MINIMAL MARKOV MODELS *

BY JESÚS E. GARCÍA[†,‡] AND VERÓNICA A. GONZÁLEZ-LÓPEZ[†,‡]

Universidade Estadual de Campinas [‡]

In this work we introduce a new and richer class of finite order Markov chain models and address the following model selection problem: find the Markov model with the minimal set of parameters (minimal Markov model) which is necessary to represent a source as a Markov chain of finite order. Let us call M the order of the chain and A the finite alphabet, to determine the minimal Markov model, we define an equivalence relation on the state space A^M, such that all the sequences of size M with the same transition probabilities are put in the same category. In this way we have one set of $(|A| - 1)$ transition probabilities for each category, obtaining a model with a minimal number of parameters. We show that the model can be selected consistently using the Bayesian information criterion.

1. Introduction. In this work we consider discrete stationary processes over a finite alphabet A. Markov chains of finite order are widely used to model stationary processes with finite memory. A problem with full Markov chains models of finite order M is that the number of parameters $(|A|^M(|A| - 1))$ grows exponentially with the order M, where $|A|$ denotes the cardinal of the alphabet A. Another characteristic is that the class of full Markov chains is not very rich, fixed the alphabet A there is just one model for each order M and in practical situations could be necessary a more flexible structure in terms of number of parameters. For an extensive discussion of those two problems se Buhlmann P. and Wyner A. [1]. A richer class of finite order Markov models introduced by Rissanen J. [6] and Buhlmann P. and Wyner A. [1] are the variable length Markov chain models (VLMC) which are mentioned in section 2.3. In the VLMC class, each model is identified by a prefix tree $\mathcal{T}$ called context tree. For a given model with a context tree $\mathcal{T}$, the final number of parameters for the model is $|\mathcal{T}|(|A| - 1)$ and depending on the tree, this produce a parsimonious model. In Csiszár, I. and Talata, Z. [4] is proved that the bayesian information criterion (BIC) can be used to consistently choose the VLMC model in an efficient way using the context tree weighting (CTW) algorithm.

In this paper we introduce a larger class of finite order Markov models, and we address the problem of model selection inside this class, showing that the model can be selected consistently using the BIC criterion. In our class, each model is determined by choosing a partition of the state space, our class of models include the full Markov chain models and

the VLMC models because a context tree can be seen as a particular partition of the state space (see for illustration the example 2.1).

In Section 2, we define the minimal Markov models and show that this models can be selected in a consistently in theorems 2.1 and 2.2. In Section 3 we show two algorithms that use the results in Section 2 to choose consistently a minimal Markov model for a sample and some simulations. Section 4 have the conclusions and Section 5 have the proofs.

2. Minimal Markov models.

2.1. *Notation.*   Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. Let us call $\mathcal{S} = A^M$ the state space. Denote the string $a_m a_{m+1} \ldots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$.

Let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$,

$$(1) \qquad P(L, a) = \sum_{s \in L} \mathrm{Prob}(X_{t-M}^{t-1} = s, X_t = a), \ \ a \in A, \ L \in \mathcal{L};$$

$$(2) \qquad P(L) = \sum_{s \in L} \mathrm{Prob}(X_{t-M}^{t-1} = s), \ \ L \in \mathcal{L}.$$

Let x_1^n be a sample of the process (X_t), $s \in \mathcal{S}$, $a \in A$ and $n > M$. We denote by $N_n(s, a)$ the number of occurrences of the string s followed by a in the sample x_1^n,

$$(3) \qquad N_n(s, a) = \left| \{t : M < t \leq n, x_{t-M}^{t-1} = s, x_t = a\} \right|,$$

the number of occurrences of s in the sample x_1^n is denoted by $N_n(s)$ and

$$(4) \qquad N_n(s) = \left| \{t : M < t \leq n, x_{t-M}^{t-1} = s\} \right|.$$

The number of occurrences of elements into L followed by a is given by,

$$(5) \qquad N_n^{\mathcal{L}}(L, a) = \sum_{s \in L} N_n(s, a), \ \ L \in \mathcal{L};$$

the accumulated number of $N_n(s)$ for s in L is denoted by,

$$(6) \qquad N_n^{\mathcal{L}}(L) = \sum_{s \in L} N_n(s), \ \ L \in \mathcal{L}.$$

2.2. *Good partitions of $\mathcal{S}$.*

DEFINITION 2.1.   *Let (X_t) be a discrete time order M Markov chain on a finite alphabet A, $\mathcal{S} = A^M$ the state space. A partition $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of $\mathcal{S}$ is a good partition of $\mathcal{S}$ if for each $s, s' \in L$, $L \in \mathcal{L}$,*

$$Prob(X_t = . \,|X_{t-M}^{t-1} = s) = Prob(X_t = . \,|X_{t-M}^{t-1} = s').$$

REMARK 2.1.   *For a discrete time order M Markov chain on a finite alphabet A with $\mathcal{S} = A^M$ the state space, $\mathcal{L} = \mathcal{S}$ is a good partition of $\mathcal{S}$.*

If $\mathcal{L}$ is a good partition of $\mathcal{S}$, we define for each category $L \in \mathcal{L}$

$$(7) \qquad P(a|L) = \text{Prob}(X_t = a|X_{t-M}^{t-1} = s) \ \ \forall a \in A,$$

where s is some element into L. As a consequence, if we write $P(x_1^n) = \text{Prob}(X_1^n = x_1^n)$, we obtain

$$(8) \qquad P(x_1^n) = P(x_1^M) \prod_{L \in \mathcal{L}, a \in A} P(a|L)^{N_n^{\mathcal{L}}(L,a)}.$$

In the same way that Csiszár, I. and Talata, Z. [4] we will define our BIC criterion using a modified maximum likelihood. We will call maximum likelihood to the maximization of the second term in the equation (8) for the given observation. For the sequence x_1^n, will be

$$(9) \qquad \text{ML}(\mathcal{L}, x_1^n) = \prod_{L \in \mathcal{L}, a \in A} \left(\frac{r_n(L,a)}{r_n(L)} \right)^{N_n^{\mathcal{L}}(L,a)},$$

where

$$(10) \qquad r_n\Big(L,a\Big) = \frac{N_n^{\mathcal{L}}(L,a)}{n}, \quad a \in A, \ L \in \mathcal{L} \quad and \quad r_n\Big(L\Big) = \frac{N_n^{\mathcal{L}}(L)}{n}, \quad L \in \mathcal{L}.$$

The BIC is given by the next definition

DEFINITION 2.2.   *Given a sample x_1^n, of the process (X_t), a discrete time order M Markov chain on a finite alphabet A with $\mathcal{S} = A^M$ the state space and $\mathcal{L}$ a good partition of $\mathcal{S}$. The BIC of the model (9) is given by*

$$BIC(\mathcal{L}, x_1^n) = \ln\left(ML(\mathcal{L}, x_1^n)\right) - \frac{(|A|-1)|\mathcal{L}|}{2} \ln(n).$$

2.3. *Good partitions and context trees.*

Let (X_t) be a finite order Markov chain taking values on A and $\mathcal{T}$ a set of sequences of symbols from A such that no string in $\mathcal{T}$ is a suffix of another string in $\mathcal{T}$, for each $s \in \mathcal{T}$, $d(\mathcal{T}) = \max\Big(l(s), s \in \mathcal{T}\Big)$ where $l(s)$ denote the length of the string s, with $l(\emptyset) = 0$ if the string is the empty string.

DEFINITION 2.3.   *$\mathcal{T}$ is a context tree for the process (X_t) if for any sequence of symbols in A, x_1^n sample of the process with $n \geq d(\mathcal{T})$, there exist $s \in \mathcal{T}$ such that*

$$Prob(X_{n+1} = a|X_1^n = x_1^n) = Prob(X_{n+1} = a|X_{n-l(s)+1}^n = s)$$

$d(\mathcal{T})$ is the depth of the tree.
The context tree is the minimal state space of the variable length Markov chain (VLMC), Buhlmann P. and Wyner A. [1]. The context tree for a VLMC with finite depth M define a good partition on the space $\mathcal{S} = A^M$ as illustrated by the next example.

EXAMPLE 2.1.   *Let be a VLMC over the alphabet $A = \{0,1\}$ with depth $M = 3$ and contexts,*

$$\{0\}, \{01\}, \{011\}, \{111\}$$

This context tree correspond to the good partition $\{L_1, L_2, L_3, L_4\}$ where
$L_1 = \{\{000\}, \{100\}, \{010\}, \{110\}\}, L_2 = \{\{001\}, \{101\}\}, L_3 = \{011\}$ and $L_4 = \{111\}$.

2.4. *Smaller good partitions.*

DEFINITION 2.4.   *Let $\mathcal{L}^{ij}$ denote the partition*

$$\mathcal{L}^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_K\},$$

where $\mathcal{L} = \{L_1, \ldots, L_K\}$ is a good partition of $\mathcal{S}$, and for $1 \leq i < j \leq K$ with $L_{ij} = L_i \cup L_j$.

Now we adapt the notation established for the partition $\mathcal{L}$ to the new partition $\mathcal{L}^{ij}$.

NOTATION 2.1.   *for $a \in A$ we write,*

$$\begin{aligned} P(L_{ij}, a) &= P(L_i, a) + P(L_j, a); \\ P(L_{ij}) &= P(L_i) + P(L_j). \end{aligned}$$

(11) $$N_n^{\mathcal{L}^{ij}}(L_{ij}, a) = N_n^{\mathcal{L}}(L_i, a) + N_n^{\mathcal{L}}(L_j, a);$$

(12) $$N_n^{\mathcal{L}^{ij}}(L_{ij}) = N_n^{\mathcal{L}}(L_i) + N_n^{\mathcal{L}}(L_j);$$

If $P(.|L_i) = P(.|L_j)$ then $\mathcal{L}^{ij}$ is a good partition and (7) remains valid for $\mathcal{L}^{ij}$, just is necessary to change $\mathcal{L}$ by $\mathcal{L}^{ij}$ in equations (8), (9) and definition (2.2).
In the following theorem, we show that the BIC criterion provides a consistent way of detecting smaller good partition.

THEOREM 2.1.   *Let (X_t) be a Markov chain with order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space. If $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ is a good partition of $\mathcal{S}$ and $L_i \neq L_j$, $L_i, L_j \in \mathcal{L}$. Then, eventually almost surely as $n \to \infty$,*

$$I_{\{BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n)\}} = 1$$

if, and only if

$$P(a|L_i) = P(a|L_j) \;\forall a \in A.$$

Where I_A is the indicator function of A, and the $\mathcal{L}^{ij}$ partition is defined under $\mathcal{L}$ by equation (2.4).

Next we extract from the previous theorem the relation that we use in the next section, in practice to find smaller good partitions.

DEFINITION 2.5.   *Let be (X_t) a Markov chain of order M, with finite alphabet A and state space $\mathcal{S} = A^M$, x_1^n a sample of the process and let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a good partition of $\mathcal{S}$,*

$$d_{\mathcal{L}}(i,j) = \frac{1}{\ln(n)} \sum_{a \in A} \left\{ N_n^{\mathcal{L}}(L_i, a) \ln\left(\frac{N_n(L_i, a)}{N_n(L_i)}\right) + N_n^{\mathcal{L}}(L_j, a) \ln\left(\frac{N_n(L_j, a)}{N_n(L_j)}\right) \right.$$

(13)
$$\left. - N_n^{\mathcal{L}^{ij}}(L_{ij}, a) \ln\left(\frac{N_n(L_{ij}, a)}{N_n(L_{ij})}\right) \right\}$$

COROLLARY 2.1.   *Under the assumptions of theorem 2.1,*

$$BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) < 0 \quad \Longleftrightarrow \quad d_{\mathcal{L}}(i,j) < \frac{(|A|-1)}{2}.$$

PROOF. From equation (14) we have the validity of the result.   □

REMARK 2.2.   *The results will remain valid if we replace the constant $\frac{(|A|-1)}{2}$ for some arbitrary constant, positive and finite value v, into the definition (2.2).*

REMARK 2.3.   *Under the assumptions of theorem 2.1, if $P(a|L_i) \neq P(a|L_j)$ for some $a \in A$, then eventually almost surely as $n \to \infty$, $BIC(\mathcal{L}, x_1^n) > BIC(\mathcal{L}^{ij}, x_1^n)$ where $\mathcal{L}^{ij}$ verified the definition (2.4).*

2.5. *Minimal good partition.*
We want to find the smaller good partition into the universe of all possible good partitions of $\mathcal{S}$. This special good partition could be defined as follows and it allows the definition of the most parsimonious model into the class considered in this paper.

DEFINITION 2.6.   *Let (X_t) be a discrete time order M Markov chain on a finite alphabet A, $\mathcal{S} = A^M$ the state space. A partition $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of $\mathcal{S}$ is the minimal good partition of $\mathcal{S}$ if, $\forall L \in \mathcal{L}$,*

$$s, s' \in L \quad \text{if, and only if } Prob(X_t = . | X_{t-M}^{t-1} = s) = Prob(X_t = . | X_{t-M}^{t-1} = s').$$

REMARK 2.4.   *For a discrete time order M Markov chain on a finite alphabet A with $\mathcal{S} = A^M$ the state space, $\exists!$ minimal good partition of $\mathcal{S}$.*

In the next example we emphasize the difference between good partitions and the minimal good partition,
The next theorem shows that for n large enough we achive the partition $\mathcal{L}^*$ which is the minimal good partition.

THEOREM 2.2.   *Let (X_t) be a Markov chain with order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space and let $\mathcal{P}$ be the set of all the partitions of $\mathcal{S}$. Define,*

$$\mathcal{L}_n^* = argmax_{\mathcal{L} \in \mathcal{P}}\{BIC(\mathcal{L}, x_1^n)\}$$

then, eventually almost surely as $n \to \infty$,

$$\mathcal{L}^* = \mathcal{L}_n^*$$

3. Minimal good partition estimation algorithm.

ALGORITHM 3.1.   *(MMM algorithm for good partitions)*
Consider x_1^n a sample of the Markov process (X_t), with order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space.
Let be $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ a good partition of $\mathcal{S}$, for each $s \in \mathcal{S}$,

1 for $i = 1, 2, \cdots, K - 1$,

    for $j = i + 1, 2, \cdots, K$,

        Calculate $d_{\mathcal{L}}(i, j)$

        $R_n^{i,j} = I_{\{d_{\mathcal{L}}(i,j) < \frac{(|A|-1)}{2}\}}$

2 If $R_n^{i,j} = 1$, define $L_{ij} = L_i \cup L_j$ and $\mathcal{L} = \mathcal{L}^{ij}$. Else $i = i + 1$, Return to step 1

The algorithm allows to define the next relation based on the sample x_1^n,

DEFINITION 3.1.   *for $r, s \in \mathcal{S}$; $r \sim_n s \iff R_n^{a(r),a(s)} = 1$.*

For n large enough, the algorithm return the minimal good partition.

COROLLARY 3.1.   *Let $\{X_t, t = 0, 1, 2, \ldots\}$ be a Markov chain with order M over a finite alphabet A, $\mathcal{S} = A^M$ and x_1^n a sample of the Markov process. $\hat{\mathcal{L}}_n$, given by the algorithm (3.1) converges almost surely eventually to $\mathcal{L}^*$, where $\mathcal{L}^*$ is the minimal good partition of $\mathcal{S}$.*

PROOF. Because $K < \infty$, for n large enough, the algorithm return the minimal good partition. □

REMARK 3.1.   *In the worst case, which correspond to an initial good partition equal to $\mathcal{S}$, we need to calculate the term $\left(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)} \right)^{N_n^{\mathcal{L}}(L,a)}$ for each $s \in \mathcal{S}$ plus $K(K-1)/2$ divisions to implement the algorithm (3.1).*

The next algorithm is a variation of the first. In this case the partitions are grow selecting the pair of elements with the minimal value of $\{d_{\mathcal{L}}(i, j)$, the algorithm stop when there is not $\{d_{\mathcal{L}}(i, j)$ lower than $(|A| - 1)/2$.

ALGORITHM 3.2.   *Consider x_1^n a sample of the Markov process (X_t), with order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space.*
Let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a good partition of $\mathcal{S}$

1 Calculate

$$(i^*, j^*) = arg \min_{i,j|1 \leq i < j \leq K} \{d_{\mathcal{L}(i,j)}\}$$

2 If $d_{\mathcal{L}(i,j)} < \frac{|A|-1}{2}$ then $\mathcal{L} = \mathcal{L}i^* j^*$, $K = K - 1$ and return to 1.
   Else end.

This algorithm is consistent and always return a partition but have a greater computational cost. Taking in consideration that the cost depend on K and that for a Markov chain of order M we consider samples of size n such that $\log(n) > M$. The two algorithms 3.1 and 3.2 have a computational cost that is linear in n (the sample size).

3.1. *Dendrograms and MMM algorithm* . In practice, when the sample size is not large enough and the algorithm 3.1 has not converged, it is possible that the algorithm will not return a partition of $\mathcal{S}$, independent of the value used in v. In that case, a better approach can be to use for each $r, s \in \mathcal{S}$ the function $d_n(r, s)$ as a similarity measure between r and s. Then $d_n(r, s)$ can be used to produce a dendrogram and then use the partition defined by the dendrogram as the partition estimator.

Also in practice it is possible that the maximum number of free parameters in our model is limited by a number K. In that case, the logic choice will be to find a value of d in the dendrogram such that the size of the partition obtained cutting the dendrogram in d is less or equal to K, the chosen model will be the one defined by that partition.

EXAMPLE 3.1.   *Consider a Markov chain of order $M = 3$ on the alphabet $A = \{0, 1, 2\}$ with classes:*

$$
\begin{aligned}
L_1 &= \{000, 100, 200, 010, 110, 210, 020, 120, 220, 022, 122, 222\}, \\
L_2 &= \{001, 101, 201, 011, 111, 211, 021, 121, 221\}, \\
L_3 &= \{012, 112, 212, 002\}, \\
L_4 &= \{102\}, \\
L_5 &= \{202\},
\end{aligned}
$$

and transition probabilities,

$$
\begin{aligned}
P(0|L_1) &= 0.2, & P(1|L_1) &= 0.3, \\
P(0|L_2) &= 0.4, & P(1|L_2) &= 0.3, \\
P(0|L_3) &= 0.4, & P(1|L_3) &= 0.1, \\
P(0|L_4) &= 0.1, & P(1|L_4) &= 0.4, \\
P(0|L_5) &= 0.3, & P(1|L_5) &= 0.5.
\end{aligned}
$$

On this example, $|A| = 3$ so the penalty constant is $1 = \frac{|A|-1}{2}$. We simulated samples of sizes $n = 5000$ and 9000, obtaining dendrograms on figure 3.1. The dendrogram for the sample size of 9000 gives the correct partition.

3.2. *Simulations.*   We implemented a simulation study for the model described on example 3.3. More precisely we simulated 1000 samples of the process for each of the sample sizes $4000, 6000, 8000$ and 10000. For each sample we calculate the values $d_n(r, s)$ and build the corresponding dendrogram (using the R-project package hclust with linkage method complete). Table 1 show the results.
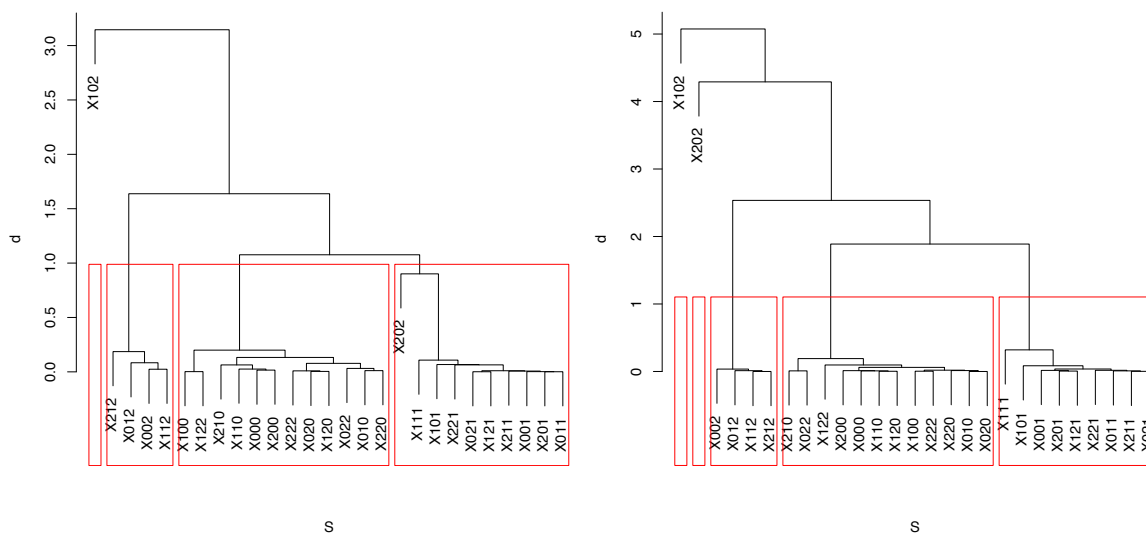
FIG 1. *The figure shows the dendrograms for the model on example 3.3 estimated using algorithm ?? for sample sizes of* 5000 *(upper picture) and* 9000 *(lower picture).*

TABLE 1
Number of errors on the partition estimated for the model on example 3.3

Sample size	Proportion of errors
4000	0.801
6000	0.495
8000	0.252
10000	0.161

3.3. *Simulations.* The VLMC corresponding to the partition on example (), have contexts:

$$
\begin{aligned}
T_1 &= \{0\}, \\
T_2 &= \{1\}, \\
T_3 &= \{12\}, \\
T_4 &= \{102\}, \\
T_5 &= \{202\}, \\
T_6 &= \{22\}, \\
T_7 &= \{002\}.
\end{aligned}
$$

We simulated 1000 samples of the process for each of the sample sizes $4000, 6000, 8000$ and 10000. Using the tree as a basic good partition, for each sample we calculate the values $d_n(L_i, L_j)$ corresponding to the algorithm (3.1) and build the corresponding partitions. Table (2) show the results.

Starting from the good partition corresponding to the context tree, the number of possible models is substantially reduced compared to those in the simulation on section (3.2) and because of that, the error rates on this simulation are much better than before.

TABLE 2
proportion of errors on the partition estimated for the model of example (3.3)

Sample size	Proportion of errors
4000	0.614
6000	0.206
8000	0.047
10000	0.007

4. Conclusions. Our main motivation to define the minimal Markov models is, in the first place, the concept of partitioning the state space in classes in which the states are equivalent, this allow us to model the redundancy that appears in many processes in the nature as in genetics, linguistics, etc. Each class in the state space has a very specific, clear and practical meaning: any sequence of symbol in the same class has the same effect on the future distribution of the process. In other words, they activate the same random mechanism to choose the next symbol on the process. We can think of the resulting minimal partition as a list of the relevant contexts for the process and their synonymous.

In second place our motivation for developing this methodology is to demonstrate that for a stationary, finite memory process it is theoretically possible to find consistently a minimal Markov model to represent this process and that this can be accomplished in practice. The utilitarian implication of the fact that the model selection process can be started from a context tree partition, is that minimal Markov models can be easily fitted to stationary sources where the VLMC models already works.

It is clear that there are applications on which the natural partition to estimate is neither the minimal nor a context tree partition. As long as the partition particular properties are well defined, we can use theorem 2.1 to estimate the minimal partition satisfying those properties.

Our theorems are still valid if we change the constant term in the penalization of the BIC criterion for any positive (and finite) number. In the case of the VLMC model, the problem of finding a better constant has been addressed in diverse works as for example Buhlmann P. and Wyner A. [1] and Galves, A., Galves, C., Garcia N. L. and Leonardi F. [5].

5. Proofs.

DEFINITION 5.1. *Let be P and Q probability distributions on A. The relative entropy between P and Q is given by,*

$$D(P(\cdot)||Q(\cdot)) = \sum_{a \in A} P(a) \ln \left(\frac{P(a)}{Q(a)} \right).$$

5.1. *Proof of theorem 2.1.*

$$BIC(\mathcal{L}, x_1^n) = \sum_{a \in A} \ln \left(\prod_{L \in \mathcal{L}} \left(\frac{r_n(L, a)}{r_n(L)} \right)^{N_n^{\mathcal{L}}(L,a)} \right) - \frac{(|A| - 1)|\mathcal{L}|}{2} \ln(n),$$

as consequence,

$$
\begin{aligned}
BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) \;=\; &\sum_{a \in A} \left\{ N_n^{\mathcal{L}}(L_i, a) \ln \left(\frac{r_n(L_i, a)}{r_n(L_i)} \right) \right. \\
& + N_n^{\mathcal{L}}(L_j, a) \ln \left(\frac{r_n(L_j, a)}{r_n(L_j)} \right) \\
& \left. - N_n^{\mathcal{L}^{ij}}(L_{ij}, a) \ln \left(\frac{r_n(L_{ij}, a)}{r_n(L_{ij})} \right) \right\} - \frac{(|A|-1)}{2} \ln(n).
\end{aligned}
\tag{14}
$$

We note that, the condition $I_{\{BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n)\}} = 1$ is true if, and only if

$$
\begin{aligned}
\sum_{a \in A} \left\{ r_n(L_i, a) \ln \left(\frac{r_n(L_i, a)}{r_n(L_i)} \right) + r_n(L_j, a) \ln \left(\frac{r_n(L_j, a)}{r_n(L_j)} \right) \right. \\
\left. - r_n(L_{ij}, a) \ln \left(\frac{r_n(L_{ij}, a)}{r_n(L_{ij})} \right) \right\} < \frac{(|A|-1) \ln(n)}{2n}.
\end{aligned}
\tag{15}
$$

Because $r_n(L, a)$ and $r_n(L)$ are non-negative, using Jensen we have that,

$$
\begin{aligned}
r_n(L_i, a) \ln \left(\frac{r_n(L_i, a)}{r_n(L_i)} \right) + r_n(L_j, a) \ln \left(\frac{r_n(L_j, a)}{r_n(L_j)} \right) \geq \\
(r_n(L_i, a) + r_n(L_j, a)) \ln \left(\frac{r_n(L_i, a) + r_n(L_j, a)}{r_n(L_i) + r_n(L_j)} \right)
\end{aligned}
$$

or equivalently,

$$
r_n(L_i, a) \ln \left(\frac{r_n(L_i, a)}{r_n(L_i)} \right) + r_n(L_j, a) \ln \left(\frac{r_n(L_j, a)}{r_n(L_j)} \right) \geq r_n(L_{ij}, a) \ln \left(\frac{r_n(L_{ij}, a)}{r_n(L_{ij})} \right),
\tag{16}
$$

with equality if and only if $\frac{r_n(L_i, a)}{r_n(L_i)} = \frac{r_n(L_j, a)}{r_n(L_j)}$, $\forall a \in A$.
As consequence, equation (16) $\Rightarrow$

$$
\begin{aligned}
\sum_{a \in A} \left\{ r_n(L_i, a) \ln \left(\frac{r_n(L_i, a)}{r_n(L_i)} \right) + r_n(L_j, a) \ln \left(\frac{r_n(L_j, a)}{r_n(L_j)} \right) \right. \\
\left. - r_n(L_{ij}, a) \ln \left(\frac{r_n(L_{ij}, a)}{r_n(L_{ij})} \right) \right\} \geq 0,
\end{aligned}
\tag{17}
$$

with equality if and only if $\frac{r_n(L_i, a)}{r_n(L_i)} = \frac{r_n(L_j, a)}{r_n(L_j)}$ $\forall a \in A$.

Considering that $\frac{(|A|-1) \ln(n)}{2n} \to 0$, as $n \to \infty$ and from the equation (15), we have that if $\lim_{n \to \infty} I_{\{BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n)\}} = 1$, then

$$
\begin{aligned}
\lim_{n \to \infty} \sum_{a \in A} \left\{ r_n(L_i, a) \ln \left(\frac{r_n(L_i, a)}{r_n(L_i)} \right) + r_n(L_j, a) \ln \left(\frac{r_n(L_j, a)}{r_n(L_j)} \right) \right. \\
\left. - r_n(L_{ij}, a) \ln \left(\frac{r_n(L_{ij}, a)}{r_n(L_{ij})} \right) \right\} \leq 0,
\end{aligned}
$$

from equation (17) and taking the limit inside the sum we obtain

$$\sum_{a \in A} \left\{ P(L_i, a) \ln \left(\frac{P(L_i, a)}{P(L_i)} \right) + P(L_j, a) \ln \left(\frac{P(L_j, a)}{P(L_j)} \right) - P(L_{ij}, a) \ln \left(\frac{P(L_{ij}, a)}{P(L_{ij})} \right) \right\} = 0,$$

using Jensen again, this means that $\frac{P(L_i,a)}{P(L_i)} = \frac{P(L_j,a)}{P(L_j)} \ \forall a \in A$, or equivalently, $P(a|L_i) = P(a|L_j) \ \forall a \in A$.

For the other half of the proof, suppose that $P(a|L_i) = P(a|L_j) \ \forall a \in A$, as a consequence we have that

$$(18) \qquad P(a|L_{ij}) = P(a|L_i) \ \forall a \in A$$

$$\begin{aligned} BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) &= \ln \left(\prod_{a \in A} \left(\frac{N_n^{\mathcal{L}}(L_i, a)}{N_n^{\mathcal{L}}(L_i)} \right)^{N_n^{\mathcal{L}}(L_i, a)} \right) \\ &+ \ln \left(\prod_{a \in A} \left(\frac{N_n^{\mathcal{L}}(L_j, a)}{N_n^{\mathcal{L}}(L_j)} \right)^{N_n^{\mathcal{L}}(L_j, a)} \right) \\ &- \ln \left(\prod_{a \in A} \left(\frac{N_n^{\mathcal{L}^{ij}}(L_{ij}, a)}{N_n^{\mathcal{L}^{ij}}(L_{ij})} \right)^{N_n^{\mathcal{L}^{ij}}(L_{ij}, a)} \right) - \frac{(|A| - 1)}{2} \ln(n). \end{aligned}$$

Now, considering that $\frac{N_n^{\mathcal{L}^{ij}}(L_{ij},a)}{N_n^{\mathcal{L}^{ij}}(L_{ij})}$ is the maximum likelihood estimator of $P(a|L_{ij})$,

$$\prod_{a \in A} \left(\frac{N_n^{\mathcal{L}^{ij}}(L_{ij}, a)}{N_n^{\mathcal{L}^{ij}}(L_{ij})} \right)^{N_n^{\mathcal{L}^{ij}}(L_{ij},a)} \geq \prod_{a \in A} P(a|L_{ij})^{N_n^{\mathcal{L}^{ij}}(L_{ij},a)}$$

$BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n)$ is bounded above by

$$\begin{aligned} &\ln \left(\prod_{a \in A} \left(\frac{N_n^{\mathcal{L}}(L_i, a)}{N_n^{\mathcal{L}}(L_i)} \right)^{N_n^{\mathcal{L}}(L_i,a)} \right) + \ln \left(\prod_{a \in A} \left(\frac{N_n^{\mathcal{L}}(L_j, a)}{N_n^{\mathcal{L}}(L_j)} \right)^{N_n^{\mathcal{L}}(L_j,a)} \right) \\ &- \ln \left(\prod_{a \in A} P(a|L_{ij})^{N_n^{\mathcal{L}^{ij}}(L_{ij},a)} \right) - \frac{(|A| - 1)}{2} \ln(n) \\ &= N_n^{\mathcal{L}}(L_i) D \left(\frac{N_n^{\mathcal{L}}(L_i, .)}{N_n^{\mathcal{L}}(L_i)} \middle\| P(.|L_i) \right) + N_n^{\mathcal{L}}(L_j) D \left(\frac{N_n^{\mathcal{L}}(L_j, .)}{N_n^{\mathcal{L}}(L_j)} \middle\| P(.|L_j) \right) - \frac{(|A| - 1)}{2} \ln(n). \end{aligned}$$

Where $D(p||q)$ is the relative entropy, given by definition (5.1). The first equality came from (18) and (11). Using proposition (**??**), proposition (**??**), for any $\delta > 0$ and n large enough,

$$(19) \qquad D \left(\frac{N_n^{\mathcal{L}}(L, .)}{N_n^{\mathcal{L}}(L)} \middle\| P(.|L) \right) \leq \sum_{a \in A} \frac{\left(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)} - P(a|L) \right)^2}{P(a|L)}$$

$$(20) \qquad \leq \sum_{a \in A} \frac{\frac{\delta \ln(n)}{N_n^{\mathcal{L}}(L)}}{P(a|L)}.$$

Then for any $\delta > 0$ and n large enough,

$$
\begin{aligned}
BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) &\leq \frac{2\delta|A|}{p}\ln(n) - \frac{(|A|-1)}{2}\ln(n) \\
&= \ln(n)\left(\frac{2\delta|A|}{p} - \frac{(|A|-1)}{2}\right)
\end{aligned}
$$

where $p = \min\{P(a|L) : a \in A, L \in \{L_i, L_j\}\}$.

In particular, taking $\delta < \frac{p(|A|-1)}{4|A|}$, for n large enough,

$$
BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) < 0.
$$

References.

[1] BUHLMANN P. and WYNER A. (1999). Variable length Markov chains. Ann. Statist. **27** 480–513.
[2] CSISZÁR, I. and SHIELDS, P. C. (2000).The consistency of the BIC Markov order estimator. Ann. Statist. **28** 1601–1619.
[3] CSISZÁR, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. IEEE Trans. Inform. Theory **48** 1616–1628.
[4] CSISZÁR, I. and TALATA, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. IEEE Trans. Inform. Theory **52** 1007–1016.
[5] GALVES ET AL. (2009). Context tree selection and linguistic rhythm retrieval from written texts. arXiv:0902.3619.
[6] RISSANEN J. (1983). A universal data compression system, IEEE Trans. Inform. Theory **29**(5) 656 – 664.

DEPARTAMENTO DE ESTATÍSTICA
INSTITUTO DE MATEMÁTICA ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
UNIVERSIDADE ESTADUAL DE CAMPINAS
RUA SERGIO BUARQUE DE HOLANDA,651
CIDADE UNIVERSITÁRIA-BARÃO GERALDO
CAIXA POSTAL: 6065
13083-859 CAMPINAS, SP, BRAZIL
E-MAIL: jg@ime.unicamp.br
            veronica@ime.unicamp.br