INFORMATICA, 1992, Vol.3, No.1, 3–20

CHANGE-POINT DETECTION AS MODEL SELECTION

Jimmy BAIKOVICIUS and László GERENCSÉR

Dept. of Electrical Engineering, McGill University, 3480 University St.
Montréal, P.Q. H3A 2A7, Canada

Dept. of Electrical Engineering, McGill University, 3480 University St.
Montréal, P.Q. H3A 2A7, Canada
and
Computer and Automation Institute of the Hungarian Academy of Sciences
Budapest, H 1502 Pf 63, Hungary

Abstract. We present a new method for solving the change-point detection problem for ARMA systems which are assumed to have a slow and non-decaying drift after the change occurs. The proposed technique is inspired by the stochastic complexity theory, which gives a basis of comparison of different models with different change-point times. Some partial results on the analysis of the estimator are stated. A simulation is included which shows that the approach exhibits surprisingly good detection capabilities.

Key words: stochastic systems, stochastic complexity, time varying systems, recursive estimation, time-varying Ljung's scheme, L-mixing processes, failure detection.

1. Introduction. Any process which presents quantitative and/or qualitative changes along its evolution and in which we would like to determine the location in time of these changes, can be said to fall into the category of change-point detection problems. Change-point detection finds its use in a variety scientific activities ranging from mathematics to economics, and certainly including engineering.

Traditionally, change-point detection problems emerged as part of the need to anticipate "catastrophic" failures of a physical system. Typical examples are in sensors and actuators, fatigue

of structures, failures in nuclear plants, earthquakes, etc. These earlier applications have evolved somehow to also include change-point detection as one of the basic feature of the modelling process of a physical system. Good illustrations are speech and ECG signals. More recently, change-point detection has entered the area of adaptive control. By making the control law also a function of the change-point estimator, the adaptation algorithm can be applied to more general systems, those which present abrupt or slowly time variant changes.

In the last two decades numerous methods have been available for tackling change-point problems. For a recent survey see Basseville (1988). Other useful references are the books of Basseville and Benveniste (1986), and Telksnys (1986). To the best of our knowledge, no change-point method for dynamical systems is given in terms of on-line, recursive, "easily" computable, and theoretically tractable algorithms. Moreover, very little has been done in the case of a change-point problem leaving the realm of time invariant systems. However, these are precisely the characteristics that any method should have if it intend to successfully solve the above applications, especially those in the area of adaptive control.

In this paper we present a general method for solving the change-point detection problem which could be applied to time variant dynamical systems. Moreover, we show that the approach is amenable for theoretical analysis and we give some partial results. We restrict our attention to the ARMA case, but the method is not confined to this particular situation.

The basis of our change-point detection method lies in the use of the stochastic complexity theory. The main idea of this theory is that we estimate the shortest code length for the entire observation sequence relative to a model class. For some early developments we refer to Rissanen (1978), Schwarz (1978) and Shibata (1980). A recent survey of the history of the evolution of the history of the evolution of the relevant ideas is given in Rissanen (1987, 1989) and Gerencsér (1991b). A major innovation in the theory of stochastic complexity has been the introduction of the concept of predictive

stochastic complexity in Rissanen (1986). The predictive stochastic complexity is defined in terms of predictive encoding, which can also be considered as universal coding procedure. It is a particularly useful concept in the context of dynamical systems since it is inherently "on-line" unlike other criteria such as AIC or BIC, (c.f. Akaike, 1970 and 1974) which are inherently off-line.

What the stochastic complexity theory actually allows us, to reduce the change-point detection problem to a model selection problem. By using the predictive stochastic complexity we are able to arrive at an overall fairly simple on-line recursive algorithm for the computation of the alarm signal. We shall illustrate all of the important aspects of our change-point detection method by means of simulations.

2. The mathematical model. We shall first describe the dynamics before the change. Let τ^* denote the location of the change-point, and (y_n) with $0 \leqslant n < \tau^*$, be the output of an ARMA system generated by

$$A^* y = C^* e, \tag{1}$$

where (e_n) is the input process. The values for e_n and y_n for $n \leqslant 0$ are assumed to be 0.

CONDITION 1. A^*, C^* are polynomials of the shift operator z^{-1} with coefficients a_i^*, c_j^* and $\deg A^* = p$ $\deg C^* = q$. We assume that the constant terms are 1 in both A^* and C^* and that A^* and C^* are stable and relative prime.

To describe the noise process let us assume that we are given a probability space $(\Omega, \mathcal{F}, P)$ and a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+)$, $n \geqslant 0$ such that $\mathcal{F}_n \subset \mathcal{F}$ is increasing and $\mathcal{F}_n^+ \subset \mathcal{F}$ is decreasing, and $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent for all n.

CONDITION 2. The driving noise process $e = (e_n)$ is a second order stationary martingale difference process with respect to $\mathcal{F}_n$, $\mathrm{E}(e_n^2|\mathcal{F}_{n-1}) = \sigma^2 = \text{const}$ a.s., and moreover e is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$. (c.f. Appendix). The process (e_n) bounded, say $|e_n| < b$ a.s. for all n. (The last part of the condition is needed in the analysis of time varying systems).

We shall now describe the dynamics after the change. Let (y_n) with $\tau^* \leqslant n \leqslant N$ be the output of an ARMA system generated by

$$(A_n^* y)_n = (C_n^* e)_n, \tag{2}$$

The notation in equation (2) is read as follows: the difference operator A_n^* acts on the process (y) and the evaluation is done at time n to get y_n. The right hand side is interpreted similarly.

CONDITION 3. A_n^* C_n^* are polynomials of the shift operator z^{-1} with coefficients $a_{i,n}^*$, $c_{j,n}^*$ and deg $A_n^* = p$, deg $C_n^* = q$. We assume that the constant terms are 1 in both A_n^* and C_n^* and that A_n^* and C_n^* are stable and relative prime.

Let θ_n^* denote the $k = p+q$-dimensional vector composed of the coefficients of A_n^* and C_n^*, that is let $\theta_n^* = (a_{n,1}^*, \ldots, a_{n,p}^*, c_{n,1}^*, \ldots, c_{n,q}^*)^T$.

CONDITION 4. We have $\sup_{n \geqslant \tau^*} |\tau_{n+1}^* - \tau_n^*| = \dot{S} < \infty$, where $\dot{S}$ is an upper bound for the rate of change of the time varying ARMA(p,q) system, and $|\cdot|$ denotes the Euclidean norm.

We denote the model class described so for by $\mathcal{M}_{\tau^*}$.

3. The encoding procedure. To compute the stochastic complexity for the model class $\mathcal{M}_{\tau^*}$ we use a predictive encoding procedure. For this purpose, we shall make use of the prediction error algorithms. (For surveys see Ljung and Söderström (1984), Caines (1988) and Söderström and Stoica (1989)). We first state the off-line procedure.

Let $D \subset \mathbf{R}^k$ be a compact domain which contains θ^* in its interior and such that for any $\theta \in D$ the corresponding polynomial A and C are stable. Then for any $\theta \in D$ compute the estimated noise process $\bar{\varepsilon}(\theta, \theta^*)$ by

$$\bar{\varepsilon}(\theta, \theta^*) = (A/C)y = (A/C)(C^*/A^*)e, \tag{3}$$

where A and C correspond to θ, and the initial conditions, i.e., the values $\bar{\varepsilon}_n$ and y_n for $n \leqslant 0$, are assumed to be 0.

The time invariant off-line prediction error estimator of θ^*, denoted by $\hat{\theta}_N^0$, is obtained as the solution of the equation

$$\frac{\partial}{\partial \theta} V_N^0(\theta, \theta^*) = V_{\theta N}^0(\theta, \theta^*) = 0, \tag{4}$$

where

$$V_N^0(\theta, \theta^*) = \frac{1}{2} \sum_{n=1}^{N} \bar{\epsilon}_n^2(\theta, \theta^*),$$

and where differentiation is taken both in the almost sure and in the M-sence. For the definition of the latter to the Appendix. More precisely, $\widehat{\theta}_N^0$ are random variables such that $\widehat{\theta}_N^0 \in D$ for all ω and if the equation has a unique solution in D then it coincides with $\widehat{\theta}_N^0$. Such a random variable exists by the measurable selection theorem.

For the time variant off-line case, we use the prediction error algorithms with forgetting given in Gerencsér (1989a), which "weighs down" past data with geometric rate. In this case the cost-function associated with this estimation method is given by

$$V_N^\lambda(\theta, \theta^*) = \sum_{n=1}^{N} (1-\lambda)^{N-m} \lambda \bar{\epsilon}_n^2(\theta, \theta^*),$$

where $0 < \lambda < 1$ is the forgetting factor. Then the so-called small gain prediction error estimator $\widehat{\theta}_N^\lambda$ of θ_N^* is given as the solution of

$$\frac{\partial}{\partial \theta} V_N^\lambda(\theta, \theta^*) = V_{\theta N}^\lambda(\theta, \theta^*) = 0. \tag{5}$$

More precisely, if a unique solution of (5) exists in D, then $\widehat{\theta}_N^\lambda$ is the $\mathcal{D}$-valued random variable representing such solution. Unfortunately, the probability of the "exceptional sets" of Ω, for which (5) has no solution, does not tend to 0 as $N \to \infty$. But this difficulty can be easily handled (Gerencsér, 1991d).

Let us now describe the encoding procedure. Let $\tilde{y}_N(\theta, \theta^*)$ denote the one-step ahead prediction of y_N assuming that the system parameter vector is θ. Then it is easily seen that the prediction error is $\bar{\epsilon}_N(\theta, \theta^*)$. This prediction is "honest" in the terminology of Rissanen (1986), i.e., to predict y_N we only use data preceding the moment N. Now, let τ, with $0 < \tau \leqslant N$, represent the possible location of the change-point. Then in order to get "good" prediction it is clear that we should use the estimates $\widehat{\theta}_n^0$, for $0 \leqslant n < \tau$, and $\widehat{\theta}_n^\lambda$, for $\tau \leqslant n \leqslant N$. Therefore we associate to the observation y_n either

the codelength $\bar{\varepsilon}_n^2(\hat{\theta}_{n-1}^0, \theta^*)$ or $\bar{\varepsilon}_n^2(\hat{\theta}_{n-1}^\lambda, \theta^*)$ according to the value of τ chosen.

Let us define the cumulative prediction error

$$\bar{S}_N(\tau) = \sum_{n=1}^{\tau-1} \bar{\varepsilon}_n^2(\hat{\theta}_{n-1}^0, \theta^*) + \sum_{n=\tau}^{N} \bar{\varepsilon}_n^2(\bar{\varepsilon}_{n-1}^\lambda, \theta), \tag{6}$$

which in the Gaussian case (i.e., when ϵ is a Gaussian white noise) is identical to what Rissanen defined as the predictive stochastic complexity. $\bar{S}_N(\tau)$ is the associated total codelength for each model class $\mathcal{M}_\tau$.

A weak point in the present expression for $\bar{S}_N(\tau)$ is that it is still hard to compute. What we really need is a recursive computable criterion. This can be obtaibned by using recursive prediction error methods.

White the recursive estimation of time invariant systems has attracted much attention (c.f. Ljung and Söderström and Stoica, 1989), the recursive estimation of time varying systems has been almost completely neglected. However, a simple method for getting recursive estimators for the parameters of a time variant system has been known for some time. While simulation results show reasonable performance (c.f., e.g., Ljung and Söderström, 1984), the lack of theoretical analysis has apparently discouraged many practitioners in the field from its use. However, this drawback has been eliminated since many of the important theoretical aspects of the problem have been recently solved. An off-line estimation method has been developed and analyzed in Gerencsér (1989a), and a general time varying Ljung's scheme was presented and analyzed in Gerencsér (1988b).

The recursive prediction error algorithm is summarized as follows. Let us assume that an initial guess $\hat{\hat{\theta}}_0^\lambda \in D$ is known. Assuming that the processes $(\hat{\hat{\theta}}_n^\lambda)$ and (ε_n^λ) have been generated for $n \leqslant N-1$ we define ε_N^λ by the equation

$$(\hat{\hat{C}}_{N-1}^\lambda \varepsilon^\lambda)_N = (\hat{\hat{A}}_{N-1}^\lambda y)_N. \tag{7}$$

We set $y_n = \varepsilon_n^\lambda = 0$ for $n \leqslant 0$, and $\hat{\hat{A}}_{N-1}^\lambda$, $\hat{\hat{C}}_{N-1}^\lambda$ denote the polynomials corresponding to $\hat{\hat{\theta}}_{N-1}^\lambda$. Similarly we define $(\partial/\partial\theta)\,\varepsilon_N^\lambda$ by

$$(\hat{\hat{C}}_{N-1}^\lambda \frac{\partial}{\partial\theta}\varepsilon^\lambda)_N = -\phi_{N-1}, \tag{8}$$

where

$$\phi_{N-1} = (-y_{N-1}, \cdots - y_{N-p}, \varepsilon_{N-1}^\lambda, \cdots \varepsilon_{N-q}^\lambda)^T.$$

Let us define

$$W(\theta, \theta^*) = \lim_{n\to\infty} \frac{1}{2}\,\mathrm{E}\,(\bar{\varepsilon}_n^\lambda(\theta, \theta^*))^2,$$

which is called the asymptotic cost function. Now, let $\hat{\hat{R}}_{N-1}^\lambda$ be an estimation of the Hessian $(\partial^2/\partial\theta^2)W(\theta, \theta^*)$ with initial guess, say $\hat{\hat{R}}_0^\lambda$. Then $\hat{\hat{\theta}}_N^\lambda$, $\hat{\hat{R}}_N^\lambda$ are computed by the following recursion

$$\hat{\hat{\theta}}_N^\lambda = \hat{\hat{\theta}}_{N-1}^\lambda - \Big(\frac{1}{N} + \lambda\Big)\Big(\hat{\hat{R}}_{N-1}^\lambda\Big)^{-1} \frac{\partial}{\partial\theta}\varepsilon_N^\lambda \cdot \varepsilon_N^\lambda \cdot \tag{9}$$

$$\hat{\hat{R}}_N^\lambda = \hat{\hat{R}}_{N-1}^\lambda + \Big(\frac{1}{N} + \lambda\Big)\Big(\Big(\frac{\partial}{\partial\theta}\varepsilon_N^\lambda\Big)\Big(\Big(\frac{\partial}{\partial\theta}\varepsilon_N^\lambda\Big)^T - \hat{\hat{R}}_{N-1}^\lambda\Big)\Big). \tag{10}$$

These tentative values will have to be adjusted if they violate a boundedness condition, which is beyond the scope of the present paper to be presented here. We refer the reader to Gerencsér (1988a) for a complete and rigorous description of the present algorithm. In (9) and (10) the choise of $1/N + \lambda$ for the forgetting factor is chosen so as to reduce the uncertainty due to initial conditions at the start of the recursive algorithm, and to track the time varying parameters afterwards.

For $\lambda = 0$, denote (ε_n^0) the resulting prediction error precess, which corresponds to the time invariant recursive prediction error algorithm. Then, we express the total codelength for each model class $\mathcal{M}_\tau$ as

$$S_N(\tau) = \sum_{n=1}^{\tau-1} (\varepsilon_n^0)^2 + \sum_{n=\tau}^{N} (\varepsilon_n^\lambda)^2, \tag{11}$$

which is not computationally intensive as opposed to $\bar{S}_N(\tau)$. Note that the prediction error sequences do not depend on τ since the two

recursive algorithms are run in parallel for the whole time interval $[1, N]$.

4. Change-point detection as model selection. Let us assume we are given N observations of the process (y_n), and that a change-point occurs at some unknown time $\tau^* \leqslant N$. For each possible $1 \leqslant \tau \leqslant N$ we have a model class $\mathcal{M}_\tau$ with the help of which the sequence (y_n) will be encoded. Each model class $\mathcal{M}_\tau$ has an associated stochastic complexity $S_N(\tau)$, which serves as a basis for comparision between different model classes, i.e., different τ's.

Let us define

$$m_N = \min_{1 \leqslant \tau \leqslant N} S_N(\tau). \tag{12}$$

Then, the estimator of τ^* is defined by

$$\widehat{\tau} = \{\tau; S_N(\tau) = m_N\}. \tag{13}$$

Now let N represent the present time. The increments of $S_N(\tau)$ with respect to τ will be denoted by

$$u_\tau = S_N(\tau + 1) - S_N(\tau). \tag{14}$$

It is straightforward to see that

$$u_\tau = (\varepsilon_\tau^\lambda)^2 - (\varepsilon_\tau^\lambda)^2,$$

and hence the increments of $S_N(\tau)$ are independent of N. Let us now rewrite $S_N(\tau)$ as

$$S_N(\tau) = S_N(1) + \sum_{k=1}^{\tau-1} S_N(k+1) - S_N(k) = S_N(1) + \sum_{k=1}^{\tau-1} u_k,$$

and the minimization of $S_N(\tau)$ with respect to τ is equivalent to the minimization of

$$S'_N(\tau) = \sum_{k=1}^{\tau-1} u_k. \tag{15}$$

With this observation, a formal correspondence between cumulative sum method and our stochastic complexity based method is

established. For the on-line detection of the minimum of $S'_N(\tau)$ we use the so-called Page-Hinkley test (c.f. Hinkley, 1971). Let

$$m'_N = \min_{1 \leqslant \tau \leqslant N} S'_N(\tau), \tag{16}$$

and consider the stopping time, or alarm time,

$$T = \min\{N > 0;\ S'_N(N) - m'_N > h > 0\}, \tag{17}$$

where h is some constant level. Then the on-line change-point detection estimate is defined as

$$\hat{\hat{\tau}} = \max\{\tau;\ S'_T(\tau) = m'_T\}. \tag{18}$$

Clearly we have that $\hat{\tau} = \hat{\hat{\tau}}$. Note that at present time N, only two prediction errors have to be computed in order to know whether or not we have an alarm. This is an improvement with respect to the recent available method applicable to unknown dynamics, the two model method (Basseville, 1988), where the sequential estimation of the dynamics depends on an assumed value of the change-point.

5. Analysis of the change-point detection method. While a complete analysis of our change-point detection method still needs further research, we nevertheless have some very encouraging results. As a first step for the analysis we need to replace the recursive prediction error process (ε_n^λ) with its off-line version $(\bar{\varepsilon}_n(\hat{\theta}_{n-1}^\lambda, \theta^*))$.

We begin by considering the problem of false alarms. Define the increments of $\overline{S}_N(\tau)$ with respect to τ by

$$\bar{u}_\tau = \overline{S}_N(\tau+1) - \overline{S}_N(\tau) \tag{19}$$

then clearly $\bar{u}_\tau = \bar{\varepsilon}_\tau^2(\hat{\theta}_{\tau-1}^0, \theta^*) - \bar{\varepsilon}_\tau^2(\hat{\theta}_{\tau-1}^0, \theta^*)$, and hence the increments of $\overline{S}_N(\tau)$ are independent of N. Similarly as with the one-line case, we get that the minimization of $\overline{S}_N(\tau)$ is equivalent to minimizing

$$\overline{S}'_N(\tau) = \sum_{k=1}^{\tau} \bar{u}_k. \tag{20}$$

For the analysis of $\overline{S}_N(\tau)$ we have the following theorem (Gerencsér, 1991d).

Theorem 1. Under Conditions 1 and 2 and for any $\lambda > 0$ the process $\bar{u}_\tau$ is an L-mixing process, and moreover

$$\mathrm{E}\,\bar{u}_\tau = -\lambda\frac{p+q}{2} + O(\lambda^{3/2}) + O(c^\tau), \tag{21}$$

for all τ such that $\tau \leqslant \tau^*$, and with some $0 < c < 1$.

Hence if we neglect the effect of "nonstationary initial conditions" and λ is small enough, we have

$$\mathrm{E}\,(\bar{u}_\tau) < -a < 0 \qquad \text{for} \quad \tau < \tau^*. \tag{22}$$

Let us rewrite the stopping time T, for the off-line case, in a form more amenable for analysis as

$$\overline{T} = \min\left\{N > 0;\ \max_{1\leqslant k\leqslant N}\sum_{k=m}^{N}\bar{u}_k \geqslant h > 0\right\}. \tag{23}$$

Now, letting $\overline{U}_k = \bar{u}_k - E(\bar{u}_k)$, we can certainly write

$$\max_{1\leqslant m\leqslant N}\sum_{k=m}^{N}\bar{u}_k \leqslant \max_{1\leqslant m\leqslant N}\sum_{k=m}^{N}\overline{U}_k - a \triangleq \overline{U}_N^*(a).$$

The moments of the process $\overline{U}_N^*(a)$ can be estimated and it can even be shown that certain exponential moments exists under some additional conditions (c.f. Gerencsér, 1990a).

Theorem 2. Assume that $\overline{U}_n$ is a zero-mean L-mixing process such that $M_\infty(\overline{U}) < \infty$ and $\Gamma_\infty(\overline{U}) < \infty$. Set $\mu = a^2/8M_\infty(\overline{U})\Gamma_\infty(\overline{U})$, then the process $\overline{U}_N^*(a)$ is L-mixing process (in a certain restricted since), and we have for $\beta \leqslant \mu/a$

$$M_q(\exp\beta\cdot\overline{U}_N^*(a)) < 1 + \frac{1}{\mu^2}\varepsilon^{-\mu}\beta a. \tag{24}$$

The frequency of false alarms is majorated by the frequency of the event $\{\overline{U}_N^*(a) > h\}$, say

$$F_1 = \frac{1}{N}\sum_{n=1}^{N} I_{x>h}(\overline{U}_N^*(a)), \tag{25}$$

where $I_{\{x>h\}}$ is the indicator function of the set $\{x > h\}$. Since $\overline{U^*_N}(a)$ is an L-mixing process in a restricted since, we have by the law of large numbers

$$\overline{\lim} \frac{1}{N} \sum_{n=1}^{N} 1_{x>h}(\overline{U^*_N}(a)) \leqslant M_1\Big(I_{x>h}(\overline{U^*}(a))\Big). \tag{26}$$

Now, applying Markov's inequality we get

$$F_1 \leqslant \mathrm{E}\left(\exp\left(\beta \overline{U^*_N}(a)\right)/\exp\left(\beta h\right)\right). \tag{27}$$

The present form of the analysis so far is not very practical since the process $\bar{\varepsilon}^2_\tau(\hat{\theta}^\lambda_{\tau-1}, \theta^*)$ is not computable in real time. A similar deficiency was overcome in Gerencsér (1988a) by using a strong approximation result which relates off-line and on-line estimators. It is conjectured that a similar result holds for fixed gain estimators. For the time being, however, we must be satisfied with the above results.

The next aspect to be analyzed is the performance of our change-point detection method as measured by the so called detection delay. That is, the time elapsed between the change-point and the alarm time. More precisely, we would like to analyze the probability

$$F_2 = P\{T - \tau^* > \delta t > 0\}. \tag{28}$$

For this matter, we need to understand the nature of the stochastic process $T - \tau^*$, or equivalently the nature of the process $\sum_{k=\tau^*}^{T} u_k$. Using the results of Gerencsér (1991c) we have the following theorem.

Theorem 3. Under suitable conditions we have for sufficiently large N

$$\left|(\varepsilon^\lambda_N)^2 - \varepsilon^2_N\right| \leqslant \delta_{1N} + \delta_{2N} + o(1), \tag{29}$$

where (δ_{1N}) is an L-mixing process such that $\delta_{1N} = O_M(\lambda^{1/2})$, and (δ_{2N}) is a deterministic process such that $\delta_{1N} = O_M(\dot{S}/\lambda)$.

Choosing $\lambda = \dot{S}^{2/3}$ we get

$$\mathrm{E}\left|(\varepsilon^\lambda_N)^2 - \varepsilon^2_N\right| \leqslant O(\dot{S}^{1/3}) + o(1). \tag{30}$$

With this choise of λ the order of magnitude of the upper bound of the tracking error in (29) is minimized. Let us denote this choise of λ by λ_{opt}.

What we actually need is a lower bound for the tracking error $(\varepsilon_N^2)^2 - \varepsilon_N^2$, in terms of $\dot{S}$. Unfortunately this seems to be a quite difficult problem. Let us illustrate this point by the following example. Suppose we have a continuous-time process modelled by a parametric model whose true parameter is given by

$$\theta_t^* = \theta^* + \epsilon \sin(\omega t), \tag{31}$$

where $\epsilon\omega$ is large but ε is small. Then $\dot{S} = \varepsilon\omega$ is large, but there is no reason to expect that the time variant estimation method will track θ_t^*. This example indicates that $\dot{S}$ is not the best measure of the rate of change. However, we are unaware of any appropriate substitute.

6. The simulation. What follows is a simulation to illustrate all of the major aspects of the stochastic complexity approach for solving the change-point detection problem. The data used to check our algorithm is generated by a computer program that simulates a time invariant ARMA(2,1) system until a chosen change-point, and by a slowly time varying ARMA(2,1) system after and including the change-point. Both ARMA-systems are driven by a Gaussian white noise input process with mean 0 and variance 1. Only a computer realization of this process, plus the order of the ARMA systems, are assumed to be given in order to detect the time of change of the data process (y_n).

The simulation is run for a time span corresponding to 4500 observations of the process (y_n). The change-point is chosen to be $\tau^* = 4000$. In Fig. 1 the realization of the process (y_n), used for the simulation, is shown. Note that the change in the dynamics of the data process (y_n) is hardly noticeable.

The time invariant ARMA(2,1) system, generating the data process (y_n) for $n = 1, \ldots, 3999$ is given by the following equation

$$y(N) + a_1^* y(N-1) + a_2^* y(N-2) = \varepsilon(N) + c_1^* \epsilon(N-1) \tag{32}$$

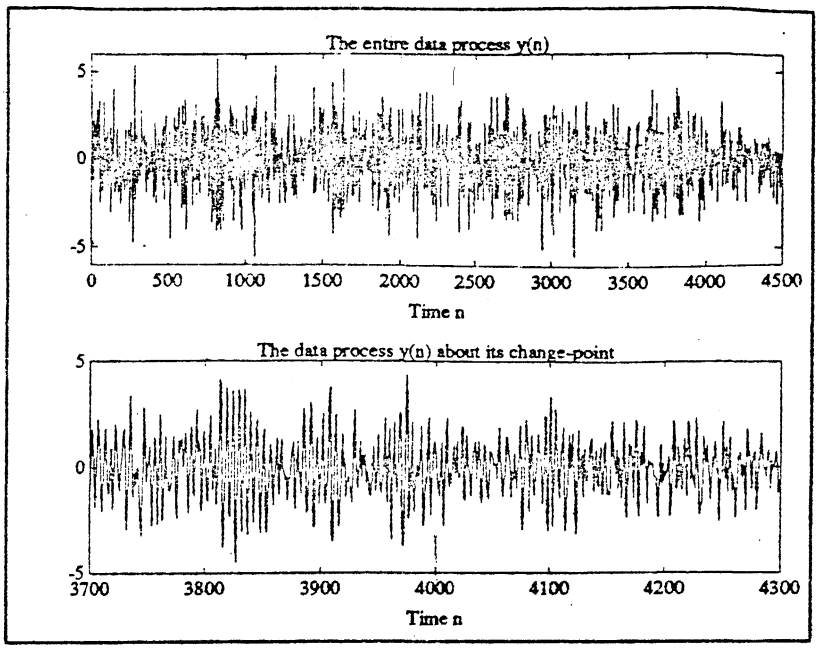

Fig. 1. The data process (y_n) with $\tau^* = 4000$.

with $a_1^* = -.7$, $a_2^* = .8$, $c_1^* = -.2$.

The slowly time varying ARMA(2,1) system, which generates the data process for (y_n) $n = 4000, \ldots, 4500$ is given by

$$y(N) + a_{N,1}^* y(N-1) + a_{N,2}^* y(N-2) = \epsilon(N) + c_{N,1}^* \epsilon(N-1)$$

where the time variant parameters $a_{n,1}^*, a_{n,2}^*$ and $c_{n,1}^*$ are obtained by linearly moving from the time invariant parameters to the parameters at final time $N_f = 4500$ set at

$$a_{N_f,1}^* = -.7 \qquad a_{N_f,2}^* = .2 \qquad c_{N_f,1}^* = -.7.$$

In other words, the poles of the ARMA system linearly move from an initial location at $.35 \pm .82i$ to a final location at $.35 \pm .28i$.

The forgetting factor of the time varying prediction error algorithm was set to the λ_{opt} value given by $\lambda_{\text{opt}} = S^{2/3} = .0113$.

Since the recursive prediction error algorithms need some time to settle down due to the unknown initial conditions, a "dead" time, set to $N_i = 3000$, is used for the on-line and the off-line detectors. The estimates of the change-point were obtained using a threshold $h = 5$.

Fig. 2 illustrates how the recursive prediction error algorithms, and the off-line and on-line detectors behave. The true and estimated AR and MA parameters are plotted using the following conventions:

i) Solid line, (——), for the true parameters.

ii) Dotted line, (.....), for the estimates from the time invariant algorithm.

iii) Broken line, (- - -), for the estimates from the time varying algorithm.

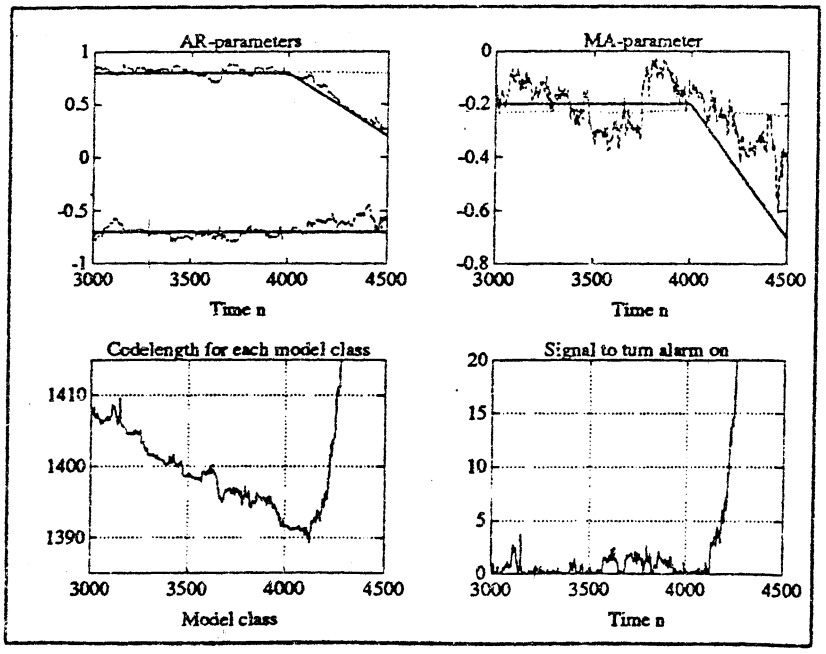

Fig. 2. Simulation with forgetting factor $\lambda_{\text{opt}} = 0.0113$.

Choosing $h = 5$ we get no false alarms, and the alarm is given at $T = 4186$, and the estimate of the change-point is $\widehat{\widehat{\tau}} = 4126$.

Table 1 shows how the estimate of the change-point $\widehat{\widehat{\tau}}$ and the stopping time T are affected by moving away from the optimal forgetting factor in both directions. No major change in the estimate of the change-point τ^* and the stopping time T is obtained. Note that modifications from the optimal forgetting factor of up to 400 % are considered. This shows the robustness of the method with respect to the parameter λ. Increasing λ may decrease the detection time, but the probability of false alarms increases. For $\lambda = .0452$ a false alarm actually occurred at $T = 3018$.

Table 1. How the estimate of the change-point $\widehat{\widehat{\tau}}$ and the stopping time T are affected by moving away from the optimal forgetting factor in both directions

SIMULATION RESULTS		
Forgetting rate λ	Change-point estimate $\widehat{\widehat{\tau}}$	Stopping time T
.0028	4126	4247
.0056	4126	4213
.0113	4126	4186
.0226	4126	4160
.0452	4126	4214

7. Appendix. We summarize a few concepts published in (Gerencsér, 1989a) and used in this paper. Let $D \subset \mathbb{R}^p$ be compact domain and let the the stochastic process (x_n) be defined on $\mathbb{Z} \times D$, where $\mathbb{Z}$ denotes the set of natural numbers.

Let $(\mathcal{F}_n)$, $n \geqslant 0$ be a family of monotone increasing σ-algebras, and $(\mathcal{F}_n^+)$, $n \geqslant 0$ be a monotone decreasing family of σ-algebras. We assume that for all $n \geqslant 0, \mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent. For $n \leqslant 0$ $\mathcal{F}_s^+ = \mathcal{F}_0^+$.

A typical example is provided by the σ-algebras

$$\mathcal{F}_n = \sigma\{e_i : i \leqslant n\}, \qquad \mathcal{F}_n^+ = \sigma\{e_i : i \geqslant n+1\},$$

where (e_i) is an i.i.d. sequence of random variables.

DEFINITION 1. We say that $(x_n(\theta))$ is M-bounded if for all $1 \leqslant q < \infty$

$$M_q(x) = \sup_{\substack{n<0 \\ \theta \epsilon D}} \mathrm{E}^{1/q} |x_n(\theta)|^q < \infty.$$

DEFINITION 2. A stochastic process (x_n), $n \geqslant 0$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ if it is $\mathcal{F}_n$-progressively measurable, M-bounded and with τ being a positive integer and

$$\gamma_r(\tau, x) = \gamma_q(\tau) = \sup_{n \geqslant \tau} \mathrm{E}^{\frac{1}{q}} |x_n - \mathrm{E}\,(x_n | \mathcal{F}_{n-\tau}^+)|^q,$$

we have

$$\Gamma_q = \Gamma_q(x) = \sum_{\tau=1}^{\infty} \gamma_q(\tau) < \infty.$$

EXAMPLE. Discrete time stationary Gaussian ARMA processes are L-mixing. (This can be seen using a state space representation).

REFERENCES

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statis. Math.*, **22**, 202–217.

Akaike, H. (1974). A new look at the statistical model identificattion. *IEEE Trans. Aut. Cont.*, **AC-19**, 716–723.

Basseville, M. (1988). Detecting changes in signals and systems – A survey. *Automatica*, **24**(3), 309–326.

Basseville, M., and A. Benveniste (Eds.) (1986). *Detection of Abrupt Changes in Signals and Dynamical Systems.* Lectures Notes in Control and Information Sciences 77, Springer, New York.

Caines, P.E. (1988). *Linear Stochastic Systems.* Wiley.

Gerencsér, L. (1988a). On Rissanen's predictive stochastic complexity for stationary ARMA-processes. Submitted to *The Annals of Statistics.*

Gerencsér, L. (1988b). Patthwise stability of random differential equations and the solution of an adaptive control related problem. *Proc. of the 4th Bad Honneff Conference on Stochastic Differential Systems.* Springer.

Gerencsér, L. (1989a). On a class of mixing processes. *Stochastics*, **26**, 165–191.

Gerencsér, L. (1989b). Rate of convergence for Ljung's scheme. McGill Research Center for Computer Vision and Intelligent Machines, TR-CIM-89-16. Submitted to Stochastic Processes and Their Applications.

Gerencsér, L. (1990a). Almost sure exponential stability of random linear differential equations. Accepted for publication in *Stochastics.*

Gerencsér, L. (1991b). Asymptotics of predictive stochastic complexity. To appear in New Directions in Time Series Analysis Proc of the 1990 IMA Workshop E.Parzen, D.Brillinger, M.Rosenblatt, M.Taqqu, J.Geweke, and P.E.Caines (Eds.), Springer.

Gerencsér, L. (1991c). The law of the cubic root. To appear in *Proc. of the 9th IFAC/IFORS Symposium on Identification and System Parameter Estimation,* Budapest.

Gerencsér, L. (1991d). On fix gain estimators of ARMA parameters. Manuscript.

Hinkley, D.V. (1971). Inference about the change-point from Cumulative Sum Tests. *Biometrika,* **58**(3), 509–523.

Ljung, L., and T.Söderström (1984). *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, Mass.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica,* **14**, 465–471.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics,* **14**, 1080–1100.

Rissanen, J. (1987). Stochastic complexity (with discussions).. *J. of the Royal Statistical Society,* Series B., Vol. **49** No. 3.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* .World Scientific

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* **6**, 461–464.

Shibata, R. (1980). Selection of the number of regression variables; A mini-max choise of generalized FPE². *Annals of Statistics,* **38**(3), 459–474.

Söderström, T., and P.Stoica (1989). *System Identification.* Prentice Hall

Telksnys, L. (Ed.) (1986). *Detection of Changes in Random Processes.* Optimization Software, Inc.

Received January 1992

J. Baikovicius was born in Montevideo, Uruguay, in 1962. He received the B.Sc and M.Sc degree in Mechanical Engineering in 1986 and 1987 respectively form the University of Minnesota, USA. Currently he is working towards the Ph.D. degree at the department of Electrical Engineering, McGill University, Montreal, Canada. He is an associate of the McGill Research Center for Intelligent Machines (McRCIM). His current research interest are linear stochastic systems, stochastic complexity, model order selection, fault detection, and adaptive control.

L. Gerencsér received his M.Sc. and Ph.D. degrees in mathematics from the the Eötvös Lóránd Science University, Budapest and his candidate of mathematical sciences degree from the Hungarian Academy of Sciences. He started his carrier in the Mathematical Institute of the Hungarian Academy of Sciences in 1969. Since 1970 he has been in what is today the Computer and Automation Institute of the Hungarian Academy of Sciences. He spent the year 1986 in the Department of Mathematics of the Chalmers University of Technology, Gothenburg, Sweden as a visiting lecturer and held the position of a visiting professor at the Department of Electrical Engineering, Mc Gill University, Montreal, Quebec, Canada between 1988 and 1991. His early research areas were graph theory, differential geometry, nonlinear programming and stochastic models of operation research. In the last decade his focus of research shifted to the statistical theory of linear stochastic systems in general, recursive estimation, strong approximation methods, stochastic complexity, continuous time systems, adaptive control, change point detection and random linear differential equations in particular. His most recent research interest is high accuracy adaptive control.