ON CROSS-VALIDATION OF CONTROLLED DYNAMIC MODELS: BAYESIAN APPROACH

Miroslav Kárný, Petr Nedoma,Václav Šmídl

Institute of Information Theory and Automation AV ČR,
P.O.Box 18, 180 00 Praha 8, Czech Republic

Abstract: The use of the model is the decisive validation step in its building. As a rule, the use of a bad model is too costly so that a model validation is an obligatory step in its learning and a naturally relevant extensive theory has been developed within statistical community. However, the available rules deal almost exclusively with independent data samples. Consequently, they are substantially disqualified for validation of *dynamic* models.
This paper provides the missing solution using Bayesian formulation and solution of the problem. The rule is elaborated for validation of the model gained via estimation within practically important exponential family.

Keywords: model validation, Bayesian estimation, dynamic models, exponential family

1. INTRODUCTION

Learning is a standard part in model building Ljung (1987); Bohlin (1991). In order to avoid costly consequences of employing inadequate model, the found model has to be validated before its final use. This led to development of an extensive theory dealing with model validation, see e.g. the review Plutowski (1996). However, the available procedures deal almost exclusively with independent data samples. Consequently, they cannot be used for validation of *dynamic* models. Just a few available exceptions deal with specific cases only Huang (2001).

A real need for systematic validation of dynamic models in led us to development of a general validation procedure based on Bayesian decision making theory Berger (1985).

After preparatory Section 2, the addressed problem is formulated and solved in Section 3. The solution is applied to estimation in dynamic exponential family, Barndorff-Nielsen (1978), in Section 4. Performance of the algorithm is illustrated on a simple example in Section 5. The paper is closed by concluding remarks, Section 6.

2. PRELIMINARIES

The paper uses the following notations: $\equiv$ is defining equality; X^* denotes a set of X-values; $\mathring{X}$ means cardinality of a finite set X^*; $f(\cdot|\cdot)$ denotes probability density function (pdf); $\propto$ means equality up to a normalizing factor; t labels discrete-time moments, $t \in t^* \equiv \{1, \ldots, \mathring{t}\}$; $\mathring{t} < \infty$ is a given learning horizon; $d_t = (y_t, u_t)$ is the data record at time t consisting of an observed system output y_t and of an optional system input u_t; x_t is an unobserved system state; $X(t)$ denotes the sequence $(X_1, \ldots, X_t)$, $X(t) \in \{d(t), y(t), u(t), x(t)\}$.

The following simplifications are also adopted.

- Names of arguments distinguish pdfs. No formal distinction is made between a random variable, its realization and an argument of a pdf.

- All integrals are definite and multivariate. The integration domain coincides with support of the pdf in its argument.

The joint pdf $f(d(\mathring{t}), x(\mathring{t})|x_0, d(0))f(x_0|d(0)) = f(d(\mathring{t}), x(\mathring{t})|x_0)f(x_0)$ of involved random variables is the most complete probabilistic description of the controlled closed loop. In it, x_0 is initial uncertain state. The symbol $d(0)$ stands for the prior information available before the choice of the first input. Habitually, $d(0)$ is considered implicitly.

The chain rule for pdfs Peterka (1981) implies the following decomposition of the above joint pdf:

$$f(d(\mathring{t}), x(\mathring{t})|x_0) = f(x_0) \times \prod_{t\in t^*} \times \tag{1}$$
$$\times \underbrace{f(y_t|u_t, d(t-1), x(t))}_{\text{observation model}} \times$$
$$\times \underbrace{f(x_t|u_t, d(t-1), x(t-1))}_{\text{state evolution model}} \times$$
$$\times \underbrace{f(u_t|d(t-1), x(t-1))}_{\text{randomized controller}}.$$

The following **assumptions** are adopted:

Observation model of y_t depends on a finite dimensional *regression vector* ψ_t, which is a function of $u_t, d_{t-1}, \ldots, d_{t-\partial}$, $\partial < \infty$, and on the system state x_t

$$f(y_t|u_t, d(t-1), x(t)) = f(y_t|\psi_t, x_t).$$

State evolution model of x_t depends on the regression ψ_t and the past system state x_{t-1}

$$f(x_t|u_t, d(t-1), x(t-1)) = f(x_t|\psi_t, x_{t-1}).$$

Randomized control providing the system input u_t is *admissible* thus exploits only the observed data history $d(t-1)$ and ignores the unobserved states $x(t-1)$

$$f(u_t|d(t-1), x(t-1)) = f(u_t|d(t-1)).$$

□

Hence, the closed loop description (1) reduces to

$$f(d(\mathring{t}), x(\mathring{t})|x_0) = \prod_{t\in t^*} f(y_t|\psi_t, x_t) \times$$
$$\times\ f(x_t|\psi_t, x_{t-1})f(u_t|d(t-1)) \tag{2}$$

and the following proposition holds.

Proposition 1. (Filtering in closed control loop). Let the pdf $f(x_0)$ be given, $d(0)$ together with u_1 determines the initial regression vector ψ_1 and the **assumptions** hold. Then, the pdf $f(x_t|d(t))$, determining the *state estimate*, the pdf $f(x_t|u_t, d(t-1))$, determining the *state prediction*, and the pdf $f(y_t|u_t, d(t-1))$, determining the *output prediction*, evolve as follows

$$\textit{Time updating} \quad f(x_t|u_t, d(t-1)) =$$
$$= \int f(x_t|\psi_t, x_{t-1})f(x_{t-1}|d(t-1))\, dx_{t-1}$$
$$\textit{Data updating} \quad f(x_t|d(t)) = \tag{3}$$
$$= \frac{f(y_t|\psi_t, x_t)f(x_t|u_t, d(t-1))}{f(y_t|u_t, d(t-1))}$$
$$\textit{Output prediction} \quad f(y_t|u_t, d(t-1)) =$$
$$= \int f(y_t|u_t, d(-1), x_t)f(x_t|u_t, d(t-1))\, dx_t.$$

Proof: Omitted. □

3. PROBLEM FORMULATION AND SOLUTION

Learning aims to find the *best model* ${}^{\lfloor o}\mathcal{M} \in \mathcal{M}^*$ of the inspected controlled system. Specification of what is meant by the "best" is possible only on a confined class of models. Without prior choise of the model class, learning is always an open-ended story. Ideally, the posterior distribution on $\mathcal{M}^*$ should be build before selecting the relevant model.

However, the set of models $\mathcal{M}^*$ is infinite dimensional and a practical construction of the prior distribution over it, as well as evaluation of its moments, is intractable. Therefore, we consider the prior to be uniform on $\mathcal{M}^*$, which implies that the maximum likelihood estimate is the best model ${}^{\lfloor o}\mathcal{M}$. The likelihood function $\mathcal{L}(d(\mathring{t}), \mathcal{M})$ of $\mathcal{M}$ coincides with the factor of $f(d(\mathring{t})|\mathcal{M})$ that depends on $\mathcal{M}$. Thus, construction of the likelihood function is implied by Proposition 1:

$$\mathcal{L}(d(\mathring{t}), \mathcal{M}) = \prod_{t\in t^*} \underbrace{f(y_t|u_t, d(t-1), \mathcal{M})}_{\text{output prediction (3)}}. \tag{4}$$

Adopting this view point, the estimation selects among various models from $\mathcal{M}^*$ the model with the highest v-likelihood (4).

Model validation is an additional test on the quality of ${}^{\lfloor o}\mathcal{M}$. Inspired by the classical model validation theory Plutowski (1996), we split all the *available data* $d(\mathring{t})$ on: (i) *learning data* ${}^{\lfloor l}d$, and (ii) *validation data* ${}^{\lfloor v}d$. The best model ${}^{\lfloor o}\mathcal{M}$ is learnt on the learning data ${}^{\lfloor l}d$ and its performance is checked on the validation data ${}^{\lfloor v}d$. The validation technique is essentially inspecting how good is the best *dynamic* model ${}^{\lfloor o}\mathcal{M}$ in extrapolating of the past to the future. Thus, the learning data ${}^{\lfloor l}d$ has to form the "prefix" part of $d(\mathring{t})$ and the validation data ${}^{\lfloor v}d$ the "suffix" part.

The results of validation strongly depend on the choise of the splitting moment. None of the existing methods, Plutowski (1996), is directly prepared for the considered dynamic models. These

models allow just cutting into contiguous sequences. Essentially, the available data up to a *cutting moment* τ are taken as learning data and the rest as validation data. This reduces the number of split ways but at the same time disqualifies the majority of the available analysis.

This motivates us to design an adequate, purely Bayesian, formulation and solution of the model validation problem. Let us consider a fixed cutting moment $\tau \in t^* \cup \{0\}$, which defines the *learning data*

$$\lfloor^l d(\tau) \equiv d(\tau), \tag{5}$$

and the *validation data*

$$\lfloor^v d(\mathring{t} \setminus \tau) \equiv (d_{\tau-\partial}, \ldots, d_{\mathring{t}}). \tag{6}$$

Consider the following hypotheses

$H_0 \equiv$ All recorded data $d(\mathring{t})$ are described by the learnt model $\lfloor^o\mathcal{M}$.

The v-likelihood of this hypothesis is obtained by performing stochastic filtering on all data giving

$$f(d(\mathring{t})|H_0) \propto \mathcal{L}(d(\mathring{t}), \lfloor^o\mathcal{M}) \equiv \mathcal{L}(d(\mathring{t}), H_0). \tag{7}$$

$H_1 \equiv$ Learning data and validation data should be described by individual models.

The v-likelihood of this hypothesis is obtained by independent filter runs on both data collections giving

$$f(d(\mathring{t})|H_1, \tau) \propto \mathcal{L}\left(\lfloor^l d(\tau), \lfloor^o\mathcal{M}\right) \times$$
$$\times\, \mathcal{L}\left(\lfloor^v d(\mathring{t} \setminus \tau), \lfloor^1\mathcal{M}\right) \equiv \mathcal{L}(d(\mathring{t}), H_1|\tau). \tag{8}$$

The model $\lfloor^1\mathcal{M}$ used on validation data may differ from $\lfloor^o\mathcal{M}$. The difficulty to find a real different competitor makes us to choose $\lfloor^1\mathcal{M} = \lfloor^o\mathcal{M}$.

Note that proportionality factor is the same in both cases as it is formed by the factor $\prod_{t\in t^*} f(u_t|d(t-1))$ describing the admissible strategy used while collecting data $d(\mathring{t})$.

This approach is graphically illustrated in Fig. 3. Estimation on the whole data $d(\mathring{t})$ yields result in the class of single models. Estimation on the split data yields result in the class of multiple (or switching) models. The latter class is, of course, richer. The first (single) model is found valid if the best estimate in the richer class has the same (close) likelihood. This case is illustrated by System 1 in Fig. 3. System 2 illustrated the case where the switching model outperforms the single model.

With no prior prejudice, $f(H_0|\tau) = f(H_1|\tau)$, the Bayes rule provides the posterior pdf $f(H_0|d(\mathring{t}), \tau)$. The learnt model can be accepted as good one

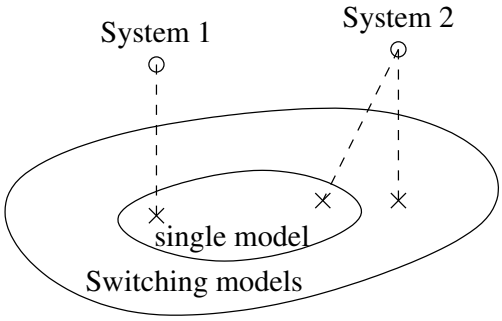

Fig. 1. Illustration of the proposed validation scheme. Elipses denotes classes of models, small circles denotes the real systems, crosses denotes models of the systems estimated within each class. Dashed lines illustrates likelihood of the models.

if the posterior pf $f(H_0|d(\mathring{t}), \tau)$ is high enough. Otherwise, we have to search for the reason why the chosen model is not reliable enough. It gives the algorithmic solution.

Algorithm 1. (Model validation for a fixed τ).

(1) Run filtering, Proposition 1, on the learning $\lfloor^l d(\tau)$, validation $\lfloor^v d(\mathring{t} \setminus \tau)$ and full $d(\mathring{t})$ data.
(2) Evaluate the v-likelihoods $\mathcal{L}\left(\lfloor^l d(\tau), H_1|\tau\right)$, $\mathcal{L}\left(\lfloor^v d(\mathring{t} \setminus \tau), H_1|\tau\right)$ and $\mathcal{L}\left(d(\mathring{t}), H_0\right)$.
(3) Using the Bayes rule, probability that the learning was suggesfull is:

$$f\left(success|d(\mathring{t}), \tau\right) \equiv f\left(H_0|d(\mathring{t}), \tau\right) =$$
$$= \frac{\mathcal{L}(d(\mathring{t}), H_0|\tau)}{\mathcal{L}(d(\mathring{t}), H_0|\tau) + \mathcal{L}(d(\mathring{t}), H_1|\tau)} \tag{9}$$

where likelihoods of both hypotheses are given by (7) and (8) respectively.
(4) The validation test is successfully passed for a given τ if $f(H_0|d(\mathring{t}), \tau)$ is close to 1. Otherwise, measures for a better learning have to be taken.

□

Results of the test depend, often strongly, on the selected cutting moment τ. Thus, it makes sense to validate learning for various cutting moments $\tau \in \tau^* \subset t^*$. we are making a pair of decisions $(\hat{H}, \tau)$ based on the available data $d(\mathring{t})$ We select $\tau \in \tau^*$ and accept $(\hat{H} = H_0)$ or reject $(\hat{H} = H_1)$ the hypothesis H_0 that the learnt model is valid.

We solve this static decision task and select the optimal decision $\lfloor^o\hat{H}$ on inspected hypotheses and optimal cutting time moment $\lfloor^o\tau$ as a minimizer of the expected loss. We assume, for simplicity, that the losses caused by a wrong acceptance and rejection are identical, say (without loss of generality) 1. The loss function is therefore chosen as

$$\mathcal{Z}(H,\hat{H},\tau)=\left[1-\delta\left(\hat{H}(\tau)-H\right)\right]$$

. where $\delta(\cdot)$ is Kronecker delta for discrete arguments and Dirac delta in continuous case. The optimal decisions is then:

$$^{\lfloor o}\hat{H},\ ^{\lfloor o}\tau=\mathrm{Arg}\min_{\hat{H},\tau^*}\mathcal{E}_{f(H|d(\mathring{t}))}\left\{\mathcal{Z}(H,\hat{H},\tau)\right\} \quad (10)$$

Proposition 2. (Optimal cutting). Let $0\in\tau^*$. Then, the optimal decision $^{\lfloor o}\hat{H}$ about the inspected hypotheses H_0, H_1 and the optimal cutting $^{\lfloor o}\tau$, that minimize the expected loss in (10), are given by the following rule.

$$\begin{aligned}&\text{Compute } {}^{\lfloor 0}\tau\in\mathrm{Arg}\max_{\tau\in\tau^*}f(H_0|d(\mathring{t}),\tau)\\ &\qquad {}^{\lfloor 1}\tau\in\mathrm{Arg}\min_{\tau\in\tau^*}f(H_1|d(\mathring{t}),\tau) \qquad (11)\\ &\text{Select } {}^{\lfloor o}\hat{H}=H_0,\ {}^{\lfloor o}\tau={}^{\lfloor 0}\tau \text{ if}\\ &\qquad f(H_0|d(\mathring{t}),\ {}^{\lfloor 0}\tau)\geq f(H_1|d(\mathring{t}),\ {}^{\lfloor 1}\tau)\\ &\quad {}^{\lfloor o}\hat{H}=H_1,\ {}^{\lfloor o}\tau={}^{\lfloor 1}\tau \text{ if}\\ &\qquad f(H_0|d(\mathring{t}),\ {}^{\lfloor 0}\tau)<f(H_1|d(\mathring{t}),\ {}^{\lfloor 1}\tau).\end{aligned}$$

Proof: Let us consider the set of cutting moments $\tau_0^*\equiv\left\{\tau\in\tau^*:\ f(H_0|d(\mathring{t}),\tau))\geq 0.5\right\}$. This finite set is non-empty, as for $\tau=0$ $f(H_0|d(\mathring{t}),\tau)=0.5$. For a fixed $\tau\in\tau_0^*$, the decision $\hat{H}=H_0$ minimizes the expected loss and the achieved minimum is expectation over $d(\mathring{t})$ of $1-f(H_0|d(\mathring{t}),\tau)$. This values is minimized by $^{\lfloor 0}\tau$. On the non-empty set, $\tau_1^*\{\tau\in\tau^*:\ f(H_0|d(\mathring{t}),\tau)\leq 0.5\}$, the achieved minimum is expectation over $d(\mathring{t})$ of $1-f(H_1|d(\mathring{t}),\tau)$, which is minimized by $^{\lfloor 1}\tau$. The smaller of these values determines, which of these pairs defines the optimum. □

Practical applications of the above test strongly depend on the set τ^* of the considered cutting moments. The finest possible choice is $\tau^*=t^*$. Exhaustive search is too demanding for extensive data sets. Search for the minimizer by a version of golden-cut rule, by a random choice or by a systematic inspection on a small predefined grid can be applied. The predefined grid seems to be the simplest and still relevant variant as minor changes in τ make little physical sense.

4. APPLICATION TO ESTIMATION

This section applies the obtained result to parameter estimation. This is the special case of filtering with time invariant state $x_t=x_{t-1}\equiv\Theta\in\Theta^*$ $\Leftrightarrow f(x_t|\psi_t,x_{t-1})=\delta(x_t-\Theta)$, which is a formal time-evolution model for time-invariant state. In this case, the time-updating step, Proposition 1, becomes identity and the pdf $f(\Theta|d(t))$, describing parameter estimates, is only evolved via the data updating.

Moreover, models in *dynamic exponential family (EF)* are considered, for which the observation model is traditionally called *parameterized model*. Introducing the *data vector* $\Psi_t\equiv[y_t,\psi_t]$, we can write members $\mathcal{M}$ of the exponential family in the form

$$f(y_t|u_t,d(t-1),\Theta)=A(\Theta)\exp\left\langle B(\Psi_t),C(\Theta)\right\rangle,$$

where $A(\Theta)\geq 0$ and $\langle\cdot,\cdot\rangle$ is a scalar product on the involved array functions $B(\Psi_t)$, $C(\Theta)$ of compatible dimensions.

Estimation of this family, i.e. computation of the posterior pdfs $f(\Theta|d(t))$, $t\in t^*$, reduces to algebraic updating of sufficient statistics

$$V_t=V_{t-1}+B(\Psi_t),\ \nu_t=\nu_{t-1}+1 \qquad (12)$$

that determine the *reproducing form of the posterior pdf*

$$\begin{aligned}f(\Theta|d(t),\mathcal{M})&=\frac{A^{\nu_t}(\Theta)\exp\left\langle V_t,C(\Theta)\right\rangle}{\mathcal{L}(V_t,\nu_t,\mathcal{M})} \qquad (13)\\ \mathcal{L}(V_t,\nu_t,\mathcal{M})&\equiv\int A^{\nu_t}(\Theta)\exp\left\langle V_t,C(\Theta)\right\rangle d\Theta.\end{aligned}$$

The reproduction is achieved when using the *conjugate prior pdf* that has the above form for $t=0$ and whose statistics V_0, ν_0 determine the initial conditions in (12).

With the introduced notations, the posterior probability (9) of the hypothesis H_0 that modelling is successful gets the form $f(H_0|d(\mathring{t}),\tau)=$

$$=\left(1+\frac{\mathcal{L}\left({}^{\lfloor l}V_{\mathring{t}_l},\ {}^{\lfloor l}\nu_{\mathring{t}_l},\mathcal{M}\right)\mathcal{L}\left({}^{\lfloor v}V_{\mathring{t}_v},\ {}^{\lfloor v}\nu_{\mathring{t}_v},\mathcal{M}\right)}{\mathcal{L}\left(V_{\mathring{t}},\nu_{\mathring{t}},\mathcal{M}\right)\mathcal{L}\left(V_0,\nu_0,\mathcal{M}\right)}\right)^{-1}. \quad (14)$$

In (14), $(V_{\mathring{t}},\nu_{\mathring{t}})$, $\left({}^{\lfloor l}V_{\mathring{t}_l},\ {}^{\lfloor l}\nu_{\mathring{t}_l}\right)$ and $\left({}^{\lfloor v}V_{\mathring{t}_v},\ {}^{\lfloor v}\nu_{\mathring{t}_v}\right)$ are pairs of sufficient statistics collected on all, learning and validation data, respectively. Let us stress that the collection always starts with the initial values (V_0,ν_0) determining the prior conjugate pdf.

For presentation simplicity, let us consider the fixed grid as the set of possible cutting moments

$$\tau^*=\{\tau_1=0<\tau_2,\ldots,\tau_{\mathring{\tau}-1}<\tau_{\mathring{\tau}}=\mathring{t}\}.$$

Then, the combination of the formula (14) and Proposition 2 provides the following algorithm.

Algorithm 2. (Estimation with validation in EF).
Initial phase

- Select a model from exponential family and structure of its regression vector.
- Select the prior statistics V_0, ν_0.

Collection of statistics

For $i = 1, \ldots, \mathring{\tau}$

Set $\Delta_i = 0_{\dim(V)},\ \rho_i = 0$

For $t = 1, \ldots, \mathring{t}$

If $t \in (\tau_i, \tau_{i+1}]$

$\Delta_i = \Delta i + B(\Psi_t),\ \rho_i = \rho_i + 1$

end If

end of the cycle over t

$\lfloor^l V_{i;\tau_i} = \lfloor^l V_{i-1;\tau_{i-1}} + \Delta_i$

$\lfloor^l \nu_{i;\tau_i} = \lfloor^l \nu_{i-1;\tau_{i-1}} + \rho_i$

end of the cycle over i

Validation

Set $\lfloor^1\tau = \lfloor^0\tau = 0,\ \lfloor^1 p = \lfloor^0 p = 0.5$

$C = \mathcal{L}(\lfloor^l V_{\mathring{\tau}} + V_0, \lfloor^l \nu_{\mathring{\tau}} + \nu_0, \mathcal{M})\mathcal{L}(V_0, \nu_0, \mathcal{M})$

For $i = \mathring{\tau}, \ldots, 1$

$\lfloor^v V_{i-1;\mathring{\tau}_{i-1}} = \lfloor^v V_{i;\mathring{\tau}_i} + \Delta_i$

$\lfloor^v \nu_{i-1;\mathring{\tau}_{i-1}} = \lfloor^v \nu_{i;\mathring{\tau}_i} + \rho_i$

Evaluate $\lfloor^l \mathcal{L}_i \equiv \mathcal{L}(\lfloor^l V_{i\mathring{\tau}_i} + V_0, \lfloor^l \nu_{i\mathring{\tau}_i} + \nu_0, \mathcal{M})$

$\lfloor^v \mathcal{L}_i \equiv \mathcal{L}(\lfloor^v V_{i\mathring{\tau}_i} + V_0, \lfloor^v \nu_{i\mathring{\tau}_i} + \nu_0, \mathcal{M})$

Evaluate $f(H_0|d(\mathring{t}), \tau_i) = \left(1 + \dfrac{\lfloor^l \mathcal{L}_i \lfloor^v \mathcal{L}_i}{C}\right)^{-1}$

If $f(H_0|d(\mathring{t}), \tau_i) > \lfloor^0 p$

Set $\lfloor^0 p = f(H_0|d(\mathring{t}), \tau_i),\ \lfloor^0\tau = \tau_i$

else If $f(H_0|d(\mathring{t}), \tau_i) < \lfloor^1 p$

$\lfloor^1 p = f(H_0|d(\mathring{t}), \tau_i),\ \lfloor^1\tau = \tau_i$

end If

end else

end of the cycle over i

If $1 - \lfloor^0 p < \lfloor^1 p$

accept the model $\mathcal{M}$ learnt on $d(\mathring{t})$ (!)

else

reject the model $\mathcal{M}$.

5. EXAMPLE

Te method was tested on two simple AR models: (i) 4th order AR model with stationary parameters, and (ii) 2nd order AR model with slowly varying parameters. Data generated by each model (200 samples for each) are displayed in Fig. 5.

Validation procedure (Algorithm 2) was performed on a uniform splitting grid with cutting points afret each 40 samples. Note, from (12), that the hypothesis are tested by comparison of their likelihood. However, on many samples, difference of the hypothesis likelihood may be large, which results in the posterior distribution on one of the hypothesis is almost one or zero, making it unsuitable for presentation. Therefore, we will present the results in logarithmic form:

$$\Delta_\tau = \log \mathcal{L}\left(d(\mathring{t}), H_0|\tau\right) - \log \mathcal{L}\left(d(\mathring{t}), H_1|\tau\right).$$

This formula expresses the difference of the likelihoods in terms of orders. Intuitively, from (9), for

$$\begin{array}{ll} \text{for } \Delta_\tau = 0, & f(H_0|d(\mathring{t}, \tau) = 0.5, \\ \text{for } \Delta_\tau \to \infty, & f(H_0|d(\mathring{t}, \tau) \to 1, \\ \text{for } \Delta_\tau \to -\infty, & f(H_0|d(\mathring{t}, \tau) \to 0. \end{array}$$

Results of the validation procedure are listed in Table 5. In the first line, the model for estimation was chosen from the same class as the one used for simulation. As expected, hypotheses H_0 is confirmed with probability almost equal to one. In the second line, the model for estimation was chosen from a class different from the simulated one. In this case, validation results significantly differ in different cutting points τ. Hypotheses H_0 is still confirmed with high probability. In the last line, non-stationary AR model was etimated using a stationary AR model of the same order. As expected, hypotheses H_1 is confirmed with probability almost equal to one.

Results of the first and the last experiment confirm a common sense expectation. However, results of the second experiment are harder to interpret. Intuitively, one might expect a validation method to reject (incorrect) models. However, keep in mind that the method does not perform an exhaustive search over all models. In fact, the method tests—at different time moments—whether the validation data can bring some information that is not yet accumulated in the model.

This suggests that the presented loss function (3) is not the only alternative and other loss function may be investigated for more reliable validation.

More examples and Monte-Carlo simulations will be provided in the final version of this paper.

6. CONCLUDING REMARKS

We have proposed a method for cross-validation of an estimated dynamic model on a finite data set. The method splits data between the learning and the validation part and uses Bayesian approach to test hypotheses: (i) the learning data sufficiently represent the whole data set within the given class of models, with (ii) the validation data brings new information that is not absorbed by the model.

The results of validation may signifficantly differ for different splitting alternatives. Therefore,

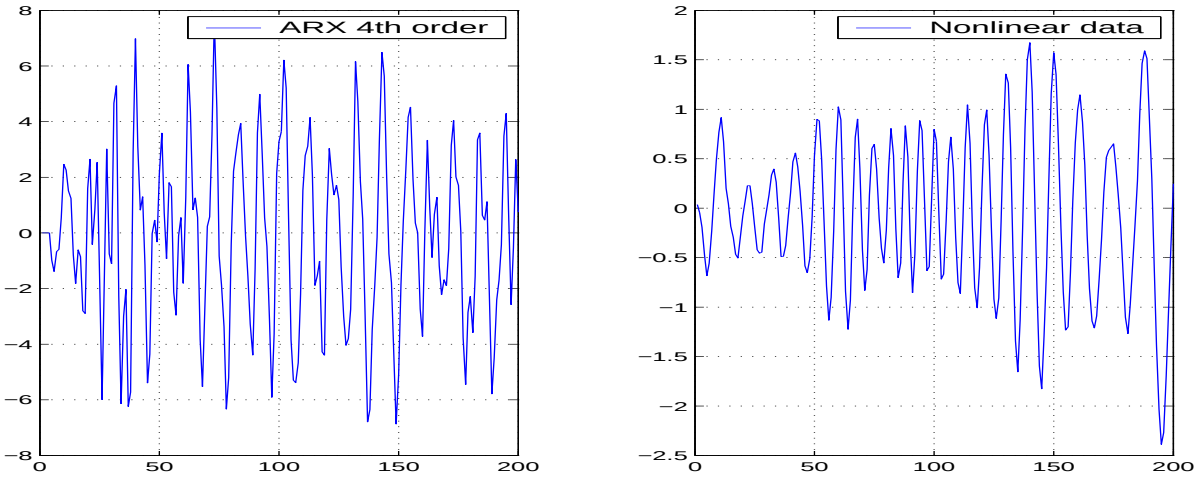

Fig. 2. Data generated by stationary AR(4) model (left), and data generated by non-stationary AR(2) model.

simulated system	estimated model	difference of log-likelihoods, Δ_τ (5) $\tau_1 = 40$	$\tau_2 = 80$	$\tau_3 = 120$	$\tau_4 = 160$	accepted hypotheses
st. AR(4)	AR(4)	43.7	46.6	46.5	51.4	$\lfloor^o\tau = \tau_4$, $\hat{H} = H_0$
st. AR(4)	AR(2)	12.7	13.3	7.4	−3.5	$\lfloor^o\tau = \tau_2$, $\hat{H} = H_0$
non-st. AR(2)	AR(2)	−27.7	−48.5	−40.7	13.6	$\lfloor^o\tau = \tau_2$, $\hat{H} = H_1$

Table 1. Results of the validation algorithm on simulated examples.

the problem was formulated for multiple splitting times and a loss function was introduced for Bayesian selection of the optimal decision.

Application of the method to estimation in the exponential family models yiedls a computationally tractable algorithm that allows—in one sweep—to investigate multiple cutting points.

The presented criteria (i.e. the loss function) was chosen as symmetric for simplicity. Typically, in many practical examples, the loss associated with choise of the wrong model is higher than the loss associated with rejection of the simpler, yet sufficient model. Further research in this direction is needed.

Acknowledgements

This research has been supported by AV ČR S1075351, 1E100750401, GA ČR 102/03/P010.

REFERENCES

O. Barndorff-Nielsen. *Information and exponential families in statistical theory.* Wiley, New York, 1978.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

T. Bohlin. *Interactive System Identification: Prospects and Pitfalls.* Springer-Verlag, Berlin, Heidelberg, New York, 1991.

B.A. Huang. On-line closed-loop model validation and detection of abrupt parameter changes. *JOURNAL OF PROCESS CONTROL*, 11(6): 699–715, 2001.

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

V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

M.E.P. Plutowski. Surwey: Cross-validation in theory and practice. Research report, Department of Computational Science Research, David Sarnoff Research Center, Princeton, New Jersey, USA, 1996.