Change detection of Hidden Markov Models

László Gerencsér and Gábor Molnár-Sáska

***Abstract*— Let $\widehat{\theta}_N$ denote the maximum-likelihood estimator of the true parameter θ^* of a Hidden Markov Model with fixed gain or forgetting rate λ. We establish an explicit formula for the error term $\widehat{\theta}_N - \theta^*$. Using the representation of the error term we investigate the effect of parameter uncertainty on the performance of an adaptive encoding procedure. Using this result and ideas from the theory of stochastic complexity a change point detection method for HMM-s will be developed.**

I. INTRODUCTION

Hidden Markov Models have become a basic tool for modelling stochastic systems with a wide range of applicability in such diverse areas as robotics, telecommunication, econometrics and protein research. For a general introduction see [16].

The estimation of the dynamics of a Hidden Markov Model is a basic problem in applications. A key element in the statistical analysis of HMM-s is a strong law of large numbers for the log-likelihood function. The first basic result in this area is due to Baum and Petrie for finite state Markov chains with finite-range read-outs [2]. This result has been extended to continuous read-outs by Leroux in [13] and LeGland and Mevel in [10].

A useful connection between the estimation theory of HMM-s and linear stochastic systems is established in [7] via the theory of L-mixing processes. The concept of L-mixing has been developed in [3]. It is a generalization of what is known as exponentially stable processes, introduced by Caines and Rissanen in [15] and Ljung [14].

The purpose of this paper is to extend the results of [7] to fixed gain estimation procedures, and to apply these results to change detection.

Fixed gain off-line estimators of ARMA parameters are investigated in [5]. The performance of adaptive predictors of ARMA processes using fixed gain estimation was studied in [4]. The results of the latter paper together with ideas of the theory of stochastic complexity were used to develop a change-point detection method for ARMA processes in [1]. The same route is followed for HMM-s. In section II the basic concepts of Hidden Markov Models is given. In section III we summarize the results of [7]. In section IV we discuss fixed gain estimation of HMM-s, while in section V we study adaptive encoding and change-detection for HMM-s. Simulation result will also be presented. Basic definitions and results necessary for the paper can be found in the Appendix.

This work was supported by the Natural Science and Research Foundation of Hungary (OTKA) under Grant T 047193

L. Gerencsér is with MTA SZTAKI, Computer and Automation Institute of the Hungarian Academy of Sciences, 1111 Budapest, Hungary gerencser@sztaki.hu

G. Molnár-Sáska is with MTA SZTAKI, Computer and Automation Institute of the Hungarian Academy of Sciences, 1111 Budapest, Hungary molnarsg@sztaki.hu

II. HIDDEN MARKOV MODELS

We consider Hidden Markov Models with a general state space $\mathcal{X}$ and a general observation or read-out space $\mathcal{Y}$. Both are assumed to be Polish spaces, i.e. they are complete, separable metric spaces.

Definition 2.1: The pair (X_n, Y_n) is a Hidden Markov process if (X_n) is a homogenous Markov chain, with state space $\mathcal{X}$ and the observations (Y_n) are conditionally independent and identically distributed given (X_n).

In the discrete case we will use the following notations

$$P(Y_k = y | X_k = x) = b^{*x}(y), \quad B^*(y) = \mathrm{diag}(b^{*i}(y)),$$

where $i = 1, \dots, N$, and $*$ indicates that we take the true value of the corresponding unknown quantity. Let Q^* be the transition matrix of the unobserved Markov process (X_n), i.e.

$$Q^*_{ij} = P(X_{n+1} = j | X_n = i).$$

A key quantity in estimation theory is the predictive filter defined by

$$p^{*j}_{n+1} = P(X_{n+1} = j | Y_n, \dots, Y_0).$$

Writing $p^*_{n+1} = (p^{*1}_{n+1}, \dots, p^{*N}_{n+1})^T$, the filter process satisfies the Baum-equation

$$p^*_{n+1} = \pi(Q^{*T} B^*(Y_n) p^*_n), \tag{1}$$

where π is the normalizing operator: for $x \geq 0, \ x \neq 0$ set $\pi(x)^i = x^i / \sum_j x^j$, see [2]. Here $p^{*j}_0 = P(X_0 = j)$.

In practice, the transition probability matrix Q^* and the initial probability distribution p^*_0 of the unobserved Markov chain (X_n) and the conditional probabilities $b^{*i}(y)$ of the observation sequence (Y_n) are possibly unknown. For this reason we consider the Baum-equation in a more general sense

$$p_{n+1} = \pi(Q^T B(Y_n) p_n), \tag{2}$$

with initial condition $p_0 = q$, where Q is a stochastic matrix, p_n is a probability vector on $\mathcal{X}$, and $B(y) = \mathrm{diag}(b^i(y))$ is a collection of conditional probabilities.

We will take an arbitrary probability vector q as initial condition, and the solution of the Baum equation will be denoted by $p_n(q)$.

Continuous read-outs will be defined by taking the following conditional densities:

$$P(Y_n \in dy | X_n = x) = b^{*x}(y) \lambda(dy),$$

where λ is a fixed nonnegative, σ-finite measure. Let

$$B^*(y) = \mathrm{diag}(b^{*i}(y)),$$

where $i = 1, \ldots, N$.

Let denote the parameter of the model by θ. θ is usually assumed to be the elements of the transition probability matrix, the read-out probabilities in discrete case and the parameters of the read-out densities in continuous case.

III. ESTIMATION OF HMM-S

Consider a Hidden Markov Process (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is continuous, a measurable subset of R^d. Assume that the transition probability matrix and the conditional read-out densities are positive, i.e. $Q^* > 0$ and $b^{*i} > 0$ for all i, y.

Let the invariant distribution of $\mathcal{X}$ be ν and the invariant distribution of $\mathcal{X} \times \mathcal{Y}$ be π. Then

$$\pi^i(dy) = \nu_i b^{*i}(y)\lambda(dy), \tag{3}$$

where π^i denotes the components of π. Furthermore let the running value of the transition probability matrix Q and the running value of the conditional read-out densities be also positive, i.e. $Q > 0$, $b^i(y) > 0$, respectively.

With the notation $p_n^i = P(X_n = i | Y_{n-1}, \ldots, Y_0)$ we have

$$p_{n+1} = \pi(Q^T B(Y_n) p_n) = f(Y_n, p_n).$$

We use capital letters for random variables and lower cases for their realizations, i.e. X is a random variable and x is a realization of X. The only exception is p, where the meaning depends on the context.

The logarithm of the likelihood function is

$$\sum_{k=1}^{n-1} \log p(y_k | y_{k-1}, \ldots y_0, \theta) + \log p(y_0, \theta).$$

Here the k-th term for $k \geq 1$ can be written as

$$\log \sum_i b^i(y_k) P(i | y_{k-1}, \ldots, y_0, \theta) = \log \sum_i b^i(y_k) p_k^i.$$

Now write

$$g(y, p) = \log \sum_i b^i(y) p^i, \tag{4}$$

then we have

$$\log p(y_N, \ldots, y_0, \theta) = \sum_{k=1}^{N} g(y_k, p_k) + \log p(y_0, \theta). \tag{5}$$

We look for a characterization of (X_n, Y_n, p_n).

Theorem 3.1: Consider a Hidden Markov Model (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is continuous, a measurable subset of R^d. Let $Q, Q^* > 0$ and $b^i(y), b^{*i}(y) > 0$ for all i, y. Let the initialization of the process (X_n, Y_n) be random, where the Radon-Nikodym derivate of the initial distribution π_0 w.r.t the stationary distribution π is bounded, i.e.

$$d\pi_0 / d\pi \leq K. \tag{6}$$

Assume that for all $i, j \in \mathcal{X}$

$$\int |\log b^j(y)|^q b^{*i}(y) \lambda(dy) < \infty. \tag{7}$$

Then the process $g(Y_n, p_n)$ is L-mixing.

To analyze the asymptotic properties of the right hand side of (5) Theorem 3.1 seems to be relevant. Under the conditions of Theorem 3.1 $g(y, p)$ is an L-mixing process and the law of large numbers is valid for such processes, see [3]. This implies the existence of the limit of (5).

Consider now a *finite state-finite read-out* HMM. This case follows from Theorem 3.1, but the integrability condition (7) is simplified due to the discrete measure.

Theorem 3.2: Consider the Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Let $Q, Q^* > 0$ and $b^i(y) \geq \delta > 0$, $b^{*i}(y) \geq \delta > 0$ for all i, y. Then with a random initialization on $\mathcal{X} \times \mathcal{Y}$ we have that $g(Y_n, p_n)$ is an L-mixing process.

IV. ESTIMATION WITH FORGETTING

Consider a finite state-finite read-out HMM with $Q^* > 0$ and $b^{*i}(y) > 0$ for all i, y, parameterized by θ, where $|\mathcal{X}| = N$ and $|\mathcal{Y}| = M$ and θ containing the elements of the transition probability matrix and the read-out probabilities. Thus θ is an $N^2 + NM - 2N$ dimensional vector with coordinates $0 < \theta^i < 1$. Furthermore let the ML estimate of the true parameter θ^* be denoted by $\hat{\theta}_N$.

If the dynamics changes slowly in time, then we should adapt to the actual system. But then the estimation procedure must be modified: instead of cumulating past data we must gradually forget them. Forgetting past data is technically realized by using exponential forgetting in the off-line case.

Following a standard device in system-identification let now $\widehat{\theta}_N = \widehat{\theta}_N(\lambda)$ be the estimator of θ^* obtained by minimizing

$$\sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda \log p(y_n | y_{n-1}, \ldots, y_0; \theta), \tag{8}$$

with $0 < \lambda < 1$. Here λ is the so-called forgetting factor: small value of λ means slow forgetting. Combining Theorem 3.1 with the techniques of [5] we get:

Theorem 4.1: Under the conditions of Theorem 3.2 we have

$$\hat{\theta}_N(\lambda) - \theta^* =$$

$$-I(\theta^*)^{-1} \sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda \frac{\partial}{\partial \theta} \log p(Y_n | Y_{n-1}, \ldots Y_0, \theta^*) + r,$$

where $0 < \alpha < 1$, $r = O_M(\lambda) + O_M(\alpha^N)$, and $I(\theta^*)$ is the Fischer-information matrix.

Theorem 4.1 implies that the covariance matrix is

$$E(\widehat{\theta}_{n-1} - \theta^*)(\widehat{\theta}_{n-1} - \theta^*)^T = \frac{\lambda}{2} I(\theta^*)^{-1} + O(\lambda^{3/2}) + o(1). \tag{9}$$

V. CHANGE-DETECTION OF HIDDEN MARKOV MODELS

We consider change-detection problems for Hidden Markov Models following [1]. For this we first note that the negative of the log-likelihood can be interpreted as a codelength, modulo a constant, which is obtained when encoding the data sequence $(y_N, \ldots, y_1)$ with a prescribed accuracy, using the assumed joint density $p(y_N, \ldots, y_0; \theta)$. This interpretation of the likelihood is a central idea of the theory of stochastic complexity. Thus we interpret

$$C_n(y_n; \theta) \stackrel{\Delta}{=} -\log p(y_n | y_{n-1}, \ldots, y_0; \theta),$$

as a codelength. A key result in the theory of the stochastic complexity can be extended for the present case (see [6]).

Let the score function be

$$\varphi_n(\theta) = \frac{\partial}{\partial \theta} \log p(y_n | y_{n-1}, \ldots, y_0, \theta)$$

We also use lower cases for the random variable $\frac{\partial}{\partial \theta} \log p(Y_n | Y_{n-1}, \ldots, Y_0, \theta)$. In the following let

$$r = \dim \theta.$$

Theorem 5.1: Under the conditions of Theorem 3.2 we have

$$\mathrm{E}(C_n(Y_n, \widehat{\theta}_{n-1}(\lambda)) - C_n(Y_n, \theta^*) =$$
$$\frac{1}{2} r\lambda + O(\lambda^{3/2-c''}) + o(1),$$

with an arbitrary small $c'' > 0$.

The faster the forgetting is i.e. the closer λ is to 1, the more we loose in encoding performance.

Sketch of the proof: Consider the second order Taylor-series expansion of $\log p(y_n | y_{n-1}, \ldots, y_0, \widehat{\theta}_{n-1})$ around θ^*. To simplify the expressions we drop λ from the arguments and we use the notation $p_n(\theta)$ for the conditional probability $p(y_n | y_{n-1}, \ldots, y_0, \theta)$ expressing the dependence on the parameter θ. We also use the score function defined above.

$$\log p_n(\widehat{\theta}_{n-1}) - \log p_n(\theta^*) = \varphi_n(\theta^*)(\widehat{\theta}_{n-1} - \theta^*) +$$
$$(\widehat{\theta}_{n-1} - \theta^*)^T \left(\frac{\partial^2}{\partial \theta^2} \log p_n \right) (\theta^*)(\widehat{\theta}_{n-1} - \theta^*) +$$
$$O_M(\lambda^{3/2}) + o_M(1),$$

where term $o_M(1)$ is due to the nonstationary initial condition and we took into account that $(\widehat{\theta}_{n-1} - \theta^*) = O_M(\lambda^{1/2})$.

Now consider the expectation of the above equality. Using that $\varphi_n(\theta^*)$ is a martingale difference, see for definition [8], the expectation of the first term is 0. In the second term write $\left(\frac{\partial^2}{\partial \theta^2} \log p_n \right) (\theta^*)$ as

$$\left(\frac{1}{p_n} (\frac{\partial^2 p_n}{\partial \theta^2})(\theta^*) \right) - (\varphi_n(\theta^*))^T (\varphi_n(\theta^*)).$$

The expectation of the first term is 0, so

$$E\left(\log p_n(\widehat{\theta}_{n-1}) - \log p_n(\theta^*) \right) =$$
$$E\left((\widehat{\theta}_{n-1} - \theta^*)^T (\varphi_n(\theta^*))^T (\varphi_n(\theta^*)) (\widehat{\theta}_{n-1} - \theta^*) \right) +$$
$$O(\lambda^{3/2}) + o(1) \qquad (10)$$

Noting that $\varphi_n(\theta^*)$ depends on the past weakly, while $(\widehat{\theta}_{n-1} - \theta^*)$ depends on the past strongly for small λ, we can use the following cutting argument. Choose a positive integer $d = -c \log \lambda$ and consider the following approximations:

$$\varphi_n^+(\theta^*) = E(\varphi_n(\theta^*) | \mathcal{F}_{n-d}^+)$$

and

$$\widehat{\theta}_{n-1}^- - \theta^* = -I(\theta^*)^{-1} \sum_{i=1}^{n-d} (1-\lambda)^{n-i-1} \lambda \varphi_i(\theta^*).$$

It is easy to see that

$$\varphi_n^+(\theta^*) - \varphi_n(\theta^*) = O_M(\alpha^d)$$

with some $0 < \alpha < 1$, thus

$$\varphi_n^+(\theta^*) - \varphi_n(\theta^*) = O_M(\lambda^{c'})$$

with some $c' > 0$ and

$$\widehat{\theta}_{n-1}^- - \widehat{\theta}_{n-1} = O_M(\lambda^{1-c''}),$$

with all $c'' > 0$ (see Theorem 4.1 of the present paper and Theorem 1.1 of [3]). Furthermore $\varphi_n^+(\theta^*)$ and $\widehat{\theta}_{n-1}^- - \theta^*$ are independent. Approximate (10) by

$$E\left((\widehat{\theta}_{n-1}^- - \theta^*)^T (\varphi_n^+(\theta^*))^T (\varphi_n^+(\theta^*)) (\widehat{\theta}_{n-1}^- - \theta^*) \right).$$

This can be written as

$$\mathrm{Tr}\left(E((\widehat{\theta}_{n-1}^- - \theta^*)(\widehat{\theta}_{n-1}^- - \theta^*)^T) E(\varphi_n^+(\theta^*))^T (\varphi_n^+(\theta^*))) \right).$$

The error of this approximation is $O(\lambda^{3/2-c''})$, for sufficienlty large c.

Combining the above approximation with (9) we have

$$E(\widehat{\theta}_{n-1}^- - \theta^*)(\widehat{\theta}_{n-1}^- - \theta^*)^T = \frac{\lambda}{2} I(\theta^*)^{-1} + O(\lambda^{3/2-c''}) + o(1).$$

Furthermore noting that

$$E(\varphi_n(\theta^*))^T (\varphi_n(\theta^*)) = I(\theta^*) + O(\alpha^i),$$

we get

$$E(\varphi_n^+(\theta^*))^T (\varphi_n^+(\theta^*)) = I(\theta^*) + O(\alpha^i) + O(\lambda^{c'}).$$

Thus we get that

$$\mathrm{E}(C_n(y_n, \widehat{\theta}_{n-1}(\lambda)) - C_n(y_n, \theta^*) =$$
$$\frac{1}{2} r\lambda + O(\lambda^{3/2-c''}) + o(1)$$

for all $c'' > 0$.

An easy consequence of Proposition 5.1 is the following.

Proposition 5.1: Consider two different forgetting factors $0 < \lambda_1 < \lambda_2 < 1$. Then we have

$$\mathrm{E}(C_n(y_n, \widehat{\theta}_{n-1}(\lambda_1)) - C_n(y_n, \widehat{\theta}_{n-1}(\lambda_2))) \simeq$$

$$\frac{1}{2} r(\lambda_1 - \lambda_2) < 0.$$

Proposition 5.1 has been useful in the design of a new model selection criterion. However for the theoretical analysis of the new method is not powerful enough. For this purpose we need a sample path characterization of the prediction error process. Let the cumulative error be

$$S_N(\lambda) = \sum_{n=1}^{N} (C_n(y_n, \widehat{\theta}_{n-1}(\lambda)) - C_n(y_n, \theta^*))$$

Theorem 5.2: Assume that the conditions of Theorem 3.2 are satisfied. Then we have

$$\limsup_{N\to\infty} |\frac{1}{N} S_N(\lambda) - \frac{\lambda}{2} r| \leq C\lambda^{3/2}$$

Sketch of the proof: The proof of this theorem is based on the fact that under the conditions of the theorem the estimator error process $\widehat{\theta}_n - \theta^*$ has an L-mixing version ([5]). Then $C(y_n, \widehat{\theta}_n)$ and $C_n(y_n, \widehat{\theta}_{n-1}(\lambda)) - C_n(y_n, \theta^*)$ are also L-mixing. Using the law of large numbers for the latter process we get the proof.

We state a similar easy consequence as above.

Proposition 5.2: Let $0 < \lambda_1 < \lambda_2 < 1$ be two different forgetting factors. Then we have

$$\limsup_{N\to\infty} |\frac{1}{N} S_N(\lambda_1) - \frac{1}{N} S_N(\lambda_2) - \frac{\lambda_1 - \lambda_2}{2} r| \leq C\lambda_2^{3/2}$$

Assume now that a jump in the parameter occurs at τ: the true value of θ is θ_1 for $n \leq \tau$ and it is θ_2 for $n \geq \tau+1$, i.e.

$$\theta^* := \begin{cases} \theta_1, & \text{if} \quad n \leq \tau \\ \theta_2, & \text{if} \quad n \geq \tau \end{cases}$$

Let $0 < \lambda_1 < \lambda_2 < 1$. Then from Proposition 5.2 we have for $N \leq \tau$

$$S_N(\lambda_1) - S_N(\lambda_2) \approx \frac{\lambda_1 - \lambda_2}{2} Nr.$$

On the other hand at the time of change the performance of the estimator with faster forgetting, i.e. with λ_2 expected to be better. Hence, consider the following algorithm for detecting the change.

The algorithm: Let $d(N) := S_N(\lambda_1) - S_N(\lambda_2)$ and set

$$d_N^* = \min_{n \leq N} d(n).$$

An alarm is generated if $d(N) - d_N^* > \epsilon$, where $\epsilon > 0$ is a prescribed treshold value. This type of algorithm is called *Hinkley detector* in the literature, see [9].

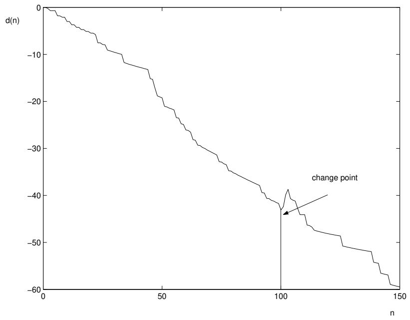

Fig. 1. We have generated a binary HMM. The change has occured at step 100.

VI. CONCLUSION

We have established the fixed gain off-line estimator of HMM-s. We have computed the effect of parameter uncertainty on the performance of an adaptive encoding procedure. The techniques used were adopted from the theory of linear stochastic systems and stochastic complexity. Based on these technical results a change point detection method for HMM-s has been developed and tested.

VII. APPENDIX

A key property of the Baum equation is its exponential stability with respect to the initial condition. This has been established in [10] for continuous read-outs. Here we state the result for HMM-s with positive transition probability matrix:

Proposition 7.1: Assume that $Q > 0$ and $b^x(y) > 0$ for all x, y. Let q, q' be any two initializations. Then

$$\|p_n(q) - p_n(q')\|_{TV} \leq C(1-\delta)^n \|q - q'\|_{TV}, \qquad (11)$$

where $\| \quad \|_{TV}$ denotes the total variation norm and $0 < \delta < 1$.

If Q is only primitive, i.e. $Q^r > 0$ with some positive integer $r > 1$, then (11) holds with a random C.

Now we are going to introduce a class of processes called L-mixing processes which have been used extensively in the statistical analysis of linear stochastic systems, see [3].

Definition 7.1: A stochastic process (X_n) $(n \geq 0)$ taking its values in an Euclidean space is M-bounded if for all $q \geq 1$

$$M_q = \sup_{n\geq 0} E^{1/q} \|X_n\|^q < \infty.$$

If (X_n) is M-bounded we shall also write $X_n = O_M(1)$. Similarly if c_n is a positive sequence we write $X_n = O_M(c_n)$ if $X_n / c_n = O_M(1)$.

Let $(\mathcal{F}_n)$ and $(\mathcal{F}_n^+)$ be two sequences of monoton increasing and monoton decreasing σ-algebras, respectively such that $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent for all n.

Definition 7.2: A stochastic process (X_n) taking its values in a finite-dimensional Euclidean space is L-mixing, if it is M-bounded and with

$$\gamma_q(\tau) = \sup_{n \geq \tau} E^{1/q} \| X_n - E(X_n | \mathcal{F}^+_{n-\tau}) \|^q$$

we have

$$\Gamma(q) = \sum_{\tau=0}^{\infty} \gamma_q(\tau) < \infty.$$

The parametric version of M-boundedness and L-mixing property can also be defined, see [3] for details.

REFERENCES

[1] J. Baikovicius and L. Gerencsér. Change point detection in a stochastic complexity framework. In *Proc. of the 29-th IEEE CDC*, volume 6, pages 3554–3555, 1990.

[2] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37:1559–1563, 1966.

[3] L. Gerencsér. On a class of mixing processes. *Stochastics*, 26:165–191, 1989.

[4] L. Gerencsér. Predictive stochastic complexity associated with fixed gain estimators. In C. Praagman J.W. Niueuwenhuis and H.L. Trentelman, editors, *Proceedings of the 2nd European Control Conference, Groningen*, pages 1673–1677, 1993.

[5] L. Gerencsér. Fixed gain off-line estimators of ARMA parameters. *Journal of Mathematical Systems, Estimation and Control*, 4(2):249–252, 1994. Retreival code for full electronic manuscript: 66945.

[6] L. Gerencsér. On Rissanen's predictive stochastic complexity for stationary ARMA processes. *Statistical Planning and Inference*, 41:303–325, 1994.

[7] L. Gerencsér, G. Molnár-Sáska, Gy. Michaletzky, and G. Tusnaďy. New methods for the statistical analysis of Hidden Markov Models. *IEEE Trans. on Automatic Control*, 2003. submitted.

[8] P. Hall, C.C. Heyde. Martingale Limit Theory. *IEEE Trans. on Automatic Control*, 2003.

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

[10] F. LeGland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals and Systems*, 13:63–93, 2000.

[11] F. LeGland and L. Mevel. Recursive identification in HMM's. In *Proc. of the 36th IEEE CDC*, 537-542, 1997.

[12] F. LeGland and L. Mevel. Fault detection in HMM's: a local asymptotic approach. In *Proc. of the 39th IEEE CDC*, 4686-4690, 2000.

[13] B. G. Leroux. Maximum-likelihood estimation for hidden Markov-models. *Stochastic Processes and their Applications*, 40:127–143, 1992.

[14] L. Ljung. On consistency and identifiability. *Mathematical Programming Study*, 5:169–190., 1976.

[15] J. Rissanen and P.E. Caines. The strong consistency of maximum likelihood estimators for ARMA processes. *Ann. Statist.*, 7:297 – 315., 1979.

[16] J. H. van Schuppen. Lecture notes on stochastic systems. Technical report. Manuscript.