b(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau)}u_{(\tau)} \tag{150}$$

$$V_{aa(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau-1)}^2 + \epsilon_{(t)}^2 \tag{151}$$

$$V_{bb(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} u_{(\tau)}^2 + \epsilon_{(t)}^2 \tag{152}$$

$$V_{ab(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau-1)}u_{(\tau)} \tag{153}$$

$$\epsilon_{(t)}^2 = \frac{\sigma_e^2}{(t-1)\sigma_\theta^2} \tag{154}$$

Notice that $\epsilon_{(t)}^2$, appearing in (151) and (152), is a very small number due to large $\sigma_\theta^2$ and is getting even smaller with growing $t$.

The required aposterior probability distribution is obtained by substitution of (147) into the general formula (109), the evaluation of which requires some algebraic rearrangement of the exponent.

$$p(\theta|D^{(t)}) = \kappa_{(t)}\exp\{-\frac{t-1}{\sigma_e^2}(\theta-\hat{\theta}_{(t)})^T V_{\theta\theta(t)}(\theta-\hat{\theta}_{(t)})\} \tag{155}$$

where

$$\theta = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$V_{\theta\theta(t)} = \begin{bmatrix} V_{aa(t)} & V_{ab(t)} \\ V_{ab(t)} & V_{bb(t)} \end{bmatrix} \tag{156}$$

$$\hat{\theta}_{(t)} = \begin{bmatrix} \hat{a}_{(t)} \\ \hat{b}_{(t)} \end{bmatrix} = V_{\theta\theta}^{-1} \begin{bmatrix} V_{a(t)} \\ V_{b(t)} \end{bmatrix} \tag{157}$$

$$\kappa_{(t)} = (2\pi)^{-1} \frac{t-1}{\sigma_e^2} |V_{\theta\theta(t)}|^{\frac{1}{2}}$$

Apparently, the maximum of the probability density (155) lies in the point $\theta = \hat{\theta}_{(t)}$, i.e. for $a = \hat{a}_{(t)}$, $b = \hat{b}_{(t)}$.

$$\hat{a}_{(t)} = \frac{V_{a(t)}V_{bb(t)} - V_{ab(t)}V_{b(t)}}{\Delta_{(t)}} \tag{158}$$

$$\hat{b}_{(t)} = \frac{V_{b(t)}V_{aa(t)} - V_{ab(t)}V_{a(t)}}{\Delta_{(t)}} \tag{159}$$

$$\Delta_{(t)} = |V_{\theta\theta_{(t)}}| = V_{aa(t)}V_{bb(t)} - V_{ab(t)}^2 \tag{160}$$

In order to characterize the shape of the aposterior distribution more fully it is appropriate to determine also the eigenvalues and the directions of the corresponding eigenvectors of the matrix

$$\begin{bmatrix} \frac{\partial^2 p(a,b|D^{(t)})}{\partial^2 a}, & \frac{\partial^2 p(a,b|D^{(t)})}{\partial a \partial b} \\ \frac{\partial^2 p(a,b|D^{(t)})}{\partial a \partial b}, & \frac{\partial^2 p(a,b|D^{(t)})}{\partial^2 b} \end{bmatrix}$$

for $a = \hat{a}_{(t)}$, $b = \hat{b}_{(t)}$.

In the given simple case this matrix is proportional to $V_{\theta\theta(t)}$, (156), its eigenvalues

$$\lambda_{(t),1,2} = \frac{1}{2}\left[V_{aa(t)} + V_{bb(t)} \pm \sqrt{(V_{aa(t)} - V_{bb(t)})^2 + 4V_{ab(t)}^2}\right] \tag{161}$$

determine (by their reciprocal values) the main axes of the ellipse

$$(\theta - \hat{\theta}_{(t)})^T V_{\theta\theta(t)} (\theta - \hat{\theta}_{(t)}) = constant$$

and the eigenvector, corresponding to the smaller eigenvalue, determines the direction along which this ellipse is situated.

$$b - \hat{b}_{(t)} = k_{(t)}(a - \hat{a}_{(t)}) \tag{162}$$

$$k_{(t)} = \frac{V_{bb(t)} - V_{aa(t)}}{2V_{ab(t)}} - \sqrt{(\frac{V_{bb(t)} - V_{aa(t)}}{2V_{ab(t)}})^2 + 1} \tag{163}$$

Clearly, the straight line (162) and the ratio of the two eigenvalues (161) determine the possible "ridge" we are interested in.

The predictive probability distribution can be obtained by substitution of (144) and (155) into general formula (93)

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = (2\pi)^{-\frac{1}{2}} \sigma_{y(t+1)}^{-1} \exp[-\frac{(y_{(t+1)} - \hat{y}_{(t+1)})^2}{2\sigma_{y(t+1)}^2}] \tag{164}$$

where the conditional mean value is

$$\hat{y}_{(t+1)} = \hat{a}_{(t)} y_{(t)} + \hat{b}_{(t)} u_{(t+1)} \tag{165}$$

and the conditional variance is

$$\sigma_{y(t+1)}^2 = \sigma_e^2 (1 + \xi_{(t+1)}) \tag{166}$$

$$\zeta_{(t+1)} = [y_{(t)}, u_{(t+1)}] \frac{V_{\theta\theta}^{-1}}{t-1} \begin{bmatrix} y_{(t)} \\ u_{(t+1)} \end{bmatrix} = \tag{167}$$

$$= \frac{1}{t-1}[\frac{y_{(t)}^2}{V_{aa(t)}} + \frac{V_{aa(t)}}{\Delta(t)}(u_{(t+1)} - \frac{V_{ab(t)}}{V_{aa(t)}}y_{(t)})^2]$$

Notice, that the variance of the prediction may strongly depend on the input $u_{(t+1)}$ applied.

Now, we shall evaluate these characteristics for the given feedback (145). Two distinct cases will be considered in particular. First, the case a purely deterministic feedback, i.e. $\sigma_v^2 = 0$, will be analyzed for any finite $t$. Second, it will be shown what happens if $\sigma_v^2 > 0$ and $t$ is relatively large.

Substitution of the deterministic feedback, i.e. (145) for $v_{(\tau)} = 0$, into the formulae (150), (152) and (153) gives

$$V_{b(t)} = c\phi_{y(t)}$$
$$V_{bb(t)} = c^2\rho_{y(t)} + \epsilon_{(t)}^2$$
$$V_{ab(t)} = c\rho_{y(t)}$$

where $\rho_{y(t)}$ is the sample variance of the output

$$\rho_{y(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau-1)}^2 \tag{168}$$

and $\phi_{y(t)}$ is the sample auto-correlation

$$\phi_{y(t)} = \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau)}y_{(\tau-1)} \tag{169}$$

With the notation (168) and (169) we also have

$$V_{a(t)} = \phi_{y(t)}$$
$$V_{aa(t)} = \rho_{(t)} + \epsilon_{(t)}^2$$

and the characteristics of the aposterior distribution for the unknown parameters $\underline{a}$ and $\underline{b}$ are

$$\hat{a}_{(t)} = \frac{\phi_{y(t)}}{(c^2+1)\rho_{y(t)} + \epsilon_{(t)}^2} \tag{170}$$

$$\hat{b}_{(t)} = c\hat{a}_{(t)} \tag{171}$$
$$\lambda_{(t)1} = (1+c^2)\rho_{y(t)} + \epsilon_{(t)}^2$$
$$\lambda_{(t)2} = \epsilon_{(t)}^2$$
$$k_{(t)} = -\frac{1}{c}$$

Notice that the ratio $\frac{\lambda_{(t)2}}{\lambda_{(t)1}}$ is extremely small and tends to zero with growing $t$. This means that the aposterior probability is concentrated at the straight line (162), as sketched in Fig. 6, and is distributed along this line almost uniformly. Any point on this line is practically equally probable as the point of theoretical maximum (170) and (171).

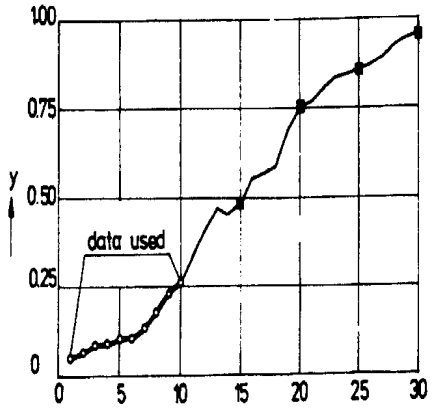



Fig. 6 Effect of a time-invariant feed-back on estimation of parameters

This fact has a drastic influence on the predictive distribution (164), the variance, of which, according to (166) and (167), is (small terms of higher order neglected)

$$\sigma^2_{y(t+1)} = \sigma^2_e(1 + \frac{y^2_{(t)}}{(t-1)\rho_{y(t)}}) + \sigma^2_\theta(u_{(t+1)} - cy_{(t)})^2$$

This clearly shows that even a very small deviation from the deterministic feedback $u_{(t+1)} = cy_{(t)}$ will cause an immense increase of variance in the prediction of the future output. However, for the feedback fixed, as it was in the past, the prediction is still reliable.

Now, we shall briefly discuss the case when the variance of the feedback noise $\sigma^2_v$ is nonzero. For a very flat prior distribution, i.e. for $\sigma^2_\theta \to \infty$ the point in which the maximum of the aposterior probability density is located, is equal to the prediction-error least-square point estimate and at the same time to the maximum likelihood estimate of the unknown parameters $\underline{a}$ and $\underline{b}$. It can be shown that, for arbitrarily small but nonzero $\sigma^2_v$, this point estimate is consistent, i.e. for $t \to \infty$, $\hat{a}_{(t)} \to \underline{a}, \hat{b}_{(t)} \to \underline{b}$ with probability one. However, this theoretical result is of little practical value if the variance of the feedback noise, $\sigma^2_v$, is small relatively to the sample variance of the output $\rho_{y(t)}$. We shall show that in such a case the situation is not much different from the previous one, even for a large data size.

From (145) it follows

$$\frac{1}{t-1}\sum_{\tau=2}^{t} u^2_{(\tau)} = c^2\frac{1}{t-1}\sum_{\tau=2}^{t} y^2_{(\tau-1)} + \frac{1}{t-1}\sum_{\tau=2}^{t} v^2_{(\tau)} + 2c\frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau-1)}v_{(\tau)} \tag{172}$$

For large $t$ it holds with good approximation

$$\frac{1}{t-1}\sum_{\tau=2}^{t} v^2_{(\tau)} \approx \sigma^2_v, \quad \frac{1}{t-1}\sum_{\tau=2}^{t} y_{(\tau-1)}v_{(\tau)} \approx 0, \quad \epsilon^2_{(t)} = \frac{\sigma^2_e}{(t-1)\sigma^2_a} \approx 0$$

Then $V_{aa(t)}$, defined by (151), is equal to the sample variance of the output (166)

$$V_{aa(t)} = \rho_{y(t)} \tag{173}$$

and (172) can be written in the following way

$$V_{bb(t)} = \rho_{y(t)}(c^2 + \delta_{(t)}) \tag{174}$$

where

$$\delta_{(t)} = \frac{\sigma^2_v}{\rho_{y(t)}} \tag{175}$$

Similarly, from the equation of the feedback (145) follows

$$V_{ab(t)} = c\rho_{y(t)} \tag{176}$$

Substitution of (173), (174) and (176) into the general formulae derived above gives

$$\frac{\lambda_{2(t)}}{\lambda_{1(t)}} = \frac{1 - \sqrt{1 - (\frac{2\delta_{(t)}}{1+c^2+\delta_{(t)}})^2}}{1 + \sqrt{1 - (\frac{2\delta_{(t)}}{1+c^2+\delta_{(t)}})^2}}$$

$$k_{(t)} = \frac{1}{2c}(c^2 + \delta_{(t)} - 1 - \sqrt{(c^2 + \delta_{(t)} + 1)^2 - 4\delta_{(t)}})$$

and for small $\delta_{(t)}$

$$\frac{\lambda_{2(t)}}{\lambda_{1(t)}} \approx (\frac{\delta_{(t)}}{1+c})^2 \tag{177}$$

$$k_{(t)} \approx -\frac{1}{c}(1 - \frac{\delta_{(t)}}{1+c^2}) \tag{178}$$

The small ratio (177) indicates that the parameters are weakly identified along the line the direction of which is determined by (178). The prediction variance (166) is given in this case for any $\delta_{(t)}$ by the formula

$$\sigma^2_{y(t+1)} = \sigma^2_e[1 + \frac{y^2_{(t)}}{(t-1)\rho_{y(t)}} + \frac{(u_{(t+1)} - cy_{(t)})^2}{(t-1)\sigma^2_v}] \tag{179}$$

which again shows, that the prediction may be very uncertain when the input $u_{(t+1)}$ deviates from the feedback law applied during the identification experiment.

It may be recommended to compare this analysis with a similar example given by [20], Example 4.1

## 4.6 Real-Time Estimation and Prediction

The recursive relation for real-time estimation can be obtained by setting $t_1 = t - 1$ into the general formula (94)

$$p(\theta|D^{(t)}) = \frac{p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta)\ p(\theta|D^{(t-1)})}{\int p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta)\ p(\theta|D^{(t-1)})d\theta} \tag{180}$$

According to (141), the denominator in (180) can be understood as Bayesian one-step-ahead prediction, which shows that the real-time estimation of unknown parameters can be decomposed into the following two steps.

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = \int p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta)\ p(\theta|D^{(t-1)})d\theta \tag{181}$$

$$p(\theta|D^{(t)}) = \frac{p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta)}{p(y_{(t)}|u_{(t)}, D^{(t-1)})}\ p(\theta|D^{(t-1)}) \tag{182}$$

In the first step (181) the one-step-ahead prediction (probability distribution for the next output $y_{(t)}$) is determined using the old "estimate" of parameters $p(\theta|D^{(t-1)})$. When the new input $\underline{u}_{(t)}$ is decided and the true value of the new output $\underline{y}_{(t)}$ is obtained then this new data pair $\underline{D}_{(t)} = \{\underline{u}_{(t)}, \underline{y}_{(t)}\}$ substituted into the probability distribution $p(y_{(t)}|u_{(t)}, D^{(t-1)})$ gives just one number which is used to calculate, for every possible value $\theta$ of the unknown parameters $\underline{\theta}$ the factor

$$g(\theta) = \frac{p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta)}{p(y_{(t)}|u_{(t)}, D^{(t-1)})} \tag{183}$$

Multiplication by this factor updates the probability distribution for the unknown parameters according to the second step of the recursion (182).

The functional recursion (181) and (182) applies for any process model defining the conditional probability distribution $p(y_{(t)}|u_{(t)}, D^{(t-1)}, \theta)$ but, in general, it may be very difficult to perform this calculation numerically as the entire probability distribution over all possible values of the unknown parameters $\theta$ (a numerical table which can be of very high dimension) has to be recalculated for each $t$. However, if such a form of the probability distribution $p(\theta|D^{(t-1)})$ can be found that remains unchanged, up to a finite set of its parameters, when $t$ is growing, then the functional recursion can be reduced to an algebraic recursion, performed only on the parameters of the distribution, which considerably simplifies the calculation. The probability distributions having this property are sometimes called reproducing, or one says that they form a conjugate family of distributions [27], [7]. This special, but practically important class of probability distributions will be treated in the next subsection.

If the purpose of system identification is to predict the output and the model parameters are not of direct interest, then it is natural to pose the question whether and under what conditions it would be possible to omit the estimation of model parameters and to update directly the conditional probability distribution for the next output. To answer this question let us express $p(\theta|D^{(t)})$ the probability distribution for the output prediction (93) through the general formula (109)

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \frac{\int p(y_{(t+1)}|u_{(t+1)}, D^{(t)}, \theta)\tilde{L}_{(t)}(\theta, D^{(t)})p(\theta|D^{(t)})d\theta}{\int \tilde{L}_{(t)}(\theta, D^{(t)})p(\theta|D^{(t_0)})d\theta} \tag{184}$$

However, according to the definition of the conditional likelihood (110) it holds

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}, \theta)\tilde{L}_{(t)}(\theta, D^{(t)}) = \tilde{L}_{(t+1)}(\theta, D^{(t+1)})$$

which suggests that the predictive probability distribution (184) can be expressed as a ratio of two integrals

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \frac{I_{(t+1)}(D^{(t+1)})}{I_{(t)}(D^{(t)})} \tag{185}$$

where

$$I_{(t)}(D^{(t)}) = \int \tilde{L}_{(t)}(D^{(t)}, \theta)p(\theta|D^{(t_0)})d\theta \tag{186}$$

Of course, the upper integral in (185) has to be understood as a function of $y_{(t+1)}$ and $u_{(t+1)}$ which have not been determined yet, and therefore it may be more appropriate to write (185) in the following form

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \frac{I_{(t+1)}(y_{(t+1)}, u_{(t+1)}, D^{(t)})}{I_{(t)}(D^{(t)})} \tag{187}$$

This clearly shows that, in order to be able to eliminate the parameter estimation in real time, we must be able to express the integral (186) as a function of the last output $y_{(t)}$ and $u_{(t)}$. We shall show that this can be done analytically for a certain class of models.

Later on, in Section 6, we shall see that the integrals (186) play a fundamental role system classification.

## 4.7 Sufficient Statistic and Self-Reproducing Forms of Probability Distributions

When operating with large amount of data it may be advantageous, and sometimes even necessary, to contract the data into a set of quantities of smaller dimension.

$$V_{(t)} = V_{(t)}(D^{(t)}) \tag{188}$$

Such a contraction is called a statistic and it is said that the statistic is sufficient for some random quantity, say $\underline{x}$, if it carries the same information about this quantity as the data themselves, i.e. if it holds

$$p(x|D^{(t)}) = p(x|V_{(t)}) \tag{189}$$

Saying that there exists a sufficient statistic for the random variable $\underline{x}$ we shall always mean that for the given model relating the random variable $\underline{x}$ with the data $D^{(t)}$, there exists a contraction (188) fulfilling (189) the dimension of which remains fixed when $t$ is growing without limit.

As pointed out by [4] the likelihood function, and hence the aposterior probability distribution, often can be calculated, for the given known (fixed) data, with almost the same ease when the sufficient statistic does not exist as if it happened to exist. Therefore, the Bayesian analysis does not suffer so much from this artificial constraint compared to other non-Bayesian methods. However, this is only partially true. If the calculation has to be performed in real time, when the data set is persistently growing, then, due to limitations on the computing time and the memory required (as in adaptive controllers), the existence of a sufficient statistic may become a crucial question.

Dealing with system identification for the purpose of prediction and control, we are interested not only in the sufficient statistic for the set of unknown parameters $\underline{\theta}$ but also in the sufficient statistic for the predicted output $\underline{y}_{(t+1)}$. These two statistics are, in general, different and, as we shall see, the existence of the former does not imply the existence of the latter. Fortunately, for a certain class of system models both these sufficient statistics exist. For this type of models such functional forms of probability distributions can be found, both for the unknown parameters and for the predicted output, that are reproduced when they are updated according to the recursion (181) and (182). This makes it possible to reduce the functional recursion to an algebraic recursion operating only on a finite set of parameters of these self-reproducing probability distributions. Practically it means that in such cases it is possible to design real-time identification algorithms which require only a finite and fixed size of the memory and do not lose any useful information. In the sequel we shall consider a class of models which fulfill this requirement.

## 4.8 Generalized Multi-variate Regression Model

Let us consider a class of models which can be given the following form

$$f_{(\tau)} = P^T z_{(\tau)} + e_{(\tau)}, \qquad \tau > t_0 \tag{190}$$

where $\{e_{(\tau)};\ \tau = t_0+1, t_0+2, \ldots, t\}$ is a sequence of $\nu$-dimensional normally distributed random variables which are independent of the foregoing input $u_{(\tau)}$ and also of all past outputs and inputs $D^{(\tau-1)}$, and can be assumed to have an unknown but constant covariance $(\nu \times \nu)$-matrix $R$

$$p(e_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) = p(e_{(\tau)}) = (2\pi)^{-\frac{\nu}{2}} |R|^{-\frac{1}{2}} \exp\{-\frac{1}{2} e_{(\tau)}^T R^{-1} e_{(\tau)}\} \tag{191}$$

The $\nu$-vector $f_{(\tau)}$ in (190) is a known vector-valued function of $y_{(\tau)}$ and possibly also of $u_{(\tau)}$ and $D^{(\tau-1)}$

$$f_{(\tau)} = f_{(\tau)}(y_{(\tau)}, u_{(\tau)}, D^{(\tau-1)}) = f_{(\tau)}(D^{(\tau)}) \tag{192}$$

but it is assumed that, for fixed $u_{(\tau)}$ and $D^{(\tau-1)}$ the transformation (192) between the $\nu$-dimensional output vector $y_{(\tau)}$ and $f_{(\tau)}$ is regular (one-to-one)with the Jacobian

$$J_{f(\tau)}(y_{(\tau)}, u_{(\tau)}, D^{(\tau-1)}) = J_{f(\tau)}(D^{(\tau)}) = \begin{bmatrix} \frac{\partial f_{(\tau)1}}{\partial y_{(\tau)1}}, & \ldots, & \frac{\partial f_{(\tau)\nu}}{\partial y_{(\tau)1}} \\ \vdots & & \vdots \\ \frac{\partial f_{(\tau)1}}{\partial y_{(\tau)\nu}}, & \ldots, & \frac{\partial f_{(\tau)\nu}}{\partial y_{(\tau)\nu}} \end{bmatrix} \tag{193}$$

The $\rho$-vector $z_{(\tau)}$ in (190) is a known vector-valued function of the input $u_{(\tau)}$ and of the known past history of inputs and outputs $D^{(\tau-1)}$

$$z_{(\tau)} = z_{(\tau)}(u_{(\tau)}, D^{(\tau-1)}) \tag{194}$$

The $(\rho \times \nu)$-matrix $P$ of regression coefficients is assumed to be unknown but constant.

The form (190) covers a rather broad class of models. For instance, for the regression model from Example 3.2 we have $f_{(\tau)} = y_{(\tau)}$ and if the regression coefficients are brought together into the single matrix $P$ and arranged as follows

$$P^T = [B_0, A_1, B_1, \ldots, A_n, B_n, c] \tag{195}$$

then the vector $z_{(\tau)}$ is

$$z_{(\tau)}^T = [u_{(\tau)}^T, y_{(\tau-1)}^T, u_{(\tau-1)}^T, \ldots, y_{(\tau-n)}^T, u_{(\tau-n)}^T, 1] \tag{196}$$

and $t_0 = n$.

Similarly, for the incremental regression model from Example 3.3 we have $f_{(\tau)} = y_{(\tau)} - y_{(\tau-1)}$

$$P_{(\tau)}^T = [B_0, A_1, B_1, \ldots, A_n, B_n] \tag{197}$$

$$z_{(\tau)}^T = [\Delta u_{(\tau)}, \Delta y_{(\tau-1)}, \Delta u_{(\tau-1)}, \ldots, \Delta y_{(\tau-n)}, \Delta u_{(\tau-n)}] \tag{198}$$

and $t_0 = n + 1$.

A continuous, nonlinear and non-stationary process which, when sampled, falls into this class will be given in Example 4.5

The model ARMA from Example 3.4 can also be converted into the form (190), but only when the coefficients $C_i (i = 1, 2, \ldots, n)$ and the initial conditions are known which is rarely a real case. However, it is possible to choose a finite number of different values for these parameters, to consider them in parallel as known and to calculate the aposterior probability distribution on this set of models as it will be shown in Section 6 where we shall deal with Bayesian system classification.

It is slightly more convenient to consider the precision matrix (131), $\Omega = R^{-1}$ as an unknown parameter instead of the covariance matrix $R$.Thus, the set of parameters in the model structure (190), which will be considered as unknown, is

$$\theta = \{P, \Omega\} \tag{199}$$

As, for any but fixed $u_{(\tau)}$ and $D^{(\tau-1)}$ the transformation between the random variables $e_{(\tau)}$ and $y_{(\tau)}$, determined by (190) and (192), is regular, it holds

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta) = \tag{200}$$

$$= J_{f(\tau)}(D^{(\tau)})(2\pi)^{-\frac{\nu}{2}}|\Omega|^{\frac{1}{2}} \exp\{-\frac{1}{2}[f_{(\tau)} - p^T z_{(\tau)}]^T \Omega [f_{(\tau)} - p^T z_{(\tau)}]\}$$

For all models which can be brought into the form of the conditional probability density function (200) it is possible to give explicit formulae for all aposterior probability distributions which might be of interest. We shall present these results. Instead of going through all tedious technical details of their derivation it will be, perhaps, more convenient for the reader if we only briefly outline how these results can be obtained referring to general lemmas which are summarized in Appendix A. We also would like to stress in advance that, whatever complex these analytical results may seem, they can be evaluated numerically very easily using a square-root filter the FORTAN subroutine for which is given in Appendix B.

To obtain the conditional likelihood function in a compact form it is suitable to rewrite the exponent in (200) as follows (apply A1 from Lemma 1).

$$[f_{(\tau)} - p^T z_{(\tau)}]^T \Omega [f_{(\tau)} - p^T z_{(\tau)}] = d_{(\tau)}^T \begin{bmatrix} I_\nu \\ -p \end{bmatrix} \Omega \begin{bmatrix} I_\nu \\ -p \end{bmatrix}^T d_{(\tau)} = \tag{201}$$

$$= tr\left(\Omega \begin{bmatrix} -I_\nu \\ p \end{bmatrix}^T d_{(\tau)} d_{(\tau)}^T \begin{bmatrix} -I_\nu \\ p \end{bmatrix}\right)$$

where

$$d_{(\tau)} = \begin{bmatrix} f_{(\tau)} \\ z_{(\tau)} \end{bmatrix} \tag{202}$$

and $I_\nu$ is a unit matrix of dimension $\nu$.

The conditional likelihood function (110), obtained as a product of factors (200) for $\tau = t_0 + 1, \ldots, t$ is (apply (A2) in reversed direction)

$$\tilde{L}_{(t)}(P, \Omega, D^{(t)}) = (2\pi)^{-\frac{\nu(t-t_0)}{2}} |\Omega|^{\frac{t-t_0}{2}} \times \tag{203}$$

$$\exp\{-\frac{1}{2}\mathrm{tr}\left(\Omega \begin{bmatrix} -I_\nu \\ P \end{bmatrix}^T \tilde{V}_{(t)} \begin{bmatrix} -I_\nu \\ P \end{bmatrix}\right)\} \prod_{\tau=t_0+1}^{t} J_{f(\tau)}(D^{(\tau)})$$

where $\tilde{V}_{(t)}$ is a matrix of dimension $(\nu + \rho) \times (\nu + \rho)$ composed in the following way.

$$\tilde{V}_{(t)} = \sum_{\tau=t_0+1}^{t} d_{(\tau)} d_{(\tau)}^T \tag{204}$$

If the prior probability density for the unknown parameters is chosen in the form

$$p(P, \Omega | D^{(t_0)}) = p(P, \Omega) = \alpha_{(t_0)} |\Omega|^{\frac{\theta_{(t_0)}}{2}} \exp\{-\frac{1}{2}\mathrm{tr}\left(\Omega \begin{bmatrix} -I_\nu \\ P \end{bmatrix}^T V_{(t_0)} \begin{bmatrix} -I_\nu \\ P \end{bmatrix}\right)\} \tag{205}$$

then as it can be easily seen from the general formula (109), the aposterior probability distribution will have the same form

$$p(P, \Omega | D^{(t)}) = \alpha_{(t)} |\Omega|^{\frac{\theta_{(t)}}{2}} \exp\{-\frac{1}{2}\mathrm{tr}\left(\Omega \begin{bmatrix} -I_\nu \\ P \end{bmatrix}^T V_{(t)} \begin{bmatrix} -I_\nu \\ P \end{bmatrix}\right)\} \tag{206}$$

where

$$\theta_{(t)} = \theta_{(t_0)} + (t - t_0) = \theta_{(t-1)} + 1 \tag{207}$$

$$V_{(t)} = V_{(t_0)} + \tilde{V}_{(t)} = V_{(t-1)} + d_{(t)} d_{(t)}^T \tag{208}$$

and $\alpha_{(t)}$ is the normalizing factor which does not depend on the unknown model parameters $P$, $\Omega$. Hence, the form (206) is self-reproducing with only two parameters $\theta_{(t)}$ and $V_{(t)}$ which can be updated according to the second equalities in (207) and (208)

If the matrix $V_{(t)}$, the sufficient statistic, is partitioned in the following way

$$V_{(t)} = \begin{bmatrix} V_{f(t)} & V_{zf(t)}^T \\ V_{zf(t)} & V_{z(t)} \end{bmatrix} \tag{209}$$

where $V_{f(t)}$ is matrix of dimension $\nu \times \nu$, $V_{zf(t)}$ of dimension $\rho \times \nu$, $V_{z(t)}$ of dimension $\rho \times \rho$, then, using the Lemma 3 in Appendix A, the aposterior probability density (206) can be given form

$$p(P, \Omega | D^{(t)}) = \alpha_{(t)} |\Omega|^{\frac{\theta_{(t)}}{2}} \exp\{-\frac{1}{2}\mathrm{tr}[\Omega([P - \hat{P}_{(t)}]^T V_{z(t)} [P - \hat{P}_{(t)}] + \Lambda_{(t)})]\} \tag{210}$$

where

$$\hat{P}_{(t)} = C_{(t)} V_{zf(t)} \tag{211}$$

$$C_{(t)} = V_{z(t)}^{-1} \tag{212}$$

$$\Lambda_{(t)} = V_{f(t)} - V_{zf(t)}^T C_{(t)} V_{zf(t)} \tag{213}$$

The maximum of the aposterior probability density (210) lies in the point $P = \hat{P}_{(t)}$, $\Omega = \theta_{(t)} \Lambda_{(t)}^{-1}$ (Lemma 7). This density can be written also in the following factorized form.

$$p(P, \Omega | D^{(t)}) = p(P | \Omega, D^{(t)}) p(\Omega | D^{(t)})$$

where

$$p(P|\Omega, D^{(t)}) = \eta_{(t)}|\Omega|^{\frac{\rho}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega[P - \hat{P}_{(t)}]^T V_{z(t)}[P - \hat{P}_{(t)}])\} \tag{214}$$

$$p(\Omega|D^{(t)}) = \gamma_{(t)}|\Omega|^{\frac{\theta_{(t)}-\rho}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega\Lambda_{(t)})\} \tag{215}$$

The normalizing factors $\eta_{(t)}$ and $\gamma_{(t)}$, when required can be determined using Lemmas 8 and 9

$$\eta_{(t)} = (2\pi)^{-\frac{\rho\nu}{2}}|V_{z(t)}|^{\frac{\nu}{2}} \tag{216}$$

$$\gamma_{(t)} = \frac{|\Lambda_{(t)}|^{\frac{\theta_{(t)}-\rho+\nu+1}{2}}}{(2^{\theta_{(t)}-\rho+\nu+1}\pi^{\frac{\nu-1}{2}})^{\frac{\nu}{2}}\prod_{j=1}^{\nu}\Gamma(\frac{\theta_{(t)}-\rho+\nu+2-j}{2})} \tag{217}$$

By $\eta_{(t)}$ and $\gamma_{(t)}$ also the normalizing factor $\alpha_{(t)}$ in (206) and (210) is determined

$$\alpha_{(t)} = \eta_{(t)} \times \gamma_{(t)} \tag{218}$$

The marginal probability distribution for the unknown precision matrix $\Omega$, (215) is the so-called Wishart distribution ([7], par.5.5).The marginal distribution for the unknown matrix $P$ of regression coefficients can be obtained by integrating out the unknown matrix $\Omega$ according to Lemma 9

$$p(P|D^{(t)}) = \int p(P, \Omega|D^{(t)})d\Omega = \tag{219}$$

$$\alpha_{(t)} \int |\Omega|^{\frac{\theta_{(t)}}{2}} \exp\{-\frac{1}{2}\mathrm{tr}[\Omega([P - \hat{P}_{(t)}]^T V_{z(t)}[P - \hat{P}_{(t)}] + \Lambda_{(t)})]\}d\Omega$$

$$p(P|D^{(t)}) = \beta_{(t)}|I_\nu + \Lambda_{(t)}^{-1}[P - \hat{P}_{(t)}]^T V_{z(t)}[P - \hat{P}_{(t)}]|^{-\frac{\theta_{(t)}+\nu+1}{2}}$$

where the normalizing factor $\beta_{(t)}$ is

$$\beta_{(t)} = \frac{\prod_{j=1}^{\nu}\Gamma(\frac{\theta_{(t)}+\nu+2-j}{2})}{\pi^{\frac{\rho}{2}}\prod_{j=1}^{\nu}\Gamma(\frac{\theta_{(t)}-\rho+\nu+2-j}{2})}\frac{|V_{z(t)}|^{\frac{\nu}{2}}}{|\Lambda_{(t)}|^{\frac{\rho}{2}}} \tag{220}$$

Now, we shall derive recursive relations, which make it possible to update the characteristics $\hat{P}_{(t)}$, $C_{(t)}$ and $\Lambda_{(t)}$, directly, instead of calculating them for each $t$ according to formulae (211) to (213).

From the relations (208) and (204), which define the matrix $V_{(t)}$, it is easy to see that for the evolution of its sub-matrices the following relations hold.

$$V_{f(t)} = V_{f(t-1)} + f_{(t)}f_{(t)}^T \tag{221}$$

$$V_{z(t)} = V_{z(t-1)} + z_{(t)}z_{(t)}^T \tag{222}$$

$$V_{zf(t)} = V_{zf(t-1)} + z_{(t)}f_{(t)}^T \tag{223}$$

From (211) and (223) we have

$$\hat{P}_{(t)} = C_{(t)}(V_{zf(t-1)} + z_{(t)}f_{(t)}^T) = C_{(t)}(C_{(t-1)}^{-1}\hat{P}_{(t-1)} + z_{(t)}f_{(t)}^T) \tag{224}$$

but from (212) and (222) we also have

$$C_{(t-1)}^{-1} = C_{(t)}^{-1} - z_{(t)}z_{(t)}^T$$

which, substituted into (224) where

$$\hat{P}_{(t)} = \hat{P}_{(t-1)} + C_{(t)}z_{(t)}\hat{e}_{(t)}^T \tag{225}$$

where

$$\hat{e}_{(t)} = f_{(t)} - \hat{P}_{(t-1)}^T z_{(t)} \tag{226}$$

As $C_{(t)}$ has been introduced as the inverse of $V_{z(t)}$, the recursion for this characteristic can be obtained by application of (412) from Lemma 4 to (222)

$$C_{(t)} = C_{(t-1)} - \frac{1}{1+\zeta_{(t)}} C_{(t-1)} z_{(t)} z_{(t)}^T C_{(t-1)} \tag{227}$$

where

$$\zeta_{(t)} = z_{(t)}^T C_{(t-1)} z_{(t)} \tag{228}$$

Multiplying (227) by $z_{(t)}$ we obtain

$$C_{(t)} z_{(t)} = C_{(t-1)} z_{(t)} (1 - \frac{\zeta(t)}{1+\zeta(t)}) = \frac{1}{1+\zeta(t)} C_{(t-1)} z_{(t)} \tag{229}$$

Substitution of (229) into (225) gives an alternative recursion for $\hat{P}_{(t)}$

$$\hat{P}_{(t)} = \hat{P}_{(t-1)} + \frac{1}{1+\zeta(t)} C_{(t-1)} z_{(t)} \hat{e}_{(t)}^T \tag{230}$$

In a similar way it is also possible to derive the recursion for the characteristic $\Lambda_{(t)}$

$$\Lambda_{(t)} = \Lambda_{(t-1)} + \frac{1}{1+\zeta(t)} \hat{e}_{(t)} \hat{e}_{(t)}^T \tag{231}$$

where $\hat{e}_{(t)}$ is defined by (226).

The recursion (230) accompanied by (227) and (228) is well known from recursive least-square estimation. The above given Bayesian interpretation clearly shows its probabilistic meaning.

The matrix $V_{(t)}$, or equivalently the triad $\{\hat{P}_{(t)}, C_{(t)}, \Lambda_{(t)}\}$, is the sufficient statistic for the unknown parameters. Unfortunately, it does not mean yet that it is not necessary to keep the entire past input-output history in the memory of the computing device. To update this sufficient statistic according to (208) or (230), (227) and (231) the values of functions $f_{(t)}$, (192) and $z_{(t)}$, (194), must be determined which, in general, may depend on all past input-output data. Hence, if the requirement of a finite and fixed memory size has to be met and no information is allowed to be lost, then an additional condition must be fulfilled. There must exist a state $s_{(t-1)}$ such that

$$f_{(t)}(y_{(t)}, u_{(t)}, D^{(t-1)}) = f_{(t)}(y_{(t)}, u_{(t)}, s_{(t-1)}) \tag{232}$$

$$z_{(t)}(u_{(t)}, D^{(t-1)}) = z_{(t)}(u_{(t)}, s_{(t-1)}) \tag{233}$$

and at the same time the evolution of this state

$$s_{(t)} = \phi(s_{(t-1)}, y_{(t)}, u_{(t)}) \tag{234}$$

must not depend on the unknown parameters, i.e. the function $\phi(.)$, as well as the initial conditions for (234), must be a priori given. The regression model (Example 3.2) and the incremental regression model (Example 3.3) are the simplest cases of this kind.

Now, the prediction of the output $y_{(t+1)}$ will be considered. The probability distribution $p(y_{(t+1)}|u_{(t+1)}, D^{(t)})$ can be found applying either the general formula (93) or (187) The latter is somewhat more convenient in our case. Using Lemmas 8 and 9 from Appendix A it is possible to find an explicit analytical expression for the integral (186)

$$I_{(t)}(D^{(t)}) = \alpha_{(t_0)} (2^{\theta_{(t_0)}+\nu+1} \pi^{\frac{\nu-1}{2}-t+t_0+\rho})^{\frac{\nu}{2}} \times \tag{235}$$

$$\times \prod_{\tau=t_0+1}^{t} J_{f(\tau)}(D^{(\tau)}) \prod_{j=1}^{\nu} \Gamma(\frac{\theta_{(t)}-\rho+\nu+2-j}{2}) \times |V_{z(t)}|^{-\frac{\nu}{2}} |\Lambda_{(t)}|^{-\frac{\theta_{(t)}-\rho+\nu+1}{2}}$$

Lemma 5 from Appendix A gives us the possibility to evaluate the expression (235) recursively. According to (415) we have

$$|V_{z(t+1)}| = |V_{z(t)} + z_{(t+1)} z_{(t+1)}^T| = |V_{z(t)}| \times (1 + \zeta_{(t+1)}) \tag{236}$$

$$|\Lambda_{(t+1)}| = |\Lambda_{(t)} + \frac{1}{1+\zeta_{(t+1)}}\hat{e}_{(t+1)}\hat{e}^T_{(t+1)}| = |\Lambda_{(t)}|(1 + \frac{\hat{e}^T_{(t+1)}\Lambda_t^{-1}\hat{e}_{(t+1)}}{1+\zeta_{(t+1)}}) \tag{237}$$

The substitution of these relations into (235) for the time-index $t+1$ and $\theta_{(t+1)} = \theta_{(t)} + 1$ gives

$$I_{(t+1)}(D^{(t+1)}) = I_{(t)}(D^{(t)})\frac{\kappa_{(t+1)}}{(1 + \frac{\hat{e}^T_{(t+1)}\Lambda^{-1}_{(t)}\hat{e}_{(t+1)}}{1+\zeta_{(t+1)}})^{\frac{\theta_t-\rho+2+\nu}{2}}} J_{f(t+1)}(D^{(t+1)}) \tag{238}$$

where

$$\kappa_{(t+1)} = \frac{\Gamma(\frac{\theta_{(t)}-\rho+\nu+2}{2})}{\pi^{(\frac{\nu}{2})}\Gamma(\frac{\theta_{(t)}-\rho+2}{2})}(1+\zeta_{(t+1)})^{-\frac{\nu}{2}}|\Lambda_{(t)}|^{-\frac{1}{2}} \tag{239}$$

According to (187) this recursion relation directly yields the probability distribution for $y_{(t+1)}$ given $u_{(t+1)}$ and $D^{(t)}$ but not the parameters $P$ and $\Omega$

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = J_{f(t+1)}\frac{\kappa_{(t+1)}}{(1 + \frac{\hat{e}^T_{(t+1)}\Lambda^{-1}_{(t)}\hat{e}_{(t+1)}}{1+\zeta_{(t+1)}})^{\frac{\theta_{(t)}-\rho+2+\nu}{2}}} \tag{240}$$

where $J_{f(t+1)}$ is the Jacobian (193), in general a function of $y_{(t+1)}$, $u_{(t+1)}$ and $D^{(t)}$ the particular form of which is given by the particular form of the function $f_{(t+1)}$ (232), and $\hat{e}_{(t+1)}$ is also a function of $y_{(t+1)}$, $u_{(t+1)}$ and the past input-output history as follows from its definition (226)

$$\hat{e}_{(t+1)} = f_{(t+1)} - \hat{p}^T_{(t)}z_{(t+1)} \tag{241}$$

As, for fixed $u_{(t+1)}$ and $D^{(t)}$, the transformation between the predicted output $y_{(t+1)}$ and the random variable $\hat{e}_{(t+1)}$ is regular, we may write

$$p(\hat{e}_{(t+1)}|u_{(t+1)}, D^{(t)}) = \frac{\kappa_{(t+1)}}{(1 + \frac{\hat{e}^T_{(t+1)}\Lambda^{-1}_{(t)}\hat{e}_{(t+1)}}{1+\zeta_{(t+1)}})^{\frac{\theta_{(t)}-\rho+2+\nu}{2}}} \tag{242}$$

which is the so-called t-distribution (see e.g. [7], par. 5.6) with a zero mean and the covariance matrix

$$E[\hat{e}_{(t+1)}\hat{e}^T_{(t+1)}|u_{(t+1)}, D^{(t)}] = \frac{1+\zeta_{(t+1)}}{\theta^{-\rho}_{(t)}}\Lambda_{(t)} \tag{243}$$

It is well know that with growing $\theta_{(t)}$ the $t$-distribution (242) very rapidly converges to the normal distribution which suggests a reasonable approximation when required.

The rest of this discussion on generalized multi-variate regression model will be devoted to some numerical aspects and practical hints. We have given two alternatives how the main characteristics, by which the aposterior distributions are determined, can be updated in real time:

(i) updating the symmetric matrix $V_{(t)}$ according to (208)

$$V_{(t)} = V_{(t-1)} + d_{(t)}d^T_{(t)} \tag{244}$$

where $d_{(t)}$ is the data vector of dimension $(\nu+\rho)$ defined by (202), and calculation of $\hat{P}_{(t)}$, $C_{(t)}$ and $\Lambda_{(t)}$ for each $t$ from its sub-matrices (see the partitioning (209)) using the formulae (211) to (213)

$$C_{(t)} = V^{-1}_{z(t)} \tag{245}$$

$$\hat{P}_{(t)} = C_{(t)}V_{zf(t)} \tag{246}$$

$$\Lambda_{(t)} = V_{f(t)} - V^T_{zf(t)}\hat{P}_{(t)} \tag{247}$$

(ii) direct updating of $\hat{P}_{(t)}$, $\Lambda_{(t)}$ and $C_{(t)}$ according to the recursion (226), (228), (230), (231) and (227)

$$g_{(t)} = C_{(t-1)} z_{(t)} \tag{248}$$

$$\zeta_{(t)} = z_{(t)}^T g_{(t)} \tag{249}$$

$$\hat{e}_{(t)} = f_{(t)} - \hat{P}_{(t-1)}^T z_{(t)} \tag{250}$$

$$\hat{P}_{(t)} = \hat{P}_{(t-1)} + \frac{1}{1+\zeta_{(t)}} g_{(t)} \hat{e}_{(t)}^T \tag{251}$$

$$\Lambda_{(t)} = \Lambda_{(t-1)} + \frac{1}{1+\zeta_{(t)}} \hat{e}_{(t)} \hat{e}_{(t)}^T \tag{252}$$

$$C_{(t)} = C_{(t-1)} - \frac{1}{1+\zeta_{(t)}} g_{(t)} g_{(t)}^T \tag{253}$$

The most awkward operation in the first alternative is the inversion (245). Moreover, both the $(\rho \times \rho)$-matrix $C_{(t)}$ and the $(\nu \times \nu)$-matrix $\Lambda_{(t)}$, must be positive definite to give the correct sense but, in ill-conditioned situations (e.g. redundant parameters or ill-excited system), they may lose this important property due to rounding errors, especially when the calculation is performed on a digital computer with reduced word length. In the second alternative the inversion is performed implicitly in (253) but similar difficulties may still occur. For these reasons we shall give a third possibility which is less suitable for theoretical considerations but exhibits an outstanding numerical stability. Its square-root nature is reflected in higher precision compared to above given alternatives (up to double precision in ill-conditioned problems) and guarantees the positive definiteness of covariance matrices. It can be recommended both real-time and one-shot identification.

Consider a lower triangular matrix $G_{(t)}$ defined as the Choleski square root of the matrix $V_{(t)}^{-1}$

$$V_{(t)}^{-1} = G_{(t)} G_{(t)}^T \tag{254}$$

If the matrix $G_{(t)}$ partitioned similarly to (209)

$$G_{(t)} = \begin{bmatrix} G_{f(t)} & 0 \\ G_{zf(t)} & G_{z(t)} \end{bmatrix} \tag{255}$$

then, using Lemma 6 from Appendix A it is easy to verify that the following relations hold

$$\Lambda_{(t)}^{-1} = G_{f(t)} G_{f(t)}^T \tag{256}$$

$$C_{(t)} = G_{z(t)} G_{z(t)}^T \tag{257}$$

$$\hat{P}_{(t)} G_{f(t)} = -G_{zf(t)} \tag{258}$$

As $G_{f(t)}$ is triangular and only of dimension $\nu$ (a single number in the case of a single output, $\nu = 1$) the matrix equation (258) can be solved very easily. As the triangular sub-matrices $G_{z(t)}$ and $G_{f(t)}$ are Choleski square roots of $C_{(t)}$ and $\Lambda_{(t)}^{-1}$, respectively, also other characteristics can be calculated very simply. For instance, the scalar $\zeta_{(t)}$ defined by (228) can be calculated as a sum of squares

$$\zeta_{(t)} = z_{(t)}^T C_{(t-1)} z_{(t)} = ||z_{(t)}^T G_{z(t-1)}||^2 \tag{259}$$

Similarly

$$\hat{e}_{(t)}^T \Lambda_{(t-1)}^{-1} \hat{e}_{(t)} = || \begin{bmatrix} f_{(t)} \\ z_{(t)} \end{bmatrix}^T \begin{bmatrix} G_{f(t)} \\ G_{zf(t)} \end{bmatrix} ||^2 \tag{260}$$

Due to the triangular forms of the matrices $G_{z(t)}$ and $G_{f(t)}$ it holds

$$|V_{z(t)}|^{\frac{1}{2}} = |G_{z(t)}|^{-1} = \frac{1}{\prod_{i=1}^{\rho} G_{z(t)ii}} \tag{261}$$

$$|\Lambda_{(t)}|^{\frac{1}{2}} = |G_{f(t)}|^{-1} = \frac{1}{\prod_{i=1}^{\nu} G_{f(t)ii}} \tag{262}$$

These square roots of determinants appear on various places, namely in the integral (235) which will play an important role in systems classification in Section 6

All this shows that it is advantageous to have a square-root filter at disposal which updates directly the triangular matrix $G_{(t)}$. Such a filter is given in the form of the FORTRAN subroutine REFIL in Appendix B. Its derivation has been given by [21] and is reported by Strejc in chapter 4 of this text. A standard usage of subroutine REFIL ($G$, $D$, $N$, $SIG2$, $VG$, $IN$) is as follows. ($\phi^2$ is the forgetting factor which will be introduced in the next Section, here $\phi^2 = 1$).

CALL state: $G = G_{(t-1)}$, $D = d_{(t)}$, $N = \nu + \rho$, $SIG2 = \phi^2$,
$VG$ arbitrary, $IN$ = dimension of $G$ in the main program
RETURN state: $G = G_{(t)}$, $SIG2 = \phi^2 + ||d_{(t)}^T G_{(t)}||^2$, $VG = G_{(t-1)} G_{(t-1)}^T d_{(t)}$,
the remaining parameters unchanged

When subroutine is used to update only the triangular matrix $G_{z(t-1)}$, then

CALL state: $G = G_{z(t-1)}$, $D = z_{(t)}$, $N = \rho$, $SIG2 = \phi^2$
RETURN state: $G = G_{z(t)}$, $SIG2 = \phi^2 + \zeta_{(t)}$, $VG = g_{(t)}$

where $\zeta_{(t)}$ is the scalar (249) and $g_{(t)}$ is the vector (248). It means that the outputs $SIG2$ and $VG$ can be used to update the point estimates of the regression coefficient $\hat{P}_{(t-1)}$ according to the formulae (251) and (250).

As shown in [25] the subroutine REFIL, when applied to the regression model from Example 3.2 or to the incremental regression model from Example 3.3, yields, at the same time, all lower order models.

Example 4.5 An autonomous system will be considered the continuous output process of which can be described by a scalar stochastic differential equation

$$\frac{dy(\tau)}{d\tau} = y(\tau)[1 - y(\tau)][c + \delta(\tau)] \tag{263}$$

where $c$ is an unknown constant and $\delta(\tau)$ is a stochastic term $t$ the properties of which will be specified later on. Here, and only here in this Example, $\tau$ denotes a continuous time the unit of which is chosen to be one year. The model (263) can be used to describe the process of competition between two competitors, which may be, for instance, two technologies satisfying the same or similar need. Then $y(\tau)$ may be interpreted as the fractional market share occupied by the superior newcomer while $[1 - y(\tau)]$ is the market share occupied by the looser. Hence, the output $y(\tau)$ must lie in the interval

$$0 < y(\tau) < 1 \tag{264}$$

For a more detailed discussion of substitution processes, mainly from macro-economic viewpoint, the reader is referred to [23] where also more complex and multi-variate cases are considered.

It is assumed that a certain, rather small set of samples of the output

$$y^{(t)} = \{y(\tau_1), y(\tau_2), \ldots, y(\tau_t)\}$$

is available. The problem is to estimate the model parameters and, above all, to forecast the future course of the substitution, i.e. to determine the probability distribution $p(y_{(t+1)}|y^{(t)})$ where $y_{(t+1)} = y(\tau_{t+1})$ and $\tau_{t+1} > \tau_t$

First, we shall show that, under certain assumptions concerning the stochastic term $\delta_{(\tau)}$, the process (263), when sampled, can be brought into the form of a generalized regression model (190). This will permit us to apply the general results derived above.

When a new variable

$$x(\tau) = \ln \frac{y(\tau)}{1 - y(\tau)} \tag{265}$$

is introduced instead of the output $y(\tau)$, the stochastic differential equation (263) gets a very simple form

$$\frac{dx(\tau)}{d\tau} = c + \delta(\tau) \tag{266}$$

Integration of (266) over the time interval $< \tau_{i-1}, \tau_i >$ gives

$$x_{(i)} - x_{(i-1)} = cT_i + \epsilon_{(i)} \tag{267}$$

where

$$x_{(i)} = x(\tau_i)$$

$$T_i = \tau_i - \tau_{i-1} \tag{268}$$

$$\epsilon_{(i)} = \int_{T_{i-1}}^{T_i} \delta(\tau) d\tau \tag{269}$$

If the sampling interval $T_i$ is large enough then it is reasonable to assume that the integrals (269) form a sequence of independent normally distributed random quantities with zero mean and with the variance proportional to the sampling interval:

$$E[\epsilon_{(i)}] = 0, \quad E[\epsilon_{(i)}^2] = T_i \sigma^2$$

$$p(\epsilon_{(i)} | \epsilon^{(i-1)}) = p(\epsilon_{(i)}) = N(0, T_i \sigma^2)$$

In other words, this assumption means that $\epsilon_{(i)}$ is considered as an increment of a normal Wiener-Levy process sometimes called integrated white noise or Brownian motion. For a more detailed justification of this assumption see reference cited.

The relation between the sampled output $y_{(i)} = y(\tau_i)$ and the quantity $x_{(i)}$ is given by (265) from which follows

$$y_{(i)} = \frac{1}{1 + \exp(-x_{(i)})}$$

Making use of (267) the following stochastic difference equation for the sampled output is obtained

$$y_{(i)} = \frac{1}{1 + \frac{1 - y_{(i-1)}}{y_{(i-1)}} \exp(-cT_i - \epsilon_{(i)})} \tag{270}$$

However, for our purposes the simple relation (267) is more suitable. If it is divided by $\sqrt{T_i}$ then it gets the form of a generalized regression model (190)

$$f_{(i)} = c\sqrt{T_i} + e_{(i)} \tag{271}$$

where

$$e_{(i)} = \frac{1}{\sqrt{T_i}} \epsilon_{(i)}$$

has an unknown but constant variance

$$Ee_{(i)}^2 = \sigma^2$$

and

$$f_{(i)} = \frac{1}{\sqrt{T_i}}(x_{(i)} - x_{(i-1)}) = \frac{1}{\sqrt{T_i}}(\ln \frac{y_{(i)}}{1 - y_{(i)}} - \ln \frac{y_{(i-1)}}{1 - y_{(i-1)}}) \tag{272}$$

Hence, all results which have been derived for the general case (190) can be applied by setting

$$P = c, \ \Omega = \frac{1}{\sigma^2} = \omega, \ z_{(i)} = \sqrt{T_i}, \ \nu = \rho = 1, \ t_0 = 1 \tag{273}$$

$$d_{(i)}^T = [f_i, \sqrt{T_i}] \tag{274}$$

It only remains to choose a suitable prior distribution for the unknown parameters $\theta = \{c, \omega\}$. As we do not assume any prior information about the unknown parameters we shall choose the improper prior for $\omega$ according to (133)

$$p(\omega|y_{(1)}) = p(\omega) = \omega^{-1}$$

and the uniform improper prior distribution for the parameter $c$

$$p(c|y_{(1)}, \omega) = p(c) = 1$$

Hence, the prior joint probability distribution assumed is

$$p(c, \omega|y_{(1)}) = \omega^{-1}$$

From comparison of this choice with the general form of the self-reproducing prior distribution (205) we have

$$\theta_{(1)} = -2, \qquad V_{(1)} = 0$$

and from (207)

$$\theta_{(t)} = t - 3$$

The matrix $V_{(t)}$ (208) of dimension $(2 \times 2)$ has just three different entries

$$V_{f(t)} = \sum_{i=2}^{t} f_{(i)}^2 = \sum_{i=2}^{t} \frac{(x_{(i)} - x_{(i-1)})^2}{T_i}$$

$$V_{zf(t)} = \sum_{i=2}^{t} \sqrt{T_i} f_{(i)} = \sum_{i=2}^{t} (x_{(i)} - x_{(i-1)}) = x_{(t)} - x_{(1)}$$

$$V_{z(t)} = \sum_{i=2}^{t} T_i = \tau_t - \tau_1$$

According to (211) and (212) the characteristics of aposterior distributions are

$$\hat{P}_{(t)} = \hat{c}_{(t)} = \frac{x_{(t)} - x_{(1)}}{\tau_t - \tau_1} \tag{275}$$

$$C_{(t)} = \frac{1}{\tau_t - \tau_1} \tag{276}$$

$$\Lambda_{(t)} = \sum_{i=2}^{t} \frac{(x_{(i)} - x_{(i-1)})^2}{T_i} - \frac{(x_{(t)} - x_{(1)})^2}{\tau_t - \tau_1} = \sum_{i=2}^{t} \frac{1}{T_i} (x_{(i)} - x_{(i-1)} - \hat{c}_{(t)} T_i)^2 \tag{277}$$

When these characteristics are substituted into (215) the following aposterior marginal probability density for $\omega$ is obtained

$$p(\omega|y^{(t)}) = \frac{\Lambda_{(t)}^{\frac{t-2}{2}}}{2^{\frac{t-2}{2}} \Gamma(\frac{t-2}{2})} \omega^{\frac{t-4}{2}} \exp\{-\frac{1}{2} \Lambda_{(t)} \omega\}$$

After the transformation $\sigma = \omega^{-\frac{1}{2}}$ we have

$$p(\sigma|y^{(t)}) = \frac{\Lambda_{(t)}^{\frac{t-2}{2}}}{2^{\frac{t-4}{2}} \Gamma(\frac{t-2}{2})} \sigma^{-(t-1)} \exp\{-\frac{\Lambda_{(t)}}{2\sigma^2}\} \tag{278}$$

For the parameter $c$ the formula (219) gives

$$p(c|y^{(t)}) = \frac{\Gamma(\frac{t-1}{2})}{\sqrt{\pi} \Gamma(\frac{t-2}{2})} \sqrt{\frac{\tau_t - \tau_1}{\Lambda_{(t)}}} \left[1 + \frac{\tau_t - \tau_1}{\Lambda_{(t)}} (c - \hat{c}_{(t)})^2\right]^{-\frac{t-1}{2}} \tag{279}$$

In this simple case the Jacobian (193) degenerates into a single derivative

$$J_{f(i)} = \frac{\partial f_{(i)}}{\partial y_{(i)}} = \frac{1}{\sqrt{T_i} y_{(i)}(1 - y_{(i)})}$$

and by substitution into the formula (240) the probability distribution for forecasting the process for the time span

$$T_{t+1} = \tau_{t+1} - \tau_t$$

ahead is obtained

$$p(y_{(t+1)}|y^{(t)}) = \frac{\Gamma(\frac{t-1}{2})}{\sqrt{\pi}\Gamma(\frac{t-2}{2})\sqrt{(1+\zeta_{(t+1)})T_{t+1}\Lambda_{(t)}}} \times \frac{[y_{(t+1)}(1-y_{(t+1)})]^{-1}}{\left[1 + \frac{(\ln\frac{y_{(t+1)}}{1-y_{(t+1)}} - \ln\frac{y_{(t)}}{1-y_{(t)}} - \hat{c}_{(t)}T_{t+1})^2}{(1+\zeta_{(t+1)})T_{t+1}\Lambda_{(t)}}\right]^{\frac{t-1}{2}}} \tag{280}$$

where

$$\zeta_{(t+1)} = T_{(t+1)}C_{(t)} = \frac{\tau_{t+1} - \tau_t}{\tau_t - \tau_1}$$

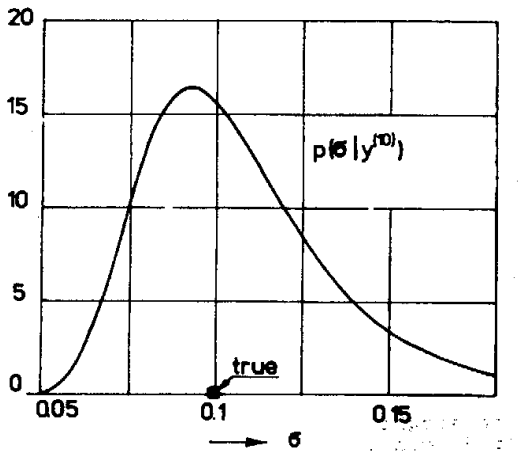
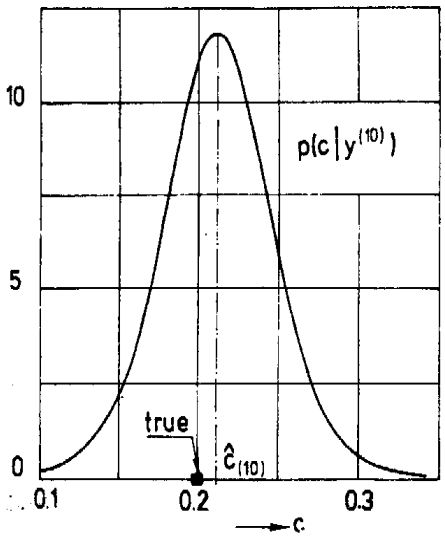


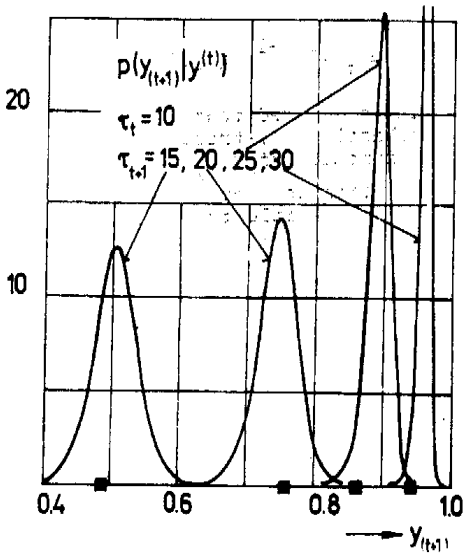

Fig. 7 Substitution process-first 10 data used for parameter estimation and for forecasting

To demonstrate these analytical results we shall apply them to simulated data in order to be able to confront the probability distributions with true values which are assumed to be unknown. Examples of real substitution processes can be found in the reference cited.

The process simulated with $c = 0.2$, $[year^{-1}]$ and $\sigma = 0.1$ is plotted in Fig. 7. Only the first 10 outputs, sampled with the period of 1 year, have been used to estimate the model parameters and to forecast the future evolution of the substitution.




1. Fig. 8 Probability distribution for the unknown parameter $\sigma$ after 10 observations



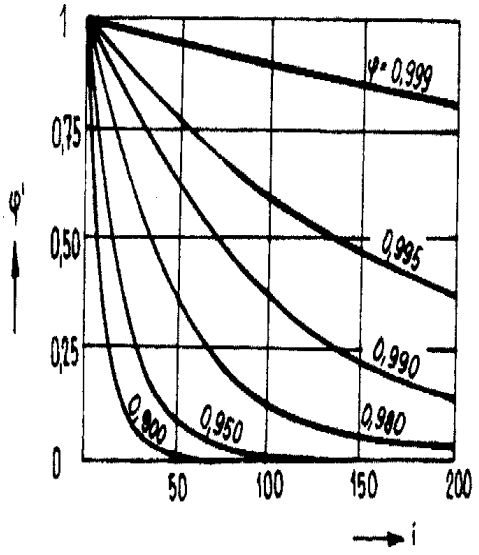

1. Fig. 9 Probability distribution for the unknown parameter $c$ after 10 observations

The numerical values of characteristics calculated according to formulae (275) and (277) are $\hat{c}_{(10)} = 0.2115$ and $\Lambda_{(10)} = 0.0770$. Marginal probability densities (278) and (279) for the unknown parameters $\sigma$ and $c$ are plotted in Fig. 8 and Fig. 9, respectively.

The forecast has been calculated for

$$\begin{aligned} \tau_{t+1} = &\quad 15, \quad 20, \quad 25, \quad 30 \\ T_{t+1} = &\quad 5, \quad 10, \quad 15, \quad 20 \end{aligned}$$

The probability distributions for the forecasted outputs, given by the formula (280), are plotted in Fig. 10 where also the corresponding outputs obtained in the simulation experiment are denoted by squares on the $y_{(t+1)}$-axis. Notice that the precision of forecasting is the higher the closer to one the predicted output lies. This can be intuitively explained by the fact that the superior newcomer, after a certain period of time, will penetrate the market with almost certainty whatever the disturbances are.Notice also that the maxima of the aposterior probability densities do not lie in the points which would be obtained from the formula (270) by setting the stochastic term $\epsilon_{(i)}$ to zero and replacing $c$ by $\hat{c}_{(i)}$




1. Fig. 10 Forecasting of the substitution process for 5, 10, 15, and 20 years ahead; true outputs obtained are denoted by squares

# 5 Time-Varying Parameters and Adaptivity

As it has been discussed in Section 3 the position taken up throughout this chapter is that of an outer observer whose objective is to predict and/or to control the output of the system observed. Therefore the system model is understood as any mathematical description of the input-output relation which defines the family of conditional probability distributions

$$\{p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}); \quad \tau > t_0\}$$

through a finite set of parameters. If the set or subset of parameters, denoted by $\theta$, is unknown then the model, or – better to say – the model structure, defines only the conditional probability distributions (81)

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta) \tag{281}$$

Until now it has been assumed that the model parameters are unknown but time - invariant constants. In this Section we shall remove this assumption and instead of $\theta$ in (281) we shall consider a quantity $q_{(\tau)}$, possibly multi-variable which can be interpreted as time-varying internal quantity which cannot be directly observed. Thus, instead of (281) we now have

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, q_{(\tau)}) \tag{282}$$

and it is assumed that this conditional probability distribution is known (as a function of $q_{(\tau)}$ ) for all $\tau > t_0 \geq 0$.

The question we want to clarify is: How is it possible to determine the conditional probability distribution

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) \tag{283}$$

which is required for prediction of the process output, and what must be known or given, in addition to (282), in order to be able to perform this task?

We shall again assume that the natural conditions of control are fulfilled, i.e., that the controller, when making decision concerning the input $u_{(\tau)}$, does not use more information about the unknown quantity $q_{(\tau)}$ than it is contained in the past history of the input-output data $D^{(\tau-1)}$. This means that the following analogy of (91) holds for any $\tau > t_0$

$$p(u_{(\tau)}|D^{(\tau-1)}, q_{(\tau)}) = p(u_{(\tau)}|D^{(\tau-1)}) \tag{284}$$

When the basic operation (13) is applied to the joint probability distribution $p(q_{(\tau)}, u_{(\tau)}|D^{(\tau-1)})$ in two different possible ways the following relation is obtained

$$p(q_{(\tau)}|u_{(\tau)}, D^{(\tau-1)})p(u_{(\tau)}|D^{(\tau-1)}) = p(u_{(\tau)}|D^{(\tau-1)}, q_{(\tau)})p(q_{(\tau)}|D^{(\tau-1)})$$

From this relation it follows that if (284) holds then also

$$p(q_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) = p(q_{(\tau)}|D^{(\tau-1)}) \tag{285}$$

The sought-for probability distribution (283) can be obtained from the known distribution (282) by eliminating the unknown quantity $q_{(\tau)}$. This can be done by using the two basic operations (12), (13) and the equality (285)

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) = \int p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, q_{(\tau)})p(q_{(\tau)}|D^{(\tau-1)})dq_{(\tau)} \tag{286}$$

The second factor of the integrated function on the right-hand side of (286), namely the conditional probability distribution

$$p(q_{(\tau)}|D^{(\tau-1)}) \tag{287}$$

is a quantitative description of the statistician's uncertainty about the unknown internal quantity $q_{(\tau)}$ at the time instant when the $u_{(\tau)}$ has to be determined and can be understood as the Bayesian "estimate" of $q_{(\tau)}$ based on the past input-output data $D^{(\tau-1)}$

Now it will be shown how the conditional probability distributions (283) and (287) can be calculated recursively in real time. To derive this recursion we shall assume that the probability distribution (287) is known for $\tau = t$ and we shall calculate it for $\tau = t + 1$.

Knowing the probability distribution (287) for $\tau = t$ we can calculate the probability distribution of the next output $\underline{y}_{(t)}$ for any $\underline{u}_{(t)}$ according to (286)

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = \int p(y_{(t)}|u_{(t)}, D^{(t-1)}, q_{(t)})p(q_{(t)}|D^{(t-1)})dq_{(t)} \tag{288}$$

When the input $\underline{u}_{(t)}$ is applied and the true value of the new output $\underline{y}_{(t)}$ is observed the new data pair $\underline{D}_{(t)} = \{\underline{u}_{(t)}, \underline{y}_{(t)}\}$ is obtained. In the next step of the recursion the new data $\underline{D}_{(t)}$ has to be incorporated into the information about the system observed.

Applying the basic operation (13) in two different ways the following relation can be written.

$$p(y_{(t)}, q_{(t)}|u_{(t)}, D^{(t-1)}) = p(y_{(t)}|u_{(t)}, D^{(t-1)}, q_{(t)})p(q_{(t)}|u_{(t)}, D^{(t-1)}) =$$
$$= p(q_{(t)}|D^{(t)})p(y_{(t)}|u_{(t)}, D^{(t-1)})$$

From the second equality in this relation and from the equality (285) reflecting the natural conditions of control it follows

$$p(q_{(t)}|D^{(t)}) = \frac{p(y_{(t)}|u_{(t)}, D^{(t-1)}, q_{(t)})}{p(y_{(t)}|u_{(t)}, D^{(t-1)})} p(q_{(t)}|D^{(t-1)}) \tag{289}$$

This operation can be understood as updating of the probability distribution for the unknown quantity $q_{(t)}$ with respect to the new data pair $\underline{D}_{(t)} = \{\underline{u}_{(t)}, \underline{y}_{(t)}\}$. Notice the analogy between (289) and (182).

To complete the recursion it remains to perform the re-calculation

$$p(q_{(t)}|D^{(t)}) \to p(q_{(t+1)}|D^{(t)}) \tag{290}$$

which could be omitted when $q_{(t+1)} = q_{(t)} = \theta$ was an unknown constant. We employ again the two basic operations (12) and (13), this time in the following way

$$p(q_{(t+1)}|D^{(t)}) = \int p(q_{(t+1)}, q_{(t)}|D^{(t)}) dq_{(t)}$$

$$p(q_{(t+1)}|D^{(t)}) = \int p(q_{(t+1)}|q_{(t)}, D^{(t)}) p(q_{(t)}|D^{(t)}) dq_{(t)} \tag{291}$$

The relation (291) completes the recursion. However, this last step of the recursion requires that the conditional probability distribution

$$p(q_{(\tau+1)}|q_{(\tau)}, D^{(\tau)}) \tag{292}$$

be available for all $\tau$. Hence, the overall system model, which makes it possible to solve the problem of prediction and control of the system output in a consistent way within the Bayesian statistics, must define both the conditional probability distributions (282) and (292) for all $\tau \geq 1$. The functional recursion, defined by the relations (288), (289) and (291) solves conceptually the problem of prediction and starts with the prior probability distribution $p(q_{(1)})$ (formally $p(q_{(1)}|D^{(0)}) = p(q_{(1)})$), which is a model of the statistician's prior uncertainty about $q_{(1)}$ when no input-output data are available. Later on we will demonstrate this conceptual solution on particular examples.

## 5.1 Bayesian Viewpoint on Adaptivity

The conceptual solution of the prediction problem given above may help to throw the "Bayesian light" upon the problem of adaptivity in order to clear up the possibilities and limitations of the Bayesian approach to the design of adaptive systems.

Usually, a system is called adaptive if it is able to accumulate the experience about the properties of its environment and to exploit this experience for improvement of its performance. The adaptive system of our interest is the predictor or controller and its environment is the process to be predicted or controlled. If the properties of the process can be described by a mathematical model of a given structure and quantitatively characterized by the parameter values of this structure then the Bayesian statistics can be understood as a tool which makes it possible to rationalize and formalize the experience accumulation. From this point of view several different types of adaptivity can be distinguished.

(I) The properties of the process, quantitatively expressible by the parameter values of a process model, do not vary in time but they are unknown. This case has been investigated in Section 4. The adaptive system designed for such a situation can be called self-adjusting or self-tuning.

(II) The model parameters do vary in time but a probabilistic model of their variations, fully defining the conditional probability distributions (292), is a priori known. In such a case the both models, defining (282) and (292), can be aggregated into a single model, which is more complex but with all parameters known. In this way the problem of adaptivity is transformed into a problem which does not contain the unknown parameters, see e.g. [3]. The one-step-ahead Bayesian predictor for this case is given by recursive relations (288), (289) and (291).

(III) The model of parameter variations is known up to a finite set of constant but unknown parameters. By model aggregation this case be reduced to case (I)

(IV) A suitable model structure for parameter variations is not available, but it is known, or it can be assumed, that the parameters vary "relatively slowly". As the conditional probability distribution (292) is not defined, the case cannot be solved within the consistent Bayesian theory based on two operations (12) and (13). It is the last step (290) of the above given recursion which cannot be performed exactly. Yet, as it will be shown later on in a separate subsection, there exist a heuristic but rationally based extension of Bayesian theory which makes it possible to overcome this difficulty and which leads to the well-known and well-tried technique of "exponential forgetting".

It is worthy to note that unknown but constant parameters (I) can be considered as a special case of time-varying parameters (II) when the model of parameter variations is

$$q_{(\tau+1)} = q_{(\tau)}$$

or in terms of conditional probability distributions

$$p(q_{(\tau+1)}|q_{(\tau)}, D^{(\tau)}) = p(q_{(\tau+1)}|q_{(\tau)}) = \delta(q_{(\tau+1)} - q_{(\tau)})$$

where $\delta(.)$ is either the Dirac $\delta$ - function if $q$ is continuous or the Kronecker's $\delta$ if q is discrete [8]. Then the last operation (291) in the general recursion gives

$$p(q_{(t+1)}|D^{(t)}) = p(q_{(t)}|D^{(t)}) = p(\theta|D^{(t)}) \tag{293}$$

In this, perhaps somewhat artificial, way all cases which can be handled within the consistent Bayesian theory can be reduced to the case (II) when the adaptive problem does not contain any unknown parameters. It justifies the following statement which is due to [11]: "It seems that any systematic formulation of the adaptive control problem leads to a meta-problem which is not adaptive".

## 5.2 State Estimation and Output Prediction

The above given recursion (288), (289) and (291) has been derived for natural conditions of control (284) but no particular assumptions have been made concerning the finite dimensional internal quantity $\underline{q}_{(\tau)}$. Therefore, the recursion also holds if $\underline{q}_{(\tau)}$ is the state $\underline{x}_{(\tau)}$ of the system, i.e., if $\underline{q}_{(t)} = \underline{x}_{(t)}$. Then, given the state $\underline{x}_{(t)}$ and $\underline{u}_{(t)}$, neither the output $\underline{y}_{(t)}$ nor the next state $\underline{x}_{(t+1)}$ depend on the past history of the system and it holds

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}, x_{(t)}) = p(y_{(t)}|u_{(t)}, x_{(t)}) \tag{294}$$

$$p(x_{(t+1)}|x_{(t)}, D^{(t)}) = p(x_{(t+1)}|x_{(t)}, u_{(t)}) \tag{295}$$

In such a case the recursion (288), (289) and (291) gets the following form

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) = \int p(y_{(t)}|u_{(t)}, x_{(t)}) p(x_{(t)}|D^{(t-1)}) dx_{(t)} \tag{296}$$

$$p(x_{(t)}|D^{(t)}) = \frac{p(y_{(t)}|u_{(t)}, x_{(t)}) p(x_{(t)}|D^{(t-1)})}{p(y_{(t)}|u_{(t)}, D^{(t-1)})} \tag{297}$$

$$p(x_{(t+1)}|D^{(t)}) = \int p(x_{(t+1)}|u_{(t)}, x_{(t)}) p(x_{(t)}|D^{(t)}) dx_{(t)} \tag{298}$$

where the conditional probability distributions

$$p(x_{(t+1)}|x_{(t)}, u_{(t)}),\ p(y_{(t)}|u_{(t)}, x_{(t)})$$

[8] In multi-variate case the Kronecker's $\delta(z) = 1$ if all elements of z are equal to zero and $\delta(z) = 0$ at least one of the elements of z is nonzero. Similarly for multi-variate $\delta$-function

are defined by the state space model of the system. A special case (linear and Gaussian) of the recursion (296) to (298) is the Kalman filter as demonstrated in the following Example. This Example also may help to understand the statement made by Kalman (1965): "The Kalman-Bucy filter is in essence a conditional probability computer".

Example 5.1 (Kalman filter). Consider a system with $\nu$-dimensional output $\underline{y}_{(\tau)}$ and $\mu$ -dimensional input $\underline{u}_{(\tau)}$. Suppose that, on the basis of a physical analysis of the system, a $n$-dimensional state $\underline{x}_{(\tau)}$ is defined on the system and the system is described by the state-space model

$$x_{(t+1)} = Ax_{(t)} + Bu_{(t)} + w_{(t)} \tag{299}$$

$$y_{(t)} = Cx_{(t)} + Du_{(t)} + e_{(t)} \tag{300}$$

where $A$, $B$, $C$, $D$ are known matrices of appropriate dimensions. The discrete white noises $\{\underline{w}_{(\tau)}\}$ and $\{\underline{e}_{(\tau)}\}$ assumed to be mutually independent and normally distributed with zero mean values and known covariances

$$E[w_{(t)}\ w_{(t)}^T] = R_w, \quad E[e_{(t)}\ e_{(t)}^T] = R_e \tag{301}$$

For the sake of simplicity (to avoid degenerate and singular cases) it is assumed that the covariance matrices $R_w$ and $R_e$ are positive definite. Hence,

$$p(w_{(t+1)}|x_{(t)}, w_{(t)}) = p(w_{(t+1)}) \sim \mathcal{N}(0, R_w) \tag{302}$$

$$p(e_{(t)}|x_{(t)}, u_{(t)}) = p(e_{(t)}) \sim \mathcal{N}(0, R_e) \tag{303}$$

For given $\underline{x}_{(t)}$ and $\underline{u}_{(t)}$ the transformation between the random variables $\underline{w}_{(t)}$ and $\underline{x}_{(t+1)}$ defined by the state equation (299) is one-to-one with the Jacobian equal to 1 and consequently

$$p(x_{(t+1)}|x_{(t)}, u_{(t)}) \sim \mathcal{N}(Ax_{(t)} + Bu_{(t)} + R_w) \tag{304}$$

Similarly the output equation (300) and the distribution (303) define

$$p(y(t)|u_{(t)}, x_{(t)}) \sim \mathcal{N}(Du_{(t)} + Cx_{(t)} + R_e) \tag{305}$$

The purpose of this Example is to show that the following statement is true: If the conditional probability distribution $p(x_{(t)}|D^{(t-1)})$ is assumed to be normal then also $p(y_{(t)}|u_{(t)}, D^{(t-1)}), p(x_{(t)}|D^{(t)})$ and $p(x_{(t+1)}|D^{(t)})$ are normal – the normality is reproduced– and the functional recursion (296) to (298) can be reduced to an algebraic recursion operating only on conditional mean values and covariances. This algebraic recursion is the well known Kalman filter. The Bayesian view yields the precise probabilistic meaning of each step and of each number which appear in this recursion.

Let $\hat{x}_{(t|t-1)}$ and $S_{(t|t-1)}$ be the mean value and the covariance matrix of the conditional distribution

$$p(x_{(t)}|D^{(t-1)}) \sim \mathcal{N}(\hat{x}_{(t|t-1)}, S_{(t|t-1)}) \tag{306}$$

First, let us consider the product

$$\begin{aligned} p(y_{(t)}|u_{(t)}, x_{(t)})p(x_{(t)}|D^{(t-1)}) &= (2\pi)^{-\frac{\nu+n}{2}}|R_e|^{-\frac{1}{2}}|S_{(t|t-1)}|^{-\frac{1}{2}}\times \\ \times \exp\{-\frac{1}{2}[&(y_{(t)} - Du_{(t)} - Cx_{(t)})^T R_e^{-1}(y_{(t)} - Du_{(t)} - Cx_{(t)}) + \\ &+(x_{(t)} - \hat{x}_{(t|t-1)})^T S_{(t|t-1)}^{-1}(x_{(t)} - \hat{x}_{(t|t-1)})]\} \end{aligned} \tag{307}$$

which enters both (296) and (297). The exponent in (307) is a sum of two quadratic forms both of which contain the vector $x_{(t)}$ which has to be integrated out according to (296). To facilitate this integration we shall rearrange the exponent in such a way that it will consist of two quadratic forms but only one of them will depend on $x_{(t)}$. This can be done by completion of squares for $x_{(t)}$ (Lemma 3 in Appendix A)

and by an algebraic rearrangement if the remainder using matrix inversion lemma (Lemma 4 in Appendix A).

$$(y_{(t)} - Du_{(t)} - Cx_{(t)})^T R_e^{-1}(y_{(t)} - Du_{(t)} - Cx_{(t)}) + \tag{308}$$

$$+(x_{(t)} - \hat{x}_{(t|t-1)})^T S_{(t|t-1)}^{-1}(x_{(t)} - \hat{x}_{(t|t-1)}) =$$

$$= (x_{(t)} - \hat{x}_{(t|t)})^T S_{(t|t)}^{-1}(x_{(t)} - \hat{x}_{(t|t)}) + (y_{(t)} - \hat{y}_{(t|t-1)})^T R_{y(t|t-1)}^{-1}(y_{(t)} - \hat{y}_{(t|t-1)})$$

where

$$S_{(t|t)}^{-1} = S_{(t|t-1)}^{-1} + C^T R_e^{-1} C \tag{309}$$

$$\hat{x}_{(t|t)} = S_{(t|t)}[S_{(t|t-1)}^{-1}\hat{x}_{(t|t-1)} + C^T R_e^{-1}(y_{(t)} - Du_{(t)})] \tag{310}$$

$$\hat{R}_{y(t|t-1)} = R_e + CS_{(t|t-1)}C^T \tag{311}$$

$$\hat{y}_{(t|t-1)} = Du_{(t)} + C\hat{x}_{(t|t-1)} \tag{312}$$

When the product (307) with the exponent rearranged according to (308) is substituted into (296) the integration (296) can be easily performed using Lemma 8 from Appendix A.

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) =$$

$$= (2\pi)^{-\frac{p}{2}}|R_e|^{-\frac{1}{2}}|S_{(t|t-1)}|^{-\frac{1}{2}}|S_{(t|t)}|^{\frac{1}{2}}\times$$

$$\times \exp\{-\frac{1}{2}(y_{(t)} - Du_{(t)} - C\hat{x}_{(t|t-1)})^T\times$$

$$\times R_{y(t|t-1)}(y_{(t)} - Du_{(t)} - C\hat{x}_{(t|t-1)})\}$$

Applying the relation (414) from Lemma 5 in Appendix A to (309) we obtain

$$|S_{(t|t)}|^{-1} = \frac{|S_{(t|t-1)}^{-1}|}{|R_e|}|R_e + CS_{(t|t-1)}C^T|$$

$$|S_{(t|t-1)}|^{-1}|R_e|^{-1}|S_{(t|t)}| = |R_{y(t|t-1)}|^{-1}$$

Hence, the first step of the recursion (296) gives

$$p(y_{(t)}|u_{(t)}, D^{(t-1)}) \sim \mathcal{N}(Du_{(t)} + C\hat{x}_{(t|t-1)}, R_y(t|t-1)) \tag{313}$$

As the result of the second step (297) we obtain in a straightforward way

$$p(x_{(t)}|D^{(t)}) \sim \mathcal{N}(\hat{x}_{(t|t)}, S_{(t|t)}) \tag{314}$$

The third step of the recursion (298) can be performed essentially in the same way as the first step. For the product in the integrand of (298) we have from (304) and (314)

$$p(x_{(t+1)}|x_{(t)}, u_{(t)})p(x_{(t)}|D^{(t)}) = (2\pi)^{-n}|R_w|^{-\frac{1}{2}}|S_{(t|t)}|^{-\frac{1}{2}}\times \tag{315}$$

$$\times \exp\{-\frac{1}{2}[(x_{(t+1)} - Ax_{(t)} - Bu_{(t)})^T R_w^{-1}(x_{(t+1)} - Ax_{(t)} - Bu_{(t)}) +$$

$$+(x_{(t)} - \hat{x}_{(t|t)})^T S_{(t|t)}^{-1}(x_{(t)} - \hat{x}_{(t|t)})]\}$$

The exponent of (315) can be rearranged as follows

$$(x_{(t+1)} - Ax_{(t)} - Bu_{(t)})^T R_w^{-1}(x_{(t+1)} - Ax_{(t)} - Bu_{(t)}) + \tag{316}$$

$$+(x_{(t)} - \hat{x}_{(t|t)})^T S_{(t|t)}^{-1}(x_{(t)} - \hat{x}_{(t|t)}) =$$

$$= (x_{(t+1)} - \hat{x}_{(t+1|t)})^T S_{(t+1|t)}^{-1}(x_{(t+1)} - \hat{x}_{(t+1|t)}) +$$

$$+(x_{(t)} - z_{(t)})^T (S^{-1}_{(t|t)} + A^T R_w^{-1} A)(x_{(t)} - z_{(t)})^T$$

where

$$\hat{x}_{(t+1|t)} = A\hat{x}_{(t|t)} \tag{317}$$

$$S_{(t+1|t)} = R_w + AS_{(t|t)}A^T \tag{318}$$

The auxiliary vector $z_{(t)}$ does not need to be known, it is sufficient to know is determined by the relation

$$(A^T R_w^{-1} A + S^{-1}_{(t|t)}) z_{(t)} = A^T R_w^{-1}(x_{(t+1)} - Bu_{(t)}) + S^{-1}_{(t|t)}\hat{x}_{(t|t)}$$

With this rearrangement of the exponent in (315) the integration (298) according to Lemma 8 from Appendix A gives

$$p(x_{(t+1)}|D^{(t)}) = (2\pi)^{-\frac{n}{2}} |R_w|^{-\frac{1}{2}} |S_{(t|t)}|^{-\frac{1}{2}} |S^{-1}_{(t|t)} + A^T R_w^{-1} A|^{-\frac{1}{2}} \times \tag{319}$$

$$\times \exp\{-\frac{1}{2}(x_{(t+1)} - \hat{x}_{(t+1|t)})^T S_{(t+1|t)}(x_{(t+1)} - \hat{x}_{(t+1|t)})\}$$

However, from Lemma 5 we also have

$$|(S^{-1}_{(t|t)} + A^T R_w^{-1} A)| = \frac{|R_w + AS_{(t|t)}A^T|}{|S_{(t|t)}||R_w|} + \frac{|S_{(t+1|t)}|}{|S_{(t|t)}||R_w|}$$

which verifies that the normalizing factor in (319) is correct

$$p(x_{(t+1)}|D^{(t)}) \sim \mathcal{N}(\hat{x}_{(t+1|t)}, S_{(t+1|t)}) \tag{320}$$

and also closes the recursion.

Summing up we can see that the normal forms of conditional probability distributions (313) and (314) and (320) are reproduced and their mean values and covariances evolve according to the algebraic relations (309), (310), (311), (312), (317) and (318). It only remains to bring this algebraic recursion into a more convenient form. This can be done by using the matrix inversion lemma as the main tool. One of many possibilities is

$$\hat{y}_{(t|t-1)} = Du_{(t)} + C\hat{x}_{(t|t-1)} \tag{321}$$

$$\hat{R}_{y(t|t-1)} = R_e + CS_{(t|t-1)}C^T \tag{322}$$

$$S_{(t|t)} = S_{(t|t-1)} - S_{(t|t-1)}C^T R^{-1}_{y(t|t-1)} CS_{(t|t-1)} \tag{323}$$

$$\hat{x}_{(t|t)} = \hat{x}_{(t|t-1)} + S_{(t|t)}C^T R_e^{-1}(y_{(t)} - \hat{y}_{(t|t-1)}) \tag{324}$$

$$\hat{x}_{(t+1|t)} = A\hat{x}_{(t|t)} + Bu_{(t)} \tag{325}$$

$$S_{(t+1|t)} = R_w + AS_{(t|t)}A^T \tag{326}$$

This is the Kalman filter written in somewhat more detailed way than customary. We have shown that it applies also for state estimation and output prediction of a system controlled in closed loop, under natural conditions of control of course. It starts with the mean value $\hat{x}_{(1|0)}$ and the covariance matrix $\hat{S}_{(1|0)}$ of the prior (subjective) probability distribution $p(x_{(1)})$ which is assumed to be normal and reflects the statisticians uncertainty about the state $\underline{x}_{(1)}$ when no input-output data are available.

## 5.3 Slowly Varying Parameters and Exponential Forgetting

In practical situations the assumption that a certain set of parameters is strictly time-invariant is fulfilled only approximately and/or temporarily. Moreover, any mathematical model can be only an approximate description of reality and it may well happen that for different time intervals slightly different parameters of the chosen approximate model structure can be appropriate. Therefore, it is of high practical importance to have a tool available which makes it possible to extend the results obtained for constant unknown parameters also to the case of "slowly varying parameters" or, in other words, to extend the parameter estimation to parameter tracking. The extension developed in this subsection gives a subjective probability interpretation to the technique which is known under the names "exponential forgetting", age weighting" or "discounting" and appears to be successful in practical applications. See e.g. [5], [12], [32], [9].

If the unknown model parameters are allowed to be time-varying then the distinction between the set of unknown parameters $\theta_{(t)}$ and the set of internal variables $q_{(t)}$ which cannot be directly observed, actually, disappears. Hence, the general Bayesian solution of the case of time-varying unknown parameters is given by the recursion (288), (289) and (291) with $q_{(t)} = \theta_{(t)}$. It is the relation (291) performing the re-calculation (290)

$$p(\theta_{(t)}|D^{(t)}) \rightarrow p(\theta_{(t+1)}|D^{(t)}) \tag{327}$$

which makes the difference between the cases of constant and of time-varying unknown parameters. Notice that the same data set appears in the condition parts of the both probability distributions in (327). Therefore, the information which is necessary to perform the re-calculation (327) cannot be extracted from the new data but must be given externally. This external information is the model defining the conditional probability distribution

$$p(\theta_{(t+1)}|\theta_{(t)}, D^{(t)}) \tag{328}$$

which is required in the last step (291) of the recursion.

Now, let us consider the meaning of the vague term "slowly varying parameters". Loosely speaking, it means that the true values of parameters $\underline{\theta}_{(t+1)}$ in some sense, cannot lay far from $\theta_{(t)}$. Such a situation can be modelled by the conditional probability distribution (328) for $\underline{\theta}_{(t+1)}$ which is, for any $\theta_{(t)}$ given in the condition, highly concentrated around this value $\theta_{(t)}$. It is not difficult to see that, in such a case, the last operation (291) of the recursion, namely

$$p(\theta_{(t+1)}|D^{(t)}) = \int p(\theta_{(t+1)}|\theta^{(t)}D^{(t)})p(\theta_{(t)}|D^{(t)})d\theta_{(t)} \tag{329}$$

results in a slight "flattening" of $p(\theta_{(t)}|D^{(t)})$ to obtain $p(\theta_{(t+1)}|D^{(t)})$. This observation suggests the idea that, instead of trying to find the most appropriate model for the conditional probability distribution (328) (which is not an easy task, in general), it may be simpler and more advantageous of flattening (increase of uncertainty or loss of belief in the old estimate) to perform the re-calculation (327).

Suppose that the probability distribution for $\underline{\theta}_{(t)}$ given the observed data $D^{(t)}$ is a known function, say $f_{(t)}(.)$ i.e.

$$p(\theta_{(t)}|D^{(t)}) = f_{(t)}(\theta_{(t)}) \tag{330}$$

Then a simple way how to perform the flattening by introducing just one new parameter is

$$p(\theta_{(t+1)}|D^{(t)}) = \alpha_{(t+1|t)}[f_{(t)}(\theta_{(t+1)})]^{\phi^2_{(t+1)}} \tag{331}$$

where $\phi_{(t+1)}, |\phi_{(t+1)}| < 1$, is the parameter which will be called the forgetting factor, and $\alpha_{(t+1|t)}$ is the normalizing factor which does not depend on the unknown parameters $\underline{\theta}_{(t+1)}$

$$\alpha_{(t+1|t)} = \frac{1}{\int [f_{(t)}(\theta_{(t+1)})]^{\phi^2_{(t+1)}} d\theta_{(t+1)}} \tag{332}$$

The question of the choice of the forgetting factor will be discussed later on. Now it will be shown how the exponential forgetting can be applied to the generalized multi-variate regression model (190) with slowly varying parameters $P_{(t)}$ and $\Omega_{(t)} = R^{-1}_{(t)}$.

Suppose that the probability distribution for $\underline{\theta}_{(t)} = \{\underline{P}_{(t)}, \underline{\Omega}_{(t)}\}$ given $D^{(t)}$ is known and has the form (206)

$$p(P_{(t)}, \Omega(t)|D^{(t)}) =$$

$$= \alpha_{(t|t)}|\Omega|^{\frac{\theta_{(t|t)}}{2}} \times \exp\{-\frac{1}{2}\mathrm{tr}\left(\Omega \begin{bmatrix} -I_\nu \\ P_{(t)} \end{bmatrix}^T V_{(t|t)} \begin{bmatrix} -I_\nu \\ P_{(t)} \end{bmatrix}\right)\} \tag{333}$$

It is easy to see that the operation (331) preserves the form of the distribution modifying only its parameters

$$p(P_{(t+1)}, \Omega(t+1)|D^{(t)}) = \alpha_{(t+1|t)}|\Omega|\frac{\theta_{(t+1|t)}}{2}\times$$

$$\times \exp\{-\frac{1}{2}\mathrm{tr}\left(\Omega \begin{bmatrix} -I_\nu \\ P_{(t+1)} \end{bmatrix}^T V_{(t+1|t)} \begin{bmatrix} -I_\nu \\ P_{(t+1)} \end{bmatrix}\right)\} \tag{334}$$

where

$$V_{(t+1|t)} = \phi^2_{(t+1)}\ V_{(t|t)} \tag{335}$$

$$\theta_{(t+1|t)} = \phi^2_{(t+1)}\ \theta_{(t|t)} \tag{336}$$

and the normalizing factor $\alpha_{(t+1|t)}$, if required, can be determined similarly to (218), (216) and (217) with obvious replacement of indices.

Going through the derivation of the formulae for the regular stationary case it is easy to verify that very similar results hold also for the case with exponential forgetting. Instead of (207) and (208) we now have

$$V_{(t+1|t)} = \phi^2_{(t)}(V_{(t|t-1)} + d_{(t)}d^T_{(t)}) \tag{337}$$

$$\theta_{(t+1|t)} = \phi^2_{(t)}(\theta_{(t|t-1)} + 1) \tag{338}$$

or equivalently

$$V_{(t|t)} = \phi^2_{(t)}V_{(t-1|t-1)} + d_{(t)}d^T_{(t)} \tag{339}$$

$$\theta(t|t) = \phi^2_{(t)}\theta_{(t-1|t-1)} + 1 \tag{340}$$

If the matrix $V_{(t+1|t)}$ is partitioned similarly to (209) we obtain by analogy to (212), (211) and (213)

$$C_{(t+1|t)} = V^{-1}_{z(t+1|t)} = \frac{1}{\phi^2_{(t+1)}}V^{-1}_{z(t|t)} = \frac{1}{\phi^2_{(t+1)}}C_{(t|t)} \tag{341}$$

$$\hat{P}_{(t+1|t)} = C_{(t+1|t)}V_{zf(t+1|t)} = C_{(t|t)}V_{zf(t|t)} = \hat{P}_{(t|t)} \tag{342}$$

$$\Lambda_{(t+1|t)} = V_{f(t+1|t)} - V^T_{zf(t+1|t)}C_{(t+1|t)}V_{zf(t+1|t)} = \tag{343}$$

$$= \phi^2_{(t+1)}\Lambda_{(t|t)}$$

The algebraic recursion (248) to (253) is modified by the exponential forgetting in the following way.

$$g_{(t)} = C_{(t-1|t-1)}z_{(t)} \tag{344}$$

$$\zeta_{(t)} = z^T_{(t)}g_{(t)} \tag{345}$$

$$e_{(t|t-1)} = f_{(t)} - \hat{P}^T_{(t-1|t-1)}z_{(t)} \tag{346}$$

$$\hat{P}_{(t|t)} = \hat{P}_{(t-1|t-1)} + \frac{1}{\phi^2_{(t)} + \zeta_{(t)}}g_{(t)}e^T_{(t|t-1)} \tag{347}$$

$$\Lambda_{(t|t)} = \phi^2_{(t)}[\Lambda_{(t-1|t-1)} + \frac{1}{\phi^2_{(t)} + \zeta_{(t)}}e_{(t|t-1)}e^T_{(t|t-1)}] \tag{348}$$

$$C_{(t|t)} = \frac{1}{\phi^2_{(t)}}[C_{(t-1|t-1)} - \frac{1}{\phi^2_{(t)} + \zeta_{(t)}}q_{(t)}q^T_{(t)}] \tag{349}$$

As discussed in Section 4 it is numerically advantageous to perform this updating using the square-root filter REFIL from Appendix B. The forgetting factor is introduced into this subroutine through its input parameter $SIG2 = \phi^2_{(t)}$. Let us recall that if the subroutine REFIL is used to update only the square root of $C_{(t-1|t-1)}$ then it supplies, as its output parameters, $SIG2 = \phi^2_{(t)} + \zeta_{(t)}$ and $VG = g_{(t)}$ which is all what is required to update $P_{(t-1|t-1)}$ and also $\Lambda_{(t-1|t-1)}$ according to (346), (347) and (348).

The conditional probability distribution for the prediction of the next output $\underline{y}_{(t+1)}$ retains its form of the transformed $t$-distribution

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = J_{f(t+1)}\kappa_{(t+1|t)}[1 + e^T_{(t+1|t)}M^{-1}_{(t+1|t)}e_{(t+1|t)}]^{-\frac{\theta_{(t+1|t)}+\rho+\nu+2}{2}} \tag{350}$$

where

$$e_{(t+1|t)} = f_{(t+1)} - \hat{P}^T_{(t|t)}z_{(t+1)} \tag{351}$$

$$M_{(t+1|t)} = (1 + z^T_{(t+1)}C_{(t+1|t)}z_{(t+1)})\Lambda_{(t+1|t)} = (\phi^2_{(t+1)} + z^T_{(t+1)}C_{(t|t)}z_{(t+1)})\Lambda_{(t|t)} \tag{352}$$

$$\kappa_{(t+1|t)} = \frac{\Gamma(\frac{\theta^{-\rho+\nu+2}_{(t+1|t)}}{2})}{\pi^{\frac{\nu}{2}}\Gamma(\frac{\theta^{-\rho+2}_{(t+1|t)}}{2})}|M_{(t+1|t)}|^{-\frac{1}{2}} \tag{353}$$

and $J_{f(t+1)}$ is the Jacobian (193). For practical applications it may be worthy to note that the conditional probability for $e_{(t+1|t)}$, defined by (351) can be, for reasonably large $\theta_{(t+1|t)}$, well approximated by the normal distribution with zero mean and the covariance matrix

$$E[e_{(t+1|t)}e^T_{(t+1|t)}|u_{(t+1)}, D^{(t)}] = \frac{1}{\theta^{-\rho}_{(t+1|t)}}M_{(t+1|t)} \tag{354}$$

Now the question of the choice of the forgetting factor will be considered. In general, there are many strategies how the value of this parameter can be selected. First, let us consider the case when the forgetting factor is time-invariant, $\phi_{(t)} = \phi < 1$ for all $t > t_0$ and is chosen as a "fiddle parameter" the purpose of which is to weaken the stationarity assumption of the abstract mathematical model in order to make the theory more realistic. In such a case it follows from the relation (339)

$$V_{(t|t)} = \sum_{i=0}^{t-t_0-1}[\phi^i d_{(t-i)}][\phi^i d_{(t-i)}]^T + \phi^{2(t-t_0)}V_{(t_0|t_0)} \tag{355}$$

This shows that the data entering the statistic $V_{(t|t)}$ are weighted according to their age. The exponential window by which the past data are weighted is plotted for different $\phi$ in Fig. 11. The figure clearly shows that a reasonable choice of $\phi$ must lie rather close to one if the system is stochastic. The value $\phi^2 \approx 0.985$ ($\phi \approx 0.992$) has appeared as a reasonable starting point for by-hand tuning of the parameter-tracking algorithm in various practical applications.

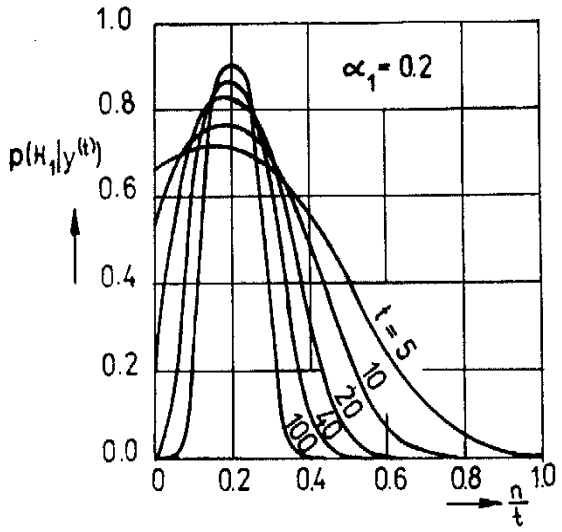



Fig. 11 Exponential forgetting of old data

As it is seen from the recursive relation (340) for constant $\phi$ the parameter $\theta_{(t|t)}$ converges to the steady state value $\theta = \frac{1}{1-\phi^2}$ which can be understood as an effective number of samples within the exponential window.

In many practical situations the parameters of a suitable model structure do not vary permanently but only from time to time when the operating conditions are changed for some, usually unknown, reasons. The exponential forgetting with a constant forgetting factor $\phi$ when applied to such situations, has the disadvantage that, on one hand, it suppresses the information about the unknown parameters which may be relevant ($\phi$ is too low) but, on the other hand, the filter may react on the changed conditions too slowly ($\phi$ is too high). The above given probabilistic interpretation of the exponential forgetting allows $\phi_{(t)}$ to be different for different $t$ and also gives, for any chosen $\phi_{(t)}$, the probability distribution for the next system output $\underline{y}_{(t)}$. This opens the possibility to verify the model in real time (ex post in each step) and to apply the forgetting when the newly observed output $\underline{y}_{(t)}$ indicates the change in model parameters. This seems to be one of simple and rationally based ways how to construct algorithms which are truly adaptive. An attempt in this direction has been made in [22], however, there are several other possibilities along this line which have not been exploited yet.

# 6 System Classification

Until now the model structure has been assumed to be given as prior information about the system studied and only a finite set of parameters $\theta$ of the given model structure has been assumed to be unknown. In this Section we will deal with a more general situation when more than one model structure have to be considered as possible.

Often the internal mechanism or physics of the system is not understood enough to specify the model structure uniquely. Suppose that the model builder is able to formulate a certain number of hypotheses about the possible model structure, he believes that one of his hypotheses is true but he does not know which one it is. If, in addition, the model builder is given a set of input-output data observed on the system he is facing the problem called by [33] the problem of system classification.

## 6.1 Model Classes and Hypotheses

Let $S_{\mathcal{M}}$ be the set of models which are considered as candidates to represent the system under study and let $\mathcal{M}$ be one particular model from this set, $\mathcal{M} \in S_{\mathcal{M}}$. In previous sections, when a single model structure was considered as possible, the system model $\mathcal{M}$ was specified by the set of constant parameters $\theta, \quad \theta \in S_\theta$, of the given model structure and the set of system models $S_{\mathcal{M}}$ was specified be the set $S_\theta$ of all possible parameter values. The present situation is more complex.

If in a particular model structure a certain set of parameters, say $\theta_i \in S_{\theta i}$ is assumed to be unknown, then the model structure and the set $S_{\theta i}$ of all parameter values which are considered as possible define the subset of models $\mathcal{C}_i \subset S_{\mathcal{M}}$ which will be called the class of system models. Then the set $S_{\mathcal{M}}$ of all system models under consideration is given as the union of all, say $N$, classes.

$$S_{\mathcal{M}} = \cup_{i=1}^{N} \mathcal{C}_i$$

In general, two different model classes may be generated by the same model structure but with two different sets of unknown parameters. A special case is when a class, say $\mathcal{C}_j$, contains just one model. This occurs when all parameters of the corresponding model structure are fixed as known; then the set of unknown parameters $\theta_j$ is empty. Another special case is when a class of models, say $\mathcal{C}_n$, is a subset of another class, say $\mathcal{C}_m$, $\mathcal{C}_n \subset \mathcal{C}_m$. This is, for instance, the case of a linear system with unknown order which can be either $m$ or $n$ and $n < m$. Hence, the classes are allowed to be overlapping - but only with probability zero. The precise meaning of and the reason for this restriction will be explained later on.

Let $\mathcal{M}_t$ be the true system model, i.e. the model equivalent to the system under study. The hypothesis that the true model $\mathcal{M}_t$ belongs to the class $\mathcal{C}_i$ will be denoted by $\mathcal{H}_i$. Using the Bayesian approach we shall describe the uncertainty of the hypotheses by a probability distribution on the set of hypotheses which are apriori considered as possibly true and we shall seek the solution of our problem of system classification in the form of the aposterior probability distribution, i.e. we are interested in the probability distribution on the set of hypotheses conditional on the input-output data observed on the system under test. Clearly, the probability of the hypothesis $\mathcal{H}_i$ is equal to the probability of the event $\mathcal{M}_t \in \mathcal{C}_i$ and the probability distribution we intend to determine is

$$p(\mathcal{H}_i|D^{(t)}) = \Pr[\mathcal{M}_t \in \mathcal{C}_i|\underline{D}^{(t)} = D^{(t)}], \tag{356}$$

$$i = 1, 2, \ldots, N$$

Perhaps, it may be helpful to the reader if we explain in more detail why we formulate the problem of system classification in terms of probability distribution and not as a decision problem, i.e. as the choice of model structure. As it will be shown, in some applications, like the prediction of the future output of an uncertain system, an explicite choice of the model structure is, actually, not required and it is possible, and conceptually correct, to calculate simultaneously with all model structures which are considered as possible, of course, with corresponding weights determined by their probabilities.

We have also another reason why we do not want to mix up the statistical data analysis with a decision. To be able to formulate the decision problem properly it would be necessary to have a particular objective in mind for which the decision is taken and to define the utility function. This is not an easy task in general. However, when the amount of data is large enough it often happens that the aposterior probability of some hypothesis, say $\mathcal{H}_k$, is much larger than the probabilities of the other hypotheses, $p(\mathcal{H}_k|D^{(t)}) >> p(\mathcal{H}_i|D^{(t)})$ for all $i \neq k$. Then, accepting the hypothesis as true we can be sure that the same decision would be obtained for a rather broad class of utility functions and we do not need to lose time and energy in trying need to lose time and energy in trying to find out the one among them which were most appropriate for the given purpose.

To be able to determine the aposterior probability distribution (356) it is necessary to define the prior probability distribution on the entire set $S_{\mathcal{M}}$ of all models. This can be done by assigning the prior probability to each of the hypotheses, $p(\mathcal{H}_i), \quad i = 1, 2, \ldots, N$, and by introducing the prior probability distribution on the set of possible parameter values within each of the hypotheses, $p(\theta_i|\mathcal{H}_i), \quad i = 1, 2, \ldots, N$. The former is the model of the statisticians prior uncertainty about the validity of his hypotheses before

the data are incorporated into his knowledge. The latter reflects the statistician's prior uncertainty about the values of unknown parameter $\theta_i$ assuming that the hypothesis $\mathcal{H}_i$ were true. Clearly, the product $p(\theta_i|\mathcal{H}_i)p(\mathcal{H}_i)$ assigns a prior probability to every subset of models within the class $\mathcal{C}_i$.

It is reasonable to formulate the hypotheses in such a way that they are mutually incompatible, i.e. that only one of them can be true at the same time. Then it must hold

$$\sum_{i=1}^{N} p(\mathcal{H}_i) = \sum_{i=1}^{N} Pr[\mathcal{M}_t \in \mathcal{C}_i] = 1$$

On the other hand, as the event $\mathcal{M}_t \in S_{\mathcal{M}}$ is assumed to be certain, it must also hold

$$Pr[\mathcal{M}_t \in S_{\mathcal{M}}] = Pr[\mathcal{M}_t \in \cup_{i=1}^{N} \mathcal{C}_i] = 1$$

These two conditions can be satisfied simultaneously only when

$$Pr[\mathcal{M} \in \mathcal{C}_i \cap \mathcal{C}_j] = 0 \quad \forall i,\ j \neq i$$

This means that a subset of models which is common for two or more classes may obtain a nonzero prior probability (and consequently also aposterior probability) only through one of the hypotheses. This is the meaning of the restriction imposed on the model classes that they can overlap only with probability zero.

If all of the hypotheses can be considered as a priori equally likely then the natural choice of the prior probability distribution on the set of hypotheses is $p(\mathcal{H}_i) = \frac{1}{N}$ for all $i$. The choice of suitable prior $p(\theta_i|\mathcal{H}_i)$ is a more crucial question than it was in parameter estimation in Section 4 and will be discussed separately later on.

## 6.2 Natural Conditions of Control in System Classification

To be able to extract all relevant information about the class of the system which is carried by the experimental input-output data it must be specified under what conditions the input is generated during the experiment. It will be again assumed that the natural conditions of control are fulfilled, i.e. that the information about the system under test which is used to generate the input $u_{(\tau)}$ is equal or less than that which is available a priori and from the data $D^{(\tau-1)}$ known when the input $u_{(\tau)}$ is decided. Consequently the sole $u_{(\tau)}$ cannot bring any additional information neither about the true class of the system nor about its parameters and it holds for all $i$

$$p(\mathcal{H}_i|u_{(\tau)}, D^{(\tau-1)}) = p(\mathcal{H}_i|D^{(\tau-1)}) \tag{357}$$

and also

$$p(\theta_i|\mathcal{H}_i, u_{(\tau)}, D^{(\tau-1)}) = p(\theta_i|\mathcal{H}_i, D^{(\tau-1)}) \tag{358}$$

The obvious relation

$$p(\mathcal{H}_i|u_{(\tau)}, D^{(\tau-1)})p(u_{(\tau)}|D^{(\tau-1)}) = p(u_{(\tau)}|D^{(\tau-1)}, \mathcal{H}_i)p(\mathcal{H}_i|D^{(\tau-1)})$$

shows that the equality (357) implies

$$p(u_{(\tau)}|D^{(\tau-1)}, \mathcal{H}_i) = p(u_{(\tau)}|D^{(\tau-1)}) \tag{359}$$

and reversely. Similarly the equality (358) is equivalent to

$$p(u_{(\tau)}|D^{(\tau-1)}, \theta_i, \mathcal{H}_i) = p(u_{(\tau)}|D^{(\tau-1)}, \mathcal{H}_i) = p(u_{(\tau)}|D^{(\tau-1)}) \tag{360}$$

Hence, the natural conditions of control can be formally introduced either by the two equalities (357) and (358) or by (360).

## 6.3 Formal Solution of the Classification Problem

For the sake of generality and for latter use the problem will be formally posed as follows:

Given $p(\mathcal{H}_i|D^{(t_s)})$ and $p(\theta_i|\mathcal{H}_i, D^{(t_s)})$ for some $t_s \geq 0$ and $i = 1, 2, \ldots, N$ calculate $p(\mathcal{H}_i|D^{(t)})$ for $t > t_s$, assuming that the natural conditions of control are satisfied.

The Bayes formula gives

$$p(\mathcal{H}_i|D^{(t)}) = \frac{p(D^{(t)}_{(t_s+1)}|D^{(t_s)}, \mathcal{H}_i)p(\mathcal{H}_i|D^{(t_s)})}{\sum_{j=1}^{N} p(D^{(t)}_{(t_s+1)}|D^{(t_s)}, \mathcal{H}_j)p(\mathcal{H}_j|D^{(t_s)})} \tag{361}$$

Applying the chain rule (19) the first factor in the numerator can be written as follows

$$p(D^{(t)}_{(t_s+1)}|D^{(t_s)}, \mathcal{H}_i) = \prod_{\tau=t_s+1}^{t} p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \mathcal{H}_i)p(u_{(\tau)}|D^{(\tau-1)}, \mathcal{H}_i)$$

When the natural conditions of control (359) are considered and the following simplification of notation is introduced

$$p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) = p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \mathcal{H}_i) \tag{362}$$

the formula (361) gets the form

$$p(\mathcal{H}_i|D^{(t)}) = \frac{\prod_{\tau=t_s+1}^{t} p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)})p(\mathcal{H}_i|D^{(t_s)})}{\sum_{j=1}^{N} \prod_{\tau=t_s+1}^{t} p_j(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)})p(\mathcal{H}_j|D^{(t_s)})} \tag{363}$$

Further, under natural conditions of control it holds

$$p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) = \int p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta_i)p_i(\theta_i|D^{(\tau-1)})d\theta_i \tag{364}$$

where, similarly to (362), we use the simplified notation

$$p_i(\theta_i|D^{(\tau-1)}) = p_i(\theta_i|D^{(\tau-1)}, \mathcal{H}_i)$$

Notice that the conditional probability distribution (364) is the Bayesian prediction of the output $\underline{y}_{(\tau)}$ given the past input-output history but not the unknown parameters $\theta_i$, all within the i-th hypothesis. For given data $D^{(t)}$ and $\tau < t$, i.e. for the observed output $\underline{y}_{(\tau)} = y_{(\tau)}$, the left-hand side of (364) is just one number - the ordinate of the conditional probability density for a continuous $\underline{y}_{(\tau)}$ or the ordinate of the conditional probability function if $\underline{y}_{(\tau)}$ is discrete or if it is an event. The products of these ordinates for each of the hypotheses determine, according to the formula (363), the relation between the prior and aposterior probability distribution on the set of hypotheses. From the formula (363) it follows that for the aposterior probability ratio for any two of the $N$ hypotheses it holds

$$\frac{p(\mathcal{H}_i|D^{(t)})}{p(\mathcal{H}_j|D^{(t)})} = \prod_{\tau=t_s+1}^{t} \frac{p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)})}{p_j(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)})} \frac{p(\mathcal{H}_i|D^{(t_s)})}{p(\mathcal{H}_j|D^{(t_s)})} \tag{365}$$

Clearly, any $N-1$ finite ratios (365) for $i \neq j$ and the condition

$$\sum_{k=1}^{N} p(\mathcal{H}_k|D^{(t)}) = 1 \tag{366}$$

determine uniquely the entire probability distribution on the set of $N$ hypotheses.

Instead of determining the aposterior probability ratio $p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)})/p_j(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)})$ for each $\underline{y}_{(\tau)} = y_{(\tau)}$ observed it may be sometimes more convenient to proceed as follows. In analogy to (94) we have for each of the hypotheses

$$p_i(\theta_i|D^{(\tau-1)}) = \frac{\prod_{k=t_s+1}^{\tau-1} p_i(y_{(k)}|u_{(k)}, D^{(k-1)}, \theta_i)p_i(\theta_i|D^{(t_s)})}{\int \prod_{k=t_s+1}^{\tau-1} p_i(y_{(k)}|u_{(k)}, D^{(k-1)}, \theta_i)p_i(\theta_i|D^{(t_s)})d\theta_i} \tag{367}$$

For the given input-output data $D^{(t)}$, $t > \tau$, introduce the values of integrals

$$I_{i(\tau|t_s)} = \int \prod_{k=t_s+1}^{\tau} p_i(y_{(k)}|u_{(k)}, D^{(k-1)}, \theta_i) p_i(\theta_i|D^{(t_s)}) d\theta_i \tag{368}$$

Then (364), when the second factor of the integrand is substituted by (367), reads

$$p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) = \frac{I_{i(\tau|t_s)}}{I_{i(\tau-1|t_s)}} \tag{369}$$

As, according to the definition (368), $I_i(t_s|t_s) = 1$ the aposterior probability ratio (365) is now obtained in the form

$$\frac{p(\mathcal{H}_i|D^{(t)})}{p(\mathcal{H}_j|D^{(t)})} = \frac{I_{i(t|t_s)}}{I_{j(t|t_s)}} \frac{p(\mathcal{H}_i|D^{(t_s)})}{p(\mathcal{H}_j|D^{(t_s)})} \tag{370}$$

and for $t_s = 0$

$$\frac{p(\mathcal{H}_i|D^{(t)})}{p(\mathcal{H}_j|D^{(t)})} = \frac{I_{i(t|0)}}{I_{j(t|0)}} \frac{p(\mathcal{H}_i)}{p(\mathcal{H}_j)} \tag{371}$$

$$I_{k(t|0)} = \int L_{k(t)}(\theta_k) p(\theta_k) d\theta_k \tag{372}$$

where

$$L_{k(t)}(\theta_k) = \prod_{\tau=1}^{t} p_k(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta_k) \tag{373}$$

is the likelihood function for the $k$-th hypothesis. These relations will be discussed in more detail later on. Now, some further general relations will be added.

Suppose that, given the input-output data $D^{(t)}$, it is required to predict the next output $\underline{y}_{(t+1)}$ for any input $\underline{u}_{(t+1)}$ applied. Then the predictive conditional probability distribution is determined as follows

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \sum_{i=1}^{N} p(y_{(t+1)}, \mathcal{H}_i|u_{(t+1)}, D^{(t)})$$

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \sum_{i=1}^{N} p_i(y_{(t+1)}|u_{(t+1)}, D^{(t)}) p(\mathcal{H}_i|D^{(t)}) \tag{374}$$

where $p_i(y_{(t+1)}|u_{(t+1)}, D^{(t)})$ is the predictive conditional probability distribution within the $i$-th hypotheses determined according to the formula (364) for $\tau = t+1$ as a function of $y_{(t+1)}$ and $u_{(t+1)}$. This clearly shows that, if the objective of the statistical data analysis is to provide rational basis for prediction or control of the output, one does not necessarily need - if one is not forced by other reasons - to take decision concerning the model structure, i.e. to chose a single hypothesis as true.

If it is possible to express the integrals (372) for $t+1$ as functions of $y_{(t+1)}$ and $u_{(t+1)}$ it may be advantageous to bring the conditional probability distribution (374) for the output prediction under uncertainty of the hypotheses into the following form.

$$p(y_{(t+1)}|u_{(t+1)}, D^{(t)}) = \frac{\sum_{i=1}^{N} I_{i(t+1|0)}(y_{(t+1)}, u_{(t+1)}) p(\mathcal{H}_i)}{\sum_{i=1}^{N} I_{i(t|0)} p(\mathcal{H}_i)} \tag{375}$$

The general formula (363) yields for $t_s = t-1$ the recursion for real time updating of the probability distribution on the hypotheses

$$p(H_i|D^{(t)}) = \frac{p_i(y_{(t)}|u_{(t)}, D^{(t-1)})}{p(y_{(t)}|u_{(t)}, D^{(t-1)})} p(H_i|D^{(t-1)}) \tag{376}$$

where the denominator of the updating factor is the ordinate of the overall predictive probability density (or probability function) (374) from the previous step for the newly observed output $\underline{y}_{(t)} = y_{(t)}$.

The purpose of the following Example is to demonstrate how the theory works in a simple and well comprehensible case.

Example 6.1 Consider an autonomous system the output of which, $\underline{y}_{(\tau)}$ is an event with just two possible outcomes, $y_{(\tau)} \in \{A, \overline{A}\}$. Suppose that any model $\mathcal{M}$ which can be considered as a candidate to describe the output process fulfills the assumption

$$p(y_{(\tau)}|y^{(\tau-1)}, \mathcal{M}) = p(y_{(\tau)}|\mathcal{M}) \quad = \quad \alpha \quad for \quad y_{(\tau)} \quad = A$$
$$= 1 - \alpha \quad for \quad y_{(\tau)} \quad = \overline{A}$$

Hence, the set $S_{\mathcal{M}}$ of all models which are a priori accepted as possible is generated by the same model structure and each model $\mathcal{M} \in S_{\mathcal{M}}$ is identified by the value of a single parameter $\alpha$, $0 \leq \alpha \leq 1$. To have a physical situation in mind the reader may consider the process of tossing a coin which may be unfair.

Suppose that we have a reason to assume that the true value of $\alpha$ is $\alpha_t = \alpha_1$ (for instance, that the tossed coin is fair, $\alpha_1 = \frac{1}{2}$), but not being quite sure that this hypothesis $\mathcal{H}_1$ is true we may wish to compare it with the other hypothesis $\mathcal{H}_2$ according to which the true value $\alpha_t$ is any other value from the interval $0 \leq \alpha \leq 1$.

Clearly the class $\mathcal{C}_1 = \{\mathcal{M} : \alpha = \alpha_1\}$ associated with the hypothesis $\mathcal{H}_1$ contains just one model the prior probability of which is directly $p(\mathcal{H}_1)$. The set $\theta_1$ is empty since under the hypothesis $\mathcal{H}_1$ all parameters are known. The class associated with the alternative hypothesis $\mathcal{H}_2$ can be chosen as $\mathcal{C}_2 = \{\mathcal{M} : 0 \leq \alpha \leq 1\}$. Apparently, $\mathcal{C}_1 \subset S_{\mathcal{M}} \equiv \mathcal{C}_2$ but if we describe the uncertainty of the unknown parameter $\alpha$ in $\mathcal{H}_2$ by any probability density $p(\alpha|\mathcal{H}_2) = p_2(\alpha)$ which is continuous for $\alpha = \alpha_1$, then the probability which is assigned to the model $\alpha = \alpha_1$ through the hypothesis $\mathcal{H}_2$ is zero and the condition that the classes $\mathcal{C}_1$ and $\mathcal{C}_\in$ may overlap only with probability zero is fulfilled.

Considering all values of $\alpha$ within the alternative hypothesis $\mathcal{H}_2$ as equally likely we choose $p_2(\alpha) = 1$. Trying to be "fair" we also choose $p(\mathcal{H}_1) = p(\mathcal{H}_2) = \frac{1}{2}$.

If the integer $n$ denotes the number of observed outputs the outcome of which was $y_{(\tau)} \equiv A$, $1 \leq \tau \leq t$, then the formulae (372) and (373) give for the alternative hypothesis $\mathcal{H}_2$

$$I_{2(t|0)} = \int_0^1 \alpha^n (1-\alpha)^{t-n} d\alpha = \frac{n!(t-n)!}{(t+1)!}$$

while for the tested hypothesis $\mathcal{H}_1$ we simply have

$$I_{1(t|0)} = \prod_{\tau=1}^{t} p_1(y_{(\tau)}) = \alpha_1^n (1-\alpha_1)^{t-n}$$

Employing the general formula (371) the aposterior probability of the tested hypothesis $\mathcal{H}_1$ is obtained as follows.

$$p(\mathcal{H}_1|y^{(t)}) = \frac{1}{1 + \frac{p(\mathcal{H}_2|y^{(t)})}{p(\mathcal{H}_1|y^{(t)})}} = \frac{1}{1 + \frac{I_{2(t|0)}}{I_{1(t|0)}} \frac{p(\mathcal{H}_2)}{p(\mathcal{H}_1)}} = (1 + \frac{n!(t-n)!}{(t+1)!\alpha_1^n (1-\alpha_1)^{t-n}})^{-1}$$

This result is illustrated in Fig.12 where it is plotted as a function of the relative frequency $n/t$ for different $t$ and $\alpha_1 = 0.2$. The complement to one in this Figure is, of course, the aposterior probability of the alternative hypothesis $p(\mathcal{H}_2|y^{(t)})$

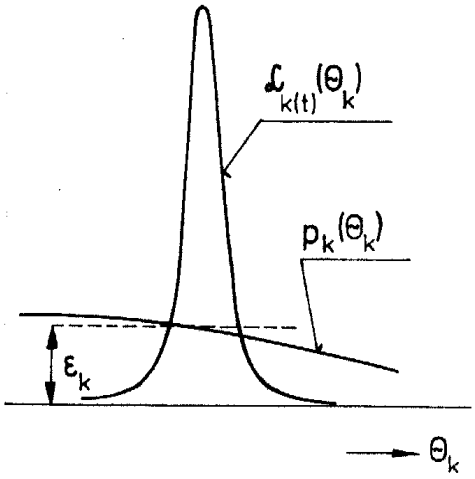



Fig. 12 Bernoullian trials - $\mathcal{H}_1 : \alpha = 0.2$ versus $\mathcal{H}_2 : \alpha \in < 0, 1 >$

## 6.4 Role of Prior in Classification

In Section 4 it has been shown that the prior distribution on the set of unknown parameters does not play a significant role in parameter estimation whenever the principle of stable estimation applies and the prior information about the possible values of unknown parameters is negligible compared to that carried by the data. Unfortunately, in system classification the problem of how to choose the prior distribution to model the situation when little is known a priori relatively to what the data can say is relatively to what the data can say is much more intricate.

To enlighten the problem let us write the general formula (370) in the following form

$$\frac{p(\mathcal{H}_i|D^{(t)})}{p(\mathcal{H}_j|D^{(t)})} = \frac{\int L_{i(t)}(\theta_i)p_i(\theta_i)d\theta_i}{\int L_{j(t)}(\theta_j)p_j(\theta_j)d\theta_j}\frac{p(\mathcal{H}_i)}{p(\mathcal{H}_j)} \tag{377}$$

and let us consider the regular case when the data $D^{(t)}$ do carry the information about the unknown parameters both in the hypothesis $\mathcal{H}_i$ and in the hypothesis $\mathcal{H}_j$ so that, for $t$ large enough, the likelihood functions $L_{k(t)}(\theta_k), \quad k = i, j$ are well peaked as sketched in Fig.13.

Then, supposing that the unknown parameters are uncertain quantities of continuous type and that the integrals over the likelihood functions exist, the right-hand side of (377) can be well approximated as follows

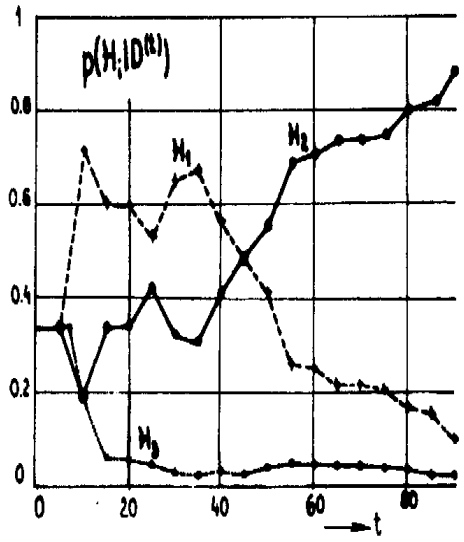



Fig.13 Likelihood function combined with a flat prior distribution

$$\frac{p(\mathcal{H}_i|D^{(t)})}{p(\mathcal{H}_j|D^{(t)})} = \frac{\int L_{i(t)}(\theta_i)d\theta_i}{\int L_{j(t)}(\theta_j)d\theta_j}\frac{\epsilon_i}{\epsilon_j}\frac{p(\mathcal{H}_i)}{p(\mathcal{H}_j)} \tag{378}$$

where $\epsilon_k$ is some value of the prior probability density $p_k(\theta_k)$ in the region where the likelihood $L_{k(t)}(\theta_k)$ is concentrated. If the set of possible values of the unknown parameters $\theta_k$ is unbounded then by flattening of the prior probability density $p_k(\theta_k)$ which must integrate to one, the value $\epsilon_k$ can be made arbitrarily small. Unlike the parameter estimation (recall the discussion of Eq. (125)) no reasonable limit of the ratio $\epsilon_i/\epsilon_j$ can be found, especially when the parameter sets $\theta_i$ and $\theta_j$ are of different nature and maybe of different dimensions. Notice that we had no difficulties of this kind in Example 6.1 because the set of possible values of the parameter $\alpha$ in the hypothesis $\mathcal{H}_2$ was bounded.

The relation (378) indicates that, theoretically, by the choice of prior, i.e. by the choice of the ratios $\epsilon_i/\epsilon_j$ one can arbitrarily influence the aposterior probability of any of compared hypotheses. Practically, the situation is not as much crucial as it might seem. As Examples will demonstrate later on, for growing $t$ the ratio of integrals over the likelihood diverges, if the hypothesis $\mathcal{H}_i$ is true (or converges to zero, if $\mathcal{H}_j$ is true) so rapidly that it dominates very soon any reasonably chosen ratio $\epsilon_i p(\mathcal{H}_i)/\epsilon_j p(\mathcal{H}_j)$ and for growing $t$ the aposterior probability of the true hypothesis will converge to one in any case. Usually, considering the physical nature of the case studied, it is not difficult to make an appropriate choice of prior. Nevertheless, it must be emphasized that for small or medium data size the choice of prior distributions on the sets of unknown parameters must be made with caution. In the following subsection a procedure will be developed which does not require an explicit choice of priors and reflects the situation when "little is known a priori".

## 6.5 Let Data Speak for Themselves

The user of the Bayesian theory may invite a procedure which makes him free of thinking much about a suitable choice of prior probability distributions if he knows little a priori compared to what the data themselves can say and which makes the theory applicable in a straightforward way. It is the objective of this subsection to develop such a procedure.

First, let us consider the problem of initial data which already appeared in parameter estimation(see Section 4). If the model structure associated with one of the hypotheses, say $\mathcal{H}_i$, has the form of

a stochastic difference equation relating the output $y_{(\tau)}$ with a finite number of delayed outputs and inputs, $D^{(\tau-1)}_{(\tau-n_i)} = \{y^{(\tau-1)}_{(\tau-n_i)}, u^{(\tau-1)}_{(\tau-n_i)}\}$, then the initial conditions – the unknown data for $\tau < 1$ – must be considered as unknown parameters if the model has to define the conditional probability distributions

$$p(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta_i, \mathcal{H}_i) = p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}, \theta_i) \tag{379}$$

right from the beginning of observation, i.e. also for $\tau = 1$. However, if we neglect a small piece of information about the system which could be extracted from the data on the basis of the prior information about the unknown initial conditions we may proceed as follows.

Let us take the very first data $D^{(n_i)}$ as known initial conditions for the model structure associated with the hypothesis $\mathcal{H}_i$. Then the set of unknown parameters is reduced, say to $\theta_i$, but the model defines the probability distributions (379) only for $\tau > n_i$ and $p_i(\theta_i|D^{(n_i)})$ has to be taken as the prior distribution for the set of unknown parameters $\theta_i$. In analogy to (111) we have

$$p_i(\theta_i|D^{(n_i)}) = p_i(\theta_i) \tag{380}$$

For $t_s = n_i$ the integral (368) is

$$I_{i(\tau|n_i)} = \int \tilde{L}_{i(\tau)}(\theta_i) p_i(\theta_i) d\theta_i \tag{381}$$

where $\tilde{L}_{i(\tau)}$ is the conditional likelihood function (110) for the $i$-th hypothesis and for the data up to and including $\tau$.

$$\tilde{L}_{i(\tau)}(\theta_i) = \prod_{k=n_i+1}^{\tau} p_i(y_{(k)}|u_{(k)}, D^{(k-1)}, \theta_i) \tag{382}$$

For $t_s = n_i$ the formula (369) gives

$$p_i(y_{(\tau)}|u_{(\tau)}, D^{(\tau-1)}) = \frac{I_{i(\tau|n_i)}}{I_{i(\tau-1|n_i)}} \tag{383}$$

and the relation (365) can be written for $t_s \geq max(n_i, n_j)$ as follows.

$$\frac{p(\mathcal{H}_i|D^{(t)})}{p(\mathcal{H}_j|D^{(t)})} = \frac{\frac{I_{i(t|n_i)}}{I_{i(t_s|n_i)}}}{\frac{I_{j(t|n_j)}}{I_{j(t_s|n_j)}}} \frac{p(\mathcal{H}_i|D^{(t_s)})}{p(\mathcal{H}_j|D^{(t_s)})} \tag{384}$$

The purpose of this rearrangement is to express the relation through the ratios of integrals in which he same prior distribution $p_i(\theta_i)$ appears. This makes it possible to introduce non-informative improper prior in the following way.

Usually, the uniform prior distribution can be taken as non-informative but, as discussed in Section 4, in some cases, like (133), an other form may be more suitable. To proceed generally let

$$p_i(\theta_i) = \phi_i(\theta_i), \qquad \theta_i \in S_{\theta i} \tag{385}$$

be the non-informative prior we want to introduce. If the set $S_{\theta i}$, on which it is defined, is unbounded the distribution (385) may be improper, i.e. does not integrate to one. Let us consider a related proper distribution defined on a bounded subset $S^*_{\theta i} \subset S_{\theta i}$

$$p_i(\theta_i) = k^*_i \quad \phi_i(\theta_i) \qquad \theta_i \in S^*_{\theta i}$$

Notice that the normalizing factor $k^*_i$ cancels in the ratio of integrals

$$\frac{I_{i(t|n)}}{I_{i(t_s|n_i)}} = \frac{\int_{S^*_{\theta i}} \tilde{L}_{i(t)}(\theta_i)\phi_i(\theta_i)d\theta_i}{\int_{S^*_{\theta i}} \tilde{L}_{i(t_s)}(\theta_i)\phi_i(\theta_i)d\theta_i} \tag{386}$$

If $t_s$ is large enough so that the integral

$$\lambda_{i(t)} = \int \tilde{L}_{i(t)}(\theta_i)\phi_i(\theta_i)d\theta_i \tag{387}$$

taken over the entire set $S_{\theta i}$ exist for $t \geq t_s$ then the limit of the ratio (386) for $S^*_{\theta i} \to S_{\theta i}$ also exist and is

$$\lim_{S^*_{\theta_i} \to S_{\theta_i}} \frac{I_{i(t|n_i)}}{I_{i(t_s|n_i)}} = \frac{\lambda_{i(t)}}{\lambda_{i(t_s)}}$$

Let $t_i$ be the first time instant for which the integral (387) exist and let us choose the indexing of the hypotheses in such a way that

$$t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_{N-1} \leq t_N$$

Then for $t_s = t_N$ and $t > t_N$ the relation (384) reads

$$\frac{p(\mathcal{H}_i|D^{(t)})}{p(\mathcal{H}_j|D^{(t)})} = \frac{\frac{\lambda_{i(t)}}{\lambda_{i(t_N)}}}{\frac{\lambda_{j(t)}}{\lambda_{j(t_N)}}} \frac{p(\mathcal{H}_i|D^{(t_N)})}{p(\mathcal{H}_j|D^{(t_N)})} \tag{388}$$

and employing the obvious relation

$$\sum_{j=1}^{N} p(\mathcal{H}_j|D^{(t)}) = 1 \tag{389}$$

the following formula is obtained for $t > t_N$

$$p(\mathcal{H}_i|D^{(t)}) = \frac{\frac{\lambda_{i(t)}}{\lambda_{i(t_N)}} p(\mathcal{H}_i|D^{(t_N)})}{\sum_{j=1}^{N} \frac{\lambda_{j(t)}}{\lambda_{j(t_N)}} p(\mathcal{H}_j|D^{(t_N)})} \tag{390}$$

For $t \leq t_n$ the data themselves (without prior information about unknown parameters) cannot correct the prior probability $p(\mathcal{H}_N)$ and it is natural to define $p(\mathcal{H}_N|D^{(t)}) = p(\mathcal{H}_N)$ for $t \leq t_N$ and similarly

$$p(\mathcal{H}_j|D^{(t)}) = p(\mathcal{H}_j) \ for \ t \leq t_j \tag{391}$$

To be able to use the formula (390) it is necessary to express the probabilities $p(\mathcal{H}_j|D^{(t_N)})$, $j < N$ through the prior probabilities (391). If $t$ is from the interval $t_k < t \leq t_{k+1}$ then $p(\mathcal{H}_j|D^{(t)}) = p(\mathcal{H}_j)$ for all $j > k$, (389) can be written as follows

$$\sum_{j=1}^{k} p(\mathcal{H}_j|D^{(t)}) = 1 - \sum_{j=k+1}^{N} p(\mathcal{H}_j), \quad t_k < t \leq t_{k+1}$$

and it holds for $i \leq k$

$$p(\mathcal{H}_i|D^{(t)}) = \frac{1 - \sum_{j=k+1}^{N} p(\mathcal{H}_j)}{\sum_{j=1}^{k} \frac{p(H_j|D^{(t)})}{p(H_i|D^{(t)})}} = \left[1 - \sum_{j=k+1}^{N} p(\mathcal{H}_j)\right] \times \tag{392}$$

$$\times \frac{\frac{\lambda_{i(t)}}{\lambda_{i(t_k)}} p(\mathcal{H}_i|D^{(t_k)})}{\frac{\lambda_{k(t)}}{\lambda_{k(t_k)}} p(\mathcal{H}_k) + \sum_{j=1}^{k-1} \frac{\lambda_{j(t)}}{\lambda_{j(t_k)}} p(\mathcal{H}_j|D^{(t_k)})}$$

Using the relations (390) and (392) it is possible to derive in a straightforward way that the following formulae hold for the non-informative prior distribution on the hypotheses $p(\mathcal{H}_k) = \frac{1}{N}, \quad k = 1, 2, \ldots, N$

$$p(\mathcal{H}_i|D^{(t)}) = \frac{\frac{\lambda_{i(t)}}{\lambda_{i(t_i)}} K_i}{\sum_{j=1}^{N} \frac{\lambda_{j(t)}}{\lambda_{j(t_j)}} K_j}, \quad t \geq t_N \tag{393}$$

where $K_k, \quad k = 1, 2, \ldots, N$, are defined recursively

$$K_1 = 1$$

$$K_k = \frac{1}{k-1} \sum_{j=1}^{k-1} \frac{\lambda_{j(t_k)}}{\lambda_{j(t_j)}} K_j \tag{394}$$

If the probability distributions for $t < t_N$ are of interest then they can be determined as follows

$$p(\mathcal{H}_i | D^{(t)}) = \frac{k}{N} \frac{\frac{\lambda_{i(t)}}{\lambda_{i(t_i)}} K_i}{\sum_{j=1}^{k} \frac{\lambda_{j(t)}}{\lambda_{j(t_j)}} K_j}, \quad t_k < t \leq t_{k+1}, \; i \leq k$$

$$p(\mathcal{H}_i | D^{(t)}) = \frac{1}{N}, \quad t_k < t \leq t_{k+1}, \; i > k$$

For large and medium data size it is usually possible to use, instead of (393) the simpler formula

$$p(\mathcal{H}_i | D^{(t)}) = \frac{\frac{\lambda_{i(t)}}{\lambda_{i(t_N)}}}{\sum_{j=1}^{N} \frac{\lambda_{j(t)}}{\lambda_{j(t_N)}}}, \quad t > t_N \tag{395}$$

which is obtained from (390) when $p(\mathcal{H}_i | D^{(t_N)}) = 1/N, \quad i = 1, 2, \ldots, N$, is chosen as the prior distribution on the set of hypotheses. This choice can be considered s somewhat unfair since it does not exploit the data $D^{(t_N)}$ to compare the hypotheses $\mathcal{H}_i, \quad i < N$, for $t = t_N$. However, for $t$ much larger than $t_N$ this defect of the simple formula (395) is, as a rule, negligible.

Dependent on the particular case and numerical means employed it may be sometimes convenient to evaluate the ratios of integrals $\lambda_{i(t)}/\lambda_{i(t_i)}$ as the product of the ordinates of the predictive probability densities (probability functions)

$$\frac{\lambda_{i(t)}}{\lambda_{i(t_i)}} = \prod_{\tau=t_i+1}^{t} p_i(y_{(\tau)} | u_{(\tau)}, D^{(\tau-1)}) \tag{396}$$

Notice that

$$p_i(y_{(t_i+1)} | u_{(t_i+1)}, D^{(t_i)})$$

is the first prediction which can be based only on data, with negligible prior information about the unknown parameters $\theta_i$.

Example 6.2 Consider a sequence of equally distributed random variables $\{\underline{y}_{(\tau)}\}$ which has been observed for $\tau = 1, 2, \ldots, t, \; D^{(t)} = y^{(t)}$. It is a priori known that the random variables are mutually independent but it is not known whether they are distributed normally or uniformly. Neither the parameters of the normal distribution (the mean $\mu$ and the variance $\sigma^2 = \omega^{-1}$) nor the bounds of the uniform distribution $(\alpha, \beta)$ are assumed to be known. The problem is to recognize, on the basis of the observed data $y^{(t)}$, which one of the two hypotheses is true.

For the first hypothesis $\mathcal{H}_1$ we have $\theta_1 = \{\mu, \omega\}$ and

$$p_1(y_{(\tau)} | y^{(\tau-1)}, \theta_1) = p_1(y_{(\tau)} | \mu, \omega) = (2\pi)^{-\frac{1}{2}} \omega^{\frac{1}{2}} \exp[-\frac{\omega}{2}(y_{(\tau)} - \mu)^2]$$

In the alternative hypothesis $\mathcal{H}_2$ the set of unknown parameters is $\theta_2 = \{\alpha, \beta\}$ and

$$p_2(y_{(\tau)} | y^{(\tau-1)}, \theta_2) = p_2(y_{(\tau)} | \alpha, \beta) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq y_{(\tau)} \leq \beta$$

$$p_2(y_{(\tau)} | \alpha, \beta) = 0 \text{ for } y_{(\tau)} < \alpha \text{ or } y_{(\tau)} > \beta$$

Clearly $n_1 = n_2 = 0$ and likelihood functions (382) are

$$L_{1(t)}(\mu, \omega) = (2\pi)^{-\frac{t}{2}} \omega^{\frac{t}{2}} \exp[-\frac{\omega}{2} \sum_{\tau=1}^{t} (y_{(\tau)} - \mu)^2]$$

$$L_{2(t)}(\alpha, \beta) = \frac{1}{(\beta - \alpha)^t} \text{ for } \alpha \leq y_{min(t)} \text{ and } \beta \geq y_{max(t)}$$

$$L_{2(t)}(\alpha, \beta) = 0 \text{ for } \alpha > y_{min(t)} \text{ or } \beta < y_{max(t)}$$

where

$$y_{min(t)} = min(y_{(1)}, y_{(2)}, \ldots, y_{(t)})$$
$$y_{max(t)} = max(y_{(1)}, y_{(2)}, \ldots, y_{(t)})$$

A suitable improper non-informative prior for the hypothesis $\mathcal{H}_1$ is (compare with (133))

$$\phi_1(\mu, \omega) = \omega^{-1}$$

In the hypothesis $\mathcal{H}_2$ the uniform distribution can be taken as non-informative prior

$$\phi_2(\alpha, \beta) = 1 \text{ for } (\beta - \alpha) \geq 0$$

$$\phi_2(\alpha, \beta) = 0 \text{ for } (\beta - \alpha) < 0$$

The integrals (387) exist for $t \geq t_1 = 2$ and $t \geq t_2 = 3$, respectively, and can be expressed analytically. (Apply Lemmas 8 and 9 from Appendix A to determine $\lambda_{1(t)}$)

$$\lambda_{1(t)} = \Gamma(\frac{t-1}{2}) t^{-\frac{t}{2}} (\pi v_{(t)})^{-\frac{t-1}{2}}$$

where $v_{(t)}$ is the sample variance

$$v_{(t)} = \frac{1}{t} \sum_{\tau=1}^{t} (y_{(\tau)} - m_{(t)})^2, \quad m_{(t)} = \frac{1}{t} \sum_{\tau=1}^{t} y_{(\tau)}$$

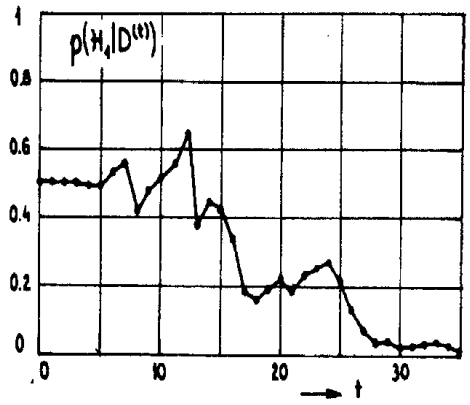



Fig. 14 Normal ($\mathcal{H}_1$) versus uniform ($\mathcal{H}_2$). Generated process was uniformly distributed, $\alpha = -1$, $\beta = 2$

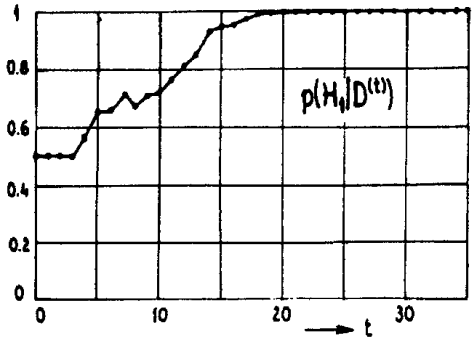



Fig. 15 Normal ($\mathcal{H}_1$) versus uniform ($\mathcal{H}_2$). Generated process was normal, $\mu = 0.5$, $\sigma^2 = \omega^{-1} = 1$

For the alternative hypothesis $\mathcal{H}_2$ we have

$$\lambda_{2(t)} = \int_{-\infty}^{y_{min(t)}} \int_{y_{max(t)}}^{\infty} \frac{1}{(\beta - \alpha)^t} d\beta d\alpha = \frac{1}{(t-1)(t-2)(y_{max(t)} - y_{min(t)})^{t-2}}$$

The formula (394) qives $K_1 = 1$, $K_2 = \lambda_{1(3)}/\lambda_{1(2)}$ and from (393) we finally obtain

$$p(H_1|y^{(t)}) = (1 + \frac{\lambda_{2(t)}}{\lambda_{1(t)}} \frac{\lambda_{1(3)}}{\lambda_{2(3)}})^{-1}$$

To illustrate this result two typical runs are shown in Fig.14 and Fig.15

## 6.6 Application to Regression-Type Model Structures

As it is seen from the formulae (393) and (394) all what is required to determine the aposterior probability distribution on the set of hypotheses under the condition that "little is known a priori" are the values of integrals (387). For the model structures of regression type (190) it is possible to give an analytical expression for these integrals. In fact, it is given by the formula (235), one only has to specify its free parameters, namely $t_0$, $\theta_{(t_0)}$ and $\alpha_{(t_0)}$ and to comb it into a form suitable for numerical evaluation. All quantities which enter this formula and may be different in different hypotheses will be indexed by $i$.

If $(\tau - n_i)$ is the time index of the most delayed output and/or input entering the known function $z_{i(\tau)}$ or $f_{i(\tau)}$ in (190) then $z_{i(n_i)}$ and $f_{i(n_i)}$ is the first couple of values of these vectors which is determined by the data and the parameter $t_0$ in (235) is $t_0 = n_i$. According to (133), let us choose the improper non-informative prior distribution on the set of unknown parameters $\theta_i = \{P_i, \Omega_i\}$ in the form

$$p_i(P_i, \Omega_i) = \phi(P_i, \Omega_i) = |\Omega_i|^{-\frac{\nu+1}{2}} \tag{397}$$

Comparison of this prior distribution with (205) gives

$$\alpha_{(t_0)} = 1, \quad \theta_{(t_0)} = -(\nu + 1), \quad V_{i(t_0)} = 0$$

The first time instant $t_i$ for which the integral $\lambda_{i(t)}$ exist (with probability one) is

$$t_i = n_i + \rho_i + \nu \tag{398}$$

For these parameters the general formula (235) gives

$$\lambda_{i(t)} = \frac{\prod_{k=1}^{\nu} \Gamma(\frac{t-t_i+k}{2})}{\pi^{(t-t_i+\frac{\nu+1}{2})\frac{\nu}{2}}} \times \tag{399}$$

$$\times \prod_{\tau=n_i+1}^{t} J_{f(\tau)} |V_{z_i(t)}|^{-\frac{\nu}{2}} |\Lambda_{i(t)}|^{-\frac{t-t_i+\nu}{2}}$$

where

$$V_{z_i(t)} = \sum_{\tau=n_i+1}^{t} z_{i(\tau)} z_{i(\tau)}^T \tag{400}$$

and $\Lambda_{i(t)}$ is defined by (213), (212) and (209) for

$$V_{i(t)} = \sum_{\tau=n_i+1}^{t} \begin{bmatrix} f_{i(\tau)} \\ z_{i(\tau)} \end{bmatrix} \begin{bmatrix} f_{i(\tau)} \\ z_{i(\tau)} \end{bmatrix}^T \tag{401}$$

To get a deeper insight into the formula (399) it may be recalled that in single-output case $\Lambda_{i(t)}$ is the sum of squares of residuals and for multiple output

$$\Lambda_{i(t)} = \sum_{\tau=n_i+1}^{t} \hat{e}_{i(\tau|t)} \hat{e}_{i(\tau|t)}^T$$

$$\hat{e}_{i(\tau|t)} = f_{i(\tau)} - \hat{P}_{i(t)}^T z_{i(\tau)}$$

where $\hat{P}_{i(t)}$ is the recent point estimate of $P_i$ defined by (211).

Again, it is numerically advantageous to operate with the lower triangular matrix $G_{i(t)}$ introduced by (254). Then, employing the relations (261) and (262) it can be easily found that

$$|V_{z_i(t)}|^{-\frac{\nu}{2}} |\Lambda_{i(t)}|^{-\frac{t-t_i+\nu}{2}} = \tag{402}$$

$$= (|G_{z_i(t)}| \times |G_{f_i(t)}|)^{\nu} \times |G_{f_i(t)}|^{t-t_i} = (\prod_{k=1}^{\rho+\nu} G_{i(t)kk})^{\nu} \times (\prod_{k=1}^{\nu} G_{i(t)kk})^{t-t_i}$$

where $G_{i(t)kk}$ are the diagonal entries of the triangular matrix $G_{i(t)}$ which can be updated in real time using the subroutine REFIL from Appendix B. To simulate the condition that "little is known apriori" the updating starts with $G_{i(n_i)} = \epsilon^{-1} I$ where $\epsilon^{-1}$ is a large number. (For data of order $10^0$ the choice $\epsilon^{-1} \geq 10^3$ is appropriate).

The gamma functions entering the formula (399) can also be calculated recursively using the relation

$$\Gamma(\frac{\tau}{2}) = \frac{\tau - 2}{2} \Gamma(\frac{\tau - 2}{2})$$

for initial conditions $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(1) = 1$. Alternatively, for $t$ large enough, the Stirling's formula can be used as a very good approximation

$$\Gamma(\frac{\tau}{2}) \approx \sqrt{2\pi} e^{-\frac{\tau}{2}} (\frac{\tau}{2})^{\frac{\tau-1}{2}}$$

Example 6.3 Consider a normal uni-variate auto-regressive process

$$y_{(\tau)} = \sum_{k=1}^{n} a_k y_{(\tau-k)} + e_{(\tau)}$$

Neither the parameters $\theta_n = \{a_1, a_2, \ldots, a_n, \sigma_e^2\}$ nor the order $n$ of the process are given, but it is known that $0 \leq n \leq N$. The problem is to estimate the order $n$ on the basis of observed data $D^{(t)} = \{y_{(1)}, y_{(2)}, \ldots, y_{(t)}\}$.

The problem can be solved as determination of the aposterior probability distribution on the set of $N$ hypotheses. Let the hypothesis $\mathcal{H}_i$ be that $n = i$. Then

$$z_{i(\tau)}^T = [y_{(\tau-1)}, y_{(\tau-2)}, \ldots, y_{(\tau-i)}] \quad \tau > i$$

$f_{i(\tau)} = y_{(\tau)}$, $J_{f(\tau)} = 1$, $t_i = 2i + 1$

Since $\nu = 1$ and

$$\Gamma(\frac{t - t_i + 1}{2}) = \Gamma(\frac{t}{2} - i) = \frac{\Gamma(\frac{t}{2})}{\prod_{k=1}^{i} (\frac{t}{2} - k)}$$

the formula (399) can be given the following form

$$\lambda_{i(t)} = \frac{\Gamma(\frac{t}{2})}{\pi^{\frac{t}{2}}} \frac{\pi^i}{\prod_{k=1}^{i} (\frac{t}{2} - k)} |V_{z_i(t)}|^{-\frac{1}{2}} \Lambda_{i(t)}^{-(\frac{t}{2}-i)}, \quad t \geq t_i$$

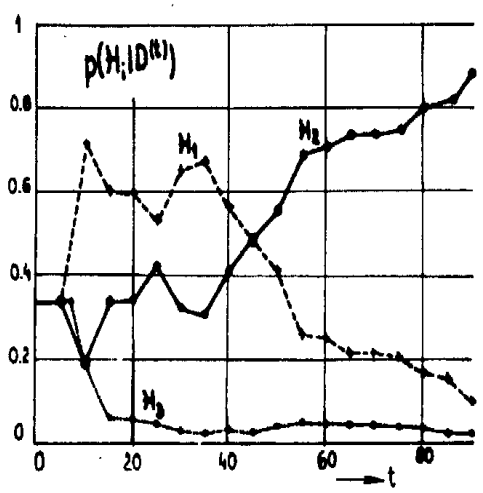



Fig.16 Uncertain order $n$ of an auto-regressive process ($\mathcal{H}_1 : n = 1$, $\mathcal{H}_2 : n = 2$, $\mathcal{H}_3 : n = 3$) simulated $n = 2$

and the formula (402) is simplified to

$$|V_{z_i(t)}|^{-\frac{1}{2}} \Lambda_{i(t)}^{-(\frac{t}{2}-i)} = \prod_{k=1}^{i+1} G_{i(t)kk} G_{i(t)11}^{t-2i-1}$$

Notice that the factor $\Gamma(\frac{t}{2})\pi^{-\frac{t}{2}}$ is common for all $\lambda_{i(t)},\ i = 1, 2, \ldots, N$, it cancels in the formula (393) for $p(\mathcal{H}_i|D^{(t)})$ and therefore does not need to be calculated for $t > t_n = 2N + 1$. Compare this results with the solution given by [16]. Algorithmic aspects are followed in more detail in [15].

For illustration, results of a simulation experiment are plotted in Fig. 16. The figure shows the evolution of the probability distribution for three hypotheses $n = i, \quad i = 1, 2, 3$, and the data generated by the model of second order $y_{(\tau)} = y_{(\tau-1)} - 0.16y_{(\tau-2)} + e_{(\tau)},\ \sigma_e^2 = 0.2$. Notice the rather small absolute value of the coefficient $a_2$ in the simulated model. This makes the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ rather close and not easy to distinguish.

Acknowledgment. The author is indebted to his colleagues Dr's. A. Halousková and M. Kárný for stimulating discussions and collaborative efforts in practical applications of the presented theory. Also a number of feed-back impulses obtained from colleagues in applied research and industry is acknowledged. Last but not least, the author's gratitude goes to Miss J. Hainová for patient and careful typing and retyping of the manuscript.

# 7 Appendices

## 7.1 Some Useful Lemmas from Matrix Algebra and Integral Calculus

Several standard operations are repeatedly met when dealing with multi-variate probability distributions which are tractable analytically. It may be helpful for the reader who is interested in applications of Bayesian statistics if we summarize them for easy reference in the form of mathematical lemmas the most of which are well known but seldom can be found on one place.

Lemma 1 The following relations hold for traces of matrix expressions

$$\mathrm{tr}(A\ B\ C) = \mathrm{tr}(C\ A\ B) = \mathrm{tr}(B\ C\ A) \tag{403}$$

$$\mathrm{tr}(A + B) = \mathrm{tr}A + \mathrm{tr}B \tag{404}$$

Proof: The relations directly follow from the definition of the trace

$$\mathrm{tr}(M) = \sum_i M_{ii} \tag{405}$$

Lemma 2 Let $\Omega$ and $A$ be positive definite matrices of dimensions $(\nu \times \nu)$ and $(\rho \times \rho)$ respectively and let $\Omega^{\frac{1}{2}}$ and $A^{\frac{1}{2}}$ be their square roots introduced so that

$$(\Omega^{\frac{1}{2}})^T\Omega^{\frac{1}{2}} = \Omega, \quad (A^{\frac{1}{2}})^T A^{\frac{1}{2}} = A$$

If $X$ is a matrix of appropriate dimensions, then

$$\mathrm{tr}(\Omega X^T A X) = ||A^{\frac{1}{2}} X (\Omega^{\frac{1}{2}})^T||^2 \tag{406}$$

where $||M||^2$ is the sum of squares of all etries $M_{ij}$, i.e. the square of Euclidean norm of the matrix M.
Proof:

$$\mathrm{tr}(\Omega X^T A X) = \mathrm{tr}[\Omega^{\frac{1}{2}} X^T (A^{\frac{1}{2}})^T A^{\frac{1}{2}} X (\Omega^{\frac{1}{2}})^T] = \mathrm{tr}[Y^T Y] = \sum_i \sum_j Y_{ij}^2$$

where

$$Y = A^{\frac{1}{2}} X (\Omega^{\frac{1}{2}})^T \tag{407}$$

Lemma 3 Let a symmetric positive definite matrix $V$ be partitioned in the following way

$$V = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix} \tag{408}$$

where $A$ and $C$ are square matrices. Further, let $\Omega$ be symmetric positive definite and $X, Y$ rectangular matrices of appropriate dimensions. Then the following relation, sometimes called completion of squares for $Y$, holds.

$$tr(\Omega \begin{bmatrix} X \\ Y \end{bmatrix}^T \begin{bmatrix} A & B^T \\ B & C \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}) = \mathrm{tr}(\Omega[X^T A X + 2Y^T B X + Y^T C Y]) = \tag{409}$$

$$= \mathrm{tr}(\Omega[Y - \hat{Y}]^T C [Y - \hat{Y}]) + \mathrm{tr}(\Omega X^T [A - B^T C^{-1} B] X)$$

where

$$\hat{Y} = -C^{-1} B X \tag{410}$$

Proof: by inspection.

Lemma 4

$$(A + BCD)^{-1} = A^{-1} - A^{-1} B (C^{-1} + DA^{-1} B)^{-1} DA^{-1} \tag{411}$$

The relation (411) is sometimes called matrix inversion lemma and holds if the required inversions exist. An often met special case is when $C$ is a scalar, $C = \frac{1}{\gamma}$ and $B = D^T = b$ is a vector:

$$(A + \frac{1}{\gamma} bb^T)^{-1} = A^{-1} - \frac{1}{\gamma + b^T A^{-1} b} A^{-1} bb^T A^{-1} \tag{412}$$

Proof: Multiply both sided of (411) by $(A + BCD)$ to obtained identity.

Lemma 5 Let $A$ and $C$ be nonsingular square matrices and $B$, $D$ rectangular matrices of appropriate dimensions. Then the following relations for determinants hold.

$$\begin{bmatrix} A & B \\ D & C \end{bmatrix} = |A| \times |C - DA^{-1} B| = |C| \times |A - BC^{-1} D| \tag{413}$$

$$|A - BC^{-1} D| = \frac{|A|}{|C|} \times |C - DA^{-1} B| \tag{414}$$

An often met special case is when $B = D^T = b$ is a vector and $C$ a scalar, $C = -\gamma^{-1}$:

$$|A + \gamma bb^T| = |A| \times (1 + \gamma b^T A^{-1} b) \tag{415}$$

Proof: see, e.g., Rao (1965, supplement to Chapter 1b)

Lemma 6 Consider a nonsingular square matrix $A$ and its inversion $B = A^{-1}$. If both these matrices are partitioned in the same way

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \tag{416}$$

then the following relations hold between the particular sub-matrices.

$$B_{11} = (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1} = A_{11}^{-1} + A_{11}^{-1} A_{12} B_{22} A_{21} A_{11}^{-1} \tag{417}$$

$$B_{22} = (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} = A_{22}^{-1} + A_{22}^{-1} A_{21} B_{11} A_{12} A_{22}^{-1} \tag{418}$$

$$B_{12} = -A_{11}^{-1} A_{12} B_{22} = -B_{11} A_{12} A_{22}^{-1} \tag{419}$$

$$B_{21} = -A_{22}^{-1} A_{21} B_{11} = -B_{22} A_{21} A_{11}^{-1} \tag{420}$$

An important special case is when $A$ is triangular. Then $B$ is also triangular. If, for instance $A_{21} = 0$, then $B_{21} = 0$ and

$$B_{11} = A_{11}^{-1}, \; B_{22} = A_{22}^{-1}, \; B_{12} = -A_{11}^{-1} A_{12} A_{22}^{-1} \tag{421}$$

Proof: by substitution of (417) to (420) into the relations $AB = I$ and $BA = I$

Lemma 7 Let $A$ be a $(\nu \times \nu)$ positive-definite matrix and let $\Omega$ be a variable matrix of the same dimensions which is restricted to be positive semi-definite. Then the maximum of the expression

$$\phi(\Omega) = |\Omega|^{\frac{\theta}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega A)\} \tag{422}$$

is reached for

$$\Omega = \Omega^* = \theta A^{-1} \tag{423}$$

and is equal to

$$\phi(\Omega^*) = |A|^{-\frac{\theta}{2}} \theta^{\frac{\nu\theta}{2}} \exp\{-\frac{\nu\theta}{2}\} \tag{424}$$

Proof: see Anderson (1958 *par.*3.2 )

Lemma 8 Let $\Omega$ and A be positive definite matrices of dimensions $(\nu \times \nu)$ and $(\rho \times \rho)$, respectively. Let $\hat{P}$ be a given $(\rho \times \nu)$ - matrix. Then

$$\int_{R^{\rho\nu}} \exp\{-\frac{1}{2}\mathrm{tr}[\Omega(P-\hat{P})^T A(P-\hat{P})]\}dP = (2\pi)^{\frac{\rho\nu}{2}} |\Omega|^{-\frac{\rho}{2}} |A|^{-\frac{\nu}{2}} \tag{425}$$

Proof: The proof is similar to that given by Anderson (1958, *par.*2.3 ) for the case when P is a vector. See [22] for details.

Lemma 9 Let $S_\Omega$ be a space of all positive definite matrices $\Omega$ of dimensions $(\nu \times \nu)$ and let $\phi$ be a given, also positive definite matrix. Then

$$\int_{S_\Omega} |\Omega|^{\frac{\theta}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega\phi)\}d\Omega = (2^{\theta+\nu+1}\pi^{\frac{\nu-1}{2}})^{\frac{\nu}{2}} |\phi|^{-\frac{\theta+\nu+1}{2}} \prod_{j=1}^{\nu} \Gamma(\frac{\theta+\nu+2-j}{2}) \tag{426}$$

where $\Gamma$ is the gamma-function.

Proof: See, e.g., [1](par. 7.2).

## 7.2 FORTRAN Subroutine REFIL

```
      SUBROUTINE REFIL(G,D,N,SIG2,VG,IN)
C
C  FUNCTION
C    SQUARE ROOT FILTER FOR REAL-TIME
C    MULTIVARIATE REGRESSION. PROCESSED
C    DATA ARE SUCCESSIVELY CONTRACTED
C    INTO A LOWER TRIANGULAR MATRIX G.
C    REFIL UPDATES THIS MATRIX WITH
C    RESPECT TO A NEW DATA VECTOR D.
C  PARAMETERS
C    INPUT:
C      G = LOWER TRIANGULAR MATRIX TO
C          BE UPDATED
C      D = VECTOR OF NEW DATA
C      N = DIMENSION OF D AND G
C   SIG2 = SQUARE OF FORGETTING FACTOR,
C          WHEN NO FORGETTING SIG2=1
C     VG = ARBITRARY N-VECTOR
C     IN = DUMMY PARAMETER TO TRANSFER
C          DIMENSION DECLARED IN MAIN
C        PROGRAM FOR ACTUAL PARAMETER G
```

```
C   OUTPUT:
C      G = UPDATED G
C   SIG2 = SIG2 + SUM OF SQUARES OF
C          G'*D
C     VG = G*G'*D
C     REMAINING PARAMETERS UNCHANGED
C  REMARK: UPPER PART OF G OVER MAIN
C          DIAGONAL NOT USED
   DIMENSION G(IN,IN), D(N), VG(N)
   SIG = SQRT(SIG2)
   PHI = SIG
   J = N
1  F = 0.
   DO 2 I = J,N
2  F = G(I,J)*D(I) + F
   A = SIG/PHI
   B = F/SIG2
   SIG2 = F*F + SIG2
   SIG = SQRT(SIG2)
   A = A/SIG
   VG(J) = G(J,J)*F
   G(J,J) =A*G(J,J)
   K = J + 1
   IF(N-K) 5, 3, 3
3  DO 4 I = K, N
   GIJ = G(I,J)
   G(I,J) = A*(GIJ - B*VG(I))
4  VG(I) = GIJ*F + VG(I)
5  J = J - 1
   IF(J) 6, 6, 1
6  CONTINUE
   RETURN
   END
```

# References


[1] T.W.Anderson: An Introduction to Multivariate Statistical Analysis, John Wiley, 1958

[2] K.J.Astrom, P.Eykhof: System Identification, Automatica, pp.123-162,1971

[3] K.J.Astrom, B.Wittenmark: Problem of Identification and Control, Journal of Mathematical Analysis and Applications,34, 90-113, 1971

[4] G.E.P.Box, G.C.Tiao: Bayesian Inference in Statistical Analysis, Addison-Wesley Publ. Comp, 1973

[5] R.G.Brown: Smoothing, Forecasting and Prediction of Discrete Time Series, Prentice-Hall, 1962

[6] R.Carnap, R.C.Jeffrey: Studies in Inductive Logic and Probability, University of California Press, 1971

[7] M.H.DeGroot: Optimal Statistical Decisions, McGraw-Hill Comp., New York, 1970

[8] W.Edwards, H.Lindman, L.J.Savage: Bayesian Statistical Inference for Psychological Research, Psychol. Rev., 70, 193-242, 1963

[9] P.Eykhof: System Identification, Parameter and State Estimation, John Wiley, 1974

[10] B. de Finneti: Theory of Probability - A Critical Introductory Treatment, Volume 1, J. Wiley, New York, 1974

[11] J.J.Florentin: Optimal Probing Adaptive Control of a Simple Bayesian System, J.Electron. Control, 11, p.571, 1962

[12] A.H.Jazwinski: Stochastic Processes and Filtering Theory, Academic Press, 1970

[13] T. Kailath: The Innovation Approach to Detection and Estimation Theory, Proc. IEEE, 58, 680-695, 1970

[14] R.E.Kalman: Linear Stochastic Filtering Theory - Reappraisal and Outlook, Proc. of the Symposium on System Theory, XV, Polytechnic Press, Brooklyn, 197-205, 1965

[15] M.Kárný : Bayesian Estimation of Model Order, Problems of Control and Information Theory, 9, 33-46, 1980

[16] R.L.Kashyap: A Bayesian Comparison of Different Classes of Dynamic Models using Empirical Data, IEEE Trans. Autom. Control, AC-22, 715-727, 1977

[17] R.L.Kashyap, A.R.Rao: Dynamic Stochastic Models from Empirical Data, Academic Press, 1976

[18] D.V.Lindley: Bayesian Statistics, a Rewiev, SIAM, Philadelphia, 1971

[19] D.V.Lindley: The Future of Statistics - a Bayesian 21st century (lecture at the Conference on Direction for Mathematical Statistic, Canada 1974, Supp. Adv. Appl. Prob., 7, 106-115, 1975)

[20] L.Ljung: On the Consistency of Prediction Error Identification Methods, In R.K.Mehra and D.G.Lainiotis (Eds.), System Identification: Advances and Case Studies, Academic Press, 1974

[21] V.Peterka: A Square Root Filter for Real-time Multivariate Regression, Kybernetika, 11, 53-67, 1975

[22] V.Peterka: Subjective Probability Approach to Real-time Identification, Proc. 4-th IFAC Symmmposium on Identification and System Parameter Estimation, Tbilisi, USSR, Vol.3, pp.83-99, 1976

[23] V.Peterka: Macrodynamics of Technological Change: Market Penetration by new technologies, Int. Inst. for Applied Systems Analysis, Laxenburg, Austria, Research Report RR-77-22, 1977

[24] V.Peterka: Experience Accumulation for Decision Making in Multivariate Time Series, Problems of Control and Information Theory, 7, 143-159, 1978

[25] V.Peterka, A.Halousková : Effective Algorithms for Real-time multivariate Regression, Proc. 4-th IFAC Symmmposium on Identification and System Parameter Estimation, Tbilisi, USSR, Vol.3, pp.100-110, 1976

[26] L.Piccinato: Predictive Distribution and Noninformative Priors, Transactions of the Sevents Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians, Vol. B, Academia, Prague, 1974

[27] H.Raiffa, R.Schleifer: Applied Statistical Decision Theory, Harvard University, Boston, 1961

[28] C.R.Rao: Linear Statistical Inference and Its Applications, John Wiley, 1965

[29] L.J.Savage: The Foundations of Statistics, John Wiley, 1954

[30] L.J.Savage: Subjective Probability and Statistical Practise, in Savage and others 1962

[31] L.J.Savage and others: The Foundation of Statistical Inference, A Dicussion, Methuen and Co Ltd, London, John Wiley and Sons Inc, New York, reprinted 1964, 1962

[32] T.J.Tarn, J.Zaborszky: A Practical Nondiverging filter, AIAA Journal, 8, 1127-1133, 1970

[33] L.A.Zadeh: From Circuit Theory to System Theory, Proc. IRE, 50, 856-865, 1962