KYBERNETIKA — VOLUME 24 (1988), NUMBER 5

# ONE-SIDED APPROXIMATION OF BAYES RULE AND ITS APPLICATION TO REGRESSION MODEL WITH CAUCHY NOISE

MIROSLAV KÁRNÝ, KATALIN M. HANGOS

The paper presents results which aim to broaden applicability of Bayesian identification to non-standard problems by using a systematic approach to the design of a feasible approximation of the optimal, but unfeasible solution.

The essence of the theory, admitting to generate global approximants in real time, is outlined. The approximation task formulated as a properly chosen problem of mathematical programming is applied to the practically important case of the regression with heavy tailed noise.

## 1. INTRODUCTION

Bayesian statistics has proved to be a suitable basis for understanding of a broad variety of identification tasks, for instance, parameter estimation [11], and tracking [8] in closed control loop or model comparison and system structure determination [4], [7]. Practical utility of the approach is, however, restricted to a relatively narrow range of numerically feasible problems. The limitations are especially urgent whenever real-time implementation is required (as for adaptive predictors and/or controllers).

The paper presents current results of the author's systematic attempts [5], [6] to provide Bayesian identification with, hopefully, well-founded method(s) to produce a feasible, real-time implementable approximation of the theoretical solution.

In the paper, the theory presented in [6] is recalled and its application to the regression model with Cauchy-distributed noise given.

The essence of the theory can be summarized as follows: A simple upper bound on the Kullback-Leibler distance between the unfeasible conditional probability density function $p(w \mid t)$ (c.p.d.f.), describing an unknown parameter $w$, and an admissible approximant $\hat{p}(w \mid t)$, is found. The bound is based on a one-sided, recursively feasible, (lower) bound on $p(w \mid t)$. Optimization of the bound leads to the rational choice of $\hat{p}(w \mid t)$ which becomes a problem of mathematical programming.

The chosen application case, regression model with Cauchy-distributed noise,

models the important estimation problem with outliers. The outlier-robust recursive estimation of regression coefficients is gained by a straightforward but nontrivial application of the present theory.

## 2. PRELIMINARIES

Necessary notions, facts and relation are summarized in the section.

### Bayesian identification

Let data items $d(t) = (y(t), u(t))$ composed of the system output $y(t)$ and the system input $u(t)$ be measured on the system at discrete time instants labelled by $t$. Let, moreover, $p(y(t) \mid t-1; u(t), w) = m(t \mid w)$ be a known model of the observed system, namely, conditional probability density function (c.p.d.f.) of the output $y(t)$ conditioned on the data measured up to and including $t-1$: $d(1 \ldots t-1) = (d(1), d(2), \ldots, d(t-1))$, on the current input $u(t)$ and a finite-dimensional constant unknown parameter $w$. To keep notational simplicity we have confined ourselves to c.p.d.f.'s with respect to Lebesgue measure.

The objective of the Bayesian identification is the determination of the c.p.d.f. $p(w \mid t) = p(w \mid d(1 \ldots t))$ under the so-called natural conditions of control [11], i.e. under the following restriction, on the class of admissible control strategies

$$p(u(t) \mid t-1; w) = p(u(t) \mid t-1), \quad t = 1, 2, 3 \ldots \tag{1}$$

Under the conditions (1), the required c.p.d.f. $p(w \mid t)$ evolves according to the following form of the Bayes rule

$$p(w \mid t) = m(t \mid w)\, p(w \mid t-1)/c(t), \quad t = 1, 2, 3 \ldots \tag{2}$$

$$c(t) = \int m(t \mid w)\, p(w \mid t-1)\, \mathrm{d}w \tag{3}$$

where $p(w \mid 0) = p_0(w)$ is a user-specified prior p.d.f.

### Unnormalized version of c.p.d.f. p(w | t)

Let us join with any c.p.d.f. $p(w \mid t)$ the class of its unnormalized versions, $\langle p(w \mid t) \rangle$. A nonnegative integrable function $l(w \mid t)$ belongs to $\langle p(w \mid t) \rangle$ iff

$$0 < \int l(w \mid t)\, \mathrm{d}w = g(t) < \infty \tag{4}$$

and

$$p(w \mid t) = l(w \mid t)/g(t). \tag{5}$$

For simplicity and because of close connections, we shall call members of $\langle p(w \mid t) \rangle$ the likelihood functions.

Clearly, $\langle p(w \mid t) \rangle$ can be viewed as an equivalence class: two likelihood functions,

say $l$, $\tilde{l}$, are equivalent if there exists a positive finite constant, say $k$, such that $l = k\tilde{l}$. The following lemma states that the equivalence class $\langle p(w \mid t-1)\rangle$ is mapped by the Bayes rule on the equivalence class $\langle p(w \mid t)\rangle$, irrespectively of the used parametrization.

**Lemma 1.** Let $l(w \mid t-1) \in \langle p(w \mid t-1)\rangle$, then for any positive finite $k$ the function

$$l(w \mid t) = k\, m(t \mid w)\, l(w \mid t-1) \in \langle p(w \mid t)\rangle \,. \tag{6}$$

The feature is invariant with respect to any regular transformation of the $w$-space.

Proof. The invariancy (useful in applications) is implied by the observation that the rules for recomputation of $p(w \mid t)$, when transforming the $w$-space, extend to their likelihoods. The rest of the simple proof is omitted. □

**Class of approximating c.p.d.f.'s**

Let us introduce the class of approximating c.p.d.f.'s, say $\hat{P}(t)$, with a generic member $\hat{p}(w \mid t)$ or $\tilde{p}(w \mid t)$. The class $\hat{P}(t)$ is characterized by the two requirements:

- $\hat{P}(t) \subset P(t)$ = c.p.d.f.'s on the $w$-space, an approximant is required to be also the c.p.d.f.,
- the functions $\hat{p}(w \mid t) \in \hat{P}(t)$ are "feasible" from the computational viewpoint.

Each approximant (being a p.d.f.) generates an equivalence class of its likelihoods. The collection of all (approximating) likelihoods will be denoted by $\langle P(t)\rangle$, $(\langle \hat{P}(t)\rangle)$, and its generic members by $l(w \mid t)$, $(\hat{l}(w \mid t), \tilde{l}(w \mid t))$, respectively.

**The Kullback-Leibler distance**

Let us choose a distance measure the need for which arises in the approximation task. The (dis)similarity measure of probability density functions

$$I(\hat{p}, p) = \int \hat{p}(w \mid t) \ln (\hat{p}(w \mid t)/p(w \mid t))\, \mathrm{d}w \tag{7}$$

called the Kullback-Leibler distance (or relative Shannon entropy) is known to have a lot of nice properties. For instance, there are estimates of an increase of the risk in any decision task due to an approximation involved [10]. They can be used to estimate the impact of the approximation on the quality of prediction and/or control.

For our treatment, the following elementary facts are needed [9]:

- the distance $I$ is always nonnegative being zero only if the compared functions differ at most on a zero-measure set,
- the distance $I = \infty$ if there is a measurable set $A$ such that $\int_A \hat{p}(w \mid t)\, \mathrm{d}w > 0$ and $p(w \mid t)$ is zero on this set.

Consequently, it holds

$$I(\hat{p}, p) < \infty \Rightarrow (\text{if } p(w \mid t) = 0 \quad \text{then} \quad \hat{p}(w \mid t) = 0) \tag{8}$$

with the exception of a set of zero measure. (As usual, the abbreviation a.e. for almost everywhere will be used to denote that some feature holds up to such a set.)

Under general conditions [12] $p(w \mid t)$ converges to the Dirac delta function concentrated on the "true" $w$. Thus $\hat{p}(w \mid t)$ guaranteeing

$$\overline{\lim_{t \to \infty}} I(\hat{p}(w \mid t), p(w \mid t)) < +\infty$$

gives a strongly consistent estimate of $w$ under the conditions referred.

**Use of arg inf(·) symbol**

Throughout the paper, the short-hand, appealling but imprecise notation

$$\hat{l} = \arg \inf_{l \in L} J(l)$$

will be used ($J$ is a functional on functions $l$ from a set $L$). The symbol $\arg\inf(\cdot)$ will be interpreted as follows:

- the existence of a minimizing argument within $L$ is generically supposed (if it is not the case, an $\varepsilon$-approximation, $\varepsilon$ a small positive number, $l_\varepsilon \in L$ for which $J(l_\varepsilon) \leqq \inf_L J(l) + \varepsilon$ is used instead),
- a specific representant is uniquely chosen if there are more minimizing functions (minimizing arguments are equivalent with respect to the task solved).

With the above agreement unneccessarily cumbersome notation is avoided.

## 3. ONE-SIDED APPROXIMATION

The section presents a one-shot approximation of the c.p.d.f. $p(w \mid t)$, the idea of which will be easily extended to the real-time-feasible approximation treated later.

The approximation would generate a $\hat{p}(w \mid t) \in \hat{P}(t)$ which is close to $p(w \mid t)$ in terms of the Kullback-Leibler distance $I$.

The need for some approximation arises whenever the "exact" c.p.d.f. $p(w \mid t)$ is too complex for a particular treatment. Consequently, the approximation itself must not require the evaluation of the c.p.d.f. $\hat{p}(w \mid t)$. Instead of the distance $I$, an upper bound on it is searched for, which can be determined without full knowledge of $p(w \mid t)$. A solution in this vein is described by the following simple but fundamental lemma.

**Lemma 2.** Let $\langle \hat{P}_p(t) \rangle$ be a nonempty subset of $\langle \hat{P}(t) \rangle$ with members having

a likelihood fulfilling the inequality

$$\bar{l}(w \mid t) \leqq p(w \mid t) \quad \text{a.e.} \tag{9}$$

Then

$$\inf_{\hat{P}(t)} I(\hat{p}, p) \leqq \inf_{\langle \hat{P}_p(t) \rangle} \left(- \ln \left(\int \bar{l}(w \mid t) \, \mathrm{d}w\right)\right). \tag{10}$$

Proof. Let $\bar{l}(w \mid t)$ be any member of $\langle \hat{P}_p(t) \rangle$; then it holds for any $\hat{p}(w \mid t) \in \hat{P}(t)$

$$\hat{p}(w \mid t) \ln \left(\hat{p}(w \mid t)/p(w \mid t)\right) \leqq \hat{p}(w \mid t) \ln \left(\hat{p}(w \mid t)/(\bar{l}(w \mid t)/\int \bar{l}(w \mid t) \, \mathrm{d}w\right) - \tag{11}$$

$$- \hat{p}(w \mid t) \ln \left(\int \bar{l}(w \mid t) \, \mathrm{d}w\right) \quad \text{a.e.}$$

By integrating the inequality (11) we have

$$I(\hat{p}, p) \leqq I(\hat{p}, \bar{l}/\int \bar{l}) - \ln \left(\int \bar{l}(w \mid t) \, \mathrm{d}w\right). \tag{12}$$

The first term on the right-hand side of the inequality (12) achieves its minimum (zero) for

$$\hat{p}(w \mid t) = \bar{p}(w \mid t) = \bar{l}(w \mid t)/\int \bar{l}(w \mid t) \, \mathrm{d}w \in \hat{P}_p(t) \subset \hat{P}(t). \tag{13}$$

The tightest bound (10) is obtained for

$$\bar{l}(w \mid t) = \operatorname*{arginf}_{\langle \hat{P}_p(t) \rangle} \left(- \ln \left(\int \bar{l}(w \mid t) \, \mathrm{d}w\right)\right). \tag{14}$$

□

**Corollary 1.** If $\hat{P}_p(t) = \hat{P}(t)$, i.e. for any $\hat{p}(w \mid t)$ there is a positive and finite $k(t)$ such that $k(t) \, \hat{p}(w \mid t) \leqq p(w \mid t)$, then

$$\inf_{\hat{P}(t)} I(\hat{p}, p) \leqq - \ln \left(\sup \int \bar{l}(w \mid t) \, \mathrm{d}w\right), \tag{15}$$

where sup is taken for $\bar{l}(w \mid t) \leqq p(w \mid t)$ a.e., $\bar{l}(w \mid t) \in \langle \hat{P}(t) \rangle$.

Lemma 2 and Corollary 1 support the following construction of the approximant: *Choose*

$$\hat{p}(w \mid t) = \bar{l}(w \mid t)/\int \bar{l}(w \mid t) \, \mathrm{d}w$$

where

$$\bar{l}(w \mid t) = \arg \sup \int \bar{l}(w \mid t) \, \mathrm{d}w, \tag{16}$$

where sup is taken for $\bar{l}(w \mid t) \leqq p(w \mid t)$ a.e., $\bar{l}(w \mid t) \in \langle \hat{P}(t) \rangle$.

**Remarks.**

(i) The proposed way of constructing the upper bound on $I(\hat{p}, p)$ using a pointwise bound on $p(w \mid t)$ can be supported by some heuristics which takes into account relations between approximation theory and statistics [2]. The decisive reason for making this choice is, however, rather pragmatic one: this is the only estimate we have found up to now which has a global character and restricts the related multivariate integration (usually unfeasible) to $\int \bar{l}(w \mid t) \, \mathrm{d}w$, which can be performed for carefully chosen class $\hat{P}(t)$, at least to the level needed for the optimization task (16).

(ii) The task (16) requires almost complete knowledge of $p(w \mid t)$, but it gives guideline how to construct a feasible (one-sided) approximation of $p(w \mid t)$ in real time. The advantage of optimization in the class $\langle \hat{P}_p(t) \rangle$ will be more transparent in the recursive case.

## 4. ONE-SIDED RECURSIVE APPROXIMATION: APPROXIMATION OF THE BAYES RULE

A recursive variant of the foregoing approximation is presented here and its alternative formulation is given.

The starting point is a simple extension of Lemma 2.

**Lemma 3.** Let

$$\hat{l}(w \mid t-1) \in \langle \hat{P}_{\langle p \rangle}(t-1) \rangle = \{\hat{l}(w \mid t-1) \in \langle \hat{P}(t-1) \rangle : \hat{l}(w \mid t-1) \leqq$$
$$\leqq l(w \mid t-1) \text{ a.e. for some } l(w \mid t-1) \in \langle p(w \mid t-1) \rangle\} .$$

A. If the set

$$\langle \hat{P}_{\langle ml \rangle}(t) \rangle = \{\hat{l}(w \mid t) \in \hat{P}(t) : \hat{l}(w \mid t) \leqq m(t \mid w)\, \hat{l}(w \mid t-1) \text{ a.e.}\} \tag{18}$$

in nonempty then

$$\inf_{\hat{p}(t)} I(\hat{p}, p) \leqq \varkappa(t) + \inf_{\hat{l}(w \mid t) \in \langle \hat{P}_{\langle ml \rangle}(t) \rangle} (-\ln (\int \hat{l}(w \mid t)\, dw)) \tag{19}$$

where $\varkappa(t)$ is a constant independent of the approximant used.

B. The likelihood minimizing upper bound (19) can be generated recursively in the time course while $\langle \hat{P}_{\langle ml \rangle}(t) \rangle$ is nonempty, using just knowledge of $p(w \mid 0) = p(w)$ and of $\hat{l}(w \mid t-1)$, $m(t \mid w)$ at each time instant.

Proof.

Case A. By definition, $\hat{l}(w \mid t-1) \in \langle \hat{P}_{\langle p \rangle}(t-1) \rangle$ means

$$\hat{l}(w \mid t-1) \leqq l(w \mid t-1) = k\, p(w \mid t-1) \tag{20}$$

for some positive finite $k$. Consequently, for any $\hat{p}(w \mid t) \in \hat{P}(t)$ and any $\hat{l}(w \mid t) \leqq$ $\leqq m(t \mid w)\, \hat{l}(w \mid t-1)$, using the Bayes rule (2), (3), we have

$$\hat{p}(w \mid t) \ln (\hat{p}(w \mid t)/p(w \mid t)) = \tag{21}$$
$$= \hat{p}(w \mid t) \ln ((\hat{p}(w \mid t)\, c(t)\, k)/(m(t \mid w)\, p(w \mid t-1)\, k)) \leqq$$
$$\leqq \hat{p}(w \mid t) \ln ((\hat{p}(w \mid t)\, c(t)\, k)/(\hat{l}(w \mid t)/\int \hat{l}(w \mid t)\, dw)) - \hat{p}(w \mid t) \ln (\int \hat{l}(w \mid t)\, dw) \text{ a.e.}$$

Integrating (21) over the entire $w$-space and proceeding as in the proof of Lemma 2 we arrive at (19) with

$$\varkappa(t) = \ln (c(t)\, k) . \tag{22}$$

Case B. It is sufficient to notice that nonnegativity of $m(t \mid w)$ justifies the im-

plication:

$$(23)\qquad \hat{l}(w \mid t) \leqq m(t \mid w)\, \hat{l}(w \mid t-1) \Rightarrow \hat{l}(w \mid t) \leqq m(t \mid w)\, l(w \mid t-1) = \\ = p(w \mid t)\, c(t)\, k$$

or equivalently

$$(24)\qquad \langle \hat{P}_{\langle ml \rangle}(t) \rangle \subset \langle \hat{P}_{\langle p \rangle}(t) \rangle .$$

The recursion has only to start with

$$(25)\qquad \hat{l}(w \mid 0) \in \langle \hat{P}_{\langle p \rangle}(0) \rangle . \qquad \square$$

The above lemma gives support for our constructive procedure which generates approximants from the class $\hat{P}(t)$ in real time:

*Choose*

$$\hat{p}(w \mid t) = \hat{l}(w \mid t) / \int \hat{l}(w \mid t)\, \mathrm{d}w$$

where

$$(26)\qquad \hat{l}(w \mid t) = \arg \sup \int \hat{l}(w \mid t)\, \mathrm{d}w ,$$

where sup is taken for $\hat{l}(w \mid t) \leqq m(t \mid w)\, \hat{l}(w \mid t-1)$ a.e., $\hat{l}(w \mid t) \in \langle \hat{P}(t) \rangle$.

**Remarks.**

(iii) The structure of $\hat{P}(t)$ guaranteeing $\langle \hat{P}_{\langle ml \rangle}(t) \rangle$ to be nonempty must be carefully chosen when selecting classes of approximants for $t = 1, 2, 3, \ldots$. Clearly, the form of the system model $m(t \mid w)$ has to be taken into account. The parts of the $w$-space where $m(t \mid w)$ falls to (is) zero must be respected. This remark is a hint how to choose $\hat{P}(t)$.

(iv) The methodological consequence of Lemmas 1 and 3, which should be noticed, is the independence of the solution of any positive renormalization of the function $m(t \mid w)$. This fact makes the lack of knowledge of $c(t)$ (3) unimportant, makes possible to choose an appropriate numerical level of evaluation and, above of all, gives an intuitive support for conjecture that the procedure (26) can converge although no forgetting is used, cf. [5].

The next lemma interprets alternatively the task (26) yielding an insight into the procedure adopted.

**Lemma 4.** The task (26) is equivalent to the "maximin" problem

$$(27)\qquad \hat{p}(w \mid t) = \arg \sup_{\bar{p}(w \mid t) \in \hat{P}(t)} \operatorname{essinf}_{w} \bar{p}(w \mid t) / \bar{p}(w \mid t)$$

where $\bar{p}(w \mid t)$ denotes "$\hat{p}(w \mid t-1)$ updated by the Bayes rule"

$$(28)\qquad \bar{p}(w \mid t) = m(t \mid w)\, \hat{p}(w \mid t-1) / \int m(t \mid w)\, \hat{p}(w \mid t-1)\, \mathrm{d}w .$$

Proof. Any $\hat{l}(w \mid t) \in \langle \hat{P}(t) \rangle$ can be written in the form

$$(29)\qquad \hat{l}(w \mid t) = \gamma(t)\, \bar{l}(w \mid t)$$

where $\gamma(t)$ is a positive finite constant and $\tilde{l}(w \mid t) \in \langle \hat{P}(t) \rangle$. Consequently, the additional parameter $\gamma(t)$ can be optimized when solving (26). An admissible pair $\gamma(t)$, $\tilde{l}(w \mid t)$ has to fulfil the inequality

$$\gamma(t)\, \tilde{l}(w \mid t) \leqq m(t \mid w)\, \hat{l}(w \mid t-1) \quad \text{a.e.} \tag{30}$$

which is equivalent to

$$\gamma(t) \leqq \operatorname*{essinf}_{w} m(t \mid w)\, \hat{l}(w \mid t-1)/\tilde{l}(w \mid t) = \Gamma(\tilde{l})\,. \tag{31}$$

The proof requires to demonstrate the equality of $\gamma(t)$ and $\Gamma(\tilde{l})$ for the optimal approximant. By contradiction, let for the optimal $\tilde{l} = \gamma^* l^*$ be $\gamma^*(t) < \Gamma(l^*)$. Then the likelihood $l^0(w \mid t) = \Gamma(l^*)\, l^*(w \mid t) \in \langle \hat{P}(t) \rangle$ and the inequality $l^0(w \mid t) \leqq \leqq m(t \mid w)\, \hat{l}(w \mid t-1)$ is fulfilled. For its integral the inequality

$$\begin{aligned} \int l^0(w \mid t)\, \mathrm{d}w &= \Gamma(l^*) \int l^*(w \mid t)\, \mathrm{d}w = \\ &= \Gamma(l^*)/\gamma^*(t) \int \tilde{l}(w \mid t)\, \mathrm{d}w > \int \tilde{l}(w \mid t)\, \mathrm{d}w \end{aligned} \tag{32}$$

holds which contradicts optimality of the pair $\gamma^*(t)$, $l^*(w \mid t)$.

Due to the optimality with respect to (26), the pair $\gamma^*(t)$, $l^*(w \mid t)$ fulfils, for any admissible $\tilde{l}(w \mid t)$, the inequality

$$\begin{aligned} \int \tilde{l}(w \mid t)\, \mathrm{d}w &\leqq \gamma^*(t) \int l^*(w \mid t)\, \mathrm{d}w = \Gamma(l^*) \int l^*(w \mid t)\, \mathrm{d}w = \\ &= \operatorname*{essinf}_{w} \bar{p}(w \mid t)/p^*(w \mid t)\, \big(\int m(t \mid w)\, \hat{l}(w \mid t-1)\, \mathrm{d}w\big)\,. \end{aligned}$$

The above bound is reached for $\tilde{l} = l^*$ and it is the highest one for the solution of the task (27).

Reversing the way of reasoning it can be found that the solution of the task (27) solves the task (26). □

**Remarks.**

(v) The task (27) benefits substantially from the structure of the space $\hat{P}(t)$; the function $\hat{p}(w \mid t)$ has to be p.d.f., i.e. nonnegative one integrating to unity. Just this structure implies that two functions $\bar{p}$, $\hat{p}$ with common support have to "cross" each other.

(vi) The optimization task (26) is generally a difficult problem with functional restrictions. The class of practically feasible problems is, however, enlarged as test cases indicate.

## 5. APPLICATION TO LINEAR REGRESSION MODEL WITH CAUCHY NOISE

The section demonstrates how the proposed theory can be used.

At first, the treated system model is introduced and necessity to approximate

$p(w \mid t)$ shown. Then the class of approximating c.p.d.f.'s is chosen and the results of optimization (26) ((27)) are summarized.

**System Model**

The linear regression model (for simplicity with single output)

$$y(t) = \Theta' z(t) + e(t) \tag{33}$$

with unknown coefficients $\Theta$, with the regressor vector $z(t)$ being a known function of $d(1 \dots t-1)$, $u(t)$ (which can be recursively updated), and with the white noise term $e(t)$, is often used as a system model when designing adaptive predictors and/or controllers. Recursive least squares are used as a standard estimation procedure. The only but substantial drawback of this procedure is its sensitivity to outliers. A lot of modifications has been proposed recently, see [3]. The Bayesian approach to this problem is based on a proper modelling of the noise term and on the consistent application of the Bayes rule.

Recursive least squares are known to arise in Bayesian identification of the regression model (33) when $e(t)$ is assumed to be normally distributed with constant (possibly unknown) dispersion $\omega^{-1}$. Normal distribution, however, assigns non-negligible probability within the interval of the width $(2 \div 3)\,\omega^{-1/2}$. If the presence of outliers is expected a more proper modelling is needed: the p.d.f. describing the noise has to fall to zero much slowly than the normal p.d.f. The Cauchy distribution

$$p(e(t) \mid \omega) = (2/\pi)\,\omega^{1/2}/(1 + \omega\, e^2(t)) \tag{34}$$

is an extreme example to which we shall apply the developed theory hoping to achieve a robust, in the discussed sense, estimation procedure of the regression coefficients.

To demonstrate the need for some approximation and to make a preliminary step in approximation, we write explicitly the exact form of the likelihood $l(w \mid t)$. Defining

$$\begin{aligned} \varepsilon(t, w) &= \omega^{1/2}(y(t) - \Theta' z(t)) = \\ &= -[y(t), z'(t)] \begin{bmatrix} -\omega^{1/2} \\ \Theta\omega^{1/2} \end{bmatrix} = -h'(t) \begin{bmatrix} -\omega^{1/2} \\ \Theta\omega^{1/2} \end{bmatrix} \end{aligned} \tag{35}$$

where

$$w = (\Theta, \omega) \quad h'(t) = [y(t), z'(t)]\,, \tag{36}$$

the system model can be written in the form (omitting the $w$-independent constant, cf. Remark (iv))

$$m(t \mid w) = \omega^{1/2}(1 + \varepsilon^2(t, w))^{-1}\,. \tag{37}$$

The likelihood functions for this model take the form

$$l(w \mid t) = \prod_{\tau=1}^{t} m(\tau \mid w)\, p(w \mid 0) = \omega^{t/2} \prod_{\tau=1}^{t} (1 + \varepsilon^2(\tau, w))^{-1}\, p(w \mid 0)\,. \tag{38}$$

The exact evaluation of $l(w \mid t)$ and consequently of the c.p.d.f. $p(w \mid t)$ requires storring the full process history; the need for approximation arises.

**Class of Approximants**

The assumed form of approximating likelihoods is

$$\hat{l}(w \mid t) = \gamma(t)\, \omega^{v(t)/2} \left(1 + \omega \begin{bmatrix} -1 \\ \Theta \end{bmatrix}' V(t) \begin{bmatrix} -1 \\ \Theta \end{bmatrix}\right)^{-v(t)} \tag{39}$$

which should be optimally "shaped" by a positive scalar factor $v(t)$ and by a positive definite matrix $V(t) > 0$. The positive scalar $v(t)$ has to evolve according to the relation

$$v(t) = v(t-1) + 1\,. \tag{40}$$

The reasons for this choice can be shortly summarized by discussing particular components in (39):

(*a*) $\gamma(t)$: is a scaling factor the role of which can be understood in deep by inspecting the proof of Lemma 4.

(*b*) $\omega^{v(t)/2}$: $\hat{l}(\cdot)$, $l(\cdot)$ should be close each other, the part of $l(\cdot)$, which can be updated recursively, should form a part of $\hat{l}(\cdot)$.

The use of the power $v(t)$ instead of $t$ just admits to incorporate some prior value, the recursion (40) corresponds simply to that of counting the number of samples $(t)$.

(c) $\left(1 + \omega \begin{bmatrix} -1 \\ \Theta \end{bmatrix}' V(t) \begin{bmatrix} -1 \\ \Theta \end{bmatrix}\right)^{-v(t)}$: see Lemma 3 and Remark (iii) stating that care has to be taken to guarantee nonemptiness of $\langle \hat{P}_{\langle ml \rangle}(t) \rangle$ for all $t$. We have to be able to find approximate likelihoods fulfilling the inequalities

$$\hat{l}(w \mid t) \leqq m(t \mid w)\, \hat{l}(w \mid t-1) \quad \text{for} \quad t = 1, 2, 3, \ldots$$

Since support of $l(t \mid w)$ forms the entire "natural" $w$-space $= \{\omega > 0,\ \Theta$ being real vector$\}$, the behaviour of $\hat{l}(w \mid t)$, when some entry of $w$ approaches to infinity, i.e. when $l(w \mid t)$ falls to zero, is decisive. Clearly, $\hat{l}(w \mid t)$ have to decay with a higher rate than $l(w \mid t)$ for $w$ tending to infinity. On the other hand tightness of the approximation requires the same rate of decrease in this area. Inspecting the form of $l(w \mid t)$ (38) as a function of $w$ (for $w$ approaching infinity) we find that $l$ behaves as the term discussed.

The following, theorem summarizes the results of the optimization for the regression model with a heavy-tailed noise.

**Theorem.** The c.p.d.f. $p(w \mid t)$ of the parameter $w = (\Theta, \omega)$, $\dim(\Theta) = \varrho$, of the "Cauchy-tailed" regression model

$$m(t \mid w) = \omega^{1/2}(1 + \omega(y(t) - \Theta' z(t))^2)^{-1} \tag{41}$$

can be optimally approximated (in the sense (27)) within the class (39) by the c.p.d.f.

$$\hat{p}(w \mid t) = \gamma(t)\,\omega^{v(t)/2}\left(1 + \omega\begin{bmatrix} -1 \\ \Theta \end{bmatrix}' V(t)\begin{bmatrix} -1 \\ \Theta \end{bmatrix}\right)^{-v(t)} \tag{42}$$

where the statistics of this distribution are recursively updated as follows

$$v(t) = v(t-1) + 1\,, \quad v(0) > 0 \quad \text{given}, \tag{43}$$

$$V(t) = (1 - \beta(t))\,V(t-1) + \beta(t)\begin{bmatrix} y(t) \\ z(t) \end{bmatrix}\begin{bmatrix} y(t) \\ z(t) \end{bmatrix}', \quad V(0) > 0 \quad \text{given}. \tag{44}$$

The weight $\beta(t)$ is a real root of the polynomial

$$\mu(\beta) = \sum_{i=0}^{3} \mu_i \beta^i \tag{45}$$

with the coefficients (time-index omitted)

$$\mu_0 = 2\,, \quad \mu_1 = -v + (v - \varrho - 1)\,\xi - (v - \varrho + 2)\,\xi_z \tag{46}$$

$$\mu_2 = \xi\xi_z + (\varrho - 1)\,\xi - (2v - \varrho)\,\xi_z\,, \quad \mu_3 = -(v-1)\,\xi\xi_z$$

where

$$\xi(t) = 1 - \begin{bmatrix} y(t) \\ z(t) \end{bmatrix}' V^{-1}(t-1)\begin{bmatrix} y(t) \\ z(t) \end{bmatrix} \tag{47}$$

$$\xi_z(t) = 1 - z(t)'\,V_z^{-1}(t-1)\,z(t) \tag{48}$$

$$V(t) = \begin{bmatrix} V_y(t) & V_{zy}'(t) \\ \underbrace{V_{zy}(t)}_{1} & \underbrace{V_z(t)}_{\varrho} \end{bmatrix}\begin{matrix} \}1 \\ \}\varrho \end{matrix} \tag{49}$$

For $v \geqq \varrho + 2$, this root lies always in the interval

$$(\beta_0,\, 2/(v-1)) \tag{50}$$

with

$$\beta_0 = 1/(v - (v-1)\,\xi^-) \tag{51}$$

$$\xi^- = \min(0, \xi)\,. \tag{52}$$

Proof. The proof is outlined in the Appendix. It is based on a chain of technical lemmas which specify the solution of the "maximin" task (27). □

**Remarks.**

(vi) Using the decomposition (49) the following well-known "least-square" quantities are defined:

Point estimate of regression coefficients

$$\hat{\Theta} = V_z^{-1} V_{zy}\,. \tag{53}$$

Point estimate of the noise dispersion

$$\hat{\omega}^{-1} = V_y - V_{zy}' V_z^{-1} V_{zy}\,. \tag{54}$$

(vii) The quantities $\hat{\Theta}$, $\hat{\omega}$ are well-known in connection with (recursive) least squares. The essence of the approach compresses, in the given case, into the choice of the weighting factor $\beta(t)$. Ordinary least squares (normed version) use

$$\beta_{LS}(t) = 1/v(t) . \tag{55}$$

A simple analysis as well as numerical experience have shown that $\beta(t)$ can be greater than (55) especially when starting the identification. The rate of convergence can be increased in this way. The claimed robustness is achieved because $\beta(t)$ can be much smaller than (55) whenever the prediction error is too high.

(viii) The computational similarity with the recursive least squares makes possible to use, almost without changes, an effective LD-factorized numerical implementation [1].

## 6. CONCLUSIONS

The reported theory is by no means complete. The results gained up to now are, however, rather promising. The proposed way of approximating the Bayes identification

- is able to produce well-motivated approximative identification algorithms working in real time,
- has a global character taking into account both deterministic and stochastic ingredients of the model,
- can be, in a rather straightforward way, extended to the problems with time-dependent parameters, i.e. to nonlinear filtering.

The presented application example supports our belief that a proper direction of the research has been chosen. By using a unified theoretical and algorithmic approach we have arrived at reasonable, outlier robust estimation of regression coefficients. The resulting algorithm can be viewed as recursive weighted least squares with well-grounded choice of the weights.

## 7. APPENDIX: PROOF OF THEOREM

In order to solve the task (26) or its equivalent (27) the dependence of the integral $\int \tilde{l}(w \mid t)\, \mathrm{d}w$ on the parameters determining $\tilde{l}(w \mid t)$ (i.e. on $v(t)$, $V(t)$) has to be found.

**Lemma 5.** For likelihoods in the class (39), it holds

$$\int \tilde{l}(w \mid t)\, \mathrm{d}w = \gamma(t)\, |V_z|^{(\nu-\varrho)/2}\, |V|^{-(\nu-\varrho+1)/2}\, \tilde{\varphi}(t) \tag{56}$$

where $\tilde{\varphi}(\cdot)$ is a term independent of $v(t)$, $V(t)$

$$V = \underbrace{\begin{bmatrix} \cdot & \cdot \\ \cdot & V_z \end{bmatrix}}_{\tilde{1}} \begin{matrix} \} 1 \\ \} \varrho \end{matrix}, \quad \varrho = \dim(\Theta) . \tag{57}$$

Proof. The proof is based essentially on the following substitution in the computed integral

$$x = V^{1/2}(t-1)\begin{bmatrix} -1 \\ \Theta \end{bmatrix}\omega^{1/2} \tag{58}$$

where $V^{1/2}(t-1)$ denotes any square root of the positive definite matrix $V(t-1)$. Further details are omitted. □

The following lemma determines the form of the optimal $\hat{l}(w \mid t)$ in the class (39), i.e. the optimal form of the matrix $V(t)$. Moreover, it makes the optimization problem more transparent by an appropriate transformation of the $w$-space. For this reason the subsequent notations will be used:

$$J = (V^{-1/2}(t-1))'\, V(t)\, V^{-1/2}(t-1) \tag{59}$$

$$f = (V^{-1/2}(t-1))'\, h(t) \tag{60}$$

$$\xi = 1 - f'f \tag{61}$$

$$\mathbf{1}' = [1, 1] \tag{62}$$

$$v = v(t)\,. \tag{63}$$

**Lemma 6.** The optimal solution of the approximation task (27), say $\hat{l}(w \mid t)$, from the class (39) can be searched for in the subclass of (39) determined by

$$V(t) = (V^{1/2}(t-1))'\, J\, V^{1/2}(t-1)\,, \tag{64}$$

where the positive definite matrix $J$ takes the form ($I$ denotes the unit matrix of an appropriate dimension)

$$J = (1-\beta)\, I + (\beta - \alpha)\, ff'/f'f\,. \tag{65}$$

The matrices $J$ are parametrized by the two-dimensional vector

$$v = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{66}$$

lying in the *open* half-plane

$$v < \mathbf{1}\,. \tag{67}$$

The computation of the "min" part of the task (27)

$$\gamma(\hat{l}) = \operatorname*{essinf}_{w} \bar{p}(w \mid t)/\hat{p}(w \mid t) \tag{68}$$

reduces to the two-dimensional minimization (over some $\varkappa' = [\varkappa_1, \varkappa_2]$)

$$\gamma(\hat{l})/\int \hat{l}(w \mid t)\, \mathrm{d}w = \operatorname*{essinf}_{0 \leqq \varkappa, \varkappa'\mathbf{1} \leqq 1} (1 - v'\varkappa)^v/(1 - [\xi, 1]\,\varkappa)\,. \tag{69}$$

Proof. The proof is based on a sequence of regular transformations of the $w$-space using the invariance of the task (27).

Starting with the substitution according to Eq. (58) the "inf" part of the optimiza-

tion task reads

$$\gamma(\bar{l}) = \operatorname*{essinf}_{x} (1 + x'Jx)^{v} / ((1 + x'x)^{v-1} (1 + x'ff'x)) \int \bar{l}(w \mid t)\, dw \tag{70}$$

where unimportant (with respect to (27)) multiplicative factor $\gamma(t-1)/\gamma(t)/\int m(t \mid w)$ . $\bar{l}(w \mid t-1)\, dw$ is omitted (cf. proof of Lemma 4) and the relations (35), (36), (39), (40), (44)–(48) are used

Before applying further transformations we shall prove that the optimal $J$ can be searched for in the form (65). The use of Cholesky (lower triangular) square root $V^{1/2}(t-1)$ in the definition (59) implies that (cf. (56))

$$\int \bar{l}(w \mid t)\, dw = \gamma(t) \mid J_z \mid^{(v-\varrho)/2} \mid J \mid^{-(v-\varrho+1)/2} \varphi(t, V(t-1)) \tag{71}$$

where $J_z$ corresponds to $V_z$ in (42) and $\varphi(\cdot, \cdot)$ to $\tilde{\varphi}(\cdot)$ in (56). By inspecting the formulae (70), (71) it is can be seen that the value of $\gamma(\bar{l})$ *is not influenced by any orthogonal transformation*

$$\mathscr{F}\colon J \to \mathscr{F} J \mathscr{F}' \tag{72}$$

*restricted to the orthogonal complement of* the subspace spanned on *the direction* $f$ (including those describing permutation of entries). Consequently, the optimal $J$ must have at most two different eigenvalues: one related to the span $f$ and another one to its complement. A symmetric matrix of such a type takes the form (65) and the admissible area (67) expresses the required definitness of $J(V(t))$.

Using the formulae proved it remains to map the real $(\varrho + 1)$ dimensional space on the unit sphere by substituting

$$\tilde{x} = x / \sqrt{(1 + x'x)} \tag{73}$$

and observing that

$$1/(1 + x'x) = 1 - \tilde{x}'\tilde{x}\,. \tag{74}$$

The remaining substitution exemplifies splitting of the $\tilde{x}$-space into the span of $f$ and its complement. Because of invariance of the task with respect to the orthogonal transformation (72), it is sufficient to take $\tilde{x}$ as a linear combination of $f$ and any orthogonal vector, say $f^c$, reducing the minimization just to the two-dimensional space of weights, say $\pm\sqrt{\varkappa_1}$, $\pm\sqrt{\varkappa_2}$, i.e.

$$\tilde{x} = \pm \sqrt{(\varkappa_1)}\, f/f'f \pm \sqrt{(\varkappa_2)}\, f^c/f^{c\prime}f^c\,, \quad f'f^c = 0\,. \tag{75}$$

The use of relations (65), (66), (73)–(75) in the task (70) proves the formula (69). □

By computing $\int \bar{l}(w \mid t)\, dw$ (with the help of (41)) for the class of approximants (39) with the kernel $V(t)$ determined according to the formula (64) and with the matrix $J$ given by (65) the following corollary can be verified.

**Corollary 2.** The approximation task (27) for the system model (37) is solved by $\hat{p}(w \mid t)$ in the class (39) with $V(t)$ (64), (65) when the parameter $\hat{v}$ (66) is found as

$$\hat{v} = \arg \sup_{v<1} \inf_{0 \leqq \varkappa, \varkappa' v \leqq 1} H(v, \varkappa) \tag{76}$$

where

$$H(v, \varkappa) = (1 - v'\varkappa)^v (1 - [\xi, 1]\varkappa)^{-1} (1 - \beta)^{-v/2} (1 - \alpha)^{-(v-\varrho+1)/2} \cdot \left(1 - v' \begin{bmatrix} \delta \\ 1 - \delta \end{bmatrix}\right)^{(v-\varrho)/2} \tag{77}$$

with

$$f' = [\underbrace{\cdot}_{1}, f_z'] \tag{78}$$

and

$$\delta = f_z' f_z / f'f \in (0, 1) . \tag{79}$$

Proof. It only suffices to recall that the square matrix $I - gg'$ where $g$ is a nonzero $n$-vector has eigenvalues equal to 1 except one equal to $1 - g'g$. □

Now, we shall exclude some approximants as nonoptimal ones using necessary conditions of optimality. The following lemma states that the infimum in (69) is attained at the boundary of the admissible area in the *generic* case.

**Lemma 7.** If

$$\alpha - \beta\xi \neq 0 \tag{80}$$

then the infinum in (69) is attained at the boundary of the region $0 \leqq \varkappa$, $\varkappa' \mathbf{1} \leqq 1$.

Proof. Differentiating the logarithm of the minimized function and setting the gradient equal to zero it can be found with some algebra that the stationary point $\hat{\varkappa}$ has to fulfil the pair of linear equations

$$A\hat{\varkappa} = B ; \quad A = \begin{bmatrix} -(v-1)\beta & \alpha - v\beta\xi \\ \xi\beta - v\alpha & -(v-1)\xi\alpha \end{bmatrix}, \quad B = \begin{bmatrix} 1 - \beta v \\ \xi - v\alpha \end{bmatrix}. \tag{81}$$

The square on the left-hand side of (80) is (positively) proportional to the determinant of $A$, so that under (80) we can find the unique stationary point by solving (81)

$$\hat{\varkappa}' = [\alpha - \xi, 1 - \beta]/(\alpha - \beta\xi) . \tag{82}$$

The restriction (67) and nonnegativity of admissible stationary points $\hat{\varkappa}$ imply

$$1 > \beta \Rightarrow \alpha - \beta\xi > 0 \Rightarrow 1 > \alpha \geqq \xi . \tag{83}$$

Taking into account the requirement $\varkappa' \mathbf{1} \leqq 1$ it can be concluded that

$$1 \geqq \varkappa' \mathbf{1} = (\alpha - \xi + 1 - \beta)/(\alpha - \beta\xi) \Rightarrow (1 - \beta)(1 - \xi) \leqq 0 \tag{84}$$

which is in contradiction to (83). Continuity of the minimized function and compactness of the admissible area of $\varkappa$ guarantee the rest. □

The following lemma restricts the domain of the optimal approximants to the "degeneracy" line

$$\alpha = \beta\xi . \tag{85}$$

**Lemma 8.** In the class (39) there is *no* optimal approximant of $\hat{p}(w \mid t)$ having the likelihood (38) while the parameter (66) is restricted by the condition (80).

Proof. Let the pair $(\hat{\nu}, \hat{\varkappa})$ determine the optimal solution of the task (76) (equivalent to (27)). Because of the "sup"-part of (76), it must hold

$$H(\nu, \hat{\varkappa}) \leqq H(\hat{\nu}, \hat{\varkappa}) \quad \text{for any} \quad \nu < \mathbf{1} . \tag{86}$$

The admissible range of $\nu$ is an open set, so that the vanishing gradient of $\ln (H(\hat{\nu}, \hat{\varkappa}))$ with respect to $\nu$ gives the necessary optimality condition of the pair $(\hat{\nu}, \hat{\varkappa})$. This condition can be rearranged into the form

$$\nu\hat{\varkappa} = (1 - \hat{\nu}'\hat{\varkappa}) \begin{bmatrix} -\dfrac{(\nu - \varrho)\,\delta}{2\left(1 - \hat{\nu}'\begin{bmatrix} \delta \\ 1-\delta \end{bmatrix}\right)} + \dfrac{(\nu - \varrho + 1)}{2\left(1 - \hat{\nu}'\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)} \\ -\dfrac{(\nu - \varrho)(1 - \delta)}{2\left(1 - \nu'\begin{bmatrix} \delta \\ 1-\delta \end{bmatrix}\right)} + \dfrac{\nu}{2\left(1 - \hat{\nu}'\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)} \end{bmatrix} = (1 - \hat{\nu}'\hat{\varkappa})\,F . \tag{87}$$

Lemma 6 implies that the optimal $\hat{\varkappa}'$ has one of the forms

$$(\varkappa_1, 0), \quad (0, \varkappa_2), \quad (\varkappa_1, 1 - \varkappa_1); \quad 0 \leqq \varkappa_1, \varkappa_2 \leqq 1 . \tag{88}$$

By checking the possibilities (73), we shall find that no such a pair exists. For instance, let $\hat{\varkappa}' = (0, \varkappa_2)$. For the optimal $\hat{\nu}, \hat{\varkappa}$, it implies (see (87))

$$F_1 = 0 . \tag{89}$$

This condition can be rearranged as

$$\mathbf{1}'\begin{bmatrix} \delta \\ (\nu - \varrho + 1)(1 - \delta) \end{bmatrix} = \hat{\nu}'\begin{bmatrix} \delta \\ (\nu - \varrho + 1)(1 - \delta) \end{bmatrix} \tag{90}$$

which is contradictory for $\hat{\nu} < \mathbf{1}$ and $\delta \in (0, 1)$ (cf. (79)). Similar reasoning can be performed for the remaining cases and it is therefore omitted. □

By combining the results of Lemmas 6, 7, 8 and using formulae (61), (65), (67), (69) and (70) we have:

**Corollary 3.** The matrix $V(t)$ determining the optimal approximant is given by the formula (64) with

$$J = (1 - \hat{\beta})\,I + \hat{\beta} f f' \tag{91}$$

and

$$\hat{\beta} = \arg \sup_{\beta < 1} \inf_{\xi^- \leqq \psi \leqq 1} \tilde{H}(\beta, \psi) \tag{92}$$

where

$$\tilde{H}(\beta, \psi) = (1 - \beta\psi)^{\nu} (1 - \psi)^{-1} (1 - \beta)^{-\nu/2} (1 - \beta\xi)^{-(\nu - \varrho + 1)/2} \cdot (1 - \beta\xi_{\varrho})^{-(\nu - \varrho)/2} \tag{93}$$

with (cf. (61), (78))

$$\xi = 1 - f'f\,, \quad \xi_z = 1 - f_z'f_z\,, \tag{94}$$

$$\xi^- = \min(0, \xi)\,. \tag{95}$$

Proof. The proof employs the mentioned formulae and the definition of the new independent, one-dimensional variable

$$\psi = (\xi, 1)\varkappa \quad \text{for which} \quad \xi^- \leqq \psi \leqq 1 \tag{96}$$

instead of $\varkappa$. □

The lemma below summarizes results of the analytical minimization of $\tilde{H}(\beta, \psi)$ (93) with respect to $\psi$.

**Lemma 9.** It holds

$$\inf_{\xi^- \leqq \psi \leqq 1} \tilde{H}(\beta, \psi) = \begin{cases} \tilde{H}(\beta, \xi^-) & \text{for} \quad \beta \leqq \beta_0 \\ v^v/(v-1)^{(v-1)}\,\beta(1-\beta)^{(v-2)/2}\,(1-\beta\xi)^{-(v-\varrho+1)/2}\,. \\ \quad .\,(1-\beta\xi_z)^{(v-\varrho)/2} & \text{for} \quad \beta \geqq \beta_0 \end{cases} \tag{97}$$

where

$$\beta_0 = 1/(v - (v-1)\,\xi^-)\,. \tag{98}$$

Proof. This simple exercise in optimization is omitted. □

The next lemma helps to avoid the branch $\beta \leqq \beta_0$ of (97) when computing the "sup"-part of (93) to determine $\hat{\beta}$.

**Lemma 10.** The optimal weight $\hat{\beta}$ in (91) is given by

$$\hat{\beta} = \arg\sup_{\beta_0 \leqq \beta < 1} \beta(1-\beta)^{(v-2)/2}\,(1-\beta\xi)^{-(v-\varrho+1)/2}\,(1-\beta\xi_z)^{(v-\varrho)/2} = \tag{99}$$

$$= \arg\sup_{\beta_0 \leqq \beta < 1} H_0(\beta)\,.$$

Proof. The function (97) is a continuous function of $\beta$ even at the point $\beta_0$. Thus, if we prove $\tilde{H}(\beta, \xi^-)$ to be an increasing function of $\beta$ for $\beta \leqq \beta_0$, we have

$$\hat{\beta} = \arg\sup_{\beta} \inf_{\psi} \tilde{H}(\beta, \psi) \geqq \beta_0\,. \tag{100}$$

To this end we shall prove the positivity of the derivative of $\ln(\tilde{H}(\beta, \xi^-))$ with respect to $\beta$ (for $\beta \leqq \beta_0$), i.e.

$$\bar{H}(\beta) = 2\,\partial\tilde{H}(\beta, \xi^-)/\partial\beta/\tilde{H}(\beta, \xi^-) = \tag{101}$$

$$= -2v\xi^-/(1-\beta\xi^-) + v/(1-\beta) + (v-\varrho+1)\,\xi/(1-\beta\xi) - (v-\varrho)\,\xi_z : \\ : (1-\beta\xi_z)\,.$$

The last term in (101) is a decreasing function of $\xi_z \leqq 1$, thus on replacing $\xi_z$ by 1 we find

$$\bar{H}(\beta) \geqq -2v\xi^-/(1-\beta\xi^-) + \varrho/(1-\beta) + (v-\varrho+1)\,\xi/(1-\beta\xi)\,. \tag{102}$$

The next estimate, proving the lemma, is obtained by observing that $\xi^- \leqq \xi$ and noting that the last term in (102) is an increasing function of $\xi$, consequently,

$$\bar{H}(\beta) \geqq -2\xi^-(v + \varrho - 1)/(1 - \beta\xi^-) + \varrho/(1 - \beta) > 0 \tag{103}$$

because $\xi^- \leqq 0$. □

The concluding "optimization" lemma states the necessary optimality conditions and specifies the interval in which the optimizer can be searched for.

**Lemma 11.** The maximizer of the function in (99)

$$H_0(\beta) = \beta(1 - \beta)^{(v-2)/2}(1 - \beta\xi)^{(v-\varrho+1)/2}(1 - \beta\xi_z)^{(v-\varrho)/2} \tag{104}$$

can be found as a real root of the polynomial

$$\mu(\beta) = \sum_{i=0}^{3} \mu_i \beta^i \tag{105}$$

with the coefficients

$$\mu_0 = 2\,, \quad \mu_1 = -v + (v - \varrho - 1)\,\xi - (v - \varrho + 2)\,\xi_z\,, \tag{106}$$
$$\mu_2 = \xi\xi_z + (\varrho - 1)\,\xi - (2v - \varrho)\,\xi_z\,, \quad \mu_3 = -(v - 1)\,\xi\xi_z\,.$$

This root has to be in the interval

$$\langle \beta_0, 2/(v - 1) \rangle\,. \tag{107}$$

Proof. If holds

$$\bar{H}_0(\beta) = 2\,\partial H_0(\beta)/\partial\beta/H_0(\beta) = \tag{108}$$
$$= 2/\beta - (v - 2)/(1 - \beta) + (v - \varrho + 1)\,\xi/(1 - \beta\xi) - (v - \varrho)\,\xi_z/(1 - \beta\xi_z)\,.$$

The numerator of (108) can be verified to have the form (105), with the coefficients (106). The last term in (108) is an increasing function of $\xi_z \leqq \xi$. Replacing $\xi_z$ by $\xi$ we get the estimate

$$\bar{H}_0(\beta) \leqq 2/\beta - (v - 2)/(1 - \beta) + \xi/(1 - \beta\xi)\,. \tag{109}$$

The last term in (109) is an increasing function of $\xi \leqq 1$ which gives the following upper bound

$$\bar{H}_0(\beta) \leqq 2/\beta - (v - 3)/(1 - \beta) = (2 - (v - 1)\,\beta)/(\beta(1 - \beta)) < 0 \tag{110}$$

for $\beta > 2/(v - 1)$, which determines the interval (107). □

**Remark.**

(ix) The upper bound of the interval (107) proved to be a reasonable starting point for an iterative gradient-type search of the optimal $\hat{\beta}$ because of the guaranteed sign of the gradient.

(Received February 2, 1988.)

## REFERENCES


[1] G. J. Biermann: Factorization Methods for Discrete Sequential Estimation. Academic Press, New York 1977.

[2] B. Harris and G. Heinell: The relation between statistical decision theory and approximation theory. In: Optimizing Method in Statistics — Proceedings of an International Conference (J. S. Rustagi, ed.). Academic Press, New York—San Francisco—London 1979.

[3] P. J. Huber: Robust Statistics. J. Wiley, New York 1981.

[4] M. Kárný: Algorithms for determining the model structure of a controlled system. Kybernetika *19* (1983), 164—178.

[5] M. Kárný nad K. M. Hangos: Approximation of Bayes rule. Preprints of the 7th IFAC/IFORS Symposium on Identification and System parameter Estimation, Vol. 1, pp. 1755—1760, York 1985.

[6] M. Kárný and K. M. Hangos: One-sided approximation of Bayes rule: theoretical background. Preprints of the 10th IFAC World Congress, Vol. 10, pp. 312—317, Munich 1987.

[7] M. Kárný and R. Kulhavý: Structure determination of regression-type models for adaptive prediction and control. In: Bayesian Analysis of Time Series and Dynamic Models (J. C. Spall, ed.). Marcel Dekker, New York 1988.

[8] R. Kulhavý: Restricted exponential forgetting in real-time identification. Automatica *23* (1987), 589—600.

[9] S. Kullback: Information Theory and Statistics. J. Wiley, New York 1959.

[10] A Perez: Information, $\varepsilon$-sufficiency and data reduction problems. Kybernetika *1* (1965), 297—323.

[11] V. Peterka: Bayesian approach to system identification. In: Trends and Progress in System Identification (P. Eykhoff, ed.), Chap. 8, pp. 239—304. (IFAC Series for Graduates, Research Workers and Practicing Engineers 1.) Pergamon Press, Oxford 1981.

[12] Y. Yashin: Martingal approach to identification of stochastic systems. Preprints of the 7th IFAC/IFORS Symposium on Identification and System parameter Estimation, Vol. 2, pp. 1755—1760, York 1985.



*Ing. Miroslav Kárný, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Science), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.*

*Dr. Katalin M. Hangos, Computer and Automation Institute — Hungarian Academy of Sciences, Kende utca 13—17, H 1502 Budapest. Hungary.*