SOME ASPECTS OF MODEL TUNING PROCEDURE: INFORMATION-THEORETIC ANALYSIS

A.I.Yashin

April 1985
WP-85-24

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

ACKNOWLEDGMENTS

I would like to thank Susanne Stock for her help in preparing the manuscript.

SOME ASPECTS OF MODEL TUNING PROCEDURE: INFORMATION-THEORETIC ANALYSIS

A.I.Yashin

1. INTRODUCTION

Computer or mathematical models are not exact representation of reality: lack of knowledge, technical restrictions and particular modeling goals make it necessary to approximate the real system in various ways. Nevertheless, the procedures by which the models are adjusted to observed data are often based on the assumptions that the real system has the same structure as the model and differs only in the values of certain parameters. These particular values usually should be included in the feasible set of the parameter values, and this fact, together with some additional conditions, usually provides the convergence property for many individual algorithms [1].

However, in reality all of these assumptions are generally false. Even if the structure of the system corresponds to the structure of the model, the real parameters values often do not belong to the presupposed feasible set. Moreover, mathematicians often consciously diminish this set in order to simplify the estimation algorithms. For instance they approximate the bounded compact set of parameter values by a set consisting of a finite number of points, thus increasing the chances that the real parameter values will be excluded.

It is therefore both remarkable and surprising to find that despite these false assumptions and approximations, the parameter estimation algorithms often still converge! The model resulting from this tuning procedure will of course not coincide with the real system, and this rises the natural question: how far is this computer model from reality?

When considering this question it is necessary to have some way of measuring the distance between individual models. One of measure of divergence was introduced by Bhattacharya [2] ; Kullback [3] also formulated some measure of information distance. However these measures were not proper metrics. Baram and Sandell [4] later introduced a modified version of Kullback measure, which have been shown to be a proper distance metric. They applied this approach to linear Gaussian systems and models; in this paper it is generalized to a wider class of systems.

2. NOTATIONS AND DEFINITION

Assume that the variety of models of the real system may be characterized by a parameter β, which takes values from the parameter set B. In view of Bayesian formulation of the problem, we will assume β to be a random variable defined on some probabilistic space (Ω, H, P). Let $\xi_n(\omega), n \geq 0$ be some random process (observation) adapted to some nondecreasing family of σ-algebras $\mathbf{H} = (H_n)_{n \geq 0}$, $H_\infty = H$ in Ω. We shall denote by $\overline{\mathbf{H}} = (\overline{H}_n)_{n \geq 0}$, $\overline{H}_\infty = \overline{H}$ the family of σ-algebras generated by the process $\xi_n, n \geq 0$, where

$$\overline{H}_n = \sigma\{\xi_m, m \leq n\}$$

is σ-algebra in Ω generated by the process ξ_t up to time t.

In the case of continuous time observation process ξ_t, $t \geq 0$ we assume the nondecreasing right continuous family of σ-algebras $\mathbf{H} = (H_t)_{t \geq 0}$ to be given, where $H_\infty = H$ and H_0 is completed by P-zero sets from H. We also introduce the family of σ-algebras $\overline{\mathbf{H}} = (\overline{H}_t)_{t \geq 0}$, where

$$\overline{H}_t = \sigma\{\xi_u, u \leq t\}$$

If the set of the parameter values is finite or denumerable we will denote by $\pi_j(n)$, (or $\pi_j(t)$) the *a posteriori* probabilities of events $\{\beta = \beta_j\}$, $j \in B$ given observations ξ_k, $k \le n$, $(\xi_u,\ u \le t\ )$.

For any $A \in \overline{H}$, $x \in B$ we denote by $\mathbf{P}^x(A)$, the family of probability measures

$$\mathbf{P}^x(A) = \mathbf{P}(A \mid \beta = x).$$

Let $\overline{\mathbf{P}}_n^x(A)$, $\overline{\mathbf{P}}^x(A)$, $x \in B$, $n \ge 0$ be the restrictions of the $\mathbf{P}^x(A)$ on σ-algebras $\overline{H}$, $\overline{H}_n$, respectively. Assume also that for any $x, y \in B$ we have $\overline{\mathbf{P}}_n^x \sim \overline{\mathbf{P}}_n^y$. Define $\mathbf{Z}_n^{x,y}$ as a Radon-Nicodim derivative

$$\mathbf{Z}_n^{x,y} = \frac{d\overline{\mathbf{P}}_n^x}{d\overline{\mathbf{P}}_n^y}$$

and let

$$\alpha_n^{x,y} = \mathbf{Z}_n^{x,y}(\mathbf{Z}_{n-1}^{x,y})^{-1}.$$

It is easy to see that if the $\overline{H}_{n-1}$-conditional distributions of ξ_n, $n \ge 0$ have densities $f^x(z \mid \overline{H}_{n-1})$, $x \in B$ then

$$\alpha_n^{x,y} = \frac{f^x(\xi_n \mid \overline{H}_{n-1})}{f_y(\xi_n \mid \overline{H}_{n-1})}$$

3. SOME BAYESIAN PARAMETER ESTIMATION ALGORITHM

Before deriving our main results, we will first consider some Bayesian parameter estimation algorithms for different observation schemes.

a) Assume that $\xi_n, n \ge 0$ is given by the formula

$$\xi_n = \vartheta_n + C_t \varepsilon_{2n},$$

where ϑ_n satisfies the recursive stochastic equation

$$\vartheta_{n+1} = \beta\vartheta_n + b_t \varepsilon_{1n+1}$$

Here $\varepsilon_{1n}, \varepsilon_{2n}$, $n \geq 0$ are the sequences of independent Gaussian random variables with zero mean and variance equal to one, and β is an unknown parameter. Assuming that β takes its values from some finite set $B_k = \{\beta_1, \beta_2, \ldots, \beta_k\}$ the *a posteriori* probabilities are

$$\pi_j(n+1) = \pi_j(n) + \pi_j(n)\left[\frac{\frac{1}{\mathbf{D}_j(n+1)}\, e^{-\frac{(\xi_{n+1}-\beta_j m_n^j)^2}{2\mathbf{D}_j^2(n)}}}{\sum_{i=1}^{k} \frac{\pi_i(n)}{\mathbf{D}_i(n+1)}\, e^{-\frac{(\xi_{n+1}-\beta_i m_n^i)^2}{2\mathbf{D}_i^2(n)}}} - 1\right]$$

where m_n^j are Kalman estimates of ϑ_n given $\{\beta = \beta_j\}$ and $\mathbf{D}_j(n)$ are functions of the conditional variance $\gamma_j(n)$ [5]

$$\mathbf{D}_j(n) = (b^2 + C^2 + \beta_j^2 \gamma_j(n))^{\frac{1}{2}}$$

b) Consider the continuous (in time) observation process ξ_t given by the stochastic differential equation

$$d\xi_t = A(\beta, \xi_t)dt + C_t\, dW_t$$

where W_t, $t \geq 0$ is the H-adapted Wiener process, β is an unknown parameter and C_t is H-adapted positive function. Assuming again that the number of parameter values is finite, we have for $\pi_j(t) = \mathbf{P}(\beta = \beta_j \mid \overline{H}_t)$ [6].

$$d\pi_j(t) = \pi_j(t)\frac{(A(\beta_j, \xi_t) - \bar{A}(\xi_0^t))}{C_t^2}(d\xi_t - \bar{A}(\xi_0^t)dt), \quad \pi_j(0) = p_j, \quad j = 1,2,\ldots k,$$

where

$$\bar{A}(\xi_0^t) = \sum_{i=1}^{\pi_i} (t) A(i, \xi_t).$$

c) Consider an observation made by a continuous-state jumping process with unknown transition intensities $\lambda^{\beta}_{i,j}$. Once again assuming a finite number of values for β we have the following equations for *a posteriori* probability $\pi_j(t)$ [7]

$$\pi_j(t) = \pi_j(0) + \sum_{u \le t} \pi_j(u-) \left[\frac{\lambda^j_{\xi_u - \cdot \xi_u}}{\bar{\lambda}_{\xi_u - \cdot \xi_u}} - 1 \right] - \int_0^t \pi_j(u)(\lambda^j_{\xi_u - \cdot \xi_u} - \bar{\lambda}_{\xi_u - \xi_u}) du$$

where

$$\bar{\lambda}_{\xi u - \cdot \xi u} = \sum_{i=1}^{k} \pi_i(u) \lambda^i_{\xi_u - \cdot \xi_u}$$

The necessary and sufficient conditions of convergence with probability one for *a posteriori* probabilities to respective indicators were given in the papers [1, 8, 9] in terms of absolute continuity and singularity of some special families of probability distributions. Papers demonstrated the applications of the general theory to various particular forms of the random processes.

4. AUXILIARY RESULTS

One of the central places in the proof of the main convergence result in [5, 9, 1] was the relation between *a posteriori* probabilities and likelihood ratio in the case of denumerable or finite number of the parameter values. More exactly the following lemma is true:

Lemma 1. *Let for any* $i = j$ *and* $n \ge o$ *measure* $\bar{P}^j_n$ *is equivalent to the measure* $\bar{P}^j_n$. *Then* $\bar{P}^j$ *-a. s. the next equality is true:*

$$\frac{\pi_i(n)}{\pi_j(n)} = \frac{p_i}{p_j} Z^{i,j}_n \tag{1}$$

The proof of this lemma follows from the definition of the likelihood ratio $Z^{i,j}_n$. The equality (1) yields that

$$\{Z^{i,j}_\infty = 0\} = \{\pi_j(\infty) = 1\} = \{\beta = \beta_j\}$$

According to the papers [1, 8, 9] this property guarantees the following result of convergence: (remind that we still deal with the case when the parameter value corresponding to the real system belongs to the feasible set of the parameter values B).

Theorem 1. *Let for any* $i = j$, $n \geq 0$, $\bar{\mathbf{P}}_n^i \sim \bar{\mathbf{P}}_n^j$. *Then the condition* $\bar{\mathbf{P}}^i \perp \bar{\mathbf{P}}^j$ *is equivalent to the condition*

$$\lim_{n\to\infty} \pi_j(n) = \mathrm{I}(\beta = \beta_j),\ \mathbf{P}\text{–}a.s.$$

The proof of this theorem is based on the property that singularity set for the measures $\bar{\mathbf{P}}^i$ and $\bar{\mathbf{P}}^j$ coincide $\mathbf{P}^j$-a. s. with the set $\{\beta = \beta_j\}$.

If the real parameter value β_k does not belong to the feasible set variables $\pi_i(n)$,$i \in B$ calculated in section 3 are already not the *a posteriori probabilities*, but some functionals of the observable process ξ_n .

Taking them as *a posteriori* probabilities, the observer expects to get the convergence one of $\pi_i(n)$,$i \in B$ (say $\pi_{i_0}(n)$) to 1 and interpret this result as if the real parameter value is equal to i_0 . However this is actually a false conclusion. The questions which arise in this relation are: When does the convergency fact for some of the $\pi_i(n)$, $i \in B$ really take place? What does it mean when $\pi_{i_0}(n)$ tends to 1 for some $i_0 \in B$? In order to answer these questions we need some auxiliary results.

Assume that the real system corresponds to a parameter value k such that $k \in B$. Introduce the function $\mathbf{I}_n^k(x,y) = \mathbf{E}_k \ln \alpha_n^{x,y}$ [4] and define the measure of distance

$$d_n(x,y) = |\mathbf{I}_n^k(x,y)|.$$

Lemma 2. *Function* $d_n(i,j)$ *is pseudo-metric. That is, the following equalities hold:*

$$d_n(x,y) = d_n(y,x)$$

$$d_n(x,x) = 0$$

$$d_n(x,k) + d_n(k,y) \geq d_n(x,y)$$

The proof of this lemma is done in [4].

Lemma 3. *For any* $x,y \in B$, $n \geq 0$ *we have*

$$\mathbf{I}_n^x(x,y) \geq 0.$$

Proof. From the definition of the $\mathbf{I}_n^x(x,y)$

$$\mathbf{I}_n^x(x,y) = \mathbf{E}_x(\ln \alpha_n^{x,y} | \overline{H}_{n-1}) = \mathbf{E}_x(\mathbf{E}_x(\ln \alpha_n^{x,y} | \overline{H}_{n-1})) = \mathbf{E}_x(\mathbf{E}_y(\psi(\alpha_n^{x,y}) | \overline{H}_{n-1})$$

where $\psi(t) = t \ln t$. According to the theorem of the mean, $\psi(\alpha_n^{x,y})$ can be represented as follows:

$$\psi(\alpha_n^{x,y}) = \psi(1) + (\alpha_n^{x,y}-1)\psi'(1) + \frac{1}{2}(\alpha_n^{x,y} - 1)^2\psi''(\delta_n^{x,y}),$$

where $\delta_n^{x,y}$ varies between $\alpha_n^{x,y}$ and 1. It is not difficult to see that

$$\mathbf{E}_y(\psi(\alpha_n^{x,y}) | \overline{H}_{n-1}) = \frac{1}{2}\mathbf{E}_y(\frac{(\alpha_n^{x,y} - 1)^2}{\delta_n^{x,y}} | \overline{H}_{n-1}) \geq 0$$

Lemma 4. *Let* $d_n(k,x) \leq d_n(k,y)$. *Then*

$$\mathbf{I}_n^k(x,y) \geq 0$$

Proof. From the definition of the $\mathbf{I}_n^k(x,y)$, we can write

$$\mathbf{I}_n^k(x,y) = \mathbf{E}_k \ln \alpha_n^{x,y} = \mathbf{E}_k \ln f^x(\xi_n | \overline{H}_{n-1}) - \mathbf{E}_k \ln f^y(\xi_n | \overline{H}_{n-1})$$

$$= -\mathbf{E}_k[\ln f^k(\xi_n | \overline{H}_{n-1}) - \ln f^x(\xi_n | \overline{H}_{n-1})] + E_k[\ln f^k(\xi_n | \overline{H}_{n-1}) - \ln f^y(\xi_n | \overline{H}_{n-1})]$$

$$= -\mathbf{E}_k \ln \alpha_n^{k,x} + \mathbf{E}_k \ln \alpha_n^{k,y}$$

From Lemma 3 for any $z \in B$

$$\mathbf{E}_k \ln \alpha_n^{k,z} \geq 0$$

and thus

$$I_n^k(x,y) \geq 0 \quad \text{if} \quad d_n(k,y) \geq d_n(k,x)$$

5. RESULTS

Assume that the process $\ln \alpha_n^{x,y}$ is ergodic, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \ln \alpha_m^{x,y} = \mathbf{E}_k \ln \alpha^{x,y} = I^k(x,y)$$

Theorem 2. *If* $d(k,x) > d(k,y)$ *then*

$$Z_n^{x,y} \to 0, \quad \mathrm{P}-a.s.$$

If it is known that $Z_n^{x,y} \to 0$ P*-a.s.*, *then*

$$d(x,y) \geq d(x,y)$$

Proof. Note that from Lemma 4, the inequality $d(k,x) > d(k,y)$ yields $I^x(x,y) < 0$ and consequently

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \ln \alpha_m^{x,y} < 0 \quad \mathrm{P}-a.s.$$

This means that

$$\ln Z_\infty^{x,y} = \sum_{m=1}^{\infty} \ln \alpha_m^{x,y} = -\infty \quad \mathrm{P}^k-a.s.$$

and consequently

$$Z_n^{x,y} \to 0 \quad \mathrm{P}^k-a.s.$$

thus proving the first part of the theorem.

In order to prove the second part of the theorem we assume that $Z_n^{x,y} \to 0$ but that $d(k,x) < d(k,y)$. This yields

$$\lim_{n\to\infty} \sum_{m=1}^{n} \ln \alpha_m^{x,y} = I^k(x,y) > 0,$$

from which

$$\ln Z_{\infty}^{x,y} = \sum_{m=1}^{\infty} \ln \alpha_m^{x,y} = \infty$$

and the theorem is proved by contradiction.

Example. Assume that the sequence ξ_n is a finite state ergodic Markov chain on any of the probability spaces $(\Omega,\overline{H},\overline{P}^i)$, $i \in B$,where B is a finite set. Let $p_{l,m}^{\beta}$, $l,m = \overline{1,k}$ be the transition probabilities for one step. It is not difficult to find (see also [8]) that $\alpha_n^{i,j}$ is given by the formula

$$\alpha_n^{i,j} = \frac{p_{\xi_{n-1},\xi_n}}{p_{\xi_{n-1},\xi_n}}$$

Well known results from the Markov chain theory (see [10] for instance) show that the process $\ln \alpha_n^{i,j}$ is ergodic. Thus if the Bayesian algorithm for $\pi_j(n)$ converges to 1 for some particular j_0 it means that this j_0 is the point from B that is the nearest (in the sense of information distance $d(k,x)$) to the real parameter value k .

REFERENCES

1. A.I. Yashin, *Bayesian Approach To Parameter Estimation: Convergence Analysis*, *WP-83-67*, International Institute For Applied Systems Analysis , Laxenburg, Austria (July 1983).
2. A. Bhattacharya, "On Measure Of Divergence Between Two Statistical Populations Defined By Probability Distributions," *Bulletin Calcutta Mathematical Society* **35**, pp.99-104 (1943).
3. S. Kullback, *Information Theory And Statistics*, Wiley, New York (1959).
4. Y. Baram and N.R. Sandell , "An Information Theoretic Approach To Dynamical System Modeling And Identification," *IEEE Transactions Automatic Control* **AC-23**(1), pp.61-66 (1978).

5. N.M. Kuznetsov, A.V. Lubkov, and A.I. Yashin, "About Consistency Of Bayesian Estimates In Adaptive Kalman Filtration Scheme ," *Automatic and Remote Control (translated from Russian)*(4), pp.47-56 (1981).

6. R.S. Liptzer and A.N. Shiryaev, *Statistics of Random Processes*, Springer-Verlag, Berlin and New York (1978).

7. A.I. Yashin, "Filtering of Jumping Processes," *Automatic and Remote Control* 5, pp.52-58 (1970).

8. A.I. Yashin, "Sostoyatelnost Bayesovskich Otcenok Parametrov (Consistency of Bayesian Parameter Estimates)," *Problemi Peredachi Informacii (in Russian)*(1), pp.62-72 (1981).

9. N.M. Kuznetsov and A.I. Yashin, "On the Conditions of the Identifiability of Partially Observed Systems," *Docladi Akademii Nauk SSSR (in Russian)* **259**(4), pp.790-793 (1981).

10. S. Karlin, *A First Course In Stochastic Processes*, Academic Press, New York and London (1968).