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 best test of quality of an estimated model is its implementation in real application. However, the use of a bad model is typically too costly. Therefore, model validation is considered as an obligatory step in model learning, and 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 approaches the problem using Bayesian formulation and solution. An algorithm for validation of models estimated within practically important exponential family is presented. Performance of the algorithms is illustrated on simulated example. *Copyright © IFAC 2005*

Keywords: validation, estimation theory, recursive estimation

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 estimated 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 exceptions are available (Huang, 2001), addressing only special cases.

A real need for systematic validation of dynamic models motivated us to develop 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 equality by definition; 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 a pdf argument.

- 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

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

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 vector 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

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

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 $f(x_t|d(t))$, is the *state estimate*, the pdf $f(x_t|u_t, d(t-1))$ determines the *state prediction*, and the pdf $f(y_t|u_t, d(t-1))$, gives the *output prediction*. They evolve as follows

Time updating $f(x_t|u_t, d(t-1)) =$

$$= \int f(x_t|\psi_t, x_{t-1}) f(x_{t-1}|d(t-1))\, dx_{t-1}$$

Data updating $f(x_t|d(t)) =$ (3)

$$= \frac{f(y_t|\psi_t, x_t) f(x_t|u_t, d(t-1))}{f(y_t|u_t, d(t-1))}$$

Output prediction $f(y_t|u_t, d(t-1)) =$

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

Proof: Omitted, see e.g. (Peterka, 1981). □

3. PROBLEM FORMULATION AND SOLUTION

Learning aims to find the *best model* $\lfloor^{o}\mathcal{M} \in \mathcal{M}^*$ of the inspected controlled system. Without explicit specification of the modelling aim, posterior distribution on the whole space $\mathcal{M}^*$ should be built before selecting the relevant model. The set of models $\mathcal{M}^*$ (1) is, however, infinite dimensional and a practical construction of the prior distribution over it, as well as evaluation of its moments, is intractable.

Therefore, the prior is considered 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 $\mathcal{M}$ is equal to the factor of $f(d(\mathring{t})|\mathcal{M})$ that depends on $\mathcal{M}$. Thus, the construction of the likelihood function is implied by Proposition 1

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

Hence, the estimation selects among various models from $\mathcal{M}^*$ the model with the highest v-likelihood (4) (likelihood on model variants).

Model validation is an additional test on the quality of $\lfloor^{o}\mathcal{M}$. Inspired by the classical model validation theory (Plutowski, 1996), all the *available data* are split $d(\mathring{t})$ into (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 essentially inspects 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 choice of the cutting moment, which splits the available data into learning and validation parts. 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 possible choices of learning and validating data. At the same time it disqualifies majority of the available analysis. This motivates us to design an adequate, purely Bayesian, solution of the model validation problem.

3.1 Validation with fixed cutting moment

Let us consider a fixed cutting moment $\tau \in t^* \cup \{0\}$, which defines

$$learning\ data \quad {}^{\lfloor l}d(\tau) \equiv d(\tau) \qquad (5)$$

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

The following hypotheses are considered.

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 results from stochastic filtering on all data giving

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

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

The corresponding v-likelihood results from independent filtering on learning and validation data giving

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

Note that the proportionality factor formed by the randomized controller (1) is common for both hypothesis.

The model ${}^{\lfloor 1}\mathcal{M}$ used on validation data may differ from ${}^{\lfloor o}\mathcal{M}$. The strength of the constructed test depends significantly on the choice of the competing model ${}^{\lfloor 1}\mathcal{M}$. It was chosen as follows: (i) ${}^{\lfloor 1}\mathcal{M}$ has the same structure as ${}^{\lfloor o}\mathcal{M}$, (ii) it is learnt on validation data, (iii) prior pdf in the validation phase is chosen as flattened version of the state estimate gained in the learning phase. Spread of the flattened pdf should be comparable to that of the prior pdf used on the learning data. This will be illustrated in detail in Section 4.

This choice intuitively meets the requirement on a real competitor: learning is exploited without fixing the results too much and thus without restricting possibility to fit the validation data in a better way.

The principle of validation is graphically illustrated in Figure 1. Estimation on the whole data $d(\mathring{t})$ yields result in the class time invariant models. Estimation on the separate data sets yields result in the class of models switched at the cutting moment. The latter class is, of course, richer but it has smaller portion of data per estimated variable at disposal. Thus, the winner is not a priori determined.

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 a good one 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.

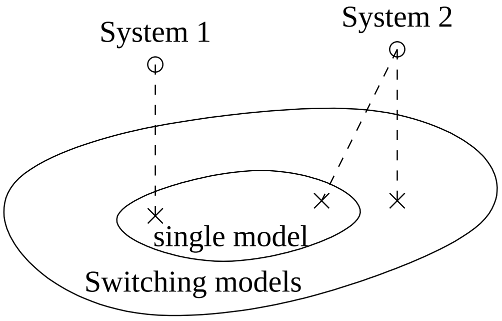

Fig. 1. Scheme of the proposed validation. Elipses denote classes of models, small circles denote alternative "positions" of the real system with respect to the model class. The crosses denote models of the systems estimated within each class. Dashed lines signify distances of the system to the best models. The hypothesis H_0 is expected to win for System 1 and H_1 for System 2.

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

(1) Select a model structure and the prior pdf.
(2) Run filtering, Proposition 1, on the learning ${}^{\lfloor l}d(\tau)$ and full $d(\mathring{t})$ data.
(3) Flatten the filtering result obtained on learning data and use it as the prior pdf for learning on validation data ${}^{\lfloor v}d(\mathring{t} \setminus \tau)$.
(4) Evaluate the v-likelihoods $\mathcal{L}\left({}^{\lfloor l}d(\tau),\ {}^{\lfloor 0}\mathcal{M}|\tau\right)$, $\mathcal{L}\left({}^{\lfloor v}d(\mathring{t} \setminus \tau),\ {}^{\lfloor 1}\mathcal{M}|\tau\right)$ and $\mathcal{L}\left(d(\mathring{t}),\ {}^{\lfloor 0}\mathcal{M}\right)$.
(5) Using the Bayes rule, probability that the learning was successful is

$$f\left(success|d(\mathring{t}),\tau\right) \equiv f\left(H_0|d(\mathring{t}),\tau\right) = \qquad (9)$$
$$= \left(1 + \frac{\mathcal{L}\left({}^{\lfloor l}d(\tau),\ {}^{\lfloor 0}\mathcal{M}|\tau\right)\mathcal{L}\left({}^{\lfloor v}d(\mathring{t} \setminus \tau),\ {}^{\lfloor 1}\mathcal{M}|\tau\right)}{\mathcal{L}\left(d(\mathring{t}),\ {}^{\lfloor 0}\mathcal{M}\right)}\right)^{-1}$$

where likelihoods of both hypotheses are given by (7) and (8) respectively.
(6) 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. □

3.2 Validation with multiple cutting moments

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 thus chosen as

$$\mathcal{Z}(H,\hat{H},\tau) = 1 - \delta\left(\hat{H}(\tau) - H\right),\ \hat{H}, H \in \{H_0, H_1\},$$

where $\delta(\cdot)$ is Kronecker delta for discrete arguments and Dirac delta in continuous case. The optimal decision ${}^{\lfloor o}\hat{H},\ {}^{\lfloor o}\tau$ minimizes expected value $\mathcal{E}[\cdot]$ taken over uncertain data $d(\mathring{t})$ and hypothesis H

$$ {}^{\lfloor o}\hat{H},\ {}^{\lfloor o}\tau \in \mathrm{Arg}\min_{\hat{H},\tau^*} \mathcal{E}\left[\mathcal{Z}(H,\hat{H},\tau)\right]. \qquad (10)$$

Proposition 2. (Optimal cutting). Let $0, \mathring{t} \in \tau^*$. 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.

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

Proof: Let us consider the set of cutting moments ${}^{\lfloor 0}\tau^* \equiv \left\{\tau \in \tau^* : f(H_0|d(\mathring{t}),\tau) \geq 0.5\right\}$. This finite set is non-empty, as for $\tau = 0$ $f(H_0|d(\mathring{t}),\tau) = 0.5$. For a fixed $\tau \in {}^{\lfloor 0}\tau^*$, the decision $\hat{H} = H_0$ leads to a smaller loss than the decision $\hat{H} = H_1$. The achieved minimum is expectation over $d(\mathring{t})$ of $1 - f(H_0|d(\mathring{t}),\tau)$. Thus, it is smallest for ${}^{\lfloor 0}\tau$ maximizing $f(H_0|d(\mathring{t}),\tau)$ on ${}^{\lfloor 0}\tau^*$.

For any fixed τ in the set ${}^{\lfloor 1}\tau^* \equiv \{\tau \in \tau^* : f(H_0|d(\mathring{t}),\tau) \leq 0.5\}$, the decision $\hat{H} = H_1$ leads to a smaller loss than the decision $\hat{H} = H_0$. The achieved minimum is expectation over $d(\mathring{t})$ of $f(H_0|d(\mathring{t}),\tau)$. Thus, it is smallest for ${}^{\lfloor 1}\tau$ minimizing $f(H_0|d(\mathring{t}),\tau)$ on ${}^{\lfloor 1}\tau^*$.

The smaller of the discussed pairs of minima determines the optimal decision pair. □

Practical applications of the above test strongly depend on the set τ^* of the considered cutting moments. The finest possible choice is $\tau^* = t^*$. The 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. Estimation is a 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 the 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 evolved only 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* $\Psi_t \equiv [y_t, \psi_t]$, the members $\mathcal{M}$ of the EF have the form

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

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 the 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*

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

The reproduction is achieved when using the *conjugate prior pdf* that has the form (13) for $t = 0$ and whose statistics $V_0,\ \nu_0$ determine the initial conditions in (12)

- for learning data, for which the recursion runs up to the cutting moment τ and gives the statistics ${}^{\lfloor l}V_\tau,\ {}^{\lfloor l}\nu_\tau$,
- for all data, for which the recursion runs over all data up to $\mathring{t}$ and gives the statistics $V_{\mathring{t}},\ \nu_{\mathring{t}}$.

Flattening of the pdf obtained on the learning data preserves the functional form (13). Its statistics at cutting moment τ that have the same spread as the prior pdf are given by formulas

$$\begin{aligned}
{}^{\lfloor v}V_\tau &= \lambda_\tau\, {}^{\lfloor l}V_\tau,\ {}^{\lfloor v}\nu_\tau = \lambda\, {}^{\lfloor l}\nu_\tau \text{ with} \\
\lambda_\tau &\equiv \frac{\nu_0}{{}^{\lfloor l}\nu_\tau} \leq 1. \qquad (14)
\end{aligned}$$

The statistics ${}^{\lfloor v}V_{\mathring{t}},\ {}^{\lfloor v}\nu_{\mathring{t}}$ on validation data ${}^{\lfloor v}d(\mathring{t}\setminus\tau)$ are obtained via recursion (12) starting from the statistics (14) at the cutting time τ.

We deal with a fixed model structure and respective models differ just by statistics. So that we can drop the argument $\mathcal{M}$ in the v-likelihood (13).

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

$$= \left(1 + \frac{\mathcal{L}\left(\lfloor^{l}V_\tau,\, \lfloor^{l}\nu_\tau\right)\mathcal{L}\left(\lfloor^{v}V_{\mathring{t}},\, \lfloor^{v}\nu_{\mathring{t}}\right)}{\mathcal{L}\left(V_{\mathring{t}}, \nu_{\mathring{t}}\right)\mathcal{L}\left(\lfloor^{v}V_\tau,\, \lfloor^{v}\nu_\tau\right)}\right)^{-1}. \quad (15)$$

Formula (15) and Proposition 2 determine validation algorithm. For presentation simplicity, we shall write it down on the fixed grid of possible cutting moments

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

Note that the relatively complex logic tries to minimize operations connected with computationally expensive accumulation of sufficient statistics $V,\ \lfloor^{l}V,\ \lfloor^{v}V$ for respective cutting moments.

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 $V_0,\ \nu_0$.
- Set $\lfloor^{l}V_0 = V_0,\ \lfloor^{l}\nu_0 = \nu_0$.

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 of If
   end   of the cycle over $t$
          $\lfloor^{l}V_{\tau_i} = \lfloor^{l}V_{\tau_{i-1}} + \Delta_i$
          $\lfloor^{l}\nu_{\tau_i} = \lfloor^{l}\nu_{\tau_{i-1}} + \rho_i$
end   of the cycle over $i$
```

Validation

```
Set $\lfloor^{1}\tau = \lfloor^{0}\tau = 0,\ \lfloor^{1}p = \lfloor^{0}p = 0.5$
Set $\lfloor^{v}V_{\mathring{\tau}} = 0_{\dim(V)},\ \lfloor^{v}\nu_{\mathring{\tau}} = 0$
        $C_{\mathring{\tau}} = \mathcal{L}(\lfloor^{l}V_{\mathring{\tau}},\, \lfloor^{l}\nu_{\mathring{\tau}})$
 For   $i = \mathring{\tau}, \ldots, 2$
        $\lfloor^{v}V_{\mathring{\tau}_{i-1}} = \lfloor^{v}V_{\mathring{\tau}_i} + \Delta_i$
        $\lfloor^{v}\nu_{\mathring{\tau}_{i-1}} = \lfloor^{v}\nu_{\mathring{\tau}_i} + \rho_i$
        $\lambda_{i-1} = \frac{\nu_0}{\lfloor^{l}\nu_{\tau_{i-1}}}$
        $\lfloor^{v}V_0 = \lambda_{i-1}\, \lfloor^{l}V_{\tau_{i-1}}$
        $\lfloor^{v}\nu_0 = \lambda_{i-1}\, \lfloor^{l}\nu_{\tau_{i-1}}$
        $C_{i-1} = \mathcal{L}(\lfloor^{v}V_0,\, \lfloor^{v}\nu_0)$
        $\lfloor^{l}\mathcal{L}_{i-1} \equiv \mathcal{L}(\lfloor^{l}V_{\mathring{\tau}_{i-1}},\, \lfloor^{l}\nu_{\mathring{\tau}_{i-1}})$
        $\lfloor^{v}\mathcal{L}_{i-1} \equiv \mathcal{L}(\lfloor^{v}V_{\mathring{\tau}_{i-1}} + \lfloor^{v}V_0,\, \lfloor^{v}\nu_{\mathring{\tau}_{i-1}} + \lfloor^{v}\nu_0)$
        $f(H_0|d(\mathring{t}), \tau_{i-1}) = \left(1 + \frac{\lfloor^{l}\mathcal{L}_{i-1}\, \lfloor^{v}\mathcal{L}_{i-1}}{C_{\mathring{\tau}}C_{i-1}}\right)^{-1}$
    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 of if
    end of 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. ILLUSTRATIVE EXAMPLE

Performance of the validation procedure described by Algorithm 2 was tested on a simulated autoregressive system of the fourth order generating 300 data, see full line in Figure 2. The validation procedure was applied on a uniform grid with distance of cutting moments equal to 10 samples.

First, Algorithm 2 was run while estimating the model of the correct fourth order. The results confirmed model validity: probabilities of the hypothesis H_0 for respective cutting moments were greater than 0.8. In fact, they practically equaled to one with exception of the initial and final non-trivial cutting moments.

Second, Algorithm 2 was run while estimating the model of the incorrect second order. The results confirmed model invalidity: probabilities of the hypothesis H_0 for several cutting moments fallen to zero. All probabilities H_0 are plotted, together with data, Figure 2 as circles connected by dotted line.

These results indicate that the validation procedure reacts appropriately on those parts of the system behavior, which are insufficiently explained by the model of insufficient order. It also confirms sensitivity of the validation with a fixed cutting moment: for instance, cutting the data at time 200 only leads to acceptance of the invalid model.

6. CONCLUDING REMARKS

A method for cross-validation of an estimated dynamic model on a finite data set was proposed. The method cuts data into the learning and the validation parts 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)

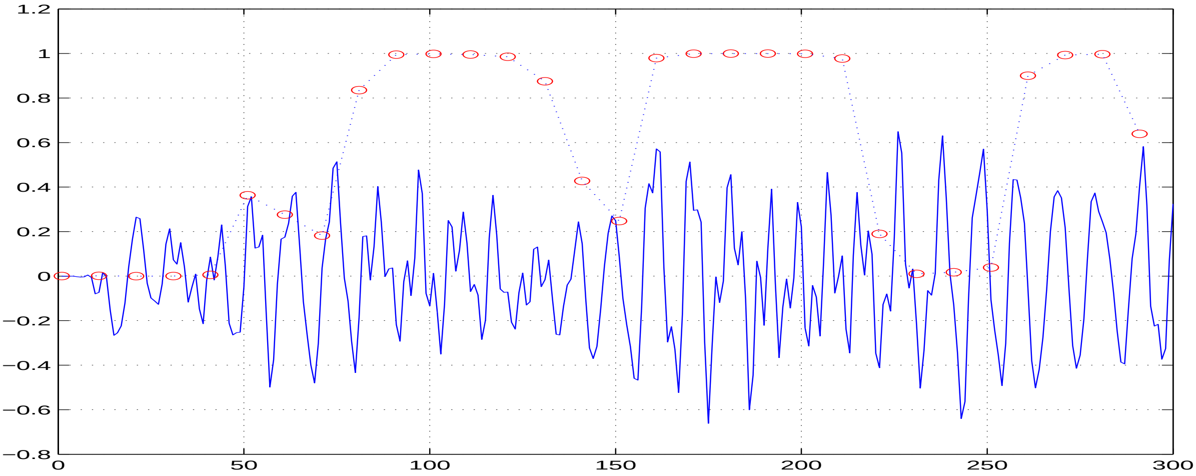

Fig. 2. System measurements (full line) and probability of the hypothesis H_0 (circles and dotted line).

the validation data brings new information that is not absorbed by the model.

The results of validation may significantly differ for different cutting alternatives. Therefore, the problem was formulated for multiple cutting times and both acceptance of the hypothesis on model validity and cutting moment were optimized within a standard Bayesian decision making set up.

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

Experience indicates that the chosen symmetric loss function might be dangerous. Typically, the loss associated with choice of the wrong model is higher than the loss associated with rejection of the simpler, yet sufficient model. Thus, the decision should be modified in this respect to make the decision rule practical. Most importantly, however, the proposed rule has to be elaborated algorithmically for widely used models like mixture models, linear state space models etc.

Acknowledgements

This research has been supported by AV ČR S1075351, 1E100750401.

REFERENCES

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

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

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

Huang, B.A. (2001). On-line closed-loop model validation and detection of abrupt parameter changes. *Journal of Process Control* **11**(6), 699–715.

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

Peterka, V. (1981). Bayesian system identification. In: *Trends and Progress in System Identification* (P. Eykhoff, Ed.). pp. 239–304. Pergamon Press. Oxford.

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