BULLETIN OF THE BRAZILIAN MATHEMATICAL SOCIETY

Bull Braz Math Soc, New Series 37(4), 561-570

Convergence rates for Markov chain order estimates using EDC criterion

C.C.Y. Dorea* and J.S. Lopes**

Abstract. The Efficient Determination Criterion (EDC) generalizes the AIC and BIC criteria and provides a class of consistent estimators for the order of a Markov chain with finite state space. In this note, we derive rates of convergence for the EDC estimates.

Keywords: Order of Markov Chain, Efficient Determination Criterion.

Mathematical subject classification: Primary: 62F12, 62M05; Secondary: 60J10.

1 Introduction

Let $X_1^n = (X_1, \cdots, X_n)$ be a sample from a multiple Markov chain $X = \{X_n\}_{n\geq 1}$ of unknown but finite order $0 \leq r \leq K < \infty$. Assume that X takes values on a finite state space E and that the transition probabilities are given by

$$p(a_{r+1}|a_1^r) = P(X_{n+1} = a_{r+1}|X_{n-r+1}^n = a_1^r) \tag{1}$$

where $a_1^r = (a_1, \cdots, a_r) \in E^r$. For $0 \leq k \leq K$ let $\hat{L}(k)$ be the maximum log-likelihood function based on the sanple X_1^n when the chain is assumed to be of order k. The approximation of Kullback-Leibler information measure by Neyman-Pearson statistics along with the asymptotic χ^2-distribution of the maximum log-likelihood ratio $2\log\frac{\hat{L}(k)}{\hat{L}(r)}$ form the basis to derive the Akaike information criterion (AIC, Akaike (1974)),

$$\mathrm{AIC}(k) = -2\log\hat{L}(k) + 2|E|^k\big(|E|-1\big), \quad \hat{r}_n = \arg\min_{0\leq k\leq K}\mathrm{AIC}(k),$$

Received 26 April 2006.

*Partially supported by CNPq, CAPES/PROCAD, FAPDF/PRONEX, FINATEC and FUNPE/UnB.

**Partially supported by CAPES.

where $|E|$ denotes the cardinality of the set E,

$$\begin{aligned} \log \hat{L}(k) &= \sum_{a_1^{k+1}} N\left(a_1^{k+1}|X_1^n\right) \log \frac{N\left(a_1^{k+1}|X_1^n\right)}{N\left(a_1^k|X_1^n\right)}, \\ \log \hat{L}(0) &= \sum_{a} N\left(a|X_1^n\right) \log \frac{N\left(a|X_1^n\right)}{n} \end{aligned} \tag{2}$$

and

$$N\left(a_1^k|X_1^n\right) = \sum_{j=1}^{n-k+1} 1\left(X_j = a_1, \cdots, X_{j+k-1} = a_k\right) \tag{3}$$

that is, the number of occurrences of a_1^k in X_1^n. If $k = 0$ define $N\left(\cdot|X_1^n\right) = n$. For the maximum log-likelihood functions the sums are over positive terms $N\left(a_1^{k+1}|X_1^n\right) > 0$ or $N\left(a|X_1^n\right) > 0$.

Katz (1981) pointed out the inconsistency of AIC estimates and the BIC estimator proposed by Tong (1975) and Schwarz (1978) has been used in place of AIC. The BIC estimator replaces 2 by $\log n$ in the penalty term and this corrects the inconsistency of AIC

$$\mathrm{BIC}(k) = -2\log \hat{L}(k) + 2|E|^k\,(|E| - 1)\log n.$$

Several authors have addressed the consistency problem for BIC, besides Katz see, for example, Finesso (1992) or Barron, Rissanen and Yu (1998). More recently, Csiszar and Shields (2000) established the strong consistency for BIC with no boundness assumption on the order r. Some works on optimal error exponents for the probability of errors can also be found in the literature: Merjav, Gutman and Ziv (1989), Finesso, Liu and Narayan (1996) and Gassiat and Boucheron (2003). More recently, Zhao, Dorea and Gonçalves (2001) proposed the EDC criterion that, under mild conditions, is a strongly consistent estimator for r. And, from which we can derive the AIC and the BIC criteria by choosing appropriately the penalty term. For $c_n > 0$ define

$$\mathrm{EDC}(k) = -2\log \hat{L}(k) + \gamma(k)c_n \quad \text{and} \quad \hat{r}_n = \arg\min_{0\le k\le K} \mathrm{EDC}(k). \tag{4}$$

It is shown that if $\gamma(k)$ in the penalty term is taken to be a strictly increasing function,

$$\frac{c_n}{n} \to 0 \quad \text{and} \quad \frac{c_n}{\log\log n} \to \infty \quad \text{as} \quad n \to \infty \tag{5}$$

then $\hat{r}_n$ is strongly consistent. In this note, by taking $\gamma(k) = 2|E|^k\,(|E| - 1)$ as in AIC and BIC cases, we present results for the rate of convergence of EDC

estimator. Our main result, Theorem 1, shows that for any initial distribution ν we have for $k \neq r$,

$$P_\nu(\mathrm{EDC}(k) - \mathrm{EDC}(r) \geq d \log\log n) \geq 1 - a \exp\{-b \log\log n\}.$$

2 Preliminaries

For $k \geq r$ let $Y_n^{(k)} = X_n^{n+k-1} = (X_n, \ldots, X_{n+k-1})$. Then $Y^{(k)} = \left\{Y_n^{(k)}\right\}_{n\geq 1}$ is a first order Markov chain on E^k with transition probabilities

$$P\left(Y_{n+1}^{(k)} = b_1^k \middle| Y_n^{(k)} = a_1^k\right) = \begin{cases} p\left(b_k \middle| a_{k-r+1}^k\right) & \text{if } a_2^k = b_1^{k-1} \\ 0 & \text{otherwise.} \end{cases}$$

We will assume that the derived first order Markov chain $Y^{(r)}$ is ergodic with stationary (equilibrium) distribution $(\pi(\cdot))$. Denoting $aa_1^{r-1} = (a, a_1, \ldots, a_{r-1})$ we have

$$\begin{aligned} \pi\left(a_1^r\right) &= \sum_{b_1^r} \pi\left(b_1^r\right) p\left(a_r | b_1^r\right) \\ &= \sum_a \pi\left(aa_1^{r-1}\right) p\left(a_r | aa_1^{r-1}\right). \end{aligned}$$

For $k \geq r$ define

$$\pi\left(a_1^k\right) = \pi\left(a_1^r\right) p\left(a_{r+1} | a_1^r\right) \cdots p\left(a_k | a_{k-r}^{k-1}\right) \tag{6}$$

then,

$$\begin{aligned} \pi\left(a_1^k\right) &= \sum_a \pi\left(aa_1^{r-1}\right) p\left(a_r | aa_1^{r-1}\right) p\left(a_{r+1} | a_1^r\right) \cdots p\left(a_k | a_{k-r}^{k-1}\right) \\ &= \sum_a \pi\left(aa_1^{k-1}\right) p\left(a_k | a_{k-r}^{k-1}\right). \end{aligned}$$

Which shows that $(\pi(\cdot))$ defined by (6) is a stationary distribution for $Y^{(k)}$. Moreover, from the ergodicity of $Y^{(r)}$ it easy to verify that $Y^{(k)}$ is also ergodic. We have just proved the following proposition,

Proposition 1. *Assume that the derived Markov chain $Y^{(r)}$ is ergodic then for $k \geq r$ the process $Y^{(k)}$ is also ergodic and possesses stationary distribution given by* (6).

Lemma 1. *Let Z be an ergodic Markov chain with values on E and transition matrix $(q(\cdot|\cdot))$. Then given any $\rho > 0$ and any initial distribution ν the associated transition counts satisfy,*

$$P_\nu\left(\left|N\left(a_1^{k+1}|Z_1^n\right) - N\left(a_1^k|Z_1^n\right) q\left(a_{k+1}|a_k\right)\right| \geq \sqrt{\rho n \log\log n}\right) \leq 4\exp\left\{-\frac{\rho}{2}\log\log n\right\}.$$

(P_ν: probability with initial distribution ν.)

Proof.

(a) We will make use of the following result from Devroye (1991): Let $\mathcal{F}_0 = \{\phi, \Omega\} \subset \mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_n$ be a sequence of nested σ-algebras. Let U be a $\mathcal{F}_n$-measurable and integrable random variable and define the Doob martingale $U_j = E(U|\mathcal{F}_j)$. Assume that there exist a $\mathcal{F}_{j-1}$-measurable random variable V_j and a constant b_j such that $V_j \leq U_j \leq V_j + b_j$. Then for any $\epsilon > 0$,

$$P\left(|U - EU| \geq \epsilon\right) \leq 4\exp\left\{-\frac{2\epsilon^2}{\sum_{k=1}^n b_j^2}\right\}.$$

(b) Given $a_1^{k+1} \in E^{k+1}$ define

$$\eta_j = 1\left(Z_j^{j+k} = a_1^{k+1}\right) - 1\left(Z_j^{j+k-1} = a_1^k\right) q(a_{k+1}|a_k)$$

and

$$U = U_n = \sum_{j=1}^{n-k} \eta_j = N\left(a_1^{k+1}|Z_1^n\right) - N\left(a_1^k|Z_1^n\right) q\left(a_{k+1}|a_k\right) + o(\delta_n).$$

The term $o(\delta_n)$ stands for

$$A_n = o(\delta_n) \quad \text{if} \quad \frac{A_n}{\delta_n} \to 0 \quad \text{for any} \quad \delta_n \to \infty. \tag{7}$$

Let $\mathcal{F}_j = \sigma(Z_1, \cdots, Z_j)$ and $U_j = E(U|\mathcal{F}_j)$. Then $U_j = 0$ for $0 \leq j \leq k$ and $U_j = \sum_{l=1}^{j-k} \eta_l$ for $k \leq j \leq n$. Since $E\{U_j\} = 0$ we have the hypotheses of (a) satisfied with $V_j = U_{j-1} - 1$ and $b_j = 2$. Result follows by noting that $\sum_{j=1}^n b_j^2 = 4n$. □

Proposition 1 along with above lemma give us:

Corollary 1. *If $Y^{(r)}$ is ergodic and $k \geq r$ then for any $a_1^{k+1} \in E^{k+1}$, any initial distribution ν and any $\rho_n > 0$*

$$P_\nu\left(\left[N\left(a_1^{k+1}|X_1^n\right) - N\left(a_1^k|X_1^n\right) p\left(a_{k+1}|a_{k-r+1}^k\right)\right]^2 \geq \rho_n n \log\log n\right) \leq 4\exp\left\{-\frac{\rho_n}{2}\log\log n\right\}. \tag{8}$$

3 Results

First, we derive some bounds for $\log \hat{L}(k) - \log \hat{L}(r)$. Let $o(\delta_n)$ as in (7) and for $k \geq r$ write the log-likelihood as

$$\log L(k) = \sum_{a_1^{k+1}} N\left(a_1^{k+1}|X_1^n\right) \log p\left(a_{k+1}|a_{k-r+1}^k\right). \tag{9}$$

Proposition 2. *If $Y^{(r)}$ is ergodic then for $k \geq r$ we have for large n $\log \hat{L}(k) - \log L(k) \geq 0$ and*

$$\log \hat{L}(k) - \log \hat{L}(r) \leq \sum_{a_1^{k+1}} \frac{\left[N\left(a_1^{k+1}|X_1^n\right) - N\left(a_1^k|X_1^n\right) p\left(a_{k+1}|a_{k-r+1}^k\right)\right]^2}{N\left(a_1^{k+1}|X_1^n\right)} + o(\delta_n). \tag{10}$$

Moreover, for $0 \leq k < r$

$$\log \hat{L}(k-1) - \log \hat{L}(k) \leq 0$$

and, if $\dfrac{c_n}{n} \to 0$, there exists a constant $\beta_r > 0$ such that for n large

$$\log \hat{L}(k) - \log \hat{L}(r) \leq -n\beta_r + o(\delta_n). \tag{11}$$

Proof.

(a) Let $k \geq r$ then

$$\begin{aligned}
\log L(k) &= \sum_{a_1^{k+1}} N\left(a_1^{k+1}|X_1^n\right) \log p\left(a_{k+1}|a_{k-r+1}^k\right) \\
&= \sum_{a_{k-r+1}^{k+1}} N\left(a_{k-r+1}^{k+1}|X_1^n\right) \log p\left(a_{k+1}|a_{k-r+1}^k\right) \\
&= \log L(r) + o(\delta_n).
\end{aligned}$$

It follows that

$$\log \hat{L}(k) - \log L(r) =$$
$$\log \hat{L}(k) - \log L(k) - [\log \hat{L}(r) - \log L(r)] + o(\delta_n).$$

To prove (10) enough to show that for n large

$$\begin{aligned} 0 &\leq \log \hat{L}(k) - \log L(k) \\ &\leq \sum_{a_1^{k+1}} \frac{\left[N\left(a_1^{k+1}|X_1^n\right) - N\left(a_1^k|X_1^n\right) p\left(a_{k+1}|a_{k-r+1}^k\right)\right]^2}{N\left(a_1^{k+1}|X_1^n\right)} + o(\delta_n). \end{aligned} \tag{12}$$

From (2) and (9) we have

$$\log \hat{L}(k) - \log L(k) = -\sum_{a_1^{k+1}} N\left(a_1^{k+1}|X_1^n\right) \log\left(1 - z_n(a_1^{k+1})\right)$$

where

$$z_n\left(a_1^{k+1}\right) = \frac{N\left(a_1^{k+1}|X_1^n\right) - N\left(a_1^k|X_1^n\right) p\left(a_{k+1}|a_{k-r+1}^k\right)}{N\left(a_1^{k+1}|X_1^n\right)}.$$

By Proposition 1 and the Law of Large Numbers for Markov chains (see, for example, Dacunha-Castelle and Duflo (1986)) we have almost surely (a.s.),

$$\frac{N\left(a_1^k|X_1^n\right)}{n} \to \pi\left(a_1^k\right) \quad \text{and} \quad \frac{N\left(a_1^{k+1}|X_1^n\right)}{n} \to \pi\left(a_1^k\right) p\left(a_{k+1}|a_{k-r+1}^k\right) \quad \text{a.s.}$$

Thus $z_n\left(a_1^{k+1}\right) \to 0$. Now, using the inequality

$$z \leq -\log(1-z) \leq z + z^2, \ |z| < \frac{1}{2}$$

and the identity $\sum_{a_{k+1}} N\left(a_1^{k+1}|X_1^n\right) z_n\left(a_1^{k+1}\right) = 0$ we get (12).

(b) Let $k < r$. Since the true order is r, for some $a_1^{r+1} \in E^{r+1}$ we must have

$$\frac{\pi\left(a_2^{r+1}\right)}{\sum_a \pi\left(a a_2^r\right) p\left(a_{r+1}|a_1^r\right)} \neq 1$$

or else $p\left(a_{r+1}|a_1^r\right)$ does not depend on a_1 for all $a_1^{r+1} \in E^{r+1}$. Let

$$\beta_r = -\sum_{a_1^{r+1}} \pi\left(a_1^r\right) p\left(a_{r+1}|a_1^r\right) \log \frac{\pi(a_2^{r+1})}{\sum_a \pi\left(a a_2^r\right) p\left(a_{r+1}|a_1^r\right)}.$$

By Jensen's inequality we have $\beta_r > 0$. For $r \geq 1$ write

$$\frac{\log \hat{L}(r-1) - \log \hat{L}(r)}{n}$$

$$\to \sum_{a_1^{r+1}} \frac{N\left(a_1^{r+1}|X_1^n\right)}{n} \log \frac{N\left(a_2^{r+1}|X_1^n\right)}{N\left(a_2^r|X_1^n\right)} \frac{N\left(a_1^r|X_1^n\right)}{N\left(a_1^{r+1}|X_1^n\right)} + o(\delta_n).$$

From the Law of Large Numbers we get

$$\frac{\log \hat{L}(r-1) - \log \hat{L}(r)}{n} =$$

$$\sum_{a_1^{r+1}} \pi\left(a_1^r\right) p\left(a_{r+1}|a_1^r\right) \log \frac{\pi\left(a_2^{r+1}\right)}{\sum_a \pi\left(a a_2^r\right) p\left(a_{r+1}|a_1^r\right)} = -\beta_r.$$

To prove (11) enough to show

$$\log \hat{L}(k-1) - \log \hat{L}(k) \leq 0.$$

and this follows using again the Jensen's inequality. □

Theorem 1. *Let $Y^{(r)}$ be an ergodic Markov chain with $|E| \geq 2$. Assume that the sequence c_n satisfies,*

$$1 \leq \liminf_{n\to\infty} \frac{c_n}{\log\log n} \quad \text{and} \quad \frac{c_n}{n} \to 0. \tag{13}$$

Then there exist $a = a(K) > 0$ and $b = b(K) > 0$ such that for any initial distribution ν and any $d < \gamma(r) - 1$ we have for $k \neq r$, $0 \leq k \leq K$,

$$P_\nu\left(\mathrm{EDC}(k) - \mathrm{EDC}(r) \geq d \log\log n\right) \geq 1 - a\exp\{-b\log\log n\}. \tag{14}$$

Proof.

(a) Let $k < r$. From (4) and (11) we have

$$\mathrm{EDC}(k) - \mathrm{EDC}(r) \geq 2n\beta_r + (\gamma(k) - \gamma(r))c_n.$$

Since $\beta_r > 0$ and c_n satisfy (13) we have for large n

$$P_\nu(\mathrm{EDC}(k) - \mathrm{EDC}(r) \geq d\log\log n) = 1$$

and (14) folows.

(b) Let $k > r$. Since $\gamma(\cdot)$ is strictly increasing, $|E| \geq 2$ and $d < \gamma(r+1) - \gamma(r)$ we have by (13)

$$\varphi_n = \left(\frac{\gamma(k) - \gamma(r)}{2}\right) c_n - \frac{d}{2} \log\log n > 0. \tag{15}$$

From (10) it follows that for

$$\psi_n\left(a_1^{k+1}\right) = \left[N\left(a_1^{k+1}|X_1^n\right) - N\left(a_1^k|X_1^n\right) p\left(a_{k+1}|a_{k-r+1}^k\right)\right]^2$$

we have

$$\begin{aligned}\left(\mathrm{EDC}(k) - \mathrm{EDC}(r) \geq d \log\log n\right) &= \left(\log \hat{L}(k) - \log \hat{L}(r) \leq \varphi_n\right) \\ &\supseteq \cup_{a_1^{k+1}} \left(\psi_n\left(a_1^{k+1}\right) > \varphi_n \frac{N\left(a_1^{k+1}|X_1^n\right)}{|E|^{k+1}}\right).\end{aligned}$$

Let $a = 4|E|^{K+1}$ and

$$b = \frac{1}{2|E|^{K+1}} \min\left\{\pi\left(a_1^{K+1}\right) : a_1^{K+1} \in E^{K+1},\ \pi\left(a_1^{K+1}\right) > 0\right\}.$$

Since $\dfrac{N\left(a_1^{k+1}|X_1^n\right)}{n|E|^{k+1}} \to \dfrac{\pi\left(a_1^{k+1}\right)}{|E|^{k+1}}$ we have

$$\left(\mathrm{EDC}(k) - \mathrm{EDC}(r) \geq d \log\log n\right) \subseteq \cup_{a_1^{k+1}} \left(\psi_n\left(a_1^{k+1}\right) > \rho_n n \log\log n\right).$$

where $\rho_n = 4b\dfrac{\varphi_n}{\log\log n}$.

By Corollary 1

$$\begin{aligned}P_\nu\left(\mathrm{EDC}(k) - \mathrm{EDC}(r) < d \log\log n\right) &\leq \sum_{a_1^{k+1}} P_\nu\left(\psi_n\left(a_1^{k+1}\right) > \rho_n n \log\log n\right) \\ &\leq 4|E|^{k+1} \exp\left\{-\frac{\rho_n}{2} \log\log n\right\} \\ &\leq a \exp\left\{-b \log\log n\right\}.\end{aligned}$$

In the last inequality we used (13) and (15) to see that

$$\frac{\rho_n}{2} = b\left((\gamma(k) - \gamma(r))\frac{c_n}{\log\log n} - d\right) \geq b(\gamma(k) - \gamma(r) - d) \geq b. \quad \square$$

Corollary 2. *Under hypotheses of Theorem* 1 *we have:*

(a) (14) *holds with BIC in place of EDC;*

(b) *the EDC estimate* (4) *is strongly consistent.*

Remark 1.

(a) Corollary 2 shows consistency under condition (13) which is weaker than condition (5) from Zhao et al. (2001).

(b) For related work on bounds for wrong determination of the order using EDC criterion see Dorea and Zhao (2006).

References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automation and Control, **19** (1974), 716–723.

[2] A. Barron, J. Rissanen and B. Yu, *The minimum description length principle in coding and modeling*, IEEE Trans. Information Theory, **44** (1998), 2743–2760.

[3] I. Csiszar and P.C. Shields, *The consistency of the BIC Markov order estimator*, Annals of Statistics, **28** (2000), 1601–1619.

[4] D. Dacunha-Castelle and M. Duflo, *Probability and Statistics*, **2** (1986), Springer-Verlag, New York.

[5] L. Devroye, *Exponential inequalities in nonparametric estimation*, In: Nonparametric Functional Estimation and Related Topics, ed. by G.G. Roussas, Kluwer Academic Publishers, 31–44 (1991).

[6] C.C.Y. Dorea and L.C. Zhao, *Exponential bounds for the probability of wrong determination of the order of a Markov chain by using the EDC criterion*, Journal of Statistical Planning and Inference, (2006).

[7] L. Finesso, *Estimation of the order of a finite Markov chain*, In: Recent Advances in Mathematical Theory of Systems, Control and Network Signals (eds. H. Kimura and S. Kodama), Mita Press, 643–645 (1992).

[8] L. Finesso, C.C. Liu P. and Narayan, *The optimal error exponent for Markov order estimation*, IEEE Trans. Information Theory, **42** (1996), 1488–1497.

[9] G. Gassiat and S. Boucheron, *Optimal error exponents in hidden Markov model order estimation*, IEEE Trans. Information Theory, **48** (2003), 964–978.

[10] R.W. Katz, *On some criteria for estimating the order of a Markov chain*, Technometrics, **23** (1981), 243–249.

[11] J.S. Lopes, *Determination of Markov chain order using EDC criterion*, Ph.D. Thesis, Department of Mathematics, University of Brasilia, Brasilia-DF, Brazil, (in portuguese), (2005).

[12] N. Merjav, M. Gutman and J. Ziv, *On the estimation of the order of a Markov chain and universal data compression*, IEEE Trans. Information Theory, **35** (1989), 1014–1019.

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

[14] L.C. Zhao, C.C.Y. Dorea and C.R. Gonçalves, *On determination of the order of a Markov Chain*, Stat. Inference for Stochastic Processes, **4** (2001), 273–282.

C.C.Y. Dorea
Departamento de Matemática
Universidade de Brasília
70910-900 Brasília, DF
BRAZIL

E-mail: cdorea@mat.unb.br

J.S. Lopes
Departamento de Matemática
Universidade Federal do Rio Grande do Norte
59072-970 Natal, RN
BRAZIL

E-mail: lopes@ccet.ufrn.br