# Simulation Results for Markov Model Seletion : AIC, BIC and EDC

Chang C.Y. Dorea, Catia R. Goncalves and Paulo A.A. Resende

***Abstract*—Higher order Markov chains, by its very definition, is the most flexible model for finitely dependent sequences of random variables. In practical settings, estimation of the dependency order is needed to identify other model parameters. Based on the penalized log-likelihood function and within nested hypotheses testing framework, several estimation alternatives have been proposed. The AIC, Akaike's entropy-based information criterion, constitutes the best known tool for model identification and has had a fundamental impact in statistical model selection. In spite of the AIC's relevance, several authors have pointed out its inconsistency that may lead to overestimation of the true order. To overcome this inconsistency, the Bayesian information criterion, BIC, was proposed by introducing in the penalty term the sample size and it is a consistent estimator for large samples. A more general approach is exhibited by the EDC, efficient determination criterion, that encompass both AIC and BIC estimates. Under proper setting, the EDC, besides being a strongly consistent estimate, is an optimal estimator. These approaches are briefly presented and compared by numerical simulation. The presented results may support decisions related to estimator's choice.**

***Index Terms*—AIC, BIC, EDC, Markov chain order.**

## I. Introduction

THE Akaike's (1974) entropy-based information criterion, AIC, was designed to be an approximately unbiased estimate of the Kullback-Leibler divergence between the fitted model relative to the true model. The fact that when mean log-likelihood ratio is used to estimate the divergence quantity, the bias introduced by the maximum likelihood estimate of the parameters needs to be corrected. For the AIC procedure the correction (penalty) term is taken to be the number of independent parameters of the model. In spite of the AIC's relevance, there was no rigorous analysis about its behaviour and showed a tendency of overestimating the true order. In fact, Katz (1981) formally derived the asymptotic distribution of AIC estimator and proved its inconsistency for the Markov Chain case, no matter how large the sample size is taken. To overcome this inconsistency, the Bayesian information criterion, BIC, was proposed by Schwarz (1978). The BIC procedure introduces in the penalty term the sample size and it is a consistent estimate. The EDC, efficient determination criterion, was introduced in Zhao et al. (2001) and encompass both the AIC and the BIC criteria

$$\text{EDC}(k) = -2\log \hat{L}(k) + \gamma(k)c_n$$

where $\gamma(\cdot)$ is a positive and strictly increasing function, $c_n \geq 0$ and $\hat{L}(k)$ is the maximum likelihood estimate. More specifically, let $X = \{X_n\}_{n\geq 1}$ be multiple Markov chain of unknown order $r$. Assume that $X$ takes value on a finite state space $E = \{1, \cdots, m\}$ and that the transition matrix $P$ has probabilities given by

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

where $a_1^r = a_1^k a_{k+1}^r = (a_1, \cdots, a_r) \in E^r$ and $X_1^n = (X_1, \cdots, X_n)$. In practical setting, given the observation $X_1^n$ from a chain $k$ we have the maximum likelihood estimate

$$\hat{L}(k) = \prod_{a^{k+1}} \hat{p}^{N(a_1^{k+1})}(a_{k+1}|a_1^k)$$

where $\hat{p}(a_{k+1}|a_1^k) = N(a_1^{k+1})/N(a_1^k)$

$$N(a_1^k) = \sum_{j=1}^{n-k+1} 1(X_j = a_1, \cdots, X_{j+k-1} = a_k),$$

that is, $N(a_1^k)$ is the number of occurrences of $a_1^k$ in $X_1^n$ and the sums are taken over positive terms $N(a_1^{k+1}) > 0$, or else, we convention $0/0 = 0 \cdot \infty = 0^0 = 0$. By assuming $\gamma(k) = m^k(m-1)$ we can derive

$$\text{AIC}(k) = -2\log \hat{L}(k) + 2m^k(m-1)$$

and

$$\text{BIC}(k) = -2\log \hat{L}(k) + m^k(m-1)\log n.$$

The corresponding estimators are

$$\hat{r}_{\text{AIC}} = \arg\min_{0\leq k\leq K} \text{AIC}(k)$$

and

$$\hat{r}_{\text{BIC}} = \arg\min_{0\leq k\leq K} \text{BIC}(k).$$

The rates of convergence from Dorea and Zhao (2006) and the results from Dorea (2008) and Resende et al. (2014) show that, under regularity conditions, the optimal choice is given by

$$\text{EDC}_{\text{opt}}(k) = -2\log \hat{L}(k) + 2m^{k+1}\log\log n$$

with

$$\hat{r}_{\text{EDC}} = \arg\min_{k\geq 0} \text{EDC}_{\text{opt}}(k).$$

With more than one alternative to estimate $r$ it is natural to seek comparison among them. Katz (1981) presented some modest numerical simulation, supported by computational resources available at the time, to compare $\hat{r}_{\text{AIC}}$ and $\hat{r}_{\text{BIC}}$. We analyse the comparative performance of $\hat{r}_{\text{AIC}}$, $\hat{r}_{\text{BIC}}$ and $\hat{r}_{\text{opt}}$. Altogether 63 cases were studied by ranging $m = 2, \cdots, 10$ and $r = 0, \cdots, 6$. For each case 1,000 models were generated and for each model, 349 samples. Since both $\hat{r}_{\text{BIC}}$ and $\hat{r}_{\text{opt}}$ possess the same asymptotic behavior, the sample size $n$ also played an important role in our analysis. The sample sizes were taken from $n = 10$ up to $n = 10^8$. Our findings show

Manuscript received June 24, 2014; revised June 24, 2014. This work was supported in part by CNPq-Brazil, CAPES-Brazil and FAPDF-Brasilia.

C.C.Y. Dorea, C.R. Gonçalves and P.A.A. Resende are with the Department of Mathematics, Universidade de Brasilia, Brasilia, DF, 70910-900 Brazil; e-mail: changdorea@unb.br; catiarg@unb.br; pa@pauloangelo.com

that, in general, $\hat{r}_{\text{opt}}$ outperforms $\hat{r}_{\text{BIC}}$. For small samples, all considered estimators have a tendency to underestimate the true order of the chain. Contrary to what Katz implicitly suggested in his numerical simulations, the probability of overestimation for $\hat{r}_{\text{AIC}}$ can be negligible in the case of complex models. These results may support the choice of which estimator to use in real situations. In the next section we gather some simulation results from the forthcoming paper Resende et al. (2014).

## II. Simulation Results

We considered 63 cases of Markov chains, varying $m = 2, \cdots, 10$ and $r = 0, \cdots, 6$. For each case we randomly generated 1,000 transition matrices and for each matrix, one large sample of length 100,000,000 and 349 "sub-samples" by considering the fragmentation from 0 to a properly chosen sample sizes. Using this technique it is possible to reuse the partial sums $N(a_1^k)$ and achieve a considerable computational gain. From the theoretical point of view, this is a reasonable approximation. The cases were chosen according to the available computational resources. The sizes of "sub-samples" where chosen empirically to properly compare the estimators. Although such numbers do not appears large, the most complex considered case, $m = 10$ and $r = 6$, has $9,000,000 = 10^6 \times (6-1)$ parameters, and the estimators couldn't fit the true order, even for samples of length 100,000,000.

**EDC vs BIC.** For small complexity $\gamma_m(r) = m^r(m-1)$ (number of free parameters), the sample sizes $n$ can be small too. Tables 1 and 2 provide simulation results for the cases $m = 4$, $r = 1$ and $m = 10$, $r = 1$, respectively. The column $n$ is the sample size, "$<$", "$=$" and "$>$" represent respectively the rates of underestimation, fitness and overestimation for each $n$.

Table 1
Distribution of hits for $m = 4$ and $r = 1$

| $n$ | EDC | | | BIC | | |
|---|---|---|---|---|---|---|
| | $<$ | $=$ | $>$ | $<$ | $=$ | $>$ |
| 10 | 98.7% | 1.3% | 0% | 99.1% | 0.9% | 0% |
| 25 | 90.2% | 9.8% | 0% | 91.4% | 8.6% | 0% |
| 68 | 50.6% | 49.4% | 0% | 60.3% | 39.7% | 0% |
| 775 | 0% | 100.0% | 0% | 0.1% | 99.9% | 0% |
| 900 | 0% | 100.0% | 0% | 0% | 100.0% | 0% |

Table 2
Distribution of hits for $m = 10$ and $r = 1$

| $n$ | EDC | | | BIC | | |
|---|---|---|---|---|---|---|
| | $<$ | $=$ | $>$ | $<$ | $=$ | $>$ |
| 218 | 99.8% | 0.2% | 0% | 100.0% | 0% | 0% |
| 425 | 40.9% | 59.1% | 0% | 100.0% | 0% | 0% |
| 450 | 28.9% | 71.1% | 0% | 99.9% | 0.1% | 0% |
| 600 | 3.1% | 96.9% | 0% | 91.1% | 8.9% | 0% |
| 775 | 0.1% | 99.9% | 0% | 48.2% | 51.8% | 0% |
| 950 | 0% | 100.0% | 0% | 15.4% | 84.6% | 0% |
| 1812 | 0% | 100.0% | 0% | 0% | 100.0% | 0% |

Table 3
Distribution of hits for $m = 4$ and $r = 3$

| $n$ | EDC | | | BIC | | |
|---|---|---|---|---|---|---|
| | $<$ | $=$ | $>$ | $<$ | $=$ | $>$ |
| 1562 | 99.9% | 0.1% | 0% | 100.0% | 0% | 0% |
| 2375 | 98.8% | 1.2% | 0% | 99.9% | 0.1% | 0% |
| 23125 | 50.2% | 49.8% | 0% | 65.1% | 34,9% | 0% |
| 9375000 | 0% | 100.0% | 0% | 0.6% | 99.4% | 0% |
| 23750000 | 0% | 100.0% | 0% | 0% | 100.0% | 0% |

Table 4
Distribution of hits for $m = 5$ and $r = 4$

| $n$ | EDC | | | BIC | | |
|---|---|---|---|---|---|---|
| | $<$ | $=$ | $>$ | $<$ | $=$ | $>$ |
| 6500 | 100.0% | 0.0% | 0% | 100.00% | 0.00% | 0% |
| 32500 | 99.8% | 0.2% | 0% | 100.0% | 0% | 0% |
| 68750 | 93.6% | 6.4% | 0% | 99.7% | 0.3% | 0% |
| 600000 | 49.8% | 50.2% | 0% | 66.4% | 33.6% | 0% |
| 1437500 | 33.5% | 66.5% | 0% | 49.9% | 50.1% | 0% |
| 16875000 | 7.0% | 93.0% | 0% | 13.8% | 86.2% | 0% |
| 100000000 | 0% | 100.0% | 0% | 3.4% | 96.6% | 0% |

In all cases, EDC exhibits a better performance than BIC. In the smaller complexity case ($m = 4$, $r = 1$) the fitness rates are similar. However, for more complex cases, EDC needed nearly half steps $n$, as compared to BIC, to achieve $50\%$ of fitness. This difference becomes bigger as larger complexities are considered. This happens because complex models require larger sample sizes that will result in larger differences between the penalty terms. Table 1 shows that the fitness of EDC and BIC are quite similar. In fact, for the very simple and atypical cases, such as $m = 2$, $r \leq 2$ or $m = 3$, $r = 1$, our simulations show that BIC performs better than EDC. It occurs because, for not too large sample size the penalty term for BIC is smaller than that for EDC.

**AIC's Performance.** Despite the inconsistency of AIC and the existence of strong consistent alternatives, this estimator have been widely used. Thus, we performed some numerical simulation with the aim to analyse AIC's behavior and to access its overestimation probabilities.

Table 5
Distribution of hits for EDC, BIC and AIC (in %)

| $r$ | $m$ | $n$ | EDC | | BIC | | AIC | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | $<$ | $=$ | $<$ | $=$ | $<$ | $=$ | $>$ |
| 1 | 3 | 10 | 80.3 | 19.7 | 73.9 | 26.1 | 63.1 | 36.9 | 0 |
| | | 22 | 72.2 | 27.8 | 67.4 | 32.6 | 39.8 | 59.9 | 0.3 |
| | | 168 | 15.0 | 85.0 | 15.8 | 84.2 | 3.3 | 93.5 | 3.2 |
| | | 1375 | 0.6 | 99.4 | 0.8 | 99.2 | 0.1 | 96.2 | 3.7 |
| | | 3125 | 0 | 100.0 | 0.1 | 99.9 | 0 | 96.2 | 3.8 |
| | | 5000 | 0 | 100.0 | 0 | 100.0 | 0 | 97.1 | 2.9 |
| 1 | 4 | 10 | 98.7 | 1.3 | 99.1 | 0.9 | 96.1 | 3.9 | 0 |
| | | 131 | 18.4 | 81.6 | 27.5 | 72.5 | 1.6 | 98.4 | 0 |
| | | 212 | 7.3 | 92.7 | 12.2 | 87.8 | 0.2 | 99.7 | 0.1 |
| | | 975 | 0 | 100.0 | 0 | 100.0 | 0 | 99.9 | 0.1 |
| 2 | 3 | 17 | 100.0 | 0 | 100.0 | 0 | 99.9 | 0.1 | 0 |
| | | 137 | 96.5 | 3.5 | 97.3 | 2.7 | 58.6 | 41.3 | 0.1 |
| | | 175000 | 1.7 | 98.3 | 3.5 | 96.5 | 0.1 | 99.8 | 0.1 |
| | | 4750000 | 0 | 100.0 | 0 | 100.0 | 0 | 99.9 | 0.1 |
| 2 | 5 | 137 | 100.0 | 0 | 100.0 | 0 | 99.7 | 0.3 | 0 |
| | | 400 | 99.9 | 0.1 | 100.0 | 0 | 67.0 | 33.0 | 0 |
| | | 650 | 99.0 | 1.0 | 100.0 | 0 | 50.1 | 49.9 | 0 |
| | | 750 | 97.0 | 3.0 | 99.9 | 0.1 | 47.0 | 53.0 | 0 |
| | | 3125 | 49.9 | 50.1 | 68.1 | 31.9 | 20.9 | 79.1 | 0 |
| | | 6000 | 35.8 | 64.2 | 49.0 | 51.0 | 13.7 | 86.3 | 0 |
| | | 106250 | 5.4 | 94.6 | 10.6 | 89.4 | 0.5 | 99.5 | 0 |
| | | 187500 | 4.1 | 95.9 | 6.4 | 93.6 | 0 | 100.0 | 0 |
| | | 837500 | 0 | 100.0 | 1.5 | 98.5 | 0 | 100.0 | 0 |
| | | 2000000 | 0 | 100.0 | 0 | 100.0 | 0 | 100.0 | 0 |
| 3 | 10 | 17500 | 100.0 | 0 | 100.0 | 0 | 99.4 | 0.6 | 0 |
| | | 81250 | 99.4 | 0.6 | 100.0 | 0 | 61.3 | 38.7 | 0 |
| | | 131250 | 91.5 | 8.5 | 100.0 | 0 | 49.2 | 50.8 | 0 |
| | | 637500 | 49.9 | 50.2 | 77.1 | 22.9 | 20.6 | 79.4 | 0 |
| | | 1812500 | 30.6 | 69.4 | 49.7 | 50.3 | 12.3 | 87.7 | 0 |
| | | 11250000 | 10.3 | 89.7 | 19.5 | 80.5 | 0.6 | 99.4 | 0 |
| | | 13750000 | 7.9 | 92.1 | 18.2 | 81.8 | 0 | 100.0 | 0 |
| | | 100000000 | 0 | 100.0 | 6.4 | 93.6 | 0 | 99.2 | 0.8 |
| 4 | 4 | 2000 | 100.0 | 0 | 100.0 | 0 | 99.8 | 0.2 | 0 |
| | | 9000 | 99.9 | 0.1 | 100.0 | 0 | 81.9 | 18.1 | 0 |
| | | 15625 | 98.4 | 1.6 | 99.9 | 0.1 | 68.7 | 31.3 | 0 |
| | | 37500 | 88.6 | 11.4 | 96.9 | 3.1 | 49.7 | 50.3 | 0 |
| | | 225000 | 49.3 | 50.7 | 64.9 | 35.1 | 20.9 | 79.1 | 0 |
| | | 475000 | 36.6 | 63.4 | 49.4 | 50.6 | 13.3 | 86.7 | 0 |
| | | 16250000 | 2.8 | 97.2 | 7.6 | 92.4 | 0 | 100.0 | 0 |
| | | 1000000000 | 0.1 | 99.9 | 0.4 | 99.6 | 0 | 100.0 | 0 |

The underestimation ($>$) columns for EDC or BIC were excluded, no cases were found. From Table 5 we can identify few characteristics for AIC's hit rates : underestimation for tiny samples; a better performance for small samples; and at stable optimal rate for very large $n$. The sample sizes for this behavior depends heavily on the complexities of the

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

considered cases. More specifically : (i) for small complexity models ($\gamma_m(r) \leq 10$) AIC tends to overestimate $r$ with small probability and performs better than EDC if the sample size is small ($n < 200$); and for large sample size ($n > 1,000$), EDC performs better than AIC; (ii) for medium or large complexity models ($\gamma_m(r) > 20$) overestimation occurs with negligible probability and AIC performs better than EDC up to medium sized sample. For instance, $\gamma_m(r) = 100$ and $n < 5,000$; $\gamma_m(r) = 800$ and $n < 500,000$.

## III. Concluding Remarks

The results show that "small" penalty terms should imply to a tendency of overestimate the true order, likewise "large" penalty terms implies underestimation. Thus, in general, $\hat{r}_{\text{AIC}} \leq \hat{r}_{\text{BIC}} \leq \hat{r}_{\text{EDC}}$.

(i) Both $\hat{r}_{\text{BIC}}$ and $\hat{r}_{\text{EDC}}$ never overestimate the true order.

(ii) $\hat{r}_{\text{BIC}}$ has a higher tendency to underestimate the order.

(iii) $\hat{r}_{\text{EDC}}$ outperforms $\hat{r}_{\text{BIC}}$; it is the most efficient consistent estimator.

(iv) For small complexity models ($\gamma_m(r) < 10$) and small sample size ($n \leq 200$), $\hat{r}_{\text{AIC}}$ performs better than $\hat{r}_{\text{EDC}}$, but may overestimate the true order.

(v) For medium or large complexity models ($\gamma_m(r) > 20$), overestimation by $\hat{r}_{\text{AIC}}$ is negligible and efficiency of $\hat{r}_{\text{AIC}}$ and $\hat{r}_{\text{EDC}}$ are comparable.

(vi) More detailed simulation comparisons as well as theoretical motivations for the estimators behavior vs small and large samples can be found in Resende et al. (2014).

## References

[1] H. Akaike, "A new look at the statistical model identification", *IEER Transactions on Automatic Control*, vol. 19, No. 6, pp. 716-723, 1974.
[2] C.C.Y. Dorea, "Optimal penalty term for EDC Markov chain estimator", *Annales de l'Inst. Stat. l'Univ. de Paris*, vol. 52, pp. 15-26, 2008.
[3] C.C.Y. Dorea and L.C. Zhao, "Bounds for the probability of wrong determination of the order of a Markov chain by using the EDC criterion", *Jour. of Statistical Planning and Inference*, vol. 136, pp. 3689-3697, 2006.
[4] R.W. Katz, "On some criteria for estimating the order of a Markov chain", *Technometrics*, vol. 23, No. 3, pp. 243-249, 1981.
[5] P.A.A. Resende, C.C.Y. Dorea and C.R. Gonçalves, "Comparing the Markov order estimators AIC, BIC and EDC", 2014 (to appear).
[6] G. Schwarz, "Estimating the dimension of a model", *Annals of Statistics*, vol. 6, No. 2, pp. 461-464, 1978.
[7] L.C. Zhao, C.C.Y. Dorea and C.R. Gonçalves, "On determination of the order of a Markov Chain", *Statistical Inference for Stochastic Processes*, vol. 4, No. 3, pp. 273-282, 2001.

ISBN: 978-988-19253-7-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)