Statistical Inference for Stochastic Processes **4:** 273–282, 2001.

On Determination of the Order of a Markov Chain

L. C. ZHAO[1], C. C. Y. DOREA[2] and C. R. GONÇALVES[3]
[1]*University of Science and Technology of China, Hefei, China*
[2,3]*Universidade de Brasília, Brasília-DF, Brazil*

Abstract. In this paper we propose a new and more general criterion (the efficient determination criterion, EDC) for estimating the order of a Markov chain. The consistency and the strong consistency of the estimates have been established under mild conditions.

AMS Mathematics Subject Classifications: 60J10, 62M05, 62F12.

Key words: Markov chain order, Bayesian information criterion, efficient determination criterion.

1. Introduction

Multiple Markov chains (MC), i.e., chains of higher order, have been used *inter alia* for fitting precipitation data in meteorological studies [4, 8, 10], for modelling repeated patterns in a DNA sequence [11, 14] and for human speech studies [5]. One key issue that is relevant to statistical problems is how to determine the order of the chain. Several alternatives to nested hypothesis testing techniques have been proposed to determine the dimension of the model. Tong [13] considered the problem of determining the order of the chain using AIC criterion proposed by Akaike [1]. Gates and Tong [8] and Chin [4] applied the AIC procedure to the MC modelling of meteorological data. Katz [10] studied the asymptotic distribution of the AIC estimator and pointed out the inconsistency of such an estimator. He suggested the Bayesian information criterion (BIC) proposed by Schwarz [12] as an alternative to the AIC criterion and showed that it is consistent for estimating the order of a MC.

In this note we propose a new and more general criterion that will allow us to choose the penalty function over a wide range and, under mild conditions, the strong consistency of such procedures can be established. This criterion was first used in Zhao et al. [15, 16] to detecting the number of signals, and it was called EDC criterion (efficient determination criterion) by Bai et al. [2]. They considered the model

$$X_t = a_1 S_t^{(1)} + \cdots + a_r S_t^{(r)} + N_t,$$

where $(S_t^{(1)}, \ldots, S_t^{(r)})$ is the signal vector, $(a_1, \ldots, a_r)$ a parameter vector, and N_t a random noise. An important problem in this setting is that of detecting the number r of signals transmitted. Simulation results comparing AIC, BIC and EDC criteria

showed that the EDC criterion outperformed the others (see Remark 2.3(c)). Assume now that the randomness is included in the r previous signals and consider the model for discrete times $t = 0, 1, \ldots$

$$X_n = a_1 X_{n-r} + \cdots + a_r X_{n-1}.$$

Clearly this yields an MC of order r and this was the basic motivation to extend the EDC criterion to estimate the order of the chain.

2. Main Results

Let $X = \{X_n\}_{n \geqslant 0}$ be a stochastic process with state space $S = \{1, \ldots, N\}$ that satisfies

$$P\left(X_n = i_n \mid X_j = i_j,\, j < n\right) = q_{i_{n-r} \ldots i_{n-1}, i_n}, \tag{2.1}$$

then X is said to be an rth order MC with (stationary) transition matrix $Q = \left(q_{i_1 \ldots i_r, i_{r+1}}\right)$. From the process X one can derive a first order MC, $Y = \{Y_n\}_{n \geqslant 0}$ by setting $Y_n = (X_n, \ldots, X_{n+r-1})$ and for $v = (i_1, \ldots, i_r)$ and $w = \left(i_1^{'}, \ldots, i_r^{'}\right)$

$$P(Y_{n+1} = w \mid Y_n = v) = \tilde{q}_{vw} = \begin{cases} q_{i_1 \ldots i_r, i_r^{'}} & \text{if} \quad i_j^{'} = i_{j+1},\ j = 1, \ldots, r-1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Clearly Y is a first order and homogeneous MC with transition matrix $\tilde{Q} = (\tilde{q}_{vw})$ and state space S^r (see, e.g. [3]). We call Y the derived process and in what follows we assume that Y is an irreducible and positive recurrent MC. Then it is well known that $\tilde{Q}$ possesses a unique stationary distribution, say $\pi = (\pi_{i_1 \ldots i_r})$. Furthermore for $v = (i_1, \ldots, i_r)$ and $w = (i_2, \ldots, i_{r+1})$, we have

$$\sum_v \pi_v \tilde{q}_{vw} = \sum_{i_1} \pi_{i_1 \ldots i_r} q_{i_1 \ldots i_r, i_{r+1}} = \pi_{i_2 \ldots i_{r+1}}. \tag{2.3}$$

Now let $X_0, X_1, \ldots, X_n$ be a sample taken from a MC of order r. Assume that the true order r is known *a priori* to be less than or equal to some fixed upper bound R. To determine the order of the chain we cast the problem in the framework of model selection. Let H_t denote the hypothesis that the MC is of order t, where $t = 0, \ldots, R$. Note that a MC of order 0 is nothing but a sequence of iid random variables. The following notation will be used

$$n_{i_1 \cdots i_s} = \sum_{k=0}^{n-s+1} I(X_k = i_1, \ldots, X_{k+s-1} = i_s) \tag{2.4}$$

and for the derived process Y

$$\tilde{n}_{v_1 \cdots v_s} = \sum_{k=0}^{n-r-s+1} I(Y_k = v_1, \ldots, Y_{k+s-1} = v_s), \tag{2.5}$$

where $I(\cdot)$ stands for the indication function. Taking the transition matrix $Q = (q_{i_1 \cdots i_t, i_{t+1}})$ as the unknown parameter matrix, under model H_t, the maximum log-likelihood function will be given by

$$\log L_t = \sum_{i_1 \cdots i_{t+1}} n_{i_1 \cdots i_{t+1}} \log \frac{n_{i_1 \cdots i_{t+1}}}{n_{i_1 \cdots i_t}}, \tag{2.6}$$

where we assume $0/0 = 0$ and $0(\infty) = 0$.

According to the AIC criterion, r is estimated by $\hat{r}$ where $\hat{r}$ is chosen so that

$$\mathrm{AIC}(\hat{r}) = \min \{\mathrm{AIC}(t), t = 0, 1, \ldots, R\} \tag{2.7}$$

and

$$\mathrm{AIC}(t) = -2 \log L_t + 2\gamma(t). \tag{2.8}$$

Here $\gamma(t)$ denotes the number of free parameters that have to be estimated under H_t. In the context of high-order MC, $\gamma(t) = N^t(t-1)$. According to the BIC criterion, r is estimated by $\hat{r}$ where $\hat{r}$ is chosen so that

$$\mathrm{BIC}(\hat{r}) = \min \{\mathrm{BIC}(t), t = 0, 1, \ldots, R\} \tag{2.9}$$

and

$$\mathrm{BIC}(t) = -2 \log L_t + \gamma(t) \log n. \tag{2.10}$$

In this paper we consider a more general criterion for the estimation of r. Where r is estimated by $\hat{r}$ such that

$$\mathrm{EDC}(\hat{r}) = \min \{\mathrm{EDC}(t), t = 0, 1, \ldots, R\} \tag{2.11}$$

and

$$\mathrm{EDC}(t) = -2 \log L_t + \gamma(t) c_n. \tag{2.12}$$

Here c_n, in the penalty term, can be taken as a sequence of positive numbers depending on n, or more generally, as a sequence of positive random variables. And, $\gamma(t)$ can be taken as any strictly increasing function of t. The advantage of this criterion is that we can flexibly choose $\{c_n\}$ in a wide range to meet various requests. For example, to assure that $\hat{r}$ is strongly consistent

$$\frac{c_n}{n} \underset{n}{\longrightarrow} 0 \text{ a.s.} \quad \text{and} \quad \frac{c_n}{\log \log n} \underset{n}{\longrightarrow} \infty \quad \text{a.s..} \tag{2.13}$$

In order for $\hat{r}$ to be a (weakly) consistent estimator of r it suffices to take c_n satisfying

$$\frac{c_n}{n} \xrightarrow[n]{p} 0 \quad \text{and} \quad c_n \xrightarrow[n]{p} \infty. \tag{2.14}$$

We have the following two main results:

THEOREM 2.1. *Assume that the derived process* $\{Y_n\}$ *is irreducible and positive recurrent and that* (2.13) *is satisfied, then* $\hat{r}$ *defined by* (2.11) *and* (2.12) *is a strongly consistent estimator of* r.

THEOREM 2.2. *Assume that the derived process* $\{Y_n\}$ *is irreducible and positive recurrent and that* (2.14) *is satisfied, then* $\hat{r}$ *defined by* (2.11) *and* (2.12) *is a consistent estimator of* r.

Remark 2.3. (a) We note that the conclusions of Theorems 2.1 and 2.2 are valid whatever initial distribution the true model has. This note also applies to all conclusions in the sequel.

(b) Carefully checking the arguments of Theorem 2.2, one can find that, if we assume that $\{c_n\}$ is uniformly bounded by a constant c instead of the condition (2.14), then $\hat{r}$ is inconsistent. Similarly, in Theorem 2.1, if we assume that

$$\frac{c_n}{n} \underset{n}{\longrightarrow} 0 \quad \text{a.s.} \quad \text{and} \quad \liminf_{n\to\infty} \frac{c_n}{\log\log n} = 0 \quad \text{a.s.},$$

instead of the condition (2.13), then $\hat{r}$ is not strongly consistent.

(c) A natural problem is how to choose the sequence c_n that will produce the most efficient estimator $\hat{r}$. This is a difficult theoretical problem that remains to be solved. But simulation results by Bai et al. [2] indicate that the frequencies of error detection were considerably smaller when the EDC criterion was used. The simulations were carried out for Gaussian signals and compared the AIC criterion $c_n^{(1)} = c_1$, the BIC criterion $c_n^{(2)} = c_2 \log n$ and the EDC criteria $c_n^{(3)} = c_3 n/\log n$ and $c_n^{(4)} = c_4\sqrt{n}$, where $c_1, \ldots, c_4$ are properly chosen constants. It was shown that when the sample size n is larger than 100, $c_n^{(3)}$ and $c_n^{(4)}$ work very well and that the frequencies of error detection decrease very rapidly with increasing sample size. In fact, the simulations clearly indicate that $c_n^{(2)}$ worked better than $c_n^{(1)}$ and that $c_n^{(3)}$ and $c_n^{(4)}$ were much better than $c_n^{(2)}$.

3. Proof of the Results

To prove the main theorems we shall need some auxiliary results. First, note that if X is an irreducible and positive recurrent MC of order 1, then it possesses a unique stationary distribution $\pi = (\pi_i)_{i\in S}$. Moreover, since every state i is positive recurrent we also have $\pi_i > 0$.

LEMMA 3.1. *Assume that* X *is an irreducible and positive recurrent MC of order* 1 *then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} I\,(X_k = i) = \pi_i, \quad \text{a.s.} \tag{3.1}$$

and for $s \geqslant 1$

$$\lim_{n\to\infty} \frac{n_{i_0\cdots i_s}}{n} = \pi_{i_0} q_{i_0 i_1} \cdots q_{i_{s-1} i_s}, \quad \text{a.s.}. \tag{3.2}$$

Moreover, for $s \geqslant 2$ if $q_{i_0 i_1} \ldots q_{i_{s-2} i_{s-1}} > 0$ then as $n \to \infty$

$$\frac{n_{i_0\cdots i_s}}{n_{i_0\cdots i_{s-1}}} - q_{i_{s-1} i_s} = O\left(\sqrt{\frac{\log\log n}{n}}\right), \quad \text{a.s.}. \tag{3.3}$$

Proof. (a) From the SLLN for MC (see, e.g. [7]) we have (3.1). To prove (3.2), write

$$U_k = I(X_k = i_0, \ldots, X_{k+s} = i_s) - q_{i_{s-1} i_s} I(X_k = i_0, \ldots, X_{k+s-1} = i_{s-1}). \tag{3.4}$$

It is easily seen that $\{U_k\}$ is a martingale difference sequence with respect to the sequence σ-algebras $\mathcal{F}_k = \sigma(X_0, \ldots, X_{k+s})$. By the martingale convergence theorem, $\sum_{k=1}^{\infty} U_k/k$ converges with probability one and by Kronecker's lemma we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} U_k = 0, \quad \text{a.s.}. \tag{3.5}$$

Now (3.2) follows from (3.1) and (3.5) by induction.

(b) If $q_{i_{s-1} i_s} = 0$ or 1 then (3.3) is trivial. Now assume that $0 < q_{i_{s-1} i_s} < 1$ and write

$$V_{ns}^2 = q_{i_{s-1} i_s}(1 - q_{i_{s-1} i_s}) \sum_{k=0}^{n-s} I\left(X_k = i_0, \ldots, X_{k+s-1} = i_{s-1}\right).$$

By (3.2), as $n \to \infty$

$$\frac{V_{ns}^2}{n} \to q_{i_{s-1} i_s}(1 - q_{i_{s-1} i_s}) \pi_{i_0} q_{i_0 i_1} \ldots q_{i_{s-2} i_{s-1}} > 0, \quad \text{a.s.}. \tag{3.6}$$

Since $S_n = \sum_{k=1}^{n} U_k$ is a zero-mean and square-integrable martingale we may use Theorem 4.7 and Theorem 4.8 (Hall and Heyde, 1980) by taking $Z_n = 1$, $X_k = U_k$ and $W_n = V_{ns}$. Condition (4.35) is trivially satisfied since $I(|U_k| > Z_k) = 0$. Condition (4.36) is verified by observing that

$$\begin{aligned} &E(U_k^2 I(|U_k| \leqslant Z_k) \mid \mathcal{F}_{k-1}) \\ &\quad = q_{i_{s-1} i_s}(1 - q_{i_{s-1} i_s}) I(X_k = i_0, \ldots, X_{k+s-1} = i_{s-1}). \end{aligned}$$

Similarly one verifies condition (4.37), and (4.38) follows from (3.6). Thus from Theorem 4.8, we have

$$\limsup_{n\to\infty} \left| \sum_{k=1}^{n} U_k / \sqrt{2 V_{ns}^2 \log\log V_{ns}^2} \right| = 1, \quad \text{a.s.} \tag{3.7}$$

and from (3.4) and (3.6) we have

$$\frac{n_{i_0\cdots i_s} - q_{i_{s-1}i_s} n_{i_0\cdots i_{s-1}}}{\sqrt{n \log\log n}} = O(1), \quad \text{a.s..}$$

And (3.3) follows using (3.2).

LEMMA 3.2. *Assume that* $p_{nl} > 0$, $p_l > 0, l = 1, \ldots, m, n = 1, 2, \ldots$. *Moreover,* $\sum_{l=1}^{m} p_{nl} = \sum_{l=1}^{m} p_l = 1$ *and*

$$\lim_{n\to\infty} p_{nl} = p_l, \quad l = 1, \ldots, m.$$

Then as $n \to \infty$ *we have*

$$\sum_{l=1}^{m} p_{nl} (\log p_{nl} - \log p_l) = \sum_{l=1}^{m} \frac{(p_{nl} - p_l)^2}{2p_{nl}} (1 + o(1)).$$

The proof is simple and is omitted.

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Suppose that r is the true order of X and that $Q = \left(q_{i_1\ldots i_r, i_{r+1}}\right)$ is the true transition matrix.

(a) Assume that $t \geqslant r \geqslant 1$ and let $s = t - r$. Define

$$\hat{q}_{i_1\ldots i_t, i_{t+1}} = \frac{n_{i_1\ldots i_{t+1}}}{n_{i_1\ldots i_t}}.$$

Note that for $v_1 = (i_1, \ldots, i_r), v_2 = (i_2, \ldots, i_{r+1}), \ldots, v_{s+2} = (i_{s+2}, \ldots, i_{t+1})$ we have for the derived process Y:

$$\hat{q}_{i_1\ldots i_t, i_{t+1}} = \frac{\tilde{n}_{v_1\ldots v_{s+2}}}{\tilde{n}_{v_1\ldots v_{s+1}}} \quad \text{with} \quad \tilde{n}_{v_1\cdots v_s} = \sum_{k=0}^{n-r-s+1} I(Y_k = v_1, \ldots, Y_{k+s-1} = v_s). \tag{3.8}$$

Since Y is an irreducible and positive recurrent chain we can apply Lemma 3.1 if $n_{i_1\ldots i_{t+1}} > 0$ for some n,

$$\lim_{n\to\infty} \hat{q}_{i_1\ldots i_t, i_{t+1}} = \tilde{q}_{v_{s+1}v_{s+2}} = q_{i_{s+1}\ldots i_t, i_{t+1}} > 0, \quad \text{a.s..} \tag{3.9}$$

From (2.6) we have,

$$\begin{aligned}
\log L_t &= \sum_{i_1\ldots i_t} n_{i_1\ldots i_t} \sum_{i_{t+1}} \hat{q}_{i_1\ldots i_t, i_{t+1}} \log \hat{q}_{i_1\ldots i_t, i_{t+1}} \\
&= \sum_{i_1\ldots i_{t+1}} n_{i_1\ldots i_{t+1}} \log q_{i_{s+1}\ldots i_t, i_{t+1}} + \\
&\quad + \sum_{i_1\ldots i_t} n_{i_1\ldots i_t} \sum_{i_{t+1}} \hat{q}_{i_1\ldots i_t, i_{t+1}} \left[\log \hat{q}_{i_1\ldots i_t, i_{t+1}} - \log q_{i_{s+1}\ldots i_t, i_{t+1}}\right].
\end{aligned}$$

Since by convention $0\,(-\infty) = 0$ in the above summation one needs to consider only those terms for which $n_{i_1\ldots i_{t+1}} > 0$. Using (3.9) and applying Lemma 3.2 with $p_{nl} = \hat{q}_{i_1\ldots i_t,l}$ and $p_l = q_{i_{s+1}\ldots i_t,l}$ we can write as $n \to \infty$

$$\begin{aligned}\frac{2}{n}\log L_t &= \frac{2}{n}\sum_{i_1\ldots i_{t+1}} n_{i_1\ldots i_{t+1}} \log q_{i_{s+1}\ldots i_t,i_{t+1}} + \\ &\quad + \frac{1}{n}\sum_{i_1\ldots i_t} n_{i_1\ldots i_t} \sum_{i_{t+1}} \frac{(\hat{q}_{i_1\ldots i_t,i_{t+1}} - q_{i_{s+1}\ldots i_t,i_{t+1}})^2}{\hat{q}_{i_1\ldots i_t,i_{t+1}}}(1+o(1)), \quad \text{a.s..}\end{aligned} \tag{3.10}$$

Note that

$$\begin{aligned}\frac{2}{n}\sum_{i_1\ldots i_{t+1}} n_{i_1\ldots i_{t+1}} \log q_{i_{s+1}\ldots i_t,i_{t+1}} &= \frac{2}{n}\sum_{i_{s+1}\ldots i_{t+1}} n_{i_{s+1}\ldots i_{t+1}} \log q_{i_{s+1}\ldots i_t,i_{t+1}} \\ &= \frac{2}{n}\sum_{i_1\ldots i_{r+1}} n_{i_1\ldots i_{r+1}} \log q_{i_1\ldots i_r,i_{r+1}}.\end{aligned} \tag{3.11}$$

Also

$$\begin{aligned}&\frac{1}{n}\sum_{i_1\ldots i_t} n_{i_1\ldots i_t} \sum_{i_{t+1}} \frac{(\hat{q}_{i_1\ldots i_t,i_{t+1}} - q_{i_{s+1}\ldots i_t,i_{t+1}})^2}{\hat{q}_{i_1\ldots i_t,i_{t+1}}} \\ &\quad = \frac{1}{n}\sum_{v_1\ldots v_{s+1}} \tilde{n}_{v_1\ldots v_{s+1}} \sum_{v_{s+2}} \left(\frac{\tilde{n}_{v_1\ldots v_{s+2}}}{\tilde{n}_{v_1\ldots v_{s+1}}} - \tilde{q}_{v_{s+1}v_{s+2}}\right)^2 \left(\frac{\tilde{n}_{v_1\ldots v_{s+2}}}{\tilde{n}_{v_1\ldots v_{s+1}}}\right)^{-1}.\end{aligned} \tag{3.12}$$

Since Y satisfies the hypothesis of Lemma 3.1, using (3.2), (3.3) and (3.9), (3.12) we can write

$$\frac{2}{n}\log L_t = \frac{2}{n}\sum_{i_1\ldots i_{r+1}} n_{i_1\ldots i_{r+1}} \log q_{i_1\ldots i_r,i_{r+1}} + O\left(\frac{\log\log n}{n}\right), \quad \text{a.s..} \tag{3.13}$$

Now assume that $t > r$. By (3.13) as $n \to \infty$

$$\begin{aligned}\mathrm{EDC}(t) - \mathrm{EDC}(r) &= -2\log L_t + 2\log L_r + (\gamma(t) - \gamma(r)c_n \\ &= O(\log\log n) + (\gamma(t) - \gamma(r))c_n, \quad \text{a.s..}\end{aligned} \tag{3.14}$$

Since $c_n/\log\log n \to \infty$ a.s. and $\gamma(t) > \gamma(r)$ we have $w.p.1$ for n large

$$\mathrm{EDC}(t) > \mathrm{EDC}(r), \quad \text{for} \quad t > r. \tag{3.15}$$

(b) If $r = 0$, then $X_0, X_1, \ldots$ are iid random variables. Put $q_i = P(X_0 = i)$ then using the same reasoning as above we have as $n \to \infty$

$$\frac{2}{n}\log L_0 = \frac{2}{n}\sum_i n_i \log\frac{n_i}{n} = \frac{2}{n}\sum_i n_i \log q_i + O\left(\frac{\log\log n}{n}\right), \quad \text{a.s..}$$

Assume that $t > 0$ then as $n \to \infty$

$$\begin{aligned}\frac{2}{n}\log L_t &= \frac{2}{n}\sum_{i_1\ldots i_{t+1}} n_{i_1\ldots i_{t+1}}\log\frac{n_{i_1\ldots i_{t+1}}}{n_{i_1\ldots i_t}}\\ &= \frac{2}{n}\sum_{i_1\ldots i_{t+1}} n_{i_1\ldots i_{t+1}}\log q_{i_{t+1}} + O\left(\frac{\log\log n}{n}\right)\\ &= \frac{2}{n}\sum_{i_{t+1}} n_{i_{t+1}}\log q_{i_{t+1}} + O\left(\frac{\log\log n}{n}\right), \quad \text{a.s..}\end{aligned}$$

So that (3.15) still holds.

(c) Now assume that $t < r$ with $r \geqslant 1$. Note that for $v_1 = (i_1, \ldots, i_r)$ and $v_2 = (i_2, \ldots, i_{r+1})$ we have

$$\begin{aligned}\frac{2}{n}\log L_r &= \frac{2}{n}\sum_{v_1,v_2}\tilde{n}_{v_1v_2}\log\frac{\tilde{n}_{v_1v_2}}{\tilde{n}_{v_1}}\\ &= \frac{2}{n}\sum_{v_1,v_2}\frac{\tilde{n}_{v_1v_2}}{\tilde{n}_{v_1}}\tilde{n}_{v_1}\log\frac{\tilde{n}_{v_1v_2}}{\tilde{n}_{v_1}}.\end{aligned}$$

Applying Lemma 3.1 to the process Y, we get

$$\begin{aligned}\lim_{n\to\infty}\frac{2}{n}\log L_r &= 2\sum_{v_1,v_2}\pi_{v_1}\tilde{q}_{v_1v_2}\log\tilde{q}_{v_1v_2}\\ &= 2\sum_{i_1\ldots i_{r+1}}\pi_{i_1\ldots i_r}q_{i_1\ldots i_r,i_{r+1}}\log q_{i_1\ldots i_r,i_{r+1}} = f_r, \quad \text{a.s..}\end{aligned} \tag{3.16}$$

Now

$$\begin{aligned}\frac{2}{n}\log L_{r-1} &= \frac{2}{n}\sum_{i_1\ldots i_r} n_{i_1\ldots i_r}\log\frac{n_{i_1\ldots i_r}}{n_{i_1\ldots i_{r-1}}}\\ &= \frac{2}{n}\sum_{i_1\ldots i_r} n_{i_1\ldots i_r}\log\frac{n_{i_1\ldots i_r}}{\sum_j n_{i_1\ldots i_{r-1}j}}.\end{aligned}$$

Using (3.1) we have

$$\lim_{n\to\infty}\frac{2}{n}\log L_{r-1} = 2\sum_{i_1\ldots i_r}\pi_{i_1\ldots i_r}\log\frac{\pi_{i_1\ldots i_r}}{\sum_j \pi_{i_1\ldots i_{r-1}j}} = f_{r-1}, \quad \text{a.s..}$$

Clearly we can write

$$f_{r-1} = 2\sum_{i_2\ldots i_{r+1}}\pi_{i_2\ldots i_{r+1}}\log\frac{\pi_{i_2\ldots i_{r+1}}}{\sum_j \pi_{i_2\ldots i_r j}},$$

and using (2.3) we have

$$f_{r-1} = 2 \sum_{i_1 \ldots i_{r+1}} \pi_{i_1 \ldots i_r} q_{i_1 \ldots i_r, i_{r+1}} \log \frac{\pi_{i_2 \ldots i_{r+1}}}{\sum_j \pi_{i_2 \ldots i_r j}}. \tag{3.17}$$

From (3.16), (3.17) and Jensen's inequality we have

$$\begin{aligned} f_{r-1} - f_r &= 2 \sum_{i_1 \ldots i_r} \pi_{i_1 \ldots i_r} \sum_{i_{r+1}} q_{i_1 \ldots i_r, i_{r+1}} \log \left[\pi_{i_2 \ldots i_{r+1}} / q_{i_1 \ldots i_r, i_{r+1}} \sum_j \pi_{i_2 \ldots i_r j} \right] \\ &\leqslant 2 \sum_{i_1 \ldots i_r} \pi_{i_1 \ldots i_r} \log \left[\sum_{i_{r+1}} \pi_{i_2 \ldots i_{r+1}} / \sum_{i_r} \pi_{i_2 \ldots i_r j} \right] = 0. \end{aligned}$$

Since each $\pi_{i_1 \ldots i_r} > 0$, equality holds only if

$$q_{i_1 \ldots i_r, i_{r+1}} = \frac{\pi_{i_2 \ldots i_{r+1}}}{\sum_j \pi_{i_2 \ldots i_r j}}. \tag{3.18}$$

But the left-hand side of (3.18) is independent of i_1 and this means that the chain X is of order $r-1$, which contradicts our assumption. It follows that

$$\lim_{n \to \infty} \frac{2}{n} (\log L_r - \log L_{r-1}) = f_r - f_{r-1} > 0, \quad \text{a.s.} \tag{3.19}$$

which implies that for $t < r$

$$\frac{2}{n} (\log L_r - \log L_t) \to f_r - f_t > 0, \quad \text{a.s..} \tag{3.20}$$

By (3.20) as $n \to \infty$

$$\begin{aligned} \frac{1}{n} (\mathrm{EDC}(t) - \mathrm{EDC}(r)) &= \frac{2}{n} (\log L_r - \log L_t) - (\gamma(r) - \gamma(t)) \frac{c_n}{n} \\ &= (f_r - f_t)(1 + o(1)) - (\gamma(r) - \gamma(t)) \frac{c_n}{n}, \quad \text{a.s..} \end{aligned}$$

Since $\frac{c_n}{n} \underset{n}{\longrightarrow} 0$ a.s. we have $w.p.1$ for large n

$$\mathrm{EDC}(t) > \mathrm{EDC}(r), \quad \text{for} \quad t < r. \tag{3.21}$$

(d) Finally, from (3.15) and (3.21) it follows that $\hat{r} = r$, a.s., and the proof of Theorem 2.1 is completed. □

Proof of Theorem 2.2. As shown in Billingsley [3] for $t > r$ as $n \to \infty$ the statistic $2(\log L_t - \log L_r)$ has a limiting distribution. It follows that

$$2(\log L_t - \log L_r) = O_p(1),$$

where O_p means bounded in probability. Since $c_n \xrightarrow{p} \infty$ and $\gamma(t) > \gamma(r)$ we have

$$\begin{aligned} \mathrm{EDC}(t) - \mathrm{EDC}(r) &= -2(\log L_t - \log L_r) + (\gamma(t) - \gamma(r))c_n \\ &= O_p(1) + (\gamma(t) - \gamma(r))c_n \xrightarrow{p} \infty. \end{aligned} \tag{3.22}$$

From (3.21) and (3.22) we conclude that $\hat{r} \xrightarrow{p} r$ as $n \to \infty$. □

Acknowledgements

Research carried out while L.C. Zhao was visiting the Department of Mathematics of University of Brasilia. He was partially supported by National Natural Science Foundation of China (19631040), Ph.D. Program Foundation of the Education Ministry of China and Special Foundation of Academic Sinica, C.C.Y. Dorea was partially supported by CNPq/PRONEX/FAPDF of Brazil, C.R. Gonçalves was partially supported by FEMAT of Brazil.

References

1. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Cont.* **19** (1974), 716–723.
2. Bai, Z. D., Krishnaiah, P. R. and Zhao, L. C.: On rates of convergence of efficient detection criteria in signal processing with white noise. *IEEE Trans. Info. Theory* **35** (1989), 380–388.
3. Billingsley, P.: Statistical Inference for Markov Processes, University of Chicago Press, Chicago, 1961.
4. Chin, E. H.: Modeling daily precipitation occurrence process with Markov chain, Water Resour. Res. **13**, 949–956.
5. Dorea, C. C. Y., Galves, A., Kira, E. and Pereira Alencar, A.: Markovian modeling of the stress contours of Brazilian and European Portuguese, *REBRAPE* **11** (2) (1971), 161–173.
6. Feller, W.: *An Introduction to Probability Theory and Its Applications*, 1, 3rd edn, Wiley, New York, 1968.
7. Freedman, D.: *Markov Chains*, Holden-Day, San Francisco, 1971.
8. Gates, P. and Tong, H.: On Markov chain modeling to some weather data, *J. App. Meteorol.* **15** (1976), 1145–1151.
9. Hall, P. and Heyde, C. C.: *Martingale Limit Theory and its Application*, Academic Press, New York, 1980.
10. Katz, R. W.: On some criteria for estimating the order of a Markov chain, *Technometrics* **23** (1981), 243–249.
11. Raftery, A. and Tavaré, S.: Estimation and modeling repeated patterns in high order Markov chains with mixture transition distribution model, *App. Statist. – J. Royal Statist. Soc.* Series C **43** (1994), 179–199.
12. Schwarz, G.: Estimating the dimension of a model, *Ann. Statist.* **6** (1978), 461–464.
13. Tong, H.: Determination of the order of a Markov chain by Akaikés information criterion, *J. App. Probab.* **12** (1975), 488-497.
14. Waterman, M. S. (ed): *Mathematical Methods for DNA Sequence*, Boca Raton, 1988.
15. Zhao, L. C., Krishaiah, P. R. and Bai, Z. D.: On detection of the number of signals in presence of white noise, *J. Multivariate Anal.* **20** (1986a), 21–25.
16. Zhao, L. C., Krishaiah, P. R. and Bai, Z. D.: On detection of the number of signals when the noise, covariance matrix is arbitrary. *J. Multivariate Anal.* **20** (1986b), 26–49.