10
Identification of Reality in Bayesian Context

Luděk Berec, Miroslav Kárný

ABSTRACT Complexity has many facets as does any general concept. The relationship between "infinitely" complex reality and restricted complexity of the artificial world of models is addressed. Particularly, the paper tries to clarify the meaning of Bayesian identification under mismodelling by answering the question, "What is the outcome of the Bayesian identification without supposing the model set considered contains the "true" system model?"
The answer relates known asympotic results to the "natural" finite-time domain of Bayesian paradigm. It serves as an interpretation "smoother" of those Bayesian identification results that quietly ignore the mismodelling present.

KEY WORDS Decision-making, model selection, Bayesian identification, approximation.

10.1 Introduction

System identification can be understood as the set of procedures which model an investigated part of reality (called object, process, plant or system) using data measured on it [8]. Modelling of the reality, often informal, is a necessary prerequisite to any prediction and/or control.

The theoretically ideal situation is that we are able to describe the objest so that the necessary simplifications have negligible influence on the quality of the decisions made. Such an ideal description is later referred to as a *true model.*

The Bayesian approach to system identification – shortly, Bayesian identification [4] – is a plausible and internally consistent paradigm. Its ability to provide consistent statements after any finite information processing is its main advantage. Seemingly, this advantage is paid by the unrealistic assumption that the unknown true model belongs to the set of the candidates considered. Under this assumption, the probability distribution constructed on this set is deductively modified by measured data. It concentrates asymptotically on the true model if the data are informative enough.

But, what is the relevance of this approach when the true model is beyond the set of its candidates?

There are definite asymptotic answers as to how the Bayesian identification behaves under some mismodelling situations, e.g. [2]. A conceptually clean understanding of its finite-time meaning is, however, lacking. This paper tries to fill this interpretation gap which is important for "beauty" of the Bayesian paradigm. The eligible answer is also of a practical significance. For instance, estimation of the control period [1] relies heavily on appropriate understanding of the problem.

To sum up, the paper answers the question:
What is the outcome of the Bayesian identification without supposing the true system model belongs into the set of considered model candidates?

It may seem that there is no connection of the paper to the notion of dimensionality. Of course there is. It is only the indirect connection but the fundamental one. In case of an affirmative answer to the question above *computations with models of reduced complexity, compared to the complexity of real systems, do have well defined sense.*

10.2 From decision-making to probability

This section explains the term "true system description" from a perspective of a wide set of decision tasks. It allows us also to introduce basic notions.

Assume we make decisions $d \in d^*$ using the knowledge $k \in k^*$ so that a real-valued non-negative loss function $\tilde{L}$ of d and of an uncertain entity $\omega \in \omega^*$ is "small". The notion "uncertain" means that the true value of the entity cannot be used for the considered decision making. Thus, a decision rule $R : k^* \to d^*$, $R \in R^*$, is searched for to "minimize" the loss function $L(R(k), \omega) \equiv \tilde{L}(d, \omega) \in [0, \infty]$. Note that:

- linguistic distinction of the terms *uncertain* and *unknown* is left out of our considerations, which cover them equally at operational level
- the available knowledge k is fixed and known at the decision time and so it is further suppressed in the notation
- x^* denotes generically a set of values of x.

The presence of the uncertain entity ω implies that the losses $L(R, \omega)$ cannot be completely ordered for different rules R. Their point-wise comparison provides a partial ordering:

Definition 10.2.1 *Let $L_{R_i}(\omega) \equiv L(R_i, \omega)$, $\omega \in \omega^*$, be the losses assigned to a pair of competitive rules R_i, $i = 1, 2$. The loss L_{R_1} is said to be*

preferable *to the* L_{R_2}, *formally* $L_{R_1} \preceq L_{R_2}$ *iff*

$$L_{R_1}(\omega) \leq L_{R_2}(\omega), \ \forall \omega \in \omega^*. \tag{10.1}$$

The loss L_{R_1} *is said to be* strictly preferable *to* L_{R_2} *if (10.1) holds and a "rich" set* $o \subset \omega^*$ *exists on which the inequality (10.1) is strict.*

The rule R_1 *is said to be* preferable *to the* R_2, $R_1 \preceq R_2$, *iff* $L_{R_1} \preceq L_{R_2}$. R_1 *is said to be* strictly preferable *to the* R_2, $R_1 \prec R_2$, *iff* $L_{R_1} \prec L_{R_2}$.

The rule $R \in R^*$ *is said to be* admissible *if there is no rule in* R^* *that is strictly preferable to it. Otherwise, it is said to be* inadmissible.

The term "rich" can be made more precise after describing the structure of the involved elements in detail. We leave it vague for the time being.

Undoubtedly, inadmissible strategies should be avoided. Their performance, judged according to L, is always worse than that of admissible ones. The partial preference ordering of strategies has to be completed in order to choose a good admissible strategy R in a systematic way. The following property helps us in the completion.

Definition 10.2.2 *A complete ordering* $\trianglelefteq$ *on the space* R^*, *admitting both strict inequality and inequality types of ordering, is said to be* strictly isotonic *with the partial ordering* $\preceq$ *iff* $R_1 \preceq R_2 \Rightarrow R_1 \trianglelefteq R_2$ *and* $R_1 \prec R_2 \Rightarrow R_1 \triangleleft R_2$.

Proposition 10.2.1 *Let the set* $(R^*, \trianglelefteq)$ *have a smallest element* R_0 *and* $\trianglelefteq$ *is strictly isotonic with respect to* $\preceq$. *Then,* R_0 *is admissible.*

Proof: By contradiction, let $R \in R^*$ be strictly preferable to R_0. Thus, $L_R(\omega) \leq L_{R_0}(\omega)$ on ω^* and this inequality becomes strict on a rich $o \subset \omega^*$. This inequality, strict isotonicity of $\trianglelefteq$ and definition of the smallest element imply the contradictory inequality $R \triangleleft R_0 \trianglelefteq R$. □

The partial ordering $\preceq$ of decision rules is defined using ordering of the associated loss functions. Under general topologic conditions [5], the complete strictly isotonic ordering $\trianglelefteq$ exists. It can be interpreted as the ordering induced by "expected" losses

$$E\{L_{R_1}\} \leq E\{L_{R_2}\} \ \Leftrightarrow \ R_1 \trianglelefteq R_2.$$

The "expectation" functional $E\{L_R\} \in [0, \infty]$ "removes" the uncertain factor ω, i.e. the value $E\{L_R\}$ depends on R only. This complete ordering on R^* will be denoted $\trianglelefteq_E$ in order to stress its connection to the functional E.

The smallest element R_0 with respect to $\trianglelefteq_E$ will be found (a posteriori) the better, the more the "expectation" E grasps the objective properties of the uncertain factor ω. Such "objective expectation" should not depend on the *set* of strategies R^* in which the optimum is searched for.

Proposition 10.2.2 *If E defining $\trianglelefteq_E$ is not strictly isotonic on R^* then there is a restriction $\tilde{R}^* \subset R^*$ on which an inadmissible minimizer exists.*

Proof: E can be recognized as not being strictly isotonic if there are R_1, $R_2 \in R^*$ such that $L_{R_1} \prec L_{R_2}$ and $E\{L_{R_1}\} \geq E\{L_{R_2}\}$. Thus, R_2 is an inadmissible minimizer (not necessarily unique) on $\tilde{R}^* = \{R_1, R_2\}$. □

Propositions 10.2.1, 10.2.2 explain why an E with the strict isotonicity property is considered further on.

The "expectation" E introduced up to now depends on the loss $\tilde{L}$. It can be called *objective* if it completes partial preference orderings $\preceq$ defined by a rich set of loss functions L^*. The following "test" set L^* is considered: ω^* be a compact set and, for a fixed $R \in R^*$, the functions in

$$L^* = \{L \equiv L_R : \omega^* \to [0, \infty)\}$$

are continuous. L^* is normed space with supreme norm $||\cdot||$.

Proposition 10.2.3 *Let $L_1, L_2, \ldots, L_n$ be in L^* and let $E\{\cdot\}$ be:*

(i) Sequentially continuous $\Leftrightarrow$ the sequence of expected values $E\{L_n\}$ be Cauchy for any point-wise convergent sequence $\{L_n\}$

(ii) Additive on loss functions with disjunct supports $\Leftrightarrow$ $E\{L_1 + L_2\} = E\{L_1\} + E\{L_2\}$ if $L_1 L_2 = 0$

(iii) Boundedly uniformly continuous $\Leftrightarrow$ $\forall(\varepsilon, \gamma) > (0,0)\ \exists \delta_{(\varepsilon,\gamma)} > 0$ such that if $||L_1|| < \gamma$, $||L_2|| < \gamma$ and $||L_1 - L_2|| < \delta_{(\varepsilon,\gamma)}$ then $|E\{L_1\} - E\{L_2\}| < \varepsilon$.

Then $E\{\cdot\}$ is representable as an integral

$$E\{L\} = \int_{\omega^*} U(L(\omega), \omega)\, \mu(d\omega). \tag{10.2}$$

μ is a finite regular Borel measure on ω^*. The function U satisfies $U(0, \omega) = 0$, is continuous in L almost everywhere (a.e.) on ω^*, bounded a.e. on ω^* for each $L \in L^*$. Moreover, sequence $U(L_n, \cdot)$ is Cauchy in the space of μ-integrable functions.

Proof: See [6], Theorem 5, Chapter 9.3, p. 479. □

This technical proposition specifies conditions under which the "expected" loss becomes the ordinary expectation of the utility function U. U is able to express decision-maker's attitude (he is risk aware or prone or indifferent). The measure μ is universal with respect to a rich class of decision tasks facing the same uncertainty.

The assumptions (i), (iii) are technical and widely acceptable (loosely speaking, a small change in the loss should not lead to a substantial change of its expected value). The additivity on loss functions with disjunct supports is the most questionable restriction. This assumption which is much

weaker than the general additivity or even linearity is, however, intuitively acceptable if the the functions L_1, L_2 are interpreted as a single loss decomposed to its restrictions on a subset of ω^* and its complement, respectively.

In the risk-indifferent case, $U(L(\omega), \omega) = L(\omega)$, the expectation (10.2) is isotonic if the measure μ is non-negative. To achieve strict isotonicity, μ has to be positive on ω^* almost everywhere if the rich set o in Definition 10.2.1 is specified as a set of non-zero Lebesgue measure. Then the expectation reduces to the standard one if the preservation of constant Ls is demanded, i.e. $\mu(\omega^*) = 1$.

For simplicity, the involved measure μ is supposed to have a Radon-Nikodym derivative $f(\omega)$ with respect to Lebesgue measure. Thus, the complete strictly isotonous ordering is induced by expected utility

$$\int_{\omega^*} U(L(\omega), \omega)\, f(\omega)\, d\omega.$$

The derivative $f(\omega)$ *has all properties of a probability density function* (pdf) and the operational equivalence *uncertain* $\equiv$ *random* becomes relevant.

Note that $f(\omega)$ has been constructed with a strategy fixed. Thus, it generally depends on it, i.e. $f(\omega) \equiv f_R(\omega)$. Particularly, this dependence distinguishes control as a special decision task.

To summarize, quite general conditions have been found under which decision tasks involving uncertainty require description of the involved uncertainties ω in probabilistic terms. Practically, the symbol ω may represent both random elements (e.g. measurement noise) and unknown constants (e.g. system gain). Here, both types of uncertainty are unified and treated as random variables. This treatment of constants (their randomization) forms basis of the Bayesian statistics. There are alternative and better justifications of the Bayesian paradigm, e.g. [4, 5]. We have, however, arrived at a pdf which corresponds to a rich class of decision problems and as such it can be called an *objective (true) pdf.*

10.3 Bayesian identification

In the complex problems met in automatic control and signal processing, the pdf describing uncertain quantities is constructed from simpler elements by standard procedures called estimation, filtering and prediction. Let us recall them here. They serve us as a starting point in presenting our main result.

Let $\omega = (u(t) = (u_1, \ldots, u_t), y(t) = (y_1, \ldots, y_t))$ be formed by input and output sequences fed into and observed on a controlled system up to some horizon t. Their relationship is uncertain (incompletely known/random) and as such it is described by the pdf $f_R(\omega) \equiv f_R(u(t), y(t))$ Let us consider the usual case that $P_{u_\tau} = [u(\tau - 1), y(\tau - 1), \text{prior knowledge}]$ is used when

choosing u_τ. Thus, the strategy (sequence of decision rules) or control law becomes $R = R(t) \equiv \{R_\tau : P^*_{u_\tau} \to u^*_\tau\}_{\tau \le t}$.

The *chain rule* for pdfs implies

$$f_R(\omega) = \prod_{\tau=1}^{t} f_R(u_\tau|P_{u_\tau}) f_R(y_\tau|u_\tau, P_{u_\tau}).$$

The conditional pdf $f_R(u_\tau|P_{u_\tau})$ determines probability of generating u_τ when the past P_{u_τ} has been observed. These pdfs describe (randomized) *control law* i.e.

$$R(t) \equiv \{f_R(u_\tau|P_{u_\tau})\}_{\tau \le t}.$$

They reflect (possibly random) rules of input selection. They are the main outcome of the supported optimization.

The conditional pdf $f_R(y_\tau|u_\tau, P_{u_\tau})$, describes probability of observing y_τ when u_τ is applied and the past P_{u_τ} observed. These pdfs describe the (random) response of the controlled system. They represent the *system model*

$$S(t) \equiv \{f_R(y_\tau|u_\tau, P_{u_\tau})\}_{\tau \le t}$$

needed for the optimal control design that minimizes expected loss. Often, $S(t)$ depends on $R(t)$ through the applied inputs only.

The considered learning systems construct models of reality indirectly by identifying a so called *parametrized system model* given by pdfs

$$f_R(y_\tau|u_\tau, P_{u_\tau}, \Theta).$$

The additional variable $\Theta \in \Theta^*$ "points" to alternative models. It is interpreted as an unknown (multivariate) parameter and has a very general structure [7]. The term unknown means that Θ is not a part of the knowledge available to control strategy, i.e. so called natural conditions of control [4] are fulfilled $f_R(u_\tau|P_{u_\tau}, \Theta) = f_R(u_\tau|P_{u_\tau})$. These pdfs specify parametrized description of the interconnection controller-system

$$f_R(u(t), y(t)|\Theta) = \prod_{\tau=1}^{t} f_R(u_\tau|P_{u_\tau}) f_R(y_\tau|u_\tau, P_{u_\tau}, \Theta)$$

through the chain rule.

Traditionally, it is supposed that a "true" parameter Θ^T exists in the considered set Θ^* of possible Θ values, i.e. the "objective" pdf $f_R(\omega)$ discussed in previous section coincides with $f_R(u(t), y(t)|\Theta^T)$. Then, a subjective prior pdf $f(\Theta) > 0$ is selected on Θ^*. It distributes (subjective) belief that particular values of Θ coincide with Θ^T. This prior pdf is corrected by the observed data. The resulting pdf is used for prediction or for construction of the system model needed for control design. The probabilistic rules employed are summarized in

Proposition 10.3.1 *The Bayesian parameter estimate (in a wide sense), i.e. the posterior pdf of the unknown Θ, evolves according to the formula*

$$f(\Theta|P_{u_{t+1}}) \propto f_R(y_t|u_t, P_{u_t}, \Theta) f(\Theta|P_{u_t}) \tag{10.3}$$

with $f(\Theta|P_{u_1}) \equiv f(\Theta)$. The symbol $\propto$ means proportionality up to a factor independent of Θ. The Bayesian prediction (in a wide sense), i.e. the predictive pdf (system model with the excluded parameter) is given by the formula

$$f_R(y_t|u_t, P_{u_t}) = \int_{\Theta^*} f_R(y_t|u_t, P_{u_t}, \Theta) f(\Theta|P_{u_t})\, d\Theta.$$

These formulae are valid under natural conditions of control [4].

Proof: In fact, the unknown parameter completes the colection of uncertain quantities to $\omega = (u(t), y(t), \Theta)$. The corresponding joint pdf is a product of "objective" and subjective factors:

$$f_R(u(t), y(t), \Theta) = f_R(u(t), y(t)|\Theta) f(\Theta).$$

Both the estimation and prediction just evaluate marginal/conditional pdfs and insert measured data. For details, see [4]. □

The formula (10.3) implies that zero prior belief keeps the posterior one at zero, irrespectively of data. Thus, we cannot learn of Θ not assumed a priori as a possible "true" parameter.

At the same time we know that, at least due to the complexity of the Nature, the *"true" parameter is out of any tractable set* Θ^*. Thus, the natural question addressed in the paper arises: *what are we doing when we apply Bayesian paradigm and at the same time face this situation?*

10.4 Bayesian paradigm revised

Recall: a set of decision tasks is considered and parametrized models specifying $f_R(u(t), y(t)|\Theta)$, $\Theta \in \Theta^*$ are selected. The triple $(u(t), y(t), \Theta)$ can be complemented by all relevant (unmodelled) influences, say $g(t)$, to the full quadruple $\omega_g = (u(t), y(t), \Theta, g(t))$ of uncertain entities in the problem. The completness means that an objective probabilistic measure $\mu(d\omega_g)$ characterized in Proposition 10.2.3 exists. For simplicity, the factor $g(t)$ representing mismodelling is assumed not to prevent us from characterizing μ by the pdf $f^T(\omega_g)$.

Obviously, the inspected losses do not depend on the *unmodelled* factor $g(t)$ (we do not know how to quantify it so that we cannot include it into our loss function). For the same reason, it cannot influence attitude to

the uncertainty risk: $U(L(\omega_g),\omega_g) = U(L(\omega),\omega),\ \omega = (u(t), y(t), \Theta)$. This fact together with the formula for expected utility imply that the marginal true pdf $f^T(\omega) \equiv f^T(u(t), y(t), \Theta) = \int_{g^*(t)} f^T(\omega, g(t))\, dg(t)$ is supposed to exist.

The parameter Θ is "man-made", it characterizes models. Its marginal (prior) distribution coincides with prior belief $f(\Theta)$ attached to the possible values $\Theta \in \Theta^*$. Thus, the true pdf can be factorized

$$f^T(\omega) \equiv f^T(u(t), y(t)|\Theta) f(\Theta)$$

The artificial nature of Θ implies that $f^T(u(t), y(t)|\Theta) = f^T(u(t), y(t))$, so, it does not depend on Θ. The true pdf of the data, $f^T(u(t), y(t))$ is unknown and need not coincide with any considered model. As a pdf, the chain rule is valid for it. Moreover, the same randomized decision rule $f(u_\tau|P_{u_\tau})$ is imposed both on the true and the model system.

Facing mismodelling, the understanding of Bayesian identification as the data correction of the prior pdf $f(\Theta)$, interpreted as a belief of the *statement* $\Theta = \Theta^T$ = *true parameter*, as described in Proposition 10.3.1, lacks meaning. Instead, the *key shift* in the paradigm, proposed here, consists of interpreting the Bayesian identification as redistribution of the prior pdf $f(\Theta)$ as the belief of the *statement* $\Theta = \Theta^{best}$ = *the pointer to the best approximant* $f(u(t), y(t)|\Theta^{best})$ *to the unknown pdf* $f^T(u(t), y(t))$. In other words, the answer to the key question of the paper is suggested as follows:

We want to learn the reality (the true system model) but what we really get, using Bayesian identification, is information on the best projection of the true model to the considered set of models candidates.

Obviously, under mismodelling we work with a sort of projection but the two questions arise:

- What type of projection we are dealing with?
- Why is the best one learnt?

These questions are answered in the next section. Some remarks should be made beforehand. Note that:

- The inpenetrable barrier between reality and artificial world of models has remained: the projection error is out of control whenever the set of considered models is fixed. It is bad news as no perpetuum mobile was proposed, but it is good news as the importance of modelling is again underlined.

- The proposed answer is really a *generalization* of the case $\Theta^T \in \Theta^*$. In this situation, the best projection coincides with the true data pdf and "classical" Bayesian interpretation remains to be valid.

- The result supports the Bayesian identification as a tool for constructing the system model as predictor generated by a parametrized model. Use of parametrized models offers the chance to select a rel-

atively rich set of models to which the true one projected. It opens the possibility of making their distance small.

10.5 What projection?

This section singles out the adequate $\Theta^{best} \in \Theta^*$. The result is motivated by Shannon-McMillan-Breiman theorem [3]. The adopted version minimizes assumptions on the true pdfs and restricts the class of parametrized models to those with finite-dimensional observable state

$$f(y_t|u_t, P_{u_t}, \Theta) = m_t^{\Theta}(\Psi_t). \tag{10.4}$$

Here, m_t^{Θ} is a known (generally, time-varying) function of a finite-dimensional data vector $\Psi_t \equiv (y_t, \psi_t)$. The "regression" vector ψ_t is a known function of ψ_{t-1} and the observed data u_t, y_{t-1}. The initial condition ψ_0 is assumed to be known, too. The functions m_t^{Θ} are positive on their domains Ψ_t^* for all fixed $\Theta \in \Theta^*$ and $t = 1, 2, \ldots$

The estimated parameter is supposed to be time-invariant. Thus, all estimators based on $u(t), y(t)$, $t = 1, 2, \ldots$ have the common aim, to estimate the same quantity. It implies that the parameter estimates found for $t \to \infty$ are the only relevant ones.

Let us fix a possible value of the parameter $\Theta \in \Theta^*$. For the model (10.4), the Bayesian estimate (10.3) can be given the form

$$\begin{aligned} f(\Theta|P_{u_{t+1}}) &\propto f(\Theta)\exp[\ln\left(\frac{f(u(t), y(t)|\Theta)}{f^T(u(t), y(t))}\right)] \\ &\propto f(\Theta)\exp\left[t\frac{1}{t}\sum_{\tau=1}^{t}\ln\left(\frac{m_\tau^{\Theta}(\Psi_\tau)}{f^T(y_\tau|u_\tau, P_{u_\tau})}\right)\right] \\ &\equiv f(\Theta)\exp[tH_t^{\Theta}]. \end{aligned} \tag{10.5}$$

The following auxiliary proposition helps us in formulating the main result of this section.

Proposition 10.5.1 *Let $m_t^{\Theta}(\Psi_t)$ be positive on $(u^*(t), y^*(t)) \equiv$ support of $f^T(u(t), y(t))$. Then, almost surely,*

$$\limsup_{t\to\infty} H_t^{\Theta} \equiv H_\infty^{\Theta} \leq 0$$

Proof: A straightforward computation shows that

$$E\left[\prod_{\tau=1}^{t}\frac{m_\tau^{\Theta}(\Psi_\tau)}{f^T(y_\tau|u_\tau, P_{u_\tau})}\right] = 1$$

where the expectation is done over the data set. The Markov inequality for positive random variables implies that, for arbitrary $\varepsilon > 0$,

$$\Pr\{\frac{1}{t}\ln\left[\prod_{\tau=1}^{t}\frac{m_\tau^\Theta(\Psi_\tau)}{f^T(y_\tau|u_\tau,P_{u_\tau})}\right]\geq\varepsilon\}\leq\exp(-t\varepsilon).$$

This inequality and Borel-Cantelli lemma give the conclusion. □

It is obvious that $H_\infty^{\Theta_0}=0$ if $m_t^{\Theta_0}$ coincides asymptotically with the true pdf. Thus, mismodelling can be expressed by the assumption

$$H_\infty^\Theta\leq -h<0 \tag{10.6}$$

for some constant $h>0$ and all $\Theta\in\Theta^*$.

Proposition 10.5.2 *Let $m_t^\Theta(\Psi_t)$ be positive on $(u^*(t),y^*(t))\equiv$ support of $f^T(u(t),y(t))$. Let us define*

$$\Theta^{*\infty}\equiv\{Arg\max_{\Theta\in\Theta^*}H_\infty^\Theta\}\cap\Theta^*.$$

If (10.6) holds then the Bayesian estimate concentrates on $\Theta^{*\infty}$ if it is non-empty.

Proof: Using formula (10.5), we can see that $f(\Theta|P_{u_t})$ behaves asymptotically as $\exp[t\{H_\infty^\Theta-\sup_{\Theta\in\Theta^*}H_\infty^\Theta\}]$. It is non-zero for maximizing arguments only. □

Note that:

- various conditions (a continuity type in Θ for m_t^Θ and compactness of Θ^*) can be impossed to guarantee $\Theta^{*\infty}\neq\emptyset$.
- if there is information on the dependence structure of f^T its strongly consistent estimators can be constructed and substituted instead of f^T. A version of the large deviation theorem is obtained that provides a constructive guideline for approximate estimation with subsufficient statistics [2].
- The quantity H_∞^Θ is tightly connected with the entropy rate notion. For instance, symbolically, $1/t\sum_{\tau=1}^t\rightarrow E$ holds under ergodicity-type assumptions [3].

10.6 Example

One of the practical examples based on the idea of the existence of mismodelling concerns the problem of the estimation of control period [1].

The control period is defined as an integer multiplier of the sampling period. The estimation of control period thus looks for the choice of an integer n (the sampling period is formally assigned the length 1). The choice of n implies by definition that during the system control, inputs are kept constant within n consequent control steps and only a single representant of the n-tuple of outputs measured within these steps is used in the feedback.

The idea of the estimation of n follows. Facing the mismodelling and trying to build an approximation of the true pdf f^T, the combination of direct two-, three- or more-steps ahead predictors can result in a better approximation of f^T than the combination of one-step ahead predictors (which seems to be the best case having no mismodelling at all, i.e. when the true model coincides with any from the set of model candidates). For details, see [1, 7].

So, together with standard model structure and parameter estimation also the number n representing the chosen n-steps ahead predictor is estimated. Such estimate $\hat{n}$ is then interpreted as the searched control period.

Using the just developed interpretation machinery, the simultaneous estimation of n, model structure and parameters themselves is just an effort to find out the best projection of the true pdf into the set of pdfs generated by used ('generalized' ARX) models [1]. We can use such estimation procedure anyway but now the result is easy to interpret and well justified.

10.7 Conclusions

The paper answers the question:
What is the outcome of the Bayesian identification without supposing that the true model belongs into the considered set of model candidates.

The answer summarizes in the statement:
We try to identify the reality (the true system model) but what we really get, using Bayesian identification, is an information on the best projection of the true model to such considered set of model candidates.

The novelty of this statement lies in its validity for any (even finite) information which is processed.

The contribution of the paper to modelling is indirect only. The statement just underlines the well known rule of thumb: use of known facts (physical laws, ability to approximate a rich set of mappings, expert's knowledge etc.) for making a set of candidates to model the reality. Such set will decide on quality of the projection, on mismodelling error.

Similarly, the answer has in fact no computational consequences, but unifies and serves as an interpretation of various previously obtained results in the Bayesian identification field.

In the paper, a mathematical treatment of this idea is presented with two particular results of interest:

- the meaning of the "true" probabilistic description is clarified and related to a set of decision tasks,
- asymptotic behaviour of Bayesian estimator is characterized.

To sum up, the paper is of a methodological nature. It tries to connect the model to reality more precisely than just saying 'model is *an* approximation of the real object'. It tries to model 'mismodelling'. In other words, it specifies the difference between what we actually get using the once chosen model and what we would get using the true model. The goal of the paper is not to say absolutely that a particular model is good or bad, but, from the set of models given in advance (which, e.g., differ by a value of one parameter of the generic model), tries to find the relatively best (with respect to this set) model to form the best projection of the true pdf to the corresponding set of model-generated pdfs. Thus, only the models specified beforehand are used for comparison.

In this way the Bayesian identification procedures used are better justified and, hopefully, a space is open for new constructions.

Acknowledgement This research was partially supported by GA AV ČR, grants No. A2075603 and A2075606.

Useful hint on Shannon-McMillan-Breiman theorem was kindly provided by Dr. I. Vajda.

10.8 References

[1] M. Kárný, *Estimation of control period for selftuners*, *Automatica*, vol. 27, no. 2, pp. 339–348, 1991, extended version of the paper presented at 11th IFAC World Congress, Tallinn.

[2] R. Kulhavý, *A Kullback-Leibler distance approach to system identification*, in *Preprints of the IFAC Symposium on Adaptive Systems in Control and Signal Processing*, Cs. Bányász, Ed., pp. 55–66. Budapest, 1995.

[3] P.H. Algoet, T.M. Cover, *A sandwich proof fo the Shannon-McMillan-Breiman theorem*, *The Annals of Probability*, vol. 16, pp. 899–909, 1988.

[4] V. Peterka, *Bayesian Approach to System Identification*, in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, 1981.

[5] T.L. Fine, *Theories of Probability*, Academic Press, New York, London, 1973.

[6] M.M. Rao, *Measure Theory and Integration*, John Wiley & Sons, New York, 1987.

[7] L. Berec, *On Model Structure Identification (A unifying view and a particular example)*, in *Preprints of the Workshop CMP'96*, L. Berec, J. Rojíček, M. Kárný, K. Warwick, Eds., pp. 121–128. Prague, 1996.

[8] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, 1987.