Non Asymptotic Performance of Some Markov Chain Order Estimators

A.R. Baigorri*
Mathematics Department
UnB

October 28, 2012

Abstract

In what follows we study non asymptotic behavior of different well known estimators AIC ([20]), BIC ([24]) and EDC ([27, 16]) in contrast with the Markov chain order estimator, named as Global Dependency Level-GDL([6]).

The estimator GDL, is based on a different principle which makes it behave in a quite different form. It is strongly consistent and more efficient than AIC(inconsistent), outperforming the well established and consistent BIC and EDC, mainly on relatively small samples.

The estimators mentioned above mainly consist in the evaluation of the Markov chain’s sample by different multivariate deterministic functions. The log likelihood approach, as in (11),

\[
L((n, a^*)_{(i,k)}) = L_1 \left( L_2 - \pi(i) x_{a^*}(i,k) \log x_{a^*}(i,k) \right)
\]

with deterministic function

\[
L((n, a^*)_{(i,j)}) = L_1 \left( L_2 - \pi(i) x_{i,j} \log x_{i,j} \right), \quad L_1 = \text{const.}, \quad L_2 = \text{const.}
\]

or, the GDL approach, as in (13),

\[
G((n, a^*)_{(i,j)}) \left( ... , x_{a^*}(s,t), ... \right) = \sum_{i=1}^{m} \sum_{j=1}^{m} G((n, a^*)_{(i,j)}) \left( ... , x_{a^*}(s,t), ... \right)
\]

*baig@unb.br
with deterministic function

\[
G_{(n, a^*_1, (i,j))}(..., x(s,t), ...) = \frac{(x(i,j) - \left[ \sum_{t=1}^{m} x(t) \right] \left[ \sum_{s=1}^{m} x(s,j) \right])^2}{(\sum_{t=1}^{m} x(t))(\sum_{s=1}^{m} x(s,j))}
\]

shall be analyzed in Section 3, exhibiting different structural properties. It will become clear the intimate differences existing between the variance of both estimators, which induce quite dissimilar performance, mainly for samples of moderated sizes.

1 Introduction

A Markov Chain is a discrete stochastic process \( X = \{X_n\}_{n \geq 0} \) with state space \( E \), cardinality \( |E| < \infty \) for which there is a \( k \geq 1 \) such that for \((x_1, ..., x_n) \in E^n, n \geq k\)

\[
P(X_1 = x_1, ..., X_n = x_n) = P(X_1 = x_1, ..., X_k = x_k) \Pi_{i=k+1}^{n} Q(x_i|x_{i-k}, ..., x_{i-1})
\]

for suitable transition probabilities \( Q(., .) \). The class of processes that holds the above condition for a given \( k \geq 1 \) will be denoted by \( M_k \), and \( M_0 \) will denote the class of i.i.d. processes. The order of a Markov Chain in \( \bigcup_{i=0}^{\infty} M_i \) is the smallest integer \( \kappa \) such that \( X = \{X_n\}_{n \geq 0} \in M_{\kappa} \).

Along the last few decades there has been a great number of research on the estimation of the order of a Markov Chains, starting with M.S. Bartlett [8], P.G. Hoel [18], I.J. Good [17], T.W. Anderson & L.A. Goodman [4], P. Billingsley [9], [10] among others, and more recently, H. Tong [26], G. Schwarz [24], R.W. Katz [19], I. Csiszar and P. Shields [13], L.C. Zhao et all [27] had contributed with new Markov chain order estimators.

Since 1973, H. Akaike [1] entropic information criterion, known as AIC, has had a fundamental impact in statistical model evaluation problems. The AIC has been applied by Tong, for example, to the problem of estimating the order of autoregressive processes, autoregressive integrated moving average processes, and Markov chains. The Akaike-Tong (AIC) estimator was derived as an asymptotic approximate estimate of the Kullback-Leibler information discrepancy and provides a useful tool for evaluating models estimated by the maximum likelihood method. Later on, Katz derived the asymptotic
distribution of the estimator and showed its inconsistency, proving that there
is a positive probability of overestimating the true order no matter how large
the sample size. Nevertheless, AIC is the most used and successful Markov
chain order estimator used at the present time, mainly because it is more
efficient than BIC for small sample.

2 Essentials on Some Estimators

2.1 Maximum Likelihood Methods

The main consistent estimator alternative, the BIC estimator, does not per-
form too well for relatively small samples, as it was pointed out by Katz [19]
and Csiszar & Shields [13]. It is natural to admit that the expansion of the
Markov Chain complexity (size of the state space and order) has significant
influence on the sample size required for the identification of the unknown
order, even though, most of the time it is difficult to obtain sufficiently large
samples.

Katz (1981) [19] obtained the asymptotic distribution of \( \hat{\kappa}_{AIC} \) and proved its
inconsistency showing the existence of a positive probability to overestimate
the order. See also Shibata (1976) [25]. On the contrary Schwarz (1978) [24]
and Zhao (2001) [27] proved strong consistency for the estimators \( \hat{\kappa}_{BIC} \) and
\( \hat{\kappa}_{EDC} \), respectively.

Clearly, for a given \( \eta \), \( AIC(\eta) \) [26], \( BIC(\eta) \) [24] and \( EDC(\eta) \) [27] [16] contain
much of the information concerning the sample’s relative dependency, never-
thless numerical simulations as well as theoretical considerations anticipates
a great deal of variability for small samples.

Let \( X^n = (X_1, \ldots, X_n) \) be a sample from a multiple stationary Markov chain
\( X = \{X_n\}_{n \geq 1} \) of unknown order \( \kappa \).

Assume that \( X \) take values on a finite state space \( E = \{1, 2, \ldots, m\} \) with
transition probabilities given by

\[
p(x_{n+1}|x^n) = P(X_{n+1} = x_{n+1}|X^n_{n-\kappa+1} = x^n_1) > 0 \tag{2}
\]

where \( x^n_1 = (x_1, \ldots, x_\kappa) = x^n_1 x^n_{j+1} \in E^c. \)
Define
\[ N(x^l_1 | X^n_1) = \sum_{j=1}^{n-l+1} 1(X_j = x_1, ..., X_{j+l-1} = x_l) \quad (3) \]
i.e. the number of occurrences of \( x^l_1 \) in \( X^n_1 \). If \( l = 0 \) we take \( N(., | X^n_1) = n \). The sums are taken over positive terms \( N(x^{l+1}_1 | X^n_1) > 0 \), or else, we convention \( 0/0 \) or \( 0/\infty \) as 0.

**Definition 2.1.** For \( a^n_1 = (a_1, ..., a_\eta) \in E^n \) and \( j \in E \), let \( X_{a^n_1} \) be the empirical random variables, extracted from the Markov chain sample \( X^n_1 = (X_1, ..., X_n) \)
\[
X_{a^n_1} : X^n_1 \rightarrow (X_{a^n_1}(1), ..., X_{a^n_1}(j), ..., X_{a^n_1}(m))
\]
\[
X_{a^n_1}(j) = \left( \frac{N(a^n_1 j | X^n_1)}{N(a^n_1 | X^n_1)} \right), \quad 1 \leq j \leq m
\]
and
\[
X_{a^n_1}(i, j) : X^n_1 \rightarrow (X_{a^n_1 i}(1), ..., X_{a^n_1 i}(j), ..., X_{a^n_1 i}(m))
\]
\[
X_{a^n_1}(i, j) = \left( \frac{N(a^n_1 i j | X^n_1)}{N(a^n_1 i | X^n_1)} \right), \quad 1 \leq i, j \leq m.
\]

Let us define for the order the log likelihood function
\[
\log \hat{L}(\eta) = \sum_{a^n_1} N(a^n_1 | X^n_1) \left( \sum_{a^n_1 j} \frac{N(a^n_1 j | X^n_1)}{N(a^n_1 i | X^n_1)} \log X_{a^n_1}(j) \right)
\]
\[
\log \hat{L}(\eta) = \sum_{a^n_1} N(a^n_1 | X^n_1) \left( \sum_{j=1}^{m} \frac{N(a^n_1 j | X^n_1)}{N(a^n_1 i | X^n_1)} \log X_{a^n_1}(j) \right)
\]
\[
\log \hat{L}(\eta) = \sum_{a^n_1} \left[ N(a^n_1 | X^n_1) \left( \sum_{j=1}^{m} X_{a^n_1}(j) \log X_{a^n_1}(j) \right) \right].
\]

4
The estimators based on likelihood estimators and penalty functions, for Markov chains of order $\kappa$ are defined, under the following hypothesis:

There exist a known $B$ so that $0 \leq \kappa \leq B$

As

$$\hat{\kappa}_{AIC} = \arg\min \{ AIC(\eta) ; \eta = 0, 1, ..., B \} \quad (7)$$

$$\hat{\kappa}_{BIC} = \arg\min \{ BIC(\eta) ; \eta = 0, 1, ..., B \} \quad (8)$$

$$\hat{\kappa}_{EDC} = \arg\min \{ EDC(\eta) ; \eta = 0, 1, ..., B \} \quad (9)$$

where

$$AIC(\eta) = -2 \log \hat{L}(\eta) + |E|^{\eta} - 2(|E| - 1) \left( \frac{\log(n)}{2} \right)$$

$$BIC(\eta) = -2 \log \hat{L}(\eta) + |E|^{\eta} - 2(|E| - 1) \left( \frac{\log \log(n)}{2(|E| - 1)} \right)$$

$$EDC(\eta) = -2 \log \hat{L}(\eta) + |E|^{\eta} - 2(|E| - 1) \left( \frac{\log \log(n)}{2(|E| - 1)} \right) \log \log(n)$$

$$AIC(\eta) \leq EDC(\eta) \leq BIC(\eta).$$

Finally, let us fix $a_1^\eta$ and consider the function

$$L_{[(n, a_1^\eta)]} : (0, 1)^{m^2} \to \mathbb{R}^+$$

defined as:

$$L_{[(n, a_1^\eta)]}(\ldots, x(i, j), \ldots) = N(a_1^\eta | X_1^\eta) \left( \sum_{i=1}^{m} X_{a_1^\eta}(i) \log X_{a_1^\eta}(i) - \sum_{i=1}^{m} \left[ \frac{N(a_1^\eta | X_1^\eta)}{N(a_1^\eta | X_1^\eta)} \sum_{j=1}^{m} X_{a_1^\eta}(i, j) \log X_{a_1^\eta}(i, j) \right] \right). \quad (10)$$

Later on in Section 3.1, we shall analyse the behavior and derivatives of $L_{[(n, a_1^\eta)(i, k)]}$ which is just a generic representation of $L_{[(n, a_1^\eta)(i, k)]}$. 

$$L_{[(n, a_1^\eta)(i, k)]} : (0, 1) \to \mathbb{R}^+$$
\[
L_{(n, \kappa_1, (i, k))} (x(i, j)) = L_1 \left( L_2 - \pi(i) \ x(i, j) \log x(i, j) \right)
\] (11)

such that

\[
L_{(n, \kappa_1, (i, k))} \left( X_{\kappa_1}(i, k) \right) \equiv L_{[(n, \kappa_1), (i, k)]} (x(i, j))
\]

where \( L_1 = N(a_{\eta_1}^i | X_1^n) \) and \( L_2 = \sum_{t=1}^{m} X_{\eta_1}(i) \log X_{\eta_1}(i) \) are assumed constants with respect to the the variables \( x(i, j) \), with \( x(i, j) = X_{\kappa_1}(i, j) \) as in (5), \( \kappa \) the Markov chain order and

\[
\pi(i) = \left[ \frac{N(a_{\eta_1}^i | X_1^n)}{N(a_{\eta_1}^i | X_1^n)} \right], \quad 1 \leq i, j \leq m.
\]

### 2.2 \( \chi^2 \)-divergence estimator

We now briefly recall this new Markov chain order’s estimator referring the reader to ([6]) for related details.

**Definition 2.2.** Let \( X_n = \{X_i\}_{i=1}^n \) be a sample of a Markov chain \( X \) of order \( \kappa \geq 0 \), \( X_{\eta_1}(i, j) \) as in (2.1), \( \eta \geq 0 \) and \( \Delta_2(X_{\eta_1}(i, j)) \) the random variable defined as follows

\[
\Delta_2(X_{\eta_1}(i, j)) =
\]

\[
\sum_{i=1}^{m} \sum_{j=1}^{m} \left( \frac{X_{\eta_1}(i, j) - \left[ \sum_{t=1}^{m} X_{\eta_1}(i, t) \right] \left[ \sum_{s=1}^{m} X_{\eta_1}(s, j) \right]}{\left( \sum_{t=1}^{m} X_{\eta_1}(i, t) \right) \left( \sum_{s=1}^{m} X_{\eta_1}(s, j) \right)} \right)^2
\]

\[
= G_{(n, \eta_1)} \sum_{i=1}^{m} \sum_{j=1}^{m} G_{(n, \eta_1)}(i, j).
\]

Assume that \( V \) is a \( \chi^2 \) random variable with \( (m - 1)^2 \) degrees of freedom where \( P \) is the continuous strictly decreasing function \( P : \mathbb{R}^+ \rightarrow [0, 1] \)

\[
P(x) = P(V \geq x), \quad x \in \mathbb{R}^+.
\]
The Local Dependency Level $LDL_n(a_1^n)$ and the Global Dependency Level $GDL_n(\eta)$, respectively, are defined as follows

$$LDL_n(a_1^n) = \frac{\Delta_2(\mathcal{X}_{a_1^n}(i,j))}{2 \log(\log(n))},$$

$$GDL_n(\eta) = \mathcal{P} \left( \sum_{a_1^n \in E^n} \left( \frac{N(a_1^n \mid \mathcal{X}_{1}^n)}{n} LDL_n(a_1^n) \right) \right) .$$

Finally, let us define the Markov chain order estimator based on the information contained in the vector $GDL_n$.

**Definition 2.3.** Given a fixed number $0 < B \in \mathbb{N}$, let us define the set $
 S = \{0, 1\}^{B+1}$ and the application $T : \mathcal{S} \rightarrow \mathbb{N}$

$$T(s) = -1 \leftrightarrow s_i = 1, \; i = 0, 1, ..., B$$

$$T(s) = \max_{0 \leq i \leq B} \{i : s_i = 0, s_{i+1} = \mathcal{P}(\mathcal{L})\}, \; s = (s_0, s_1, ..., s_B).$$

**Definition 2.4.** Let $\mathcal{X}_1^n = \{X_i\}_{i=1}^n$ be a sample for the Markov chain $\mathcal{X}$ of order $\kappa$, $0 \leq \kappa < B \in \mathbb{N}$ and $\{GDL_n(i)\}_{i=1}^B$ as above. We define the order’s estimator $\kappa_{GDL}(\mathcal{X}_1^n)$ as

$$\hat{\kappa}_{GDL}(\mathcal{X}_1^n) = T(\sigma_n) + 1$$

with $\sigma_n \in \mathcal{S}$ so that $\forall \; s \in \mathcal{S}$

$$\sum_{i=0}^{B} (GDL_n(i) - \sigma_n(i))^2 \leq \sum_{i=0}^{B} (GDL_n(i) - s(i))^2 .$$

Observe that the **Local Dependency Level** $LDL_n(a_1^n)$ entirely relies on the just defined $\chi^2$-square divergence estimator which itself is the summation of several univariate random variables $\mathcal{G}_{[(n,a_1^n)(i,j)]}$, $1 \leq i, j \leq m$. 

7
\[ G[(n, a^\eta_1) (i, j)] = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{X_{a^\eta_1}(i, j) - \left[ \sum_{t=1}^{m} X_{a^\eta_1}(i, t) \right] \left[ \sum_{s=1}^{m} X_{a^\eta_1}(s, j) \right]}{\left( \sum_{t=1}^{m} X_{a^\eta_1}(i, t) \right) \left( \sum_{s=1}^{m} X_{a^\eta_1}(s, j) \right)^2}. \] (12)

Later on in Section 3.2, we shall analyse the behavior of the deterministic function \( G[(n, a^\eta_1) (i, j)] \), \( 1 \leq i, j \leq m \) and their derivatives

\[ G[(n, a^\eta_1) (i, j)] : (0, 1)^{2m} \to \mathbb{R}^+, \quad 1 \leq i, j \leq m \]

with

\[ G[(n, a^\eta_1) (i, j)](\ldots, x(s, t), \ldots) = \frac{(x(i, j) - \left[ \sum_{t=1}^{m} x(i, t) \right] \left[ \sum_{s=1}^{m} x(s, j) \right]^2}{\left( \sum_{t=1}^{m} x(i, t) \right) \left( \sum_{s=1}^{m} x(s, j) \right)}. \] (13)

such that

\[ G[(n, a^\eta_1) (i, j)](\ldots, X_{a^\eta_1}(s, t), \ldots) \equiv G[(n, a^\eta_1) (i, j)](\ldots, x(s, t), \ldots) \]

with \( x(s, t) = X_{a^\eta_1}(i, j), \ 1 \leq i, j \leq m, \ X_{a^\eta_1}(i, j) \) as in (5).

3 Deterministic Accessory Functions

3.1 Functions Related with AIC-Estimator

Let us calculate the derivatives of the deterministic function \( L_{(n, a^\gamma)} \) as in (11), which for the sake of notational simplicity and for a fixed \( n \) and \( a^\gamma \), we’ll temporarily rename the function

\[ L = L_{(n, a^\gamma) (i, k)}, \quad L : D_L \subseteq (0, 1) \to \mathbb{R}, \]
\[ D_L = \{ x(i, k) : x(i, k) \in (0, 1) \}, \]

\[ L(x) = x(i, k) \log(x(i, k)). \]

(14)

**First Order Derivatives:**

\[ \frac{\partial L}{\partial x(i, k)} = 1 + \log(x(i, k)), \quad 1 \leq i, k \leq m - 1. \]

**Second Order Derivatives:**

\[ \frac{\partial^2 L}{\partial x^2(i, k)} = \frac{1}{x(i, k)}, \quad 1 \leq i, k \leq m - 1, \] (15)

\[ \frac{\partial^2 L}{\partial x(j, l) \partial x(i, k)} = 0, \quad 1 \leq i, k, l \leq m - 1, \] (16)

respectively.

Later on we shall obtain the gradient vector and the hessian matrix

\[ \nabla_L (A_{\alpha_1}(o)), \ H_L (A_{\alpha_1}(o)) \]

of the function \( L \) at a point

\[ A_{\alpha_1}(o) = \left( ..., E(X_{\alpha_1}(i, k)), ... \right). \]
3.2 Functions Related with GDL-Estimator

Herein, we shall consider the deterministic multivariate set of functions, as in (13),

\[
G_{(n, a_{\kappa}^i (i,k))}(x(i,k), h(i), v(k)) = \frac{(x(i,k) - h(i)v(k))^2}{h(i)v(k)}, \quad 1 \leq i, k \leq m.
\]

which, after fixing \((i, k)\), we \textit{temporarily rename} it as follows:

\[
G \equiv G_{(n, a_{\kappa}^i (i,k))}, \quad G : D_G \subseteq (0, 1)^3 \rightarrow \mathbb{R},
\]

\[
D_G = \{ x \in (0,1)^3 : x = (x, h, v) \},
\]

\[
G(x) = \frac{(x - hv)^2}{hv}.
\] (17)

\textbf{First Order Derivatives :}

\[
\frac{\partial G}{\partial x} = \frac{2x}{hv} - 2, \quad \frac{\partial G}{\partial h} = \frac{-x^2}{vh^2} + v, \quad \frac{\partial G}{\partial v} = \frac{-x^2}{hv^2} + h.
\]

\textbf{Second Order Derivatives :}

\[
\frac{\partial^2 G}{\partial x^2} = \frac{2}{hv}, \quad \frac{\partial^2 G}{\partial h^2} = \frac{-x^2}{h^3v}, \quad \frac{\partial^2 G}{\partial v^2} = \frac{-x^2}{v^3h},
\]

\[
\frac{\partial^2 G}{\partial h \partial x} = \frac{-2x}{h^2v}, \quad \frac{\partial^2 G}{\partial v \partial x} = \frac{-2x}{hv^2}, \quad \frac{\partial^2 G}{\partial v \partial h} = \frac{x^2}{h^2v^2} + 1.
\] (18) (19)

Likewise, as in the previous subsection we get the gradient vector and the hessian matrix \(\nabla_G (\Gamma_{a_i^1(o)})\), \(\mathcal{H}_G (\Gamma_{a_i^1(o)})\) of the function \(G(x,h,v)\) at a
\[
\Gamma_{a_1^\kappa}(o) = \left( E(\mathcal{X}_{a_1^\kappa}^*(i, k)), E(\mathbb{H}_{a_1^\kappa}(i)), E(\mathcal{V}_{a_1^\kappa}(k)) \right).
\]

4 Multivariate Variances

Focusing on \(G_{((n, a_1^\kappa)) (i, k)}\) for fixed \((i, k)\), \(\kappa\) the order of the Markov chain and \(G_{[(n, a_1^\kappa)] (i, k)}\) as in (13), let us recall the empirical random variables, introduced in Definition 2.1

\[
\mathcal{X}_{a_1^\kappa}(i, k) = \frac{N(i a_1^\kappa | \mathcal{X}_{a_1^\kappa}^\kappa)}{N(a_1^\kappa | \mathcal{X}_{a_1^\kappa}^\kappa)}, \quad \mathbb{H}_{a_1^\kappa}(i) = \sum_{t=1}^{m} \mathcal{X}_{a_1^\kappa}(i, t), \quad \mathcal{V}_{a_1^\kappa}(k) = \sum_{s=1}^{m} \mathcal{X}_{a_1^\kappa}(s, k).
\]

Observe that the Markov chain we are interested in, has order \(\kappa\) and it is clear that \(\mathcal{X}_{a_1^\kappa}(i, k)\) independent random variables, \(1 \leq i, k \leq m\),

are independent, with

\[
G_{[[i, a_1^\kappa] (i, j)]}, \quad 1 \leq i, k \leq m - 1, \tag{20}
\]

as well as for adequatly sample size \(n\) the random variables

\[
E(\mathbb{H}_{a_1^\kappa}(i)) \approx \frac{1}{m}, \quad E(\mathcal{V}_{a_1^\kappa}(k)) \approx m E(\mathcal{X}_{a_1^\kappa}(s, k)), \quad 1 \leq s \leq m.
\]

For the sake of notation’s simplicity, we temporarily rename

\[
G_{[(n, a_1^\kappa)] (i, j)] (x(i, k), h(i), v(k))
\]

as \(G(x, h, v)\) where its derivatives, as well as the variances, covariances and related information of \(\{ \mathcal{X}_{a_1^\kappa}(i, k), \mathbb{H}_{a_1^\kappa}(i), \mathcal{V}_{a_1^\kappa}(k) \}\) shall be as follows:
\[
E \left( \mathbb{X}_{a_{1}}(i, k) \mathbb{V}_{a_{1}}(k) \right) \cong \frac{E \left( \mathbb{X}_{a_{1}}(i, k) \right)}{E \left( \mathbb{V}_{a_{1}}(k) \right)} \cong \frac{E \left( \mathbb{X}_{a_{1}}(i, k) \right)}{\frac{1}{m} m E \left( \mathbb{X}_{a_{1}}(i, k) \right)} \cong 1,
\]

\[
\sum_{\alpha \in \{x, h, v\}} \frac{\partial G}{\partial \alpha} (x, h, v) \frac{\partial G}{\partial \beta} (x, h, v) \cong 0.
\]

\[
\text{cov} \left( \mathbb{X}_{a_{1}}(i, k), \mathbb{V}_{a_{1}}(k) \right) \cong \text{cov} \left( \mathbb{X}_{a_{1}}(i, k), \mathbb{V}_{a_{1}}(k) \right) \cong \text{cov} \left( \mathbb{H}_{a_{1}}(i), \mathbb{V}_{a_{1}}(k) \right) \cong \sigma_{x_{a_{1}}(i, k)}^{2}.
\]

Likewise,
\[
\sum_{\alpha, \beta, \gamma \in \{x, h, v\}} \frac{\partial^2 G}{\partial \alpha^2} (x, h, v) \frac{\partial^2 G}{\partial \beta \partial \gamma} (x, h, v) \cong 0.
\]

\[
\text{cov} \left( \mathbb{X}_{a_{1}}(i, k), \mathbb{X}_{a_{1}}^{2}(i, k) \right) \cong \text{cov} \left( \mathbb{X}_{a_{1}}(i, k), \mathbb{H}_{a_{1}}(i) \right) \cong \text{cov} \left( \mathbb{X}_{a_{1}}(i, k), \mathbb{V}_{a_{1}}^{2}(k) \right) \cong \sigma_{x_{a_{1}}(i, k)}^{3}.
\]

Finally,
\[
\sum_{\alpha, \beta \in \{x, h, v\}} \frac{\partial^2 G}{\partial \alpha^2} (x, h, v) \frac{\partial^2 G}{\partial \beta^2} (x, h, v) \cong \left( \frac{\partial^2 G}{\partial x^2} \right)^2 + \frac{\partial^2 G}{\partial x^2} \frac{\partial^2 G}{\partial v^2} + \frac{\partial^2 G}{\partial x^2} \frac{\partial^2 G}{\partial v \partial x} + \frac{\partial^2 G}{\partial x^2} \frac{\partial^2 G}{\partial v \partial x} + \frac{\partial^2 G}{\partial v^2} \left( \frac{\partial^2 G}{\partial v \partial x} \right)^2 = \frac{1}{h^2 v^2} \left( 4 + \frac{x^4}{v^4} + \frac{4x^3}{v^3} - \frac{8x}{v} \right) \cong 4 \frac{1}{x^2} \left( 4 + \frac{1}{m^4} + \frac{4}{m^3} - \frac{8}{m} \right) \cong \frac{4}{x^2}.
\]
\[ \text{cov}(X_{a_1^2}^2(i,k), X_{a_1^2}^2(i,k)) \cong \text{cov}(X_{a_1^2}^2(i,k), H_{a_1^2}(i)) \cong \text{cov}(X_{a_1^2}^2(i,k), V_{a_1^2}(k)) \cong \sigma_{X_{a_1^2}}^4(i,k) \]

\[ \text{cov}(H_{a_1^2}^2(i), H_{a_1^2}^2(i)) \cong \text{cov}(H_{a_1^2}^2(i), V_{a_1^2}(k)) \cong \text{cov}(V_{a_1^2}^2(k), V_{a_1^2}^2(k)) \cong \sigma_{X_{a_1^2}}^4(i,k). \]

Let us denote by \( B \in \mathbb{R}^3 \), the unit ball centered at the point

\[ \Gamma_{a_1^2}(o) = \left( E(X_{a_1^1}(i,k)), E(H_{a_1^1}(i)), E(V_{a_1^1}(k)) \right) \]

with

\[ \omega = \left( X_{a_1^1}(i,k), H_{a_1^1}(i), V_{a_1^1}(k) \right), \Delta = \omega - \Gamma_{a_1^2}(o). \]

Taylor ([5]) showed that there exist 0 < \( c_g, c_l < 1 \) such that

\[ G_{((n,a_1^2)(i,j))}(\omega) = G_{(n,a_1^2)[i,j]}(\Gamma_{a_1^2}(o)) + \nabla G_{(n,a_1^2)[i,j]}(\Gamma_{a_1^2}(o)) \cdot (\omega - \Gamma_{a_1^2}(o)) + \frac{1}{2!} (\omega - \Gamma_{a_1^2}(o)) \cdot \Delta \]

where the variance of

\[ G_{((n,a_1^2)(i,j))}(X_{a_1^1}(i,k), H_{a_1^1}(i), V_{a_1^1}(k)) \]

is

\[ \sigma_{G_{((n,a_1^2)(i,j))}}^2 = \mathbb{E} \left[ G_{(n,a_1^2)[i,j]}(\omega) - G_{(n,a_1^2)[i,j]}(\Gamma_{a_1^2}(o)) \right]^2 = \mathbb{E} \left[ \nabla G_{(n,a_1^2)[i,j]}(\Gamma_{a_1^2}(o)) \cdot \Delta + \frac{1}{2!} \Delta \cdot \Delta \cdot H_{G_{(n,a_1^2)[i,j]}(\Gamma_{a_1^2}(o) + c_g (\omega - \Gamma_{a_1^2}(o)))} \cdot \Delta \right]^2. \]
\[ \sigma^2_{G[(n, a\kappa_1)](i,k)} \approx \frac{1}{2!^2} \Phi''_g \left( \Gamma_{a\kappa_1} (\omega) + c_g (\omega - \Gamma_{a\kappa_1} (\omega)) \right) \sigma^4_{X_{a\kappa_1}(i,k)} \]

where \( \omega = \left( X_{a\kappa_1}(i, k), H_{a\kappa_1}(i), V_{a\kappa_1}(k) \right) \in \mathcal{B} \) and

\[ \Phi''_g(x(i, k), h(i), v(k)) = \left( \frac{1}{h^2(i)v^2(k)} \right) \left[ 4 + \frac{x^4(i, k)}{v^4(k)} + 4 \frac{x^3(i, k)}{v^3(k)} - 8 \frac{x(i, k)}{v(k)} \right] \]

\[ \approx \frac{4}{x(i, k)} \]

with variance for \( 1 \leq i, k \leq m - 1 \)

\[ \sigma^2_{G[(n, a\kappa_1)](i,k)} \approx \frac{\sigma^4_{X_{a\kappa_1}(i,k)}}{\left[ \Gamma_{a\kappa_1} (\omega) + c_g (\omega - \Gamma_{a\kappa_1} (\omega)) \right]^2} \]

and, by (20), the total variance

\[ \sigma^2_{G([n, a\kappa_1])} = \sum_{i=1}^{m-1} \sum_{k=1}^{m-1} \sigma^2_{G[(n, a\kappa_1)](i,k)} \]

Exactly as before we can obtain the total variance of \( L_{(n, a\kappa_1)}(X_{a\kappa_1}(i, k)) \) and defining

\[ \sigma^2_{L_{(n, a\kappa_1)}(i,k)} = \]

\[ = E \left[ L_{(n, a\kappa_1)}(i, k) \right] \left( X_{a\kappa_1}(i, k) \right) - L_{(n, a\kappa_1)}(i, k) \left( E(X_{a\kappa_1}(i, k)) \right) \right]^2 = \]

\[ = E \left[ \nabla L_{(n, a\kappa_1)}(i, k) \cdot \left( X_{a\kappa_1}(i, k) - E(X_{a\kappa_1}(i, k)) \right) + \frac{1}{2!} H_{L_{(n, a\kappa_1)}}(i, k) \left( E(X_{a\kappa_1}) + c_l [x(i, k) - E(X_{a\kappa_1})] \right) \left( X_{a\kappa_1}(i, k) - E(X_{a\kappa_1}(i, k)) \right) \right]^2. \]

\[ \sigma^2_{L_{(n, a\kappa_1)}(i,k)} \approx \left[ 1 + \ln(x(i, k)) \right]^2 \sigma^2_{X_{a\kappa_1}(i,k)} + \Phi''_g (x(i, k)) \sigma^4_{X_{a\kappa_1}(i,k)}, \ 1 \leq i, k \leq m \]

14
where \( \omega = (x(i, k), h(i), v(k)) \in B \),

\[
\Phi'(x(i, k)) = \left[ \frac{1}{x^2(i, k)} \right]
\]

with variance for \( 1 \leq i, k \leq m - 1 \)

\[
\sigma^2_{L_{(n, a \kappa_1)}(i, k)} \cong \left[ 1 + \ln(x(i, k)) \right]^2 \sigma^2_{x(i, k)}(i, k) + \frac{\sigma^2_{x(i, k)}(i, k)}{\left( E(X_{a \kappa_1}) + c_l \left[ x(i, k) - E(X_{a \kappa_1}) \right] \right)^2},
\]

and, by (20), the total variance

\[
\sigma^2_{L_{(n, a \kappa_1)}} \cong \sum_{k=1}^{m-1} \sum_{i=1}^{m-1} \sigma^2_{L_{(n, a \kappa_1)}(i, k)}.
\]

5 Conclusion

The purpose of this work was the comparative analysis of the non asymptotic behavior for the estimators \( AIC(7), BIC(8), EDC(9) \), versus the estimator defined in Definition 2.2 and named as \( \text{Global Depency Level-GDL} \), for details see (6).

The \( \text{GDL} \) uses a function different to the log likelihood function applied to the sample, which makes the estimator perform in a quite different form. It is strongly consistent and more efficient than \( \text{AIC} \) (inconsistent), outperforming the well established and consistent \( \text{BIC} \) and \( \text{EDC} \), mainly on reasonable small samples.

The estimators just mentioned are based on the composition of the empirical random variables with two different \( \text{deterministic} \) functions. The log likelihood approach, as in (11), with

\[
L_{(n, a \kappa_1)(i, k)}(x(i, k)) = \mathcal{L}_1(\mathcal{L}_2 - \pi(i) x(i, k) \log x(i, k))
\]

or, the \( \text{GDL} \) approach, as in (12), with
Since the sample only depends on the Markov chain $X_n^a$ and its size $n$, once the sample is chosen, the entirely responsible for the estimator’s variance are the following random variables:

$$\mathbb{I}_{[(n, a) \kappa_1]} = \mathcal{L}_1 \left( \mathcal{L}_2 - \pi(i) X_{a \kappa_1}(i, k) \log X_{a \kappa_1}(i, k) \right)$$

and

$$G_{[(n, a) \kappa_1]} = \sum_{k=1}^m \sum_{i=1}^m \frac{\left( X_{a \kappa_1}(i, k) - \left[ \sum_{t=1}^m X_{a \kappa_1}(i, t) [ \sum_{s=1}^m X_{a \kappa_1}(s, k) ] \right] \right)^2}{\left( \sum_{t=1}^m X_{a \kappa_1}(i, t) [ \sum_{s=1}^m X_{a \kappa_1}(s, k) ] \right)^2}$$

with variances for $1 \leq i, k \leq m - 1$

$$\sigma^2_{\mathbb{I}_{[(n, a) \kappa_1]}} \approx \left[ 1 + \ln(x(i, k)) \right]^2 \sigma^2_{X_{a \kappa_1}(i, k)} + \frac{\sigma^4_{X_{a \kappa_1}(i, k)}}{\left( E(X_{a \kappa_1}) + c_l [ x(i, k) - E(X_{a \kappa_1}) ] \right)^2},$$

and

$$\sigma^2_{G_{[(n, a) \kappa_1]}} \approx \frac{4 \sigma^4_{X_{a \kappa_1}(i, k)}}{[ \Gamma_{a \kappa_1}(o) + c_g (\omega - \Gamma_{a \kappa_1}(o)) ]^2}. $$

Finally the reader should notice that the log likelihood based estimators are heavily affected by $\log(x(i, k))$ which in cases where the Markov chain intrinsically presents empirical random variables $X_{a \kappa_1}(i, k)$ with small expectations, the fluctuating values of $x(i, k)$ converging to $E(X_{a \kappa_1}(i, k)) \approx 0$ imposes the coefficients $[ 1 + \log(x(i, k)) ]^2$ and its variance $\sigma^2_{\mathbb{I}_{[(n, a) \kappa_1]}}$ a great deal of instability or variance.
The following Appendix presents a few examples exhibiting such anomaly.

6 Appendix

6.1 Numerical Evidence

In what follows we shall compare the non-asymptotic performance, mainly for small samples, of some of the most used Markov chains order estimators.

It is quite intuitive that the random information regarding the order of a Markov chain, is spread over an exponentially growing set of empirical distributions \( \Theta \) with \(|\Theta| = n^{B+1} \), where \( B \) is the maximum integer \( \eta \), as in \( \alpha = (i_1 i_2 \ldots i_\eta) \). It seems reasonable to think that a small viable sample, i.e. samples able to retrieve enough information to estimate the chain order, should have size \( n \approx O(n^{B+1}) \). Keeping in mind that for the present numerical simulation, the maximum length to be used is \( B = 5 \), from now on the sample sizes for \(|E| = 3 \) and \(|E| = 4 \) should be \( n \approx 1.500 \) and \( n \approx 5.000 \), respectively.

The following numerical simulation, based on an algorithm due to Raftery[23], starts on with the generation of a Markov chain transition matrix, \( Q = (q_{i_1 i_2 \ldots i_\kappa; i_{\kappa+1}}) \) with entries

\[
q_{i_1 i_2 \ldots i_\kappa; i_{\kappa+1}} = \sum_{t=1}^{\kappa} \lambda_{i_t} R(i_{\kappa+1}, i_t), \quad 1 \leq i_t, i_{\kappa+1} \leq m.
\]  

(21)

where the matrix

\[
R(i, j), \quad 0 \leq i, j \leq m, \quad \sum_{i=1}^{m} R(i, j) = 1, \quad 1 \leq j \leq m
\]

and the positive numbers

\[
\{\lambda_i\}_{i=1}^{\kappa}, \quad \sum_{i=1}^{\kappa} \lambda_i = 1
\]

are arbitrarily chosen in advance.

Once the matrix \( Q = (q_{i_1 i_2 \ldots i_\kappa; i_{\kappa+1}}) \) is obtained, two hundreds replications of the Markov chain sample of size \( n \), space state \( E \) and transition matrix \( Q \)
are generated to compare $GDL(\eta)$ performance against the standards, well
known and already established order estimators just mentioned above.
Finally, after applying all estimators to each one of the replicated samples,
the final results two hundreds replications are registered in the form of tables.

**Case I: Markov Chain Examples with** $\kappa = 0$, $|E| = 3$.

Firstly, we choose the matrix \{$Q_1, Q_2, Q_3$\} to produce samples with sizes
$500 \leq n \leq 2,000$, originated from Markov chains of order $\kappa = 0$ with quite
different probability distributions.

\[
Q_1 = \begin{bmatrix}
0.33 & 0.335 & 0.335 \\
0.33 & 0.335 & 0.335 \\
0.33 & 0.335 & 0.335
\end{bmatrix},
Q_2 = \begin{bmatrix}
0.05 & 0.475 & 0.475 \\
0.05 & 0.475 & 0.475 \\
0.05 & 0.475 & 0.475
\end{bmatrix},
Q_3 = \begin{bmatrix}
0.05 & 0.05 & 0.90 \\
0.05 & 0.05 & 0.90 \\
0.05 & 0.05 & 0.90
\end{bmatrix}.
\]

| $|E| = 3$ | $\leftrightarrow$ | $\kappa = 0$ | $\leftrightarrow$ | $\lambda_i = 1/3$, $i = 1,2,3$. |
|---|---|---|---|---|
| $Q_1$ | $Q_1$ | $Q_1$ |
| $n = 500$ | $n = 1,000$ | $n = 1,500$ |
| $k$ | Aic | Bic | Educ | Gdl | Aic | Bic | Educ | Gdl | Aic | Bic | Educ | Gdl |
| 0 | 75.5% | 100% | 100% | 99% | 80% | 100% | 100% | 99.5% | 71.5% | 100% | 100% | 99% |
| 1 | 24.5% | 1% | 18% | 0.5% | 22.5% | 1% |
| 2 | 2% | 6% |
| 3 |
| 4 |

| $|E| = 3$ | $\leftrightarrow$ | $\kappa = 0$ | $\leftrightarrow$ | $\lambda_i = 1/3$, $i = 1,2,3$. |
|---|---|---|---|---|
| $Q_2$ | $Q_2$ | $Q_2$ |
| $n = 1,000$ | $n = 1,500$ | $n = 500$ |
| $k$ | Aic | Bic | Educ | Gdl | Aic | Bic | Educ | Gdl | Aic | Bic | Educ | Gdl |
| 0 | 63.5% | 100% | 100% | 99% | 63% | 100% | 100% | 99% | 59% | 100% | 100% | 99% |
| 1 | 29% | 1% | 34.5% | 1% | 37% | 1% |
| 2 | 7.5% | 2.5% | 4% |
| 3 |
| 4 |
$|E| = 3 \leftrightarrow \kappa = 0 \leftrightarrow \lambda_i = 1/3, i = 1,2,3.$

|       | $Q_3$       | $Q_3$       | $Q_3$       |
|-------|-------------|-------------|-------------|
|       | $n = 1.000$ | $n = 1.500$ | $n = 2.000$ |
| $k$   | $Aic$ | $Bic$ | $Edc$ | $Gdl$ | $Aic$ | $Bic$ | $Edc$ | $Gdl$ | $Aic$ | $Bic$ | $Edc$ | $Gdl$ |
| 0     | 43%     | 100% | 100% | 98% | 47%     | 100% | 100% | 99.5% | 96% | 46%     | 100% | 100% | 97% |
| 4     | 53%     | 2%   | 51.5% | 0.5% | 4%      | 50.5% | 2%   | 51.5% | 0.5% | 3.5%     | 2%   | 51.5% | 0.5% |
| 3     | 4%      | 1.5% | 3.5% | 1%  | 4%      | 3.5% | 1%   | 3.5% | 1%  | 3.5%     | 1%   | 3.5% | 1%  |

Notice that for a fixed sample size $n = \{500, 1.000, 1.500, 2.000\}$, the order estimator $\hat{\kappa}_{AIC}$ steadily overestimate the real order $\kappa = 0$ with the excessiveness depending on the probability distribution of the Markov chain. Differently, the order estimators $\hat{\kappa}_{BIC}$, $\hat{\kappa}_{EDC}$ and $\hat{\kappa}_{GDL}$ show consistent performance, mainly obtaining the right order, free from the influence of the sample size and the generating matrix. Regarding $\hat{\kappa}_{BIC}$ and $\hat{\kappa}_{EDC}$ improved effect, most likely depends on their correcting factor, $\frac{\log(n)}{2}$ and $\left(\frac{\log\log(n)}{2(|E|-1)}\right)$ which tend to decrease the estimated order.

For $|E| = 4$ the greater complexity of a Markov chain of order $\kappa = 3$ impose the use of larger sample size for estimators to accomplish some reliability. Finally, we choose the matrix $\{Q_6, Q_7\}$ to produce samples with size $n = 5.000$, originated from Markov chains of order $\kappa \in \{2, 3, 0\}$ like in the previous cases.

\[
Q_6 = \begin{bmatrix}
0.05 & 0.05 & 0.05 & 0.85 \\
0.05 & 0.05 & 0.85 & 0.05 \\
0.05 & 0.85 & 0.05 & 0.05 \\
0.85 & 0.05 & 0.05 & 0.05
\end{bmatrix}, \\
Q_7 = \begin{bmatrix}
0.05 & 0.05 & 0.05 & 0.85 \\
0.05 & 0.05 & 0.05 & 0.85 \\
0.05 & 0.05 & 0.05 & 0.85 \\
0.05 & 0.05 & 0.05 & 0.85
\end{bmatrix}.
\]
For the order for $|E| = 4$, $\kappa = 0$, apparently $\hat{\kappa}_{AIC}$ keeps overestimating the order in some degree, while $\hat{\kappa}_{BIC}$ as in example $\kappa = 3$ severely underestimate the order, presumably due to the excessive weight of the correcting factors $\frac{\log(n)}{2}$. On the contrary $\hat{\kappa}_{EDC}$ and $\hat{\kappa}_{GDL}$ behaves quite well in same setting.

### References

[1] H. Akaike, “A New Look at the Statistical Model Identification”, *IEEE Trans. Autom. Cont.*, vol. 19, 1998.

[2] H. Akaike & G. Kitagawa, (Eds.), “The Practice of Time Series Analysis”, *Springer-Verlag, New York*, 1974.

[3] M. S. Ali & D. Silvey, “A general class of coefficients of divergence of one distribution from another”, *Journal of the Royal Statistical Society, Ser. B*, No. 28, pp. 131-140, 1966.

[4] T. W. Anderson & Leo A. Goodman, “Statistical Inference about Markov Chains”, *Annals of Mathematical Statistics*, vol. 28, 1957.

[5] Tom M. Apostol, “Calculus”, *John Wiley & Sons, Vol. 1, Second Edition*, 1967.

[6] A.R. Baigorri, C.R. Gonçalves & P.A. Resende, “Markov Chain Order Estimation and $\chi^2$–divergence measure”, [arXiv:0910.0264v4/math.ST], 2012.
[7] A. Banerjee, X. Guo & H. Wang, “The Optimallity of Conditional Expectation as a Bregman Predictor”, IEEE Transactions on Information Theory, vol. 51, No 7, 2005.

[8] M. S. Bartlett, “The Frequency Goodness of Fit Test for Probability Chains”, Math. Proc. Camb. Phil. Soc., vol. 9, 1951.

[9] P. Billingsley, “Statistical Methods in Markov Chains”, Annals of Mathematical Statistics, Vol. 32, No 1, 1961.

[10] P. Billingsley, “Statistical Inference for Markov Chains”, University of Chicago Press, Chicago, 1961.

[11] W. Cochran, “The $\chi^2$ Test of Goodness of Fit”, Annals of Mathematical Statistics, vol. 23, No 3, 1952.

[12] I. Csiszar, “Information-type measures of difference of probability distributions and indirect observation”, Studia Sci. Math. Hungar., Vol. 2, pp. 229-318, 1967.

[13] I. Csiszar & P. Shields, “The Consistency of the BIC Markov Order Estimator”, The Annals of statistics, Vol 28, No 6, pp. 1601-1619, 2000.

[14] I. Csiszar & P. Shields, “Information Theory and Statistics: A Tutorial”, Foundations and Trends in Communications and Information Theory, Vol. 1, No 4, pp. 417-528, 2004.

[15] J.L. Doob, “Stochastic Processes(Wiley Publication in Statistics)”, John Wiley & Sons. Inc., 1966.

[16] C.C.Y. Dorea,”Optimal Penalty Term for EDC Markov Chain Order Estimator”, Annales de l’Institut de Statistique de l’Universite de Paris (l’ISUP), v. 52, p. 15-26, 2008.

[17] I.J. Good, “The Likelihood Ratio Test for Markov Chains”, Biometrica, vol. 42, (3/4), pp. 531-533, 1955.

[18] P. G. Hoel, “A test for Markov Chains”, Biometrika, vol. 41, 1954.

[19] R. W. Katz, “On Some Criteria for Estimating the Order Of Markov Chains”, Technometrics, vol. 23(3), 1981.

[20] S.P. Meyn & R.L. Tweedie, “Markov Chain and Stochastic Stability”, Springer-Verlag, London, 1993.
[21] L. Pardo, “Statistical Inference Based on Divergence Measures”, Taylor & Francis Group, LLC, New York, 2006.

[22] R Development Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL [http://www.R-project.org](http://www.R-project.org), 2008.

[23] A. E. Raftery, “A model for High-Order Markov Chains”, Journal of the Royal Statistical Society, Serie B, vol. 47, No. 3, 1985.

[24] G. Schwarz, “Estimating the Dimension of a Model”, The Annals of Statistics, vol. 6, No 2, 1978.

[25] R. Shibata, “Selection of the Order of an Autoregressive model by Akaike’s Information Criterion.”, Biometrika, vol. 6, 1976.

[26] H. Tong, “Determination of the order of a Markov Chain by Akaike’s Information Criterion”, Journal of Applied Probability, vol. 12, pp. 488-497, 1975.

[27] H. C. Zhao, C.C.Y. Dorea & C.R. Gonçalves, “On Determination of the Order of a Markov Chain”, Stat. Infer. for Stoc. Processes, vol. 4, 2001.