Markov Chain Order Estimation and \( \chi^2 - divergence \) measure

A.R. Baigorri\(^{*}\)  
Mathematics Department  
UnB  

C.R. Gonçalves \( \dagger \)  
Mathematics Department  
UnB  

P.A.A. Resende \( \ddagger \)  
Mathematics Department  
UnB

May 28, 2011

Abstract

We use the \( \chi^2 - divergence \) as a measure of diversity between probability densities and review the basic properties of the estimator \( D_2(\|.)\). In the sequence we define a few objects which capture relevant information from the sample of a Markov Chain to be used in the definition of a couple of estimators i.e. the Local Dependency Level and Global Dependency Level for a Markov chain sample. After exploring their properties we propose a new estimator for the Markov chain order. Finally we show a few tables containing numerical simulation results, comparing the performance of the new estimator with the well known and already established AIC and BIC estimators.

\( ^* \)baig@unb.br  
\( ^\dagger \)The author was partially supported by PROCAD/CAPES, CASADINHO/CAPES and PRONEX/FAPDF — catia@mat.unb.br  
\( ^\ddagger \)pa@mat.unb.br
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 $n \geq k$, $(x_1, ..., x_n) \in E^n$

$$P(X_1 = x_1, ..., X_n = x_n) = P(X_1 = x_1, ..., X_k = x_k)\prod_{i=k+1}^{\infty} Q(x_i|x_{i-k}, ..., x_{i-1})$$

for suitable transition probabilities $Q(\cdot|\cdot)$. 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 process 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 [6], P.G. Hoel [16], I.J. Good [15], T.W. Anderson & L.A. Goodman [4], P. Billingsley [7], [8] among others, and more recently, H. Tong [23], G. Schwarz [22], R.W. Katz [17], I. Csiszar and P. Shields [11], L.C. Zhao et all [24] 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 successfull Markov chain order estimator used at the present time, mainly because it is more efficient than BIC for small sample.

The main consistent estimator alternative, the BIC estimator, does not perform too well for relatively small samples, as it was pointed out by Katz [17] and Csiszar & Shields [11]. 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.

In this notes we’ll use a different entropic object called $\chi^2$–divergence, and study its behaviour when applied to samples from random variables with
multinomial empirical distributions

\[ X' = \{X_i\}_{1\leq i \leq r} \]

derived from a Markov Chain sample. Finally, we shall propose a new strongly consistent Markov Chain order estimator more efficacious than the already established AIC and BIC, which it shall be exhibited through the outcomes of several numerical simulations.

In Section 2 we succinctly review the concept of \( f \) – divergence and its properties. In Section 3, the \( \chi^2 \)-divergence estimator is defined reviewing some results concerning its convergence, as well as we briefy elaborate about the Law of Iterated Logarithm (LIL) for our particular situation. In Section 4 the Markov chain sample is brought to attention, some notation introduced and the estimators Local Dependency Level and Global Dependency Level, which are the groundsis of the consistent Markov chain order estimator, subsequently defined. Finally, in Section 4 we describe the procedures used and the results obtained in an exploratory numerical simulations.

2 Entropy and \( f \)-divergences

2.1 Definitions and Notations

An \( f \) – divergence is a function that measures the discrepancy between two probability distributions \( P \) and \( Q \). The divergence is intuitively an average of the function \( f \) of the odds ratio given by \( P \) and \( Q \).

These divergences were introduced and studied independently by Csiszar, Csiszar&Shields and Ali&Silvey among others ([10], [12], [3]) and sometimes are referred as Ali-Silvey distances.

**Definition 2.1.** Let \( P \) and \( Q \) be discrete probability densities with support \( S(P) = S(Q) = E = \{1, ... , m\} \). For \( f(t) \) convex function defined for \( t > 0 \) and \( f(1) = 0 \), the \( f \) – divergence for the distributions \( P \) and \( Q \) is defined as

\[
D_f(P\|Q) = \sum_{a \in A} Q(a) f \left( \frac{P(a)}{Q(a)} \right).
\]

Here we take \( 0f(0) = 0 \), \( f(0) = \lim_{t \to 0} f(t) \), \( 0f(\frac{0}{0}) = \lim_{t \to 0} t f(\frac{\frac{0}{0}}{0}) = a \lim_{u \to \infty} f(u) \).
For example:

\[ f(t) = t \log(t) \Rightarrow D_f(P\|Q) = D(P\|Q) = \sum_{a \in A} P(a) \log \left( \frac{P(a)}{Q(a)} \right) \]

\[ f(t) = (1 - t^2) \Rightarrow D_f(P\|Q) = \sum_{a \in A} \frac{(P(a) - Q(a))^2}{Q(a)} , \]

which are called relative entropy and \( \chi^2 \)-divergence, respectively. From now on the \( \chi^2 \)-divergence shall be denote by \( D_2(P\|Q) \).

Observe that the triangular inequality is not satisfied in general, so that \( D_2(P\|Q) \) defines no distance in the strict sense.

A basic theorem about \( f \)-divergences is the following approximation by the \( D_2(P\|Q) \).

**Theorem 2.1.** (Csiszar & Shields [12]) If \( f \) is twice differentiable at \( t=1 \) and \( f''(1) > 0 \) then for any \( Q \) with support \( S(Q) = A \) and \( P \) close to \( Q \)

\[ D_f(P\|Q) \sim \frac{f''(1)}{2} D_2(P\|Q) . \]

Formally, \( D_f(P\|Q)/D_2(P\|Q) \to f''(1)/2 \) as \( P \xrightarrow{D} Q \)

The \( \chi^2 \)-square divergence \( D_2(P\|Q) \) test is well known statistical test procedure close related to the chi-square distribution. See [19] for thorough and detailed references.

### 2.2 The \( \Delta_2 \) Estimator

Now we’ll consider a set \( \mathcal{X} = \{X_1, ..., X_r\} \) of discrete independent random variables and \( \{P_{X_i}\}_{1 \leq i \leq r} \) their probability distributions with common support \( S(P_{X_i}) = E = \{1,2, ..., m\}, \ i = 1, ..., r \).

Let \( X_i = (X_i^{(1)}, ..., X_i^{(m)}) \) be a random sample of \( X_i \) with size \( n_i \), respectively, where \( n = \sum_{i=1}^{r} n_i \). To test the distribution’s homogeneity of \( \{P_{X_i}\}_{i \leq i \leq r} \), we compare the empirical observed and expected frequencies, by means of the \( \chi^2 \)-divergency.
Definition 2.2. Let the observed empirical random variables defined as

\[ O_n^X(i, j) = |k : X_i^{(k)} = j|, \quad i = 1, ..., r, \quad j \in E, \]

the expected empirical random variables given by

\[ E_n^X(i, j) = \frac{\sum_l O_n^X(i, l) \sum_l O_n^X(l, j)}{\sum_{kl} O_n^X(k, l)}, \quad i = 1, ..., r, \quad j \in E, \]

and the respective probability functions

\[ P_{O_n^X}(i, j) = \frac{O_n^X(i, j)}{n}, \quad i = 1, ..., r, \quad j \in E, \]

\[ P_{E_n^X}(i, j) = \frac{E_n^X(i, j)}{n}, \quad i = 1, ..., r, \quad j \in E. \]

Let us define the estimator

\[ \hat{\Delta}_2(P_{O_n^X} \| P_{E_n^X}) = n \sum_{i=1}^{r} \sum_{j=1}^{m} \frac{(P_{O_n^X}(i, j) - P_{E_n^X}(i, j))^2}{P_{O_n^X}(i, j)} = n \Delta_2(P_{O_n^X} \| P_{E_n^X}). \]

Remark 2.2. The above empirical random variables \( \hat{\Delta}_2(P_{O_n^X} \| P_{E_n^X}) \) is the chi-squared hypothesis test applied to a 2-dimensional contingency table where \( O_n^X(i, .) \) are the observed frequencies of \( X_i \) and \( E_n^X(i, .) \), the expected frequencies under the assumption that \( X_i \in \mathcal{X} \), are independent identically distributed.

3 Derived Markov Chains

Let \( X_1^r = (X_1, ..., X_n) \) be sample from a multiple stationary Markov chain \( \mathcal{X} = \{X_n\}_{n \geq 1} \) of unknown order \( \kappa \). Assume that \( \mathcal{X} \) take value on a finite with state space \( E = \{1, 2, ..., m\} \) with transition probabilities given by

\[ p(x_{\kappa+1}|x_1^\kappa) = P(X_{n+1} = x_{n+1}|X_n^{n-\kappa+1} = x_1^\kappa) > 0 \quad (1) \]
where \( x_1^\kappa = x_1^\kappa x_{j+1}^\kappa = (x_1, \ldots, x_\kappa) \in E^\kappa \).

Following Doob [13], from the process \( X \) we can derive a first order MC, \( \mathcal{Y}^{(\kappa)} = \{Y_n^{(\kappa)}\}_{n \geq 0} \) by setting \( Y_n^{(\kappa)} = (X_n, \ldots, X_{n+\kappa-1}) \) so that for \( v = (i_1, \ldots, i_\kappa) \) and \( w = (i_1', \ldots, i'_\kappa) \)

\[
P(Y_{n+1}^{(\kappa)} = w|Y_n^{(\kappa)} = v) = \tilde{p}_{vw} = \begin{cases} p(i'_\kappa|i_1 \ldots i_\kappa), & i'_j = i_{j+1}, \ j = 1, \ldots, (\kappa - 1) \\ 0, & \text{otherwise.} \end{cases}
\]

Clearly \( \mathcal{Y}^{(\kappa)} \) is a first order and homogeneous MC that will be called the derived process, which by (1) is an irreducible and positive recurrent MC having unique stationary distribution, say \( \Pi^{(\kappa)} \). It is well known, see [13, Chap. 5.3], that the derived Markov Chains \( \mathcal{Y}^{(l)} \), \( l \geq \kappa \) is irreducible and aperiodic, consequently ergodic.

There exists an equilibrium (stationary) distribution \( \Pi^{(\kappa)}(.) \) satisfying for any initial distribution \( \nu \) on \( E^\kappa \)

\[
\lim_{n \to \infty} |P_{\nu}(Y_n^{(\kappa)} = x_1^\kappa) - \Pi^{(\kappa)}(x_1^\kappa)| = 0,
\]

and

\[
\Pi^{(\kappa)}(x_1^\kappa) = \sum_{z_1^\kappa} \Pi^{(\kappa)}(z_1^\kappa) p(x_\kappa|z_1^\kappa) = \sum_x \Pi^{(\kappa)}(x x_1^{\kappa-1}) p(x_\kappa|x x_1^{\kappa-1}).
\]

Likewise, for \( \mathcal{Y}^{(l)} \), \( l > \kappa \)

\[
\Pi_l(x_1^l) = \Pi^{(\kappa)}(x_1^\kappa) p(x_{\kappa+1}|x_1^{\kappa}) \ldots p(x_l|x_1^{l-1}) = \sum_x \Pi_l(x x_1^{l-1}) p(x_l|x x_1^{l-1}).
\]

which shows that \( \Pi_l \) defined above, is a stationary distribution for \( \mathcal{Y}^{(l)} \). For the sake of notation’s simplicity we’ll use, from now on

\[
\Pi(a_1^l) = \Pi_l(a_1^l), \ l \geq \kappa.
\]

Now, let us shift our attention back to \( X_1^n = (X_1, X_2, \ldots, X_n) \) and define

\[
N(x_1^l|X_1^n) = \sum_{j=1}^{n-l+1} 1(X_j = x_1, \ldots, X_{j+l-1} = x_l)
\]
i.e. the number of occurrences of \( x_i^l \) in \( X_n^a \). If \( l = 0 \) we take \( N(\cdot \mid X_n^a) = n \). The sums are taken over positive terms \( N(x_l^{l+1} \mid X_n^a) > 0 \), or else, we convention \( 0/0 \) or \( 0/\infty \) as 0.

Now we define the empirical random variables \( X_{i\alpha} \), for \( i \in E \) and \( \alpha \in E^n \).

**Definition 3.1.** For \( \alpha = (a_1, \ldots, a_\eta) = a_\eta^n \in E^n \) and \( i \in E \), let \( X_{i\alpha} \) be the random variable taking values in \( E \), extracted from the MC sample \( X_n^a \), defined as

\[
P(X_{i\alpha} = l) = \frac{N(i a_\eta^l \mid X_n^a)}{N(i a_\eta^l \mid X_n^a)}, \quad l \in E.
\]

with \( X_{i\alpha} = (X_{i\alpha}^{(1)}, \ldots, X_{i\alpha}^{(n_\alpha)}) \) its sample of size \( n_\alpha \).

As in Section 2.2, now we'll particularize \( \mathcal{X} \) to the set \( \mathcal{X}_\alpha = \{X_{i_1\alpha}, \ldots, X_{m\alpha}\} \) of discrete independent random variables aiming to the definition and application of the corresponding \( O_n^{\mathcal{X}_\alpha}(i, j) \), \( E_n^{\mathcal{X}_\alpha}(i, j) \) and \( \Delta_2(O_n^{\mathcal{X}_\alpha} \parallel E_n^{\mathcal{X}_\alpha}) \). For the sake of simplicity they'll be denoted by \( O_n^{\alpha}(i, j) \), \( E_n^{\alpha}(i, j) \) and \( \Delta_2(O_n^{\alpha} \parallel E_n^{\alpha}) \), respectively.

Observe that for \( i, j \in E \)

\[
O_n^{\alpha}(i, j) = N(i a_\eta^l \mid X_n^a)
\]

where \( O_n^{\alpha} \) is the empirical random variables that describe the \( X_{i\alpha} \), \( 1 \leq i \leq m \) observed frequencies. Likewise, we define the expected frequencies

\[
E_n^{\alpha}(i, j) = \frac{\sum_l O_n^{\alpha}(i, l) \sum_l O_n^{\alpha}(l, j)}{\sum_{kl} O_n^{\alpha}(k, l)},
\]

and the respective probability functions

\[
P_{O_n^{\alpha}}(i, j) = \frac{O_n^{\alpha}(i, j)}{N(i a_\eta^l \mid X_n^a)}, \quad i, j \in E
\]

\[
P_{E_n^{\alpha}}(i, j) = \frac{E_n^{\alpha}(i, j)}{N(i a_\eta^l \mid X_n^a)}, \quad i, j \in E.
\]

Finally

\[
\hat{\Delta}_2(P_{O_n^{\alpha}} \parallel P_{E_n^{\alpha}}) = n \sum_{i=1}^r \sum_{j=1}^m \frac{(P_{O_n^{\alpha}}(i, j) - P_{E_n^{\alpha}}(i, j))^2}{P_{O_n^{\alpha}}(i, j)}
\]

\[
= n \Delta_2(P_{O_n^{\alpha}} \parallel P_{E_n^{\alpha}}).
\]
Now we derive a version of the Law of Iterated Logarithm, significant for the establishment of subsequent results about the convergence of $\Delta_2(P_{O_\eta} \| P_{E_\eta})$.

**Lemma 3.1.** \cite{[18]}(Theorems 17.0.1 & 17.2.2) Let $X = \{X_n\}_{n>0}$ be a ergodic Markov chain with finite state space $E$ and stationary distribution $\Pi$, $g: E \rightarrow \mathbb{R}$, $S_n(g) = \sum_{j=1}^{n} g(X_j)$ and

$$\sigma_g^2 = E_\Pi (g^2(X_1)) + 2 \sum_{j=2}^{n} E_\Pi (g(X_1)g(X_j))$$

then:

(a) If $\sigma_g^2 = 0$, then a.s. $\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}}(S_n(g) - E_\Pi(S_n(g))) = 0$.

(b) If $\sigma_g^2 > 0$, then a.s.

$$\limsup_{n \rightarrow \infty} \frac{S_n(g) - E_\Pi(S_n(g))}{\sqrt{2 \sigma_g^2 n \log(\log(n))}} = 1$$

and

$$\liminf_{n \rightarrow \infty} \frac{S_n(g) - E_\Pi(S_n(g))}{\sqrt{2 \sigma_g^2 n \log(\log(n))}} = -1,$$

($E_\Pi$ : expectation with initial distribution $\Pi$; a.s. : almost surely). ♦

**Lemma 3.2.** \cite{[17]}(Lemma 2) If $Y^{(\kappa)}$ is ergodic then for $\eta \geq \kappa - 1$, $\alpha = (a_1^{\eta})$ and $i \alpha j = (i, a_1, ..., a_\eta, j) = (i a_1^{\eta} j) \in E^{\eta+2}$ we have a.s.

$$\limsup_{n \rightarrow \infty} \frac{(N(i a_1^{\eta} j | X^n) - N(i a_1^{\eta} | X^n) p(j | i a_1^{\eta}))^2}{n \log(\log(n))} = 2 \Pi(i a_1^{\eta} j)(1 - p(j | i a_1^{\eta})).$$ ♦

**Theorem 3.3.** Let us refer to \cite{[4]} for the definition of $\Delta_2(P_{O_\eta} \| P_{E_\eta})$, as well as the beginning of the present section for complementary definitions and references related to the following result:

If $\kappa \leq \eta$, there exist $\mathcal{L} < \infty$ so that for every $\alpha = (i_1^{\eta}) \in E^n$
\[
P\left( \limsup_{n \to \infty} \left[ \frac{\hat{\Delta}_2(P_{\alpha_n} \parallel P_{E_n})}{2 \log(\log(n))} \right] \leq L \right) = 1. \quad (7)
\]

If \( \eta = \kappa - 1 \), there exist \( a_1^n \& i, j, k \neq i \) such that, \( p(j \mid i a_1^n) \neq p(j \mid k a_1^n) \), consequently
\[
P\left( \limsup_{n \to \infty} \left[ \frac{\hat{\Delta}_2(P_{\alpha_n} \parallel P_{E_n})}{2 \log(\log(n))} \right] = \infty \right) = 1. \quad \blacklozenge \quad (8)
\]

**Proof**: The following proof shall be divided in the next two cases.

**Case I**: \( 0 \leq \kappa \leq \eta \).

From (24), Lemma 3.1 and by Definition 2.2 we can calculate
\[
O_n^\alpha(i, j) - E_n^\alpha(i, j) = N(i a_1^n j \mid X_1^n) - \frac{N(i a_1^n \mid X_1^n)N(a_1^n j \mid X_1^n)}{N(a_1^n \mid X_1^n)}
\]
or, in the limit
\[
\lim_{n \to \infty} \left( O_n^\alpha(i, j) - E_n^\alpha(i, j) \right)^2 = \lim_{n \to \infty} \left( N(i a_1^n j \mid X_1^n) - N(i a_1^n \mid X_1^n) p(j \mid i a_1^n) \right)^2
\]

\[
\limsup_{n \to \infty} \frac{\left( O_n^\alpha(i, j) - E_n^\alpha(i, j) \right)^2}{n \log(\log(n)) P_{E_n}(i, j)} =
\]
\[
= \limsup_{n \to \infty} \left[ \frac{\left( N(i a_1^n j \mid X_1^n) - N(i a_1^n \mid X_1^n) p(j \mid i a_1^n) \right)^2}{n \log(\log(n))} \frac{1}{P_{E_n}(i, j)} \right].
\]

Similarly
\[
\lim_{n \to \infty} P_{E_n}(i, j) = \lim_{n \to \infty} \frac{E_n^\alpha(i, j)}{N(a_1^n \mid X_1^n)} = \lim_{n \to \infty} \left( \frac{N(i a_1^n \mid X_1^n)N(a_1^n j \mid X_1^n)}{N(a_1^n \mid X_1^n)} \right) =
\]
\[
= \lim_{n \to \infty} \left( \frac{N(i a_1^n \mid X_1^n)}{n} \frac{N(a_1^n j \mid X_1^n)}{N(a_1^n \mid X_1^n)} \right) = \Pi(i a_1^n) \frac{1}{\Pi(a_1^n)} p(j \mid a_1^n) = \theta(i, j) > 0.
\]
By (1) and Lemma 3.2 we have that \( \min_{i,j} \theta(i, j) > 0 \) with

\[
\mathcal{L} = \min_{i,j} \theta(i, j) \sum_{i=1}^{m} \sum_{j=1}^{m} \Pi(i a_1^\eta j)(1 - p(j | i a_1^\eta)) \leq 1
\]

\[
P\left( \limsup_{n \to \infty} \frac{\hat{\Delta}_n(P O_\alpha \| P E_\alpha)}{2 \log(\log(n))} \leq \mathcal{L} \right) = 1.
\]

**Case II:** \( \eta = \kappa - 1 \).

In accordance with the following

\[
\lim_{n \to \infty} \frac{N(a_1^n | X_1^n)}{n} = \lim_{n \to \infty} \sum_{a \in E} \frac{N(a a_1^n | X_1^n)}{n} = \sum_{a \in E} \Pi(a a_1^n) \text{ a.s.}
\]

\[
\lim_{n \to \infty} \frac{N(i a_1^n | X_1^n)}{n} = \Pi(i a_1^n) \text{ a.s.}
\]

we can obtain, as in previous case

\[
\lim_{n \to \infty} \mathbf{P}_{E_\alpha}(i, j) = \lim_{n \to \infty} \frac{\mathbf{E}_\alpha^n(i, j)}{N(a_1^n | X_1^n)} = \lim_{n \to \infty} \frac{N(i a_1^n | X_1^n) N(a_1^n j | X_1^n)}{N(a_1^n | X_1^n) N(a_1^n j | X_1^n)} = \frac{\Pi(i a_1^n) \Pi(a_1^n j)}{\sum_{a \in E} \Pi(a a_1^n) \sum_{a \in E} \Pi(a a_1^n) \neq 0,
\]

and

\[
\lim_{n \to \infty} \mathbf{P}_{O_\alpha}(i, j) = \lim_{n \to \infty} \frac{\mathbf{O}_\alpha^n(i, j)}{N(a_1^n | X_1^n)} = \lim_{n \to \infty} \frac{N(i a_1^n j | X_1^n)}{N(a_1^n | X_1^n)} = \frac{\Pi(i a_1^n) p(j | i a_1^n)}{\sum_{a \in E} \Pi(a a_1^n) \neq 0.
\]

Clearly, if \( \eta = \kappa - 1 \), there exist \( \alpha = a_1^n \& i, j \in E \) so that

\[
\lim_{n \to \infty} (\mathbf{P}_{O_\alpha}(i, j) - \mathbf{P}_{E_\alpha}(i, j)) \neq 0
\]

since, otherwise, it should imply that

\[
p(j | i a_1^n) = \frac{\Pi(a_1^n j)}{\sum_{a \in E} \Pi(a a_1^n)}
\]
i.e. $p(j \mid i a_l^η)$ does not depend on $i \in E$, contradicting the assumption that the order $\kappa > \eta$.

$$P \left( \hat{\Delta}_2(P_{\mathcal{O}_\kappa} \| P_{\mathcal{E}_{\kappa}}) = nO(1) \right) = 1$$

and (8) is proved. ✓

### 3.1 Local and Global Dependency Level

In what follows we define the Local Dependency Level and the Global Dependency Level.

**Definition 3.2.** Let $X_n = \{X_i\}_{i=1}^n$ be a sample of a Markov chain $X$ of order $\kappa$, $\kappa \geq 0$ and $\hat{\Delta}_2(P_{\mathcal{O}_\kappa} \| P_{\mathcal{E}_{\kappa}})$, $\alpha = (a_l^η)$, $\eta \geq 0$ be as in Definition 2.2.

Let us assume that $V$ is an exponential random variable with parameter $\lambda$ where $\mathcal{P}$ is the function $\mathcal{P} : \mathbb{R}^+ \rightarrow [0, 1]$

$$\mathcal{P}(x) = P(V \geq x) = e^{-\lambda x}, \ x \in \mathbb{R}^+.$$ 

We define the Local Dependency Level $\hat{LD}_n(a_l^η)$, for $\alpha = a_l^1$ as

$$\hat{LD}_n(a_l^1) = \frac{\hat{\Delta}_2(P_{\mathcal{O}_\kappa} \| P_{\mathcal{E}_{\kappa}})}{2 \log(\log(n))},$$

and the Global Dependency Level $\hat{GD}_n(\eta)$ as

$$\hat{GD}_n(\eta) = \mathcal{P} \left( \sum_{a_l^1 \in E^\eta} \left( \frac{N(a_l^1 \mid X_l^n)}{n} \right) \hat{LD}_n(a_l^1) \right).$$

Observe that, if the hypothesis $H_0^\eta$ is true, then $\forall a_l^1, \ \eta \geq \kappa,$

$$P \left( \lim_{n \to \infty} \left( \hat{GD}_n(\eta) \right) \geq \mathcal{P}(\mathcal{L}) \right) = 1 \quad (9)$$

and for $\eta = \kappa - 1$

$$P \left( \lim_{n \to \infty} \left( \hat{GD}_n(\eta) \right) = \mathcal{P}(\infty) = 0 \right) = 1. \quad (10)$$
By (9) and (10) it is clear that, for $n$ sufficiently large,

$$P \left( \hat{GDL}_n(\eta) \approx 0 \right) = 1, \quad \eta = \kappa - 1,$$

and

$$P \left( \hat{GDL}_n(\eta) \approx \mathcal{P}(\mathcal{L}) \right) = 1, \quad \eta \geq \kappa.$$

and consequently, for a multiple stationary Markov chain $X_{n \geq 1}$ of order $\kappa$

\[
\kappa = 0 \Leftrightarrow \lim_{n \to \infty} \hat{GDL}_n(\eta) = \mathcal{P}(\mathcal{L}), \quad \eta = 0, 1, \ldots, B, \\
\kappa = \max_{0 \leq \eta \leq B} \left\{ \eta : \lim_{n \to \infty} \hat{GDL}_n(\eta) = 0 \right\} + 1.
\]

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

**Definition 3.3.** Given a fixed number $0 < B \in \mathbb{N}$, let us define the set $S = \{0, 1\}^{B+1}$ and the application $T : S \to \mathbb{N}$

\[
T(s) = -1 \iff s_i = 1, \quad i = 0, 1, \ldots, B \\
T(s) = \max_{0 \leq i \leq B} \{i : s_i = 0, s_{i+1} = \mathcal{P}(\mathcal{L})\}, \quad s = (s_0, s_1, \ldots, s_B).
\]

**Definition 3.4.** Let $X_n = \{X_i\}_{i=1}^n$ be a sample for the Markov chain $X$ 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 $\hat{\kappa}_{GDL}(X_n)$ as

$$\hat{\kappa}_{GDL}(X_n) = T(\sigma_n) + 1$$

with $\sigma_n \in S$ so that $\forall s \in S$

$$\sum_{i=0}^B \left( \hat{GDL}_n(i) - \sigma_n(i) \right)^2 \leq \sum_{i=0}^B \left( \hat{GDL}_n(i) - s(i) \right)^2.$$
By (9), (10) and (3.1) it is clear that, for \( n \) large enough, \( \{GDL_n(i)\}_{i=1}^B \) satisfies the hypothesis of therefore, the order estimator converges almost surely to its value, i.e.,

\[
P \left( \lim_{n \to \infty} \hat{\kappa}_{GDL}(X_n) = \kappa \right) = 1, \quad \kappa = 0, 1, 2, ..., B.
\] (11)

4 Numerical Simulations

Clearly, for a given \( \eta \), the random variables \( LDL_n(a_i^\eta) \) and \( GDL(\eta) \) contains much of the information concerning the sample’s relative dependency, nevertheless, numerical simulations as well as theoretical considerations anticipates a great deal of variability for small samples.

The numerical simulation starts on with the creation, based on an algorithm due to Raftery [21], of a Markov chain transition matrix, \( Q = (q_{i_1i_2...i_{\kappa};i_{\kappa+1}}) \) with entries

\[
q_{i_1i_2...i_{\kappa};i_{\kappa+1}} = \sum_{t=1}^{\kappa} \lambda_t R(i_{\kappa+1}, i_t), \quad 1 \leq i_t, i_{\kappa+1} \leq m.
\]

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

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

were arbitrarily chosen in advance.

Once the matrix \( R \) and the set of numbers \( \{\lambda_i\}_{i=1}^{\kappa} \) are selected, a Markov chain sample of size \( n \), space state \( E \) and transition matrix \( Q \) is generated. It is quite intuitive that the random information about the order of a Markov chain, is spread over an exponentially growing set of multinomial empirical distributions \( \Theta \) with \( |\Theta| = m^{B+2} \), where \( B \) is the maximum integer \( k \), as in \( \alpha = (i_1i_2...i_k) \). 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 of \( n \approx O(m^{B+2}) \).
Table 1: Markov Chain Examples with $|E| = 3$.

\[
Q_1 = \begin{bmatrix}
0.05 & 0.05 & 0.90 \\
0.05 & 0.90 & 0.05 \\
0.90 & 0.05 & 0.05
\end{bmatrix}, \quad Q_2 = \begin{bmatrix}
0.05 & 0.05 & 0.90 \\
0.05 & 0.90 & 0.05 \\
0.90 & 0.05 & 0.05
\end{bmatrix}
\]

| $|E| = 3$ ↔ $n = 1,000$ | $\kappa = 2$ | $\kappa = 3$ | $\kappa = 0$ |
|------------------------|--------|--------|--------|
| $Q_1 \leftrightarrow \lambda_i=1/2, i=1,2.$ | $Q_1 \leftrightarrow \lambda_i=1/3, i=1,...,3.$ | $Q_2 \leftrightarrow \lambda_i=1/3, i=1,...,3.$ |
| $k$ | Aic | Bic | Ede | Gdl | Aic | Bic | Ede | Gdl | Aic | Bic | Ede | Gdl |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0   |     |     |     |     |     |     |     |     |     |     |     |     |
| 1   |     |     |     |     |     |     |     |     |     |     |     |     |
| 2   | 100%| 100%| 100%| 98.5%| 99.5%| 84.5%| 40%  | 4%  |     |     |     |     |
| 3   |     | 100%| 0.5%| 15.5%| 60%  |     |     |     |     |     |     |     |
| 4   |     |     |     |     |     |     |     |     |     |     |     |     |
| 5   |     |     |     |     |     |     |     |     |     |     |     |     |
| 6   |     |     |     |     |     |     |     |     |     |     |     |     |

Whenever the sample is obtained, the order estimators GDL, AIC and BIC are calculated, based on the same Markov chain sample and the numerical results registered in the form of tables. To conclude, we would like to point out that everything related to numerical simulations were done using the remarkable free software $R^{[20]}$.

5 Conclusion

The pioneer research started with the contributions of Bartlett$^{[6]}$, Hoel$^{[16]}$, Good$^{[15]}$, Anderson & Goodman$^{[3]}$, Billingsley$^{[7]}$, $[8]}$ among others, where they developed tests of hypothesis for the detection of the order of a Markov chain.

Later on these tools were adapted and improved with the used of Penalty Functions$^{[23]}$, Katz$^{[17]}$) together with other tools created in the realm of Models Selection (Akaike$^{[1]}$, Schwarz$^{[22]}$). Since then, there have been a considerable number of subsequent contributions on this subject, several of them consisting in the enhancement of the already existing techniques.
Table 2: Markov Chain Examples with $|E| = 4$.

$$Q_3 = \begin{bmatrix} 0.05 & 0.05 & 0.05 & 0.85 \\ 0.05 & 0.05 & 0.85 & 0.05 \\ 0.05 & 0.85 & 0.05 & 0.05 \\ 0.85 & 0.05 & 0.05 & 0.05 \end{bmatrix}, \quad Q_4 = \begin{bmatrix} 0.05 & 0.05 & 0.05 & 0.85 \\ 0.05 & 0.05 & 0.05 & 0.85 \\ 0.05 & 0.05 & 0.05 & 0.85 \\ 0.05 & 0.05 & 0.05 & 0.85 \end{bmatrix}$$

In this note, we propose a new Markov chain order estimator based on a different idea which makes it behave in a quite different form. This estimator is strongly consistent and more efficient than AIC (inconsistent), outperforming the well established and consistent BIC, mainly on relatively small samples.

(Csiszar[11], Zhao et al[24]).

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] 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.

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

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

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

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

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

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

[12] 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.

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

[14] 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.

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

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

[17] R. W. Katz, “On Some Criteria for Estimating the Order Of Markov Chains”, *Technometrics*, vol. 23, 1981.
[18] S.P. Meyn & R.L. Tweedie, “Markov Chain and Stochastic Stability”, Springer-Verlag, London, 1993.

[19] L. Pardo, “Statistical Inference Based on Divergence Measures”, Taylor & Francis Group, LLC, New York, 2006.

[20] 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, 2008.

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

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

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

[24] 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.