Recursive Estimation in Hidden Markov Models*

François LeGland and Laurent Mevel
IRISA / INRIA and IRMAR
Campus de Beaulieu
35042 Rennes Cédex, France
e–mail : {legland,lmevel}@irisa.fr

Abstract

We consider a hidden Markov model (HMM) with multidimensional observations, and where the coefficients (transition probability matrix, and observation conditional densities) depend on some unknown parameter. We study the asymptotic behaviour of two recursive estimators, the recursive maximum likelihood estimator (RMLE), and the recursive conditional least squares estimator (RCLSE), as the number of observations increases to infinity. Firstly, we exhibit the contrast functions associated with the two non–recursive estimators, and we prove that the recursive estimators converge a.s. to the set of stationary points of the corresponding contrast function. Secondly, we prove that the two recursive estimators are asymptotically normal.

1 Introduction

We consider the problem of identification of a partially observed finite–state Markov chain, based on observations in $\mathbf{R}^d$. We study the asymptotic behaviour (convergence and asymptotic normality) of two recursive estimators, the recursive maximum likelihood estimator (RMLE), and the recursive conditional least squares estimator (RCLSE), as the number of observations increases to infinity. Assuming stationarity of the nonobserved Markov chain, the non–recursive MLE has been studied by Petrie [10] in the case of observations in a finite set, and by Leroux [9], and Bickel, Ritov and Rydén [3], in the case of observations in $\mathbf{R}^d$. These results have been obtained in LeGland and Mevel [8] without any stationarity assumption. The convergence of the RCLSE has been proved in Arapostathis and Marcus [1] for a special case of observations in a finite set — see also Le Gland and Mevel [6] for a study of the RMLE in the general case of observations in a finite set.

*This work was partially supported by the Commission of the European Communities, under the SCIENCE project *System Identification*, project number SC1*–CT92–0779, and under the HCM project *Statistical Inference for Stochastic Processes*, project number CHRX–CT92–0078, and by the Army Research Office, under grant DAAH04–95–1–0164.

The convergence and the asymptotic normality have been obtained by Rydén in [11] for the recursive maximum split data likelihood estimator (MSDLE) defined therein, under stationarity assumption. In this paper, we prove the convergence and the asymptotic normality of the RMLE and the RCLSE, in the case of observations in $\mathbf{R}^d$, without any stationary assumption, using geometric ergodicity results on the *approximate* prediction filter obtained in LeGland and Mevel [7]. We prove also that the RMLE is asymptotically efficient, i.e. its asymptotic covariance matrix is the Fisher information matrix.

1.1 Statistical model

Let $\{X_n\,,\,n \geq 0\}$ and $\{Y_n\,,\,n \geq 0\}$ be two sequences, defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, with values in the finite set $S = \{1, \cdots, N\}$ and in $\mathbf{R}^d$ respectively. On the space $(\Omega, \mathcal{F})$ we consider a family $(\mathbf{P}^\theta\,,\,\theta \in \Theta)$ of probability measures, with Θ convex compact subset of $\mathbf{R}^p$, such that under $\mathbf{P}^\theta$:

- The unobserved state sequence $\{X_n\,,\,n \geq 0\}$ is a Markov chain with *primitive* transition probability matrix $Q_\theta = (q_\theta^{i,j})$, i.e. for any $i, j \in S$

$$q_\theta^{i,j} = \mathbf{P}^\theta[X_{n+1} = j \mid X_n = i]\,,$$

  and initial probability distribution $p_0 = (p_0^i)$ independent of $\theta \in \Theta$, and possibly different of the *true* initial probability distribution $p_\bullet = (p_\bullet^i)$ of X_0, i.e. for any $i \in S$

$$p_0^i = \mathbf{P}^\theta[X_0 = i] \neq \mathbf{P}[X_0 = i] = p_\bullet^i\,.$$

- The observations $\{Y_n\,,\,n \geq 0\}$ are mutually independent given the sequence of states of the Markov chain, i.e.

$$\mathbf{P}^\theta[Y_n \in dy_n, \cdots, Y_0 \in dy_0 \mid X_n = i_n, \cdots, X_0 = i_0]$$

$$= \prod_{k=0}^{n} \mathbf{P}^\theta[Y_k \in dy_k \mid X_k = i_k]\,.$$

  For any $n \geq 0$, and for any $i \in S$, the conditional probability distribution of the observation

Y_n given that $(X_n = i)$, is absolutely continuous with respect to a positive and σ–finite measure λ on $\mathbf{R}^d$, i.e.

$$\mathbf{P}^\theta[Y_n \in dy \mid X_n = i] = b^i_\theta(y)\ \lambda(dy)\ ,$$

with a λ–a.e. positive density. For any $y \in \mathbf{R}^d$, let

$$b_\theta(y) = [b^1_\theta(y), \cdots, b^N_\theta(y)]^*\ ,$$

$$B_\theta(y) = \mathrm{diag}\,[b^1_\theta(y), \cdots, b^N_\theta(y)]\ .$$

Here and throughout the paper, the notation * denotes the transpose of a matrix.

Example 1.1 [conditionally Gaussian observations] Assume that for any $\theta \in \Theta$, the observations are of the form

$$Y_n = h_\theta(X_n) + V^\theta_n\ ,$$

for all $n \geq 0$, where $\{V^\theta_n\ ,\ n \geq 0\}$ is a Gaussian white noise sequence under $\mathbf{P}^\theta$, with identity covariance matrix. The mapping h_θ from S to $\mathbf{R}^d$ is equivalently defined as $h_\theta = (h^i_\theta)$ where $h^i_\theta \in \mathbf{R}^d$ for all $i \in S$. In this case, the mutual independence condition is satisfied, and

$$b^i_\theta(y) = (2\pi)^{-d/2}\ \exp\left\{\ -\tfrac{1}{2}\,|y - h^i_\theta|^2\ \right\}\ ,$$

for any $i \in S$. Here and throughout the paper, the notation $|\cdot|$ denotes the Euclidean norm.

Throughout the paper, the *true* value of the parameter will be denoted by α, and we make the following assumption :

Assumption A : For the *true* value $\alpha \in \Theta$, the transition probability matrix $Q_\alpha = (q^{i,j}_\alpha)$ is a positive matrix, i.e. $q^{i,j}_\alpha \geq \varepsilon$, for all $i, j \in S$, and for some *known* $\varepsilon > 0$.

Assumption A' : The mapping $\theta \mapsto Q_\theta$ is two–times differentiable, with bounded first and second derivatives, and Lipschitz continuous second derivative.

Assumption B : For any $y \in \mathbf{R}^d$, the mapping $\theta \mapsto b_\theta(y)$ is three–times differentiable, and we define

$$\delta^{(s)}(y) = \max_{\theta\in\Theta}\ \max_{k_1,\cdots,k_s=1,\cdots,p}\ \frac{\max\limits_{i\in S} |\partial^s_{k_1,\cdots,k_s}\, b^i_\theta(y)|}{\min\limits_{i\in S} b^i_\theta(y)}\ .$$

Definition 1.2 *For any $p \geq 0$, and any $s = 0, 1, 2, 3$,*

$$\Delta^{(s)}_p = \max_{\theta\in\Theta}\ \max_{i\in S}\ \int_{\mathbf{R}^d} [\delta^{(s)}(y)]^p\ b^i_\theta(y)\ \lambda(dy)\ ,$$

$$\Gamma_p = \max_{i\in S}\ \max_{\theta\in\Theta}\ \int_{\mathbf{R}^d} [\max_{j\in S} |\log b^j_\theta(y)|]^p\ b^i_\alpha(y)\ \lambda(dy)\ ,$$

$$\bar{Y}^p = \max_{\theta\in\Theta}\ \max_{i\in S}\ \int_{\mathbf{R}^d} |y|^p\ b^i_\theta(y)\ \lambda(dy)\ .$$

We use also the alternate notation $\Delta_p = \Delta^{(0)}_p$, $\Delta'_p = \Delta^{(1)}_p$, $\Delta''_p = \Delta^{(2)}_p$, and $\Delta'''_p = \Delta^{(3)}_p$.

Assumption B' : $\bar{Y}^2$, Δ'_2, Δ''_2, and Δ'''_2 are finite.

1.2 Prediction filters

For all $n \geq 1$, let $p^\bullet_n = (p^i_n)$ denote the *prediction filter*, i.e. the conditional probability distribution under $\mathbf{P}$ of the state X_n given observations $(Y_0, \cdots, Y_{n-1})$: for any $i \in S$

$$p^i_n = \mathbf{P}[X_n = i \mid Y_0, \cdots, Y_{n-1}]\ .$$

The random sequence $\{p^\bullet_n\ ,\ n \geq 0\}$ takes values in the set $\mathcal{P}(S)$ of probability distributions over the finite set S, and satisfies the forward Baum equation

$$p^\bullet_{n+1} = \frac{Q^*_\alpha B_\alpha(Y_n)\ p^\bullet_n}{b^*_\alpha(Y_n)\ p^\bullet_n}\ ,$$

for all $n \geq 0$. Notice that the initial condition $p^\bullet_0$ is the probability distribution $p_\bullet$ of X_0.

The true value of the parameter, and the initial probability distribution of X_0 are unknown, and for any $\theta \in \Theta$ we consider the following equation for the corresponding *approximate* prediction filter

$$p^\theta_{n+1} = \frac{Q^*_\theta B_\theta(Y_n)\ p^\theta_n}{b^*_\theta(Y_n)\ p^\theta_n} \triangleq f_\theta[Y_n, p^\theta_n]\ , \qquad (1)$$

for all $n \geq 0$, where $Q_\theta = (q^{i,j}_\theta)$ is the stochastic matrix generated by the parameter θ.

Differentiating (1) w.r.t. the κ–th component of the p–dimensional parameter θ yields

$$\partial_k p^\theta_{n+1}\ =\ \Phi_\theta[Y_n, p^\theta_n]\ \partial_k p^\theta_n + \partial_k f_\theta[Y_n, p^\theta_n]\ ,$$

where for any $y \in \mathbf{R}^d$, and any $p \in \mathcal{P}(S)$

$$\Phi_\theta[y,p]\ =\ \frac{Q^*_\theta B_\theta(y)}{b^*_\theta(y)\,p}\,[I - \frac{p\,b^*_\theta(y)}{b^*_\theta(y)\,p}]$$

$$=\ Q^*_\theta\,[I - \frac{B_\theta(y)\,p\,e^*}{b^*_\theta(y)\,p}]\,\frac{B_\theta(y)}{b^*_\theta(y)\,p}\ ,$$

$$\partial_k f_\theta[y,p]\ =\ Q^*_\theta\,[I - \frac{B_\theta(y)\,p\,e^*}{b^*_\theta(y)\,p}]\,\frac{\partial_k B_\theta(y)\,p}{b^*_\theta(y)\,p}$$

$$+\ \frac{\partial_k Q^*_\theta\,B_\theta(y)\,p}{b^*_\theta(y)\,p}\ .$$

The sequence $\partial p^\theta_n = (\partial_k p^\theta_n)$ belongs to Σ^p, where

$$\Sigma = \{w \in \mathbf{R}^N\ :\ e^*\,w = 0\}\ ,$$

and $e = (1, \cdots, 1)^*$.

It has been proved in LeGland and Mevel [7], that under the probability measure corresponding to the *true* value α of the parameter, the extended Markov chain $\{Z_n^\theta = (X_n, Y_n, p_n^\theta, \partial p_n^\theta)\,,\, n \geq 0\}$ with values in $E \stackrel{\Delta}{=} S \times \mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p$, is geometrically ergodic under suitable integrability assumptions. The existence and uniqueness of the invariant measure μ_θ for the extended Markov chain $\{Z_n^\theta\,,\, n \geq 0\}$ is used below to prove the convergence of the two recursive estimators.

2 Recursive identification algorithm

In LeGland and Mevel [8], we have proved the consistency and the asymptotic normality of the non–recursive MLE and CLSE, under suitable integrability conditions stated therein. In this section, we will present the recursive estimators, and prove their consistency.

Definition 2.1 *Let L_Θ denote the set of functions $g_\theta = (g_\theta^k)$ defined on $\mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p$, such that for any $k = 1, \cdots, p$, and any $y \in \mathbf{R}^d$, the partial mapping $(p, w) \mapsto g_\theta^k(y, p, w)$ is locally Lipschitz continuous, in the sense that*

$$|g_\theta^k(y,p,w) - g_\theta^k(y,p',w')| \leq \mathrm{Lip}(g,y)\,[\,\|w_k - w_k'\| + \|p - p'\|\,(1 + \|w_k\| + \|w_k'\|\,)\,]$$

for any $p, p' \in \mathcal{P}(S)$, and any $w = (w_k), w' = (w_k') \in \Sigma^p$, and

$$|g_\theta^k(y,p,w)| \leq \mathrm{K}(g,y)\,(1 + \|w_k\|\,)\,,$$

for any $p \in \mathcal{P}(S)$, and any $w = (w_k) \in \Sigma^p$.

Moreover, for any $k = 1, \cdots, p$, any $y \in \mathbf{R}^d$, any $p \in \mathcal{P}(S)$, and any $w = (w_k) \in \Sigma^p$, the partial mapping $\theta \mapsto g_\theta^k(y,p,w)$ is locally Lipschitz continuous, in the sense that for any $\theta, \theta' \in \Theta_\varepsilon$,

$$|g_\theta^k(y,p,w) - g_{\theta'}^k(y,p,w)| \leq \mathrm{Lip}_\Theta(g,y)\,(1 + \|w_k\|)\,\|\theta - \theta'\|\,.$$

With the above notations, we consider the recursive algorithm defined as follows : for all $k = 1, \cdots, p$

$$\begin{aligned} \widehat{\theta}_{n+1}^k &= \pi_\varepsilon(\widehat{\theta}_n^k + \gamma_{n+1}\,\Psi_n(H_{\widehat{\theta}_n}^k\,(Y_n, \widehat{p}_n, \widehat{w}_n))\,)\,, \\ \bar{\theta}_{n+1} &= \bar{\theta}_n + \frac{1}{n}\,(\widehat{\theta}_{n+1} - \bar{\theta}_n)\,, \end{aligned} \tag{2}$$

where $\gamma_n = n^{-2/3}$, $\pi_\varepsilon(\cdot)$ is the projection on the set $\Theta_\varepsilon = \{\theta \in \Theta : q_\theta^{i,j} \geq \varepsilon \text{ ,for all } i,j \in S\}$, $\Psi_n(\cdot)$ is a projection on a ball of growing radius (hence made only a finite number of times), and

$$\begin{aligned} \widehat{p}_{n+1} &= f_{\widehat{\theta}_{n+1}}[Y_n, \widehat{p}_n]\,, \\ \widehat{w}_{n+1}^k &= \Phi_{\widehat{\theta}_{n+1}}[Y_n, \widehat{p}_n]\,\widehat{w}_n^k + \partial_k f_{\widehat{\theta}_{n+1}}[Y_n, \widehat{p}_n]\,. \end{aligned}$$

We define $\widehat{Z}_n = (X_n, Y_n, \widehat{p}_n, \widehat{w}_n)$. Then

$$\mathbf{P}[\,\widehat{Z}_{n+1} \in B \mid \widehat{\theta}_0, \cdots, \widehat{\theta}_{n+1}, \widehat{Z}_0, \cdots, \widehat{Z}_n\,] = \Pi_{\widehat{\theta}_{n+1}}(\widehat{Z}_n, B)\,,$$

i.e. the algorithm (2) belongs to the class of stochastic algorithms with Markovian dynamics, see Benveniste, Métivier and Priouret [2].

2.1 The recursive MLE

It has been proved in LeGland and Mevel [8] that the log–likelihood function (suitably normalized) for the estimation of the parameter θ based on observations $(Y_0, \cdots, Y_n)$ can be expressed as an additive functional of the extended Markov chain $\{Z_n^\theta\,,\, n \geq 0\}$ as follows

$$\ell_n(\theta) = \frac{1}{n} \sum_{k=0}^{n} \log[\,b_\theta^*(Y_k)\,p_k^\theta\,]\,.$$

In addition, we have proved that the following strong law of large numbers holds :

Proposition 2.2 *Under Assumption A, and if Δ_1 and Γ_1 are finite, then for any $\theta \in \Theta$ there exists a finite constant $\ell(\theta)$ such that*

$$\ell_n(\theta) \longrightarrow \ell(\theta)\,, \qquad \mathbf{P}\text{-}a.s.$$

as $n \to \infty$, where

$$\ell(\theta) = \int_{\mathbf{R}^d \times \mathcal{P}(S)} \log[\,b_\theta^*(y)\,p\,]\;\nu_\theta(dy, dp)\,,$$

and where ν_θ denotes the marginal on $\mathbf{R}^d \times \mathcal{P}(S)$ of the invariant measure μ_θ.

Under the above assumptions, we define for any $\theta \in \Theta$ the Kullback Leibler information

$$K_{\mathrm{ML}}(\theta) = -[\ell(\theta) - \ell(\alpha)] \geq 0\,,$$

and we have proved that the true value α of the parameter belongs to the set M_{ML} of global minima of K_{ML}

$$M_{\mathrm{ML}} = \operatorname*{argmin}_{\theta \in \Theta} K_{\mathrm{ML}}(\theta) \supseteq \{\alpha\}\,.$$

The maximum likelihood estimator (MLE) is defined as

$$\widehat{\theta}_n^{\mathrm{ML}} \in \operatorname*{argmax}_{\theta \in \Theta_\varepsilon} \ell_n(\theta)\,.$$

Theorem 2.3 *Under Assumptions A, and B, and if Δ_2, Δ_2', Γ_1 and Γ_2 are finite, then any MLE sequence $\{\widehat{\theta}_n^{\mathrm{ML}}\,,\, n \geq 0\}$ converges in probability to the deterministic set M_{ML} as $n \to \infty$.*

The recursive maximum likelihood estimator (RMLE) $\widehat{\theta}_n^{\mathrm{RML}} = (\widehat{\theta}_n^k)$, is defined by the following recursive algorithm : for all $k = 1, \cdots, p$

$$\widehat{\theta}_{n+1}^k = \pi_\varepsilon(\widehat{\theta}_n^k + \gamma_{n+1}\,\Psi_n(H_{\widehat{\theta}_n^{\mathrm{RML}}}^k(Y_n, \widehat{p}_n, \widehat{w}_n))\,)\,,$$

where for any $\theta \in \Theta$, and any $k = 1, \cdots, p$, the function H_θ^k is defined by

$$H_\theta^k(y,p,w) = \frac{b_\theta^*(y)\, w_k}{b_\theta^*(y)\, p} + \frac{\partial_k\, b_\theta^*(y)\, p}{b_\theta^*(y)\, p} \ ,$$

for any $y \in \mathbf{R}^d$, any $p \in \mathcal{P}(S)$, and any $w = (w_k) \in \Sigma^p$, and belongs to the set L_Θ.

For any $\theta \in \Theta$, and any $k = 1, \cdots, p$, we define

$$\begin{aligned} h_{\mathrm{ML}}^k(\theta) &= \int_{\mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p} H_\theta^k(y,p,w)\, \lambda_\theta(dy,dp,dw) \\ &= \partial_k K_{\mathrm{ML}}(\theta) \ , \end{aligned}$$

where λ_θ denotes the marginal on $\mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p$ of the invariant measure μ_θ. We define the set of stationary points of the Kullback–Leibler information K_{ML}

$$L_{\mathrm{ML}} = \{\theta \in \Theta \ : \ h_{\mathrm{ML}}(\theta) = 0\} \ .$$

Under the following two additional assumptions :

Assumption C : The function K_{ML} is of class C^p.

Assumption D : There exists ε', such that $\varepsilon > \varepsilon' > 0$ and :

(i) $L_{\mathrm{ML}} \subset \Theta_\varepsilon$,

(ii) For any $\theta \in \Theta_{\varepsilon'} \setminus \Theta_\varepsilon$

$$h_{\mathrm{ML}}^*(\theta)\, (\pi_\varepsilon(\theta) - \theta) \geq \delta\, |\pi_\varepsilon(\theta) - \theta\,| \ ,$$

for some $\delta > 0$.

the following result holds, which is based on Delyon [4], and Delyon and Iouditski [5].

Theorem 2.4 *Under Assumptions A, A', B, B', C, and D, and if Δ_8, Δ_4', Δ_2'', Δ_1''', and Γ_2 are finite, then the RMLE sequence $\{\widehat{\theta}_n^{\mathrm{RML}}\, , n \geq 0\}$ converges $\mathbf{P}$–a.s. to the deterministic set L_{ML} as $n \to \infty$, and assuming that $\widehat{\theta}_n^{\mathrm{RML}}$ converges to α, then*

$$\sqrt{n}\, (\bar{\theta}_n^{\mathrm{RML}} - \alpha) \Rightarrow \mathcal{N}(0, I_{\mathrm{F}}^{-1}) \ ,$$

where

$$I_{\mathrm{F}} = \partial^2 K_{\mathrm{ML}}(\alpha) \ ,$$

is the Fisher information matrix.

2.2 The recursive CLSE

We define the following (suitably normalized) prediction error criterion

$$e_n(\theta) \triangleq \frac{1}{2n} \sum_{k=1}^{n} |\, Y_k - \mathbf{E}^\theta[Y_k \mid \mathcal{Y}_{k-1}]\,|^2 \ .$$

Introducing the $N \times d$ matrix $\phi_\theta = (\phi_\theta^i)$, where for any $i \in S$, the mean vector $\phi_\theta^i \in \mathbf{R}^d$ is defined by

$$\phi_\theta^i = \int_{\mathbf{R}^d} y\, b_\theta^i(y)\, \lambda(dy) \ ,$$

(assuming the integral exist), the prediction error criterion can be expressed as an additive functional of the extended Markov chain $\{Z_n^\theta\, , n \geq 0\}$ as follows

$$e_n(\theta) = \frac{1}{2n} \sum_{k=1}^{n} |Y_k - \phi_\theta^*\, p_k^\theta\,|^2 \ .$$

In addition, we have proved in LeGland and Mevel [8] that the following law of large numbers holds :

Proposition 2.5 *Under Assumptions A, B, and B', and if Δ_1 and $\bar{Y}^2$ are finite, then for any $\theta \in \Theta$ there exists a finite constant $e(\theta)$ such that*

$$e_n(\theta) \longrightarrow e(\theta) \ , \qquad \mathbf{P}\text{–}a.s.$$

as $n \to \infty$, where

$$e(\theta) = \tfrac{1}{2} \int_{\mathbf{R}^d \times \mathcal{P}(S)} |y - \phi_\theta^*\, p\,|^2\, \nu_\theta(dy,dp) \ ,$$

and where ν_θ denotes the marginal on $\mathbf{R}^d \times \mathcal{P}(S)$ of the invariant measure μ_θ.

Under the above assumptions, we define for any $\theta \in \Theta$ the prediction error contrast function

$$K_{\mathrm{CLS}}(\theta) = e(\theta) - e(\alpha) \geq 0 \ ,$$

and we have proved that the true value α of the parameter belongs to the set M_{CLS} of global minima of K_{CLS}

$$M_{\mathrm{CLS}} = \operatorname*{argmin}_{\theta \in \Theta} K_{\mathrm{CLS}}(\theta) \supseteq \{\alpha\} \ .$$

The conditional least square estimator (CLSE) is defined as

$$\widehat{\theta}_n^{\mathrm{CLS}} \in \operatorname*{argmin}_{\theta \in \Theta_\varepsilon} e_n(\theta) \ .$$

Theorem 2.6 *Under Assumptions A, B, and B', and if $\bar{Y}^4$ is finite, then any CLSE sequence $\{\widehat{\theta}_n^{\mathrm{CLS}}\, , n \geq 0\}$ converges in probability to the deterministic set M_{CLS} as $n \to \infty$.*

The recursive conditional least squares estimator (RCLSE) $\widehat{\theta}_n^{\mathrm{RCLS}} = (\widehat{\theta}_n^k)$, is defined by the following recursive algorithm : for any $k = 1, \cdots, p$

$$\widehat{\theta}_{n+1}^k = \pi_\varepsilon(\widehat{\theta}_n^k + \gamma_{n+1}\, \Psi_n(H_{\widehat{\theta}_n^{\mathrm{RCLS}}}^k(Y_n, \widehat{p}_n, \widehat{w}_n))) \ ,$$

where for any $\theta \in \Theta$, and any $k = 1, \cdots, p$, the function H_θ^k is defined by

$$H_\theta^k(y,p,w) = [\, \phi_\theta\, (y - \phi_\theta^*\, p)\,]^*\, w_k + [\, \partial_k \phi_\theta\, (y - \phi_\theta^*\, p)\,]^*\, p \ .$$

for any $y \in \mathbf{R}^d$, any $p \in \mathcal{P}(S)$, and any $w = (w_k) \in \Sigma^p$, and belongs to the set L_Θ.

For any $\theta \in \Theta$, and any $k = 1, \cdots, p$, we define

$$h^k_{\mathrm{CLS}}(\theta) = \int_{\mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p} H^k_\theta(y,p,w)\, \lambda_\theta(dy,dp,dw)$$

$$= \partial_k K_{\mathrm{CLS}}(\theta) \ ,$$

where λ_θ denotes the marginal on $\mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p$ of the invariant measure μ_θ. We define the set of stationary points of the prediction error contrast function K_{CLS}

$$L_{\mathrm{CLS}} = \{\theta \in \Theta \ : \ h_{\mathrm{CLS}}(\theta) = 0\} \ .$$

Adapting Assumptions C and D to the case of the RCLSE, we obtain the counterpart of Theorem 2.4.

Theorem 2.7 *Under Assumptions A, B, B', C, and D, and if* Δ_8, Δ'_4, Δ''_2, Δ'''_1, *and* $\bar{Y}^4$ *are finite, then the RCLSE sequence* $\{\widehat{\theta}^{\mathrm{RCLS}}_n, n \geq 0\}$ *converges* $\mathbf{P}$*–a.s. to the deterministic set* L_{CLS} *as* $n \to \infty$, *and assuming that* $\widehat{\theta}^{\mathrm{RCLS}}_n$ *converges to* α, *then*

$$\sqrt{n}\,(\bar{\theta}^{\mathrm{RCLS}}_n - \alpha) \Rightarrow \mathcal{N}(0, V^{-1}) \ ,$$

where

$$V = \partial^2 K_{\mathrm{CLS}}(\alpha)\, R^{-1}\, \partial^2 K_{\mathrm{CLS}}(\alpha) \ ,$$

and

$$R = \lim_{n \to \infty} n\, \mathbf{E}[\, \partial e_n(\alpha)\, \partial e^*_n(\alpha)\,] \ .$$

3 Numerical example

Our example consists in a three–state Markov chain observed in white noise, i.e.

$$Y_n = h(X_n) + W_n \ ,$$

where $\{W_n \, , \, n \geq 0\}$ is an $\mathcal{N}(0,1)$ i.i.d. sequence, and

$$h = (\, 5 \ \ 3 \ \ 7 \,) \ , \qquad Q = \begin{pmatrix} 0.85 & 0.08 & 0.07 \\ 0.4 & 0.5 & 0.1 \\ 0.35 & 0.05 & 0.6 \end{pmatrix} \ .$$

The initial estimates of the transition probability matrix entries are set uniformly at $1/N$, and the initial estimate of the vector h are randomly chosen in an appropriate compact set, depending on prior knowledge about the true value of the parameter. Firstly, assume we know h exactly. In Figure 1 and Figure 2, we show the respective performance of both estimators. We display the evolution of both estimators, on the basis of the same observation sequence (1 unit = 10 observations). The theoretical better rate of convergence of the RMLE over the RCLSE is obviously demonstrated, by comparaison of the evolution of both estimators. In Figure 3, we construct, for each estimator, an empirical confidence tube, based on a 50–experiment average. The tube is obtained by using the estimates of the covariance matrix of each estimator. The tube is centered on the difference between θ_n and α. The fact that the zero value belongs to both tubes confirms the convergence of both estimator to α. The diameter of both tubes goes to zero, with a rate which is proportionnal to the estimate of the corresponding asymptotic covariance matrix. The RMLE tube is much thinner than the RCLSE tube, which shows the asymptotic covariance matrix of the RMLE to be the smallest.

Secondly, we relax the assumed knowledge about h. In the experience displayed in Figure 4, we investigate the effect of knowing the vector h or not, on the estimation of the entries of the transition matrix, for both estimators. In the case of the RCLSE, the lack of knowledge about h results in a clear degradation of the performance of the estimator. On the contrary, the RMLE is less sensible to the lack of knowledge about h, and seems to be a more robust algorithm in this respect (1 unit = 10 observations). Notice that the fast estimation of h by the RMLE as seen in Figure 5, illustrates the robustness of the RMLE (compare with the very slow estimation of h by the RCLSE). More precisely, using the same confidence tube construction as above, we compare the behaviour of the RMLE, for the estimation of Q. Zero belongs to both confidence tubes, and their diameters goe to zero. In that worst situation, the RMLE seems more robust, see Figure 6. On the contrary, the behaviour of the RCLSE is less robust.

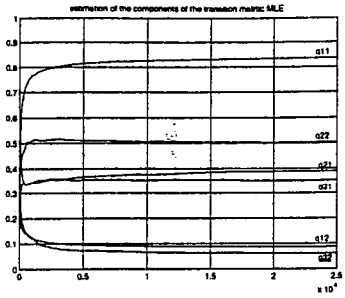
In Figure 7, we compare the adaptive prediction filter $\widehat{p}_n$ and the prediction filter p^α_n for the *true* value α of the parameter, and we see that the difference between both probability distributions goes to zero as n tends to infinity.

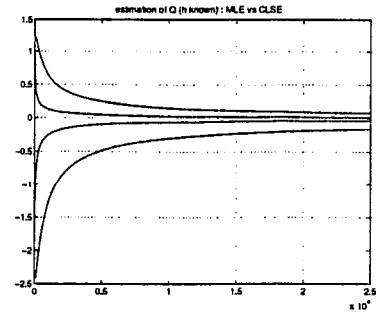
References

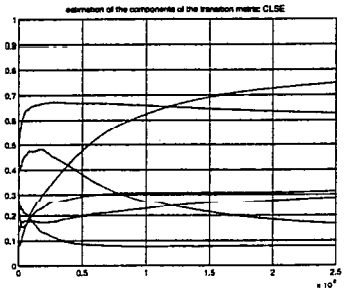
[1] A. ARAPOSTATHIS and S.I. MARCUS. Analysis of an identification algorithm arising in the adaptive estimation of Markov chains. *Mathematics of Control, Signals, and Systems*, 3(1):1–29, 1990.

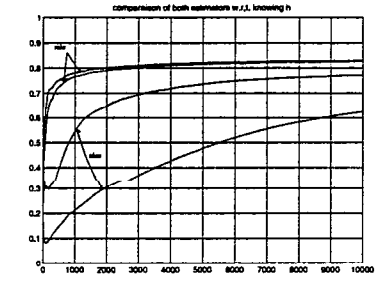
[2] A. BENVENISTE, M. MÉTIVIER, and P. PRIOURET. *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Applications of Mathematics*. Springer Verlag, New York, 1990.

[3] P.J. BICKEL, Y. RITOV, and T. RYDÉN. Asymptotic normality of the maximum–likelihood estimator for general hidden Markov models. Research Report 1997:2, Department of Mathematical Statistics, Lund University, 1997.









[4] B. DELYON. General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control*, AC–41(9):1245–1255, September 1996.

[5] B. DELYON and A. IOUDITSKI. Stochastic approximation with averaging. Publication Interne 952, IRISA, October 1995. `ftp://ftp.irisa.fr/techreports/1995/PI-952.ps.gz`.

[6] F. LE GLAND and L. MEVEL. Recursive identification of HMM's with observations in a finite set. In *Proceedings of the 34th Conference on Decision and Control, New Orleans 1995*, pages 216–221. IEEE–CSS, December 1995.

[7] F. LE GLAND and L. MEVEL. Geometric ergodicity in hidden Markov models. Publication Interne 1028, IRISA, July 1996. `ftp://ftp.irisa.fr/techreports/1996/PI-1028.ps.gz`.

[8] F. LE GLAND and L. MEVEL. Asymptotic behaviour of the maximum likelihood estimator and the conditional least squares estimator in hidden Markov models. Publication interne, IRISA, 1997. (to appear).

[9] B.G. LEROUX. Maximum–likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127–143, 1992.

[10] T. PETRIE. Probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115, 1969.

[11] T. RYDÉN. On recursive estimation for hidden Markov models. *Stochastic Processes and their Applications*, 66(1):79–96, 1997.