entropy

MDPI

Article

Consistent Estimation of Partition Markov Models

Jesús E. García and Verónica A. González-López *

Department of Statistics, University of Campinas, Rua Sérgio Buarque de Holanda, 651, Campinas, São Paulo 13083-859, Brazil; jg@ime.unicamp.br

* Correspondence: veronica@ime.unicamp.br; Tel.: +55-19-3521-6057

Academic Editor: Maxim Raginsky

Received: 1 March 2017; Accepted: 4 April 2017; Published: 6 April 2017

Abstract: The Partition Markov Model characterizes the process by a partition $\mathcal{L}$ of the state space, where the elements in each part of $\mathcal{L}$ share the same transition probability to an arbitrary element in the alphabet. This model aims to answer the following questions: what is the minimal number of parameters needed to specify a Markov chain and how to estimate these parameters. In order to answer these questions, we build a consistent strategy for model selection which consist of: giving a size n realization of the process, finding a model within the Partition Markov class, with a minimal number of parts to represent the process law. From the strategy, we derive a measure that establishes a metric in the state space. In addition, we show that if the law of the process is Markovian, then, eventually, when n goes to infinity, $\mathcal{L}$ will be retrieved. We show an application to model internet navigation patterns.

Keywords: Bayesian Information Criterion; distance measure; model selection; statistical inference in Markov processes

1. Introduction

The Markov models have received enormous visibility for being powerful tools [1–3]. In recent years, the theoretical advances have allowed for its users to identify the most suitable methods for estimating them. For instance, [4] shows that the Bayesian Information Criterion (BIC)—[5]—can be used to consistently choose a Variable Length Markov Chain model in an efficient way using the Context Tree Maximization (CTM) algorithm. See also [6,7]. In this paper, we show that the criterion BIC is also consistent for estimating a more general Markovian family (the Partition Markov Models), which includes the Variable Length Markov Chain models and the complete Markov chains. We consider a discrete stationary process with finite alphabet A of size $|A|$. Markov chains of finite order are widely used to model stationary processes with finite memory. In databases with Markovian structure, it is frequently observed a pronounced degree of redundancy, which means that different sequences of symbols have the same effect over the law of the process. For example, in datasets coming from the linguistic field, we can observe words which are synonyms. In some cases, exchanging a word for a synonym, does not change the meaning of sentences. In a more general context, there are also sequences of several words, which are equivalent in that sense. See, for instance, [8,9]. For this kind of data, a model should retrieve and use the redundancy to improve the quality of the estimate. The Partition Markov Model represents the redundancy through a partition of the state space (see [10]). See also [11] and contemporary literature of [10]. Under the assumption of this family, we address the problem of model selection, showing that the model can be selected consistently using the BIC. We show that, in order to apply the BIC criterion, it is not necessary to find a global maximum inside the set of partitions, which will be impossible even for a moderate size of the state space. Instead, it is possible to start the searching process by an initial partition; for instance, the state space itself, and then coarsen the partition, step by step. This process is associated with a metric that governs the state space.

The Partition Markov Models are being used and explored intensively: for instance, [12] combines two statistical concepts—Copulas and Partition Markov Models—with the purpose of defining a natural correction for the estimator of the transition probabilities of a multivariate Markov process. Moreover, Reference [13] presents a simulation study that identifies when this correction succeeds well. A second strategy to deal with this issue, is shown and applied to real data in [14]. The idea is to combine through a copula the partitions coming from the marginal processes and the partition coming from the multivariate process. This strategy shows excellent theoretical properties that will be essential to increasing the predictive ability of the estimation. In [14], the strategy was applied to multivariate Brazilian financial data in order to show how this new estimator allows for considering a longer past (order of the process), in the estimation of the transition probabilities, in comparison with the allowed past in the Partition Markov Models, when the data size is not large enough to ensure reliable results. The application of Partition Markov Models has also been useful to reveal important facts in other areas, as shown in [15], where these models were applied to written texts of European Portuguese, in order to identify change points from the period 16th century–19th century.

Here is a description of the issues addressed in this article, section by section. In Section 2, we introduce the concept of Markov chain with partition $\mathcal{L}$, which is a partition of the state space defined through a stochastic equivalence between the strings of the state space (see [10]). In Section 3, we describe the model selection procedure for choosing the optimal partition, which is based on the BIC criterion and on the concept of good partitions of the state space, which were also introduced in this section. We introduce a distance between the parts of a partition, and this concept defines a metric on the state space and also allows it to build efficient algorithms for estimating the optimal partition (see [10]). In Section 3, we show that the optimal partition can be obtained through the BIC criterion, eventually almost surely, when the sample size tends to infinity. Section 4 shows the application to model navigation patterns on a website. We conclude this paper with a discussion in Section 5. The proof of the results introduced in this paper are included in Appendixes A and B.

2. Preliminaries

Let (X_t) be a discrete time order M Markov chain on a finite alphabet A, with $M < \infty$. 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 $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in \mathcal{S}$, $P(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s)$. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ be a partition of $\mathcal{S}$, $\forall a \in A$ and $L \in \mathcal{L}$, define $P(L, a) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s, X_t = a)$, $P(L) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s)$. If $P(L) > 0$, we define $P(a|L) = \frac{P(L,a)}{P(L)}$. With the purpose of formulating the model, we introduce the following equivalence relation.

Definition 1. *Let (X_t) be a discrete time order M Markov chain on a finite alphabet A, with state space $\mathcal{S} = A^M$:*

(i) *$s, r \in \mathcal{S}$ are equivalent (denoted by $s \sim_p r$) if $P(a|s) = P(a|r)$ $\forall a \in A$.*

(ii) *(X_t) is a Markov chain with partition $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ if this partition is the one defined by the equivalence relationship $\sim_p$ introduced by item (i).*

The equivalence relationship defines a partition on $\mathcal{S}$. The parts of this partition are subsets of $\mathcal{S}$ with the same transition probabilities, i.e., $s, r \in \mathcal{S}$ are in different parts if, and only if, they have different transition probabilities. To understand more deeply the motivation of a model like this, we note that the full Markov chains show a restriction, in terms of point estimation, which is, for an order M model, that the number of parameters given by $|A|^M(|A| - 1)$ grows exponentially with the order M. Another limitation is that the class of full Markov chains is not very rich, since, due to the fixed the alphabet A, there is just one model for each order M and in practical situations a more flexible structure could be necessary. For an extensive discussion of these two restrictions, see [1]. A well-known and richer class of finite order Markov models, introduced by [1,2], is composed of the Variable Length Markov Chains (VLMC). 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 total number of parameters is $|\mathcal{T}|(|A|-1)$. We will see later that the Definition 1(ii) supports both: complete Markov chains and VLMC, becoming a natural extension of the two possibilities.

Remark 1. *Given a Markov chain over the alphabet $A = \{a_1, a_2, ..., a_{|A|}\}$ with partition $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ (Definition 1(ii)); in order to specify the process, it is necessary to estimate $(|A|-1)$ transition probabilities for each part in $\mathcal{L}$. Thus, the set of parameters to estimate is $\{P(a_i|L_j) : 1 \leq i < |A|, 1 \leq j \leq |\mathcal{L}|\}$ and the total number of parameters for the model is $|\mathcal{L}|(|A|-1)$. If the estimation of the transition probabilities is performed under any other conception—for instance, considering a complete Markov chain or a VLMC—the number of parameters to estimate will be higher than $|\mathcal{L}|(|A|-1)$, since they do not consider that there are strings that share transition probabilities.*

The structure of a VLMC can be expressed by a partition in the sense described before. Each model in the family of VLMC models is identified by its Context Tree, and we will use this structure to establish the relation between VLMC and Partition Markov Models (see Example 1).

Example 1. *Let (X_t) be a finite order Markov chain taking values on $A = \{0, 1\}$ 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 $\mathcal{T}$, $d(\mathcal{T}) = \max\big(l(s), s \in \mathcal{T}\big)$, where $l(s)$ is the length of the string $s \in \mathcal{T}$. Consider $d(\mathcal{T}) = 3$ and $\mathcal{T} = \{\{0\}, \{01\}, \{011\}, \{111\}\}$. Define a partition of A^3 as being $\mathcal{L} = \{L_1, L_2, L_3, L_4\}$, where $L_1 = \{\{000\}, \{100\}, \{010\}, \{110\}\}$, $L_2 = \{\{001\}, \{101\}\}$, $L_3 = \{011\}$ and $L_4 = \{111\}$:*

- *(i) Suppose $P(\cdot|s) \neq P(\cdot|s'), \forall s, s' \in \mathcal{T}$. Then, $\mathcal{L}$ verifies Definition 1(ii);*
- *(ii) Suppose $P(\cdot|s) \neq P(\cdot|s'), \forall s, s' \in \mathcal{T} \setminus \{0\}$ and $P(\cdot|\{0\}) = P(\cdot|\{01\})$. Define $L_1' = L_1 \cup L_2$, and then $\mathcal{L}' = \{L_1', L_3, L_4\}$ verifies Definition 1(ii), while $\mathcal{L}$ does not check that definition.*

In the next example, we can see a situation in which can be observed the economy in the number of parameters, achieved by a Partition Markov Model following Definition 1(ii) (see also Remark 1).

Example 2. *Let (X_t) be a finite order Markov chain taking values on $A = \{0, 1\}$ with state space A^3. Suppose that this chain follows the transition probabilities given by the Table 1.*

Table 1. Transition probabilities.

s	000	001	010	011	100	101	110	111
$P(0\|s)$	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1

Considering the process as a full chain, we have eight parameters. Then, if we look more closely, a Context Tree is enough to describe the process, with just four parameters, because $\mathcal{T} = \{\{0\}, \{01\}, \{011\}, \{111\}\}$. Moreover, if we analyze this situation from the perspective of Definition 1(ii), we note that only two parameters are needed to describe the source, since $\mathcal{L} = \{\{000, 001, 010, 100, 101, 110, 111\}, \{011\}\}$, because just the string 011 has different transition probability to 0.

3. Consistent Estimation through the Bayesian Information Criterion

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, which is $N_n(s, a) = \big|\{t : M < t \leq n, x_{t-M}^{t-1} = s, x_t = a\}\big|$. In addition, the number of occurrences of s

in the sample x_1^n is denoted by $N_n(s)$ and $N_n(s) = \left|\{t : M < t \leq n, x_{t-M}^{t-1} = s\}\right|$. The number of occurrences of elements into L followed by a and the total number of strings in L are given by

$$N_n(L,a) = \sum_{s \in L} N_n(s,a), \; N_n(L) = \sum_{s \in L} N_n(s), \; L \in \mathcal{L}. \tag{1}$$

In order to simplify the notation, we use the same notation N_n with different arguments, a string s or a part L. In addition, we note that $N_n(L)$ is a function of the partition $\mathcal{L}$. As a consequence, if we write $P(x_1^n) = \text{Prob}(X_1^n = x_1^n)$, we obtain under the assumption of a hypothetical partition $\mathcal{L}$ of $\mathcal{S}$:

$$P(x_1^n) = P(x_1^M) \prod_{L \in \mathcal{L}, a \in A} P(a|L)^{N_n(L,a)}. \tag{2}$$

The Bayesian Information Criterion (BIC) is defined through a modified maximum likelihood (see [4]). We will call maximum likelihood the maximization of the second term in the Equation (2) for a given observation x_1^n. We denote that term as $\text{ML}(\mathcal{L}, x_1^n)$,

$$\text{ML}(\mathcal{L}, x_1^n) = \prod_{L \in \mathcal{L}, a \in A} \left(\frac{N_n(L,a)}{N_n(L)}\right)^{N_n(L,a)}, \text{ with } N_n(L) \neq 0, L \in \mathcal{L}, \tag{3}$$

and the BIC is given by the next definition.

Definition 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 state space $\mathcal{S} = A^M$ and $\mathcal{L}$ a partition of $\mathcal{S}$, the BIC of the model given by Definition 1(ii), and according to the modified likelihood, Equation (3), is $BIC(\mathcal{L}, x_1^n) = \ln\left(ML(\mathcal{L}, x_1^n)\right) - \frac{(|A|-1)|\mathcal{L}|}{2} \ln(n)$.*

Remark 2. *The results of this paper will remain valid if we replace, in Definition 2, the constant $\frac{(|A|-1)}{2}$ for some arbitrary constant v, positive and finite.*

Below, we define some concepts that help, in practice, to limit the search for an ideal partition to a subset of possible partitions, with natural characteristics.

Definition 3. *Let (X_t) be a discrete time order M Markov chain on a finite alphabet A and $\mathcal{S} = A^M$ the state space. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ be a partition of $\mathcal{S}$:*

(i) *$L \in \mathcal{L}$ is a good part of $\mathcal{L}$ if $\forall s, s' \in L$, $Prob(X_t = .|X_{t-M}^{t-1} = s) = Prob(X_t = .|X_{t-M}^{t-1} = s')$, for values of $t : t > M$;*

(ii) *$\mathcal{L}$ is a good partition of $\mathcal{S}$ if for each $i \in \{1, \ldots, |\mathcal{L}|\}$, L_i verifies item (i).*

Example 3. *We will consider two situations:*

(i) *$\mathcal{L} = \mathcal{S}$ is a good partition of $\mathcal{S}$.*

(ii) *Consider the Example 1(ii), and the partition $\mathcal{L}$ is a good partition of $\mathcal{S} = \{0,1\}^3$.*

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

$$P(a|L) = \text{Prob}(X_t = a|X_{t-M}^{t-1} = s) \;\; \forall a \in A, \tag{4}$$

where $s \in L$.

We introduce a notation that will be used in the next results.

Notation 1.

(a) *Let $\mathcal{L}^{ij}$ denote the partition* $\mathcal{L}^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_{|\mathcal{L}|}, \}$ *where* $\mathcal{L} = \{L_1, \ldots, L_{|\mathcal{L}|}\}$ *is a partition of $\mathcal{S}$, and for $1 \leq i < j \leq |\mathcal{L}|$ with $L_{ij} = L_i \cup L_j$.*

(b) *For $a \in A$, we write $P(L_{ij}, a) = P(L_i, a) + P(L_j, a)$ and $P(L_{ij}) = P(L_i) + P(L_j)$. In addition,*

$$N_n(L_{ij}, a) = N_n(L_i, a) + N_n(L_j, a); \quad N_n(L_{ij}) = N_n(L_i) + N_n(L_j).$$

Note that, if $\mathcal{L}$ is a good partition and $P(\cdot|L_i) = P(\cdot|L_j)$, then $\mathcal{L}^{ij}$ is a good partition. We show a way to build partitions, from good partitions, which are candidates more suitable for checking the Definition 1(ii). This way of building partitions seeks to reduce the size of the partition, step by step.

3.1. A Metric on the State Space

The next result allows formulating the main findings of this section. Nonetheless, this result could be applied to partitions with at least two good parts, i.e., it is not necessary to have good partitions. In this section, we also define a measure to quantify the distance between the parts of a partition. This distance is based on the practical use of the next theorem and allows building an efficient algorithm for estimating the partition given by Definition 1(ii). For complementing, see [16].

Theorem 1. *Let (X_t) be a Markov chain of order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space and x_1^n a sample of the Markov process. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ be a partition of $\mathcal{S}$ and suppose that i and $j$exist, and $i \neq j$ such that L_i and L_j verified the Definition 3(i) (are good parts). Then, $P(a|L_i) = P(a|L_j)$ $\forall a \in A$ if, and only if, eventually almost surely as $n \to \infty$,*

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

where $\mathcal{L}^{ij}$ is defined under $\mathcal{L}$ by Notation 1(a).

Proof. See Appendix A.1. □

It is also possible to decide simultaneously if more than two good parts should be put together, as shown in the next corollary.

Corollary 1. *Let (X_t) be a Markov chain of order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space and x_1^n a sample of the Markov process. If $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ is a partition of $\mathcal{S}$ with K_1 good parts, denoted by $\{L_{i_k}\}_{k=1}^{K_1}$, and T is an index set, $T \subseteq \{1, \ldots, K_1\}$, then, $P(a|L_{i_k}) = P(a|L_{i_l})$ $\forall a \in A, \forall k, l \in T$ if, and only if, eventually almost surely as $n \to \infty$, $BIC(\mathcal{L}, x_1^n) < BIC(\mathcal{L}^T, x_1^n)$, where $\mathcal{L}^T$ denotes the partition which join the $|T|$ good parts in $\cup_{k \in T} L_{i_k}$ generalizing the Notation 1(a).*

Proof. Replace Equation (A3) in the proof of Theorem 1 by

$$\sum_{a \in A} \left\{ \sum_{k \in T} \frac{N_n(L_{i_k}, a)}{n} \ln\left(\frac{N_n(L_{i_k}, a)}{N_n(L_{i_k})}\right) - \frac{N_n(\cup_{k \in T} L_{i_k}, a)}{n} \ln\left(\frac{N_n(\cup_{k \in T} L_{i_k}, a)}{N_n(\cup_{k \in T} L_{i_k})}\right) \right\} < \frac{(|A|-1)(|T|-1)\ln(n)}{2n}.$$

Applying log-sum inequality, the result follows. □

Remark 3. *Let (X_t) be a Markov chain of order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space and x_1^n a sample of the Markov process. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ be a partition of $\mathcal{S}$, given $i, j \in \{1, 2, \ldots, |\mathcal{L}|\}$, $i \neq j$,* *such that L_i and L_j verified the Definition 3(i)(are good parts). 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 Notation 1(a).*

Now, we can introduce a distance in $\mathcal{L}$. This distance allows to establish a metric in the state space $\mathcal{S}$.

Definition 4. *Let (X_t) be 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 $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ be a good partition of $\mathcal{S}$*

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

The next theorem shows that $d_{\mathcal{L}}$ is a distance in $\mathcal{L}$.

Theorem 2. *Let (X_t) be a Markov chain of order M over a finite alphabet A, and $\mathcal{S} = A^M$ the state space and x_1^n a sample of the Markov process. If $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ is a good partition of $\mathcal{S}$, for each n, and for any $i, j, k \in \{1, 2, ..., |\mathcal{L}|\}$:*

(i) $d_{\mathcal{L}}(i,j) \geq 0$ *with equality if and only if* $\frac{N_n(L_i,a)}{N_n(L_i)} = \frac{N_n(L_j,a)}{N_n(L_j)}$ $\forall a \in A$;

(ii) $d_{\mathcal{L}}(i,j) = d_{\mathcal{L}}(j,i)$;

(iii) $d_{\mathcal{L}}(i,k) \leq d_{\mathcal{L}}(i,j) + d_{\mathcal{L}}(j,k)$.

Proof. See Appendix A.2. □

Some observations of practice order are appropriate at this time. Suppose the good partition of Theorem 2 is the space $\mathcal{S}$ itself, so each part of the partition is given by each string of $\mathcal{S}$. Thus, the distance (Definition 4) defines the following relation of equivalence between strings of $\mathcal{S}$, for each value n :

$$s \sim_n r \iff \frac{N_n(s,a)}{N_n(s)} = \frac{N_n(r,a)}{N_n(r)} \quad \forall a \in A, \ \ s, r \in \mathcal{S}.$$

The next result formalizes how the distance is related to the BIC criterion.

Corollary 2. *Let (X_t) be a Markov chain of order M over a finite alphabet A, with $\mathcal{S} = A^M$ the state space and x_1^n a sample of the Markov process. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ be a partition of $\mathcal{S}$, given $i, j \in \{1, 2, \ldots, |\mathcal{L}|\}$, $i \neq j$, such that L_i and L_j verified the Definition 3(i). (are good parts):*

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

Proof. From Equation (A2) in the proof of Theorem 1. □

The previous corollary provides the statistical interpretation of the distance.

3.2. Consistent Estimation of the Process's Partition

In this section, we prove that the partition following Definition 1(ii), referred to herein as minimal good partition, can be obtained by maximizing the equation introduced in Definition 2, in the space of all possible partitions of the state space. For instance, the smaller good partition in the universe of all possible good partitions of $\mathcal{S}$ is the partition defined by the equivalence relationship in Definition 1. Note that, for a discrete time order M Markov chain on a finite alphabet A, with $\mathcal{S} = A^M$ the state space, there exists one and only one minimal good partition of $\mathcal{S}$. The next theorem shows that for large enough n, we obtain, through the BIC, the minimal good partition.

Theorem 3. *Let (X_t) be a Markov chain of order M over a finite alphabet A, with $\mathcal{S} = A^M$ the state space and x_1^n a sample of the Markov process. 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 $n \to \infty$, $\mathcal{L}^* = \mathcal{L}_n^*$, where $\mathcal{L}^*$ is the minimal good partition of $\mathcal{S}$, following Definition 1(ii).

Proof. See Appendix A.3. □

From Corollary 2, algorithms can be formulated to obtain $\mathcal{L}^*$. See, for instance, Algorithm 3.1 in [10]. For large enough n, the algorithm returns the minimal good partition as shown by the next result.

Corollary 3. *Let (X_t) be a Markov chain of order M over a finite alphabet A, $\mathcal{S} = A^M$ the state space and x_1^n a sample of the Markov process. $\hat{\mathcal{L}}_n$, given by the Algorithm 3.1, [10] converges almost surely eventually to $\mathcal{L}^*$, where $\mathcal{L}^*$ is the minimal good partition of $\mathcal{S}$.*

Remark 4. *Algorithm 3.1 [10] requires as initial input a good partition. In the case in which there is not previous information about a good partition or about the length of the memory, the initial good partition can be chosen as the set of sequences, satisfying the suffix property and appearing in the sample at least B times, where B is a positive integer, which corresponds to the first part of the Context Algorithm ([2,3]).*

We note that Corollary 3 also applies to other clustering algorithms based on distances, such as single-linkage clustering. However, exploring this aspect is beyond the scope of this paper.

4. Navigation Patterns on a Web Site (MSNBC.com)

The MSNBC.com anonymous web data set consists of one million user sessions recorded in 24 h on the web site. The dataset can be retrieved from [17] The web pages on the site are divided into 17 categories: frontpage, news, tech, local, opinion, on-air, misc, weather, msn-news, health, living, business, msn-sports, sports, summary, bbs and travel.

Each category will be a letter in the alphabet A, with total size equal to 17. Each user session corresponds to a sequence of symbols from the alphabet starting with the category during which the session is initiated on the MSNBC site. The sequence of categories which the user visits defines the string that finishes when the user leaves the MSNBC site. In Table 2, we show an illustrative sample of 12 user sessions.

Table 2. Sample of 12 sessions. Each line represents the path followed by an user.

User 1	frontpage, tech, tech, frontpage.
User 2	weather, weather, weather, misc, local, weather, weather, weather.
User 3	on-air, msn-news, msn-news, msn-news, msn-news, misc, msn-news.
User 4	news.
User 5	msn-sports, sports, msn-sports.
User 6	frontpage, frontpage, frontpage.
User 7	news, business, tech, local, business, business.
User 8	frontpage.
User 9	local.
User 10	frontpage, tech, tech.
User 11	frontpage, frontpage, business, frontpage.
User 12	sports, sports, sports, sports, sports, sports.

The issue that the model can clarify is to identify the strings that can be considered equivalent, in terms of the next step of the internet surfers. This information could be extremely important in determining the profile of users, in relation to the preferences of states in A. The idealization that supports this application is that there are different sequences in the state space that share the same transition probability to the next symbol in the alphabet of the process. Sequences with such properties form a part of the minimal good partition, which completely describes the process. Furthermore, from this perspective, these sequences are equivalent to deciding the next symbol for the process. These sequences in our application are the paths used by internet surfers.

We use the distance $d_{\mathcal{L}}$ to find the minimal partition. It is known that there are several strategies based on a measure which allows us to reach this purpose, and one of them is the algorithm introduced in [10]. We used three strategies, with input given by the set of strings ($\mathcal{S}$): (i) Algorithm 3.1 introduced in [10]; (ii) Algorithm 3.1 [10] modified; and (iii) an agglomerative strategy. Option (ii) is composed of two stages. First, we join in the same part all the strings r and s of $\mathcal{S}$, which show that $d_{\mathcal{L}}(r,s) < \epsilon, and$ this process generates an initial partition that is used as input of Algorithm 3.1 [10], which is the second stage. The agglomerative strategy of (iii) explores the ability of $d_{\mathcal{L}}$ to build distances between strings of $\mathcal{S}$ and between groups of strings of $\mathcal{S}$. Thus, in this strategy, all of the distances are computed joining groups of strings L_i and L_j if $d_{\mathcal{L}}(i,j) < \frac{(|A|-1)}{2}$. We show in Table 3 that the agglomerative algorithm produces the best BIC value. We expose two cases with order $M = 3$ and $M = 2$, where the first is the order chosen as usual, $3 = \lfloor \log_{|A|}(1.0 \times 10^6) \rfloor - 1$.

Table 3. Number of parts or cardinal of $\mathcal{L}$ and BIC value of the model (Definition 2), for memories 2 and 3, respectively. In (ii), $\epsilon = \frac{(|A|-1)}{2}\frac{1}{10}$ was used. In bold we mark the highest BIC values, which indicate the best method.

	Order 3	
Method	**Number of Parts** ($|\mathcal{L}|$)	**BIC Value**
(i)	196	−2957442
(ii)	210	−2895322
(iii)	269	**−2865622**
	Order 2	
Method	**Number of Parts** ($|\mathcal{L}|$)	**BIC Value**
(i)	177	−3614825
(ii)	177	−3613655
(iii)	181	**−3611092**

Let us look at some specific situations. For example, suppose that our interest is to investigate the parts that lead to the *local* state with probability greater than 0.6. There are seven different parts (using $M = 3$ and method (iii)) that fulfill this condition, and Table 4 shows the composition of each part and its probability of being *local* the next state.

For instance, once the model was selected, in order to predict the next place that the user will visit, we first check during which of the 269 parts his/her path fall and then the corresponding probabilities are used. For example, if the user's path is *weather.weather.misc*, then the probability for the user to visit *local* is 0.7822 and this probability is shared by all three of the strings of L_5. The full partition obtained by the algorithm and the set of transition probabilities associated to each part can be obtained from [18]. We can draw some observations. For example, the transition probabilities of each part $P(a|L_i)$ have been computed, using, in general, several strings, representing a natural improvement in the calculation of the transition probabilities. The reader can find many larger parts in [18], making this observation more incisive. On the other hand, the strings listed as members of the same part (see, for instance, several situations in Table 4), must be considered stochastically equivalent.

Table 4. Parts of $\mathcal{L}$ such that $P(\text{local}|L) > 0.6$.

Part	Strings	$\mathbf{P}(\text{local}|L_i)$
L_1	msn-news.news.local, msn-news.business.local, on-air.tech.local tech.local.local, msn-news.tech.local, business.local.local on-air.local.local, msn-news.local.local	0.7257
L_2	health.news.local, health.local.local, news.local.local	0.6096
L_3	local.local.local	0.8874
L_4	misc.local.local, tech.weather.local, frontpage.opinion.misc local.news.misc, local.misc.local, misc.misc.local	0.6355
L_5	weather.local.local, local.weather.misc, weather.weather.misc	0.7822
L_6	local.local.misc, on-air.weather.misc, msn-news.weather.misc, local.misc.misc	0.7373
L_7	misc.local.misc	0.8563

5. Conclusions

The development of the partition concept in Markov processes allows for proving that, for a stationary, finite memory process and a sample large enough, it is theoretically possible to consistently find a minimal partition to represent the process and this can be accomplished in practice. In this paper, we show (in Theorem 3) that the Bayesian Information Criterion can be used to obtain a consistent estimation of the partition of a Markov process. We show that the use of this criterion is also convenient for producing a consistent strategy that allows for deciding if a candidate partition is preferable to another candidate partition (in Theorem 1). We also define a metric on the state space, allowing for the introduction of a definition of a distance between the parts of a partition (Theorem 2). The distance allows the construction of consistent estimation algorithms to identify the partition. Research in progress suggests that this measure can be harnessed for the development and implementation of robust estimation techniques, given that there are records (see [19]) of the need of these techniques for Markov processes. In summary, in this paper, in addition to responding positively to the question of whether the Bayesian Information Criterion is capable of allowing a consistent estimation of the partition of the Markov process, we also obtain that, in terms of the model selection procedure, the Bayesian Information Criterion corresponds to a distance in the state space of the Markov process.

Acknowledgments: The authors wish to express their gratitude to three referees for their helpful comments on an earlier draft of this paper.

Author Contributions: The authors of this paper jointly conceived the idea for this paper, discussed the agenda for the research, performed the theoretical and numerical calculations, and prepared each draft of the paper. The authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Previously, we defined some concepts and established some useful notation that will be used in this section.

Definition A1. *Let P and Q be probability distributions on A. The relative entropy between P and Q is given by $D(P(\cdot)||Q(\cdot)) = \sum_{a\in A} P(a)\ln\left(\frac{P(a)}{Q(a)}\right)$, with $Q(a) \neq 0, \forall a \in A$.*

From Equation (1), define

$$r_n(L,a) = \frac{N_n(L,a)}{n} \quad \textit{and} \quad r_n(L) = \frac{N_n(L)}{n}, \quad a \in A, \quad L \in \mathcal{L}. \tag{A1}$$

Because $N_n(L) \neq 0 \forall L \in \mathcal{L}$, $r_n(L) \neq 0, \forall L \in \mathcal{L}$.

Appendix A.1. Proof of Theorem 1

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

Then,

$$\begin{aligned} BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) = & \sum_{a \in A} \Big\{ N_n(L_i,a) \ln \left(\frac{r_n(L_i,a)}{r_n(L_i)} \right) \\ & + N_n(L_j,a) \ln \left(\frac{r_n(L_j,a)}{r_n(L_j)} \right) \\ & - N_n(L_{ij},a) \ln \left(\frac{r_n(L_{ij},a)}{r_n(L_{ij})} \right) \Big\} - \frac{(|A|-1)}{2} \ln(n). \end{aligned} \tag{A2}$$

We note that $BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n)$ if, and only if,

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

Because $r_n(L,a)$ and $r_n(L)$ are non-negative and applying the log-sum inequality, we obtain

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

or, equivalently,

$$r_n(L_i,a) \ln \left(\frac{r_n(L_i,a)}{r_n(L_i)} \right) + r_n(L_j,a) \ln \left(\frac{r_n(L_j,a)}{r_n(L_j)} \right) \geq r_n(L_{ij},a) \ln \left(\frac{r_n(L_{ij},a)}{r_n(L_{ij})} \right), \tag{A3}$$

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 a consequence, inequality Equation (A3) $\Rightarrow$

$$\begin{aligned} \sum_{a \in A} \Big\{ & r_n(L_i,a) \ln \left(\frac{r_n(L_i,a)}{r_n(L_i)} \right) + r_n(L_j,a) \ln \left(\frac{r_n(L_j,a)}{r_n(L_j)} \right) \\ & - r_n(L_{ij},a) \ln \left(\frac{r_n(L_{ij},a)}{r_n(L_{ij})} \right) \Big\} \geq 0, \end{aligned} \tag{A4}$$

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 (A3), we have that if $\lim_{n\to\infty} I_{\{BIC(\mathcal{L}^{ij},x_1^n)>BIC(\mathcal{L},x_1^n)\}} = 1$, where I_W is the indicator function of the set W, then

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

from Equation (A4), and, taking the limit inside the sum, we obtain

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

applying the log-sum inequality. 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

$$P(a|L_{ij}) = P(a|L_i)\ \forall a \in A, \tag{A5}$$

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

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

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

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

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

By using Assumption (A5), where $D(P(\cdot)||Q(\cdot))$ is the relative entropy, given by Definition A1, and applying the Lemma 6.3 from [4] and the Proposition A1, for any $\delta > 0$ and large enough n,

$$\begin{aligned} D\left(\frac{N_n(L,.)}{N_n(L)}\middle|\middle| P(.|L)\right) &\leq \sum_{a\in A}\frac{\left(\frac{N_n(L,a)}{N_n(L)} - P(a|L)\right)^2}{P(a|L)} \\ &\leq \sum_{a\in A}\frac{\frac{\delta\ln(n)}{N_n(L)}}{P(a|L)}. \end{aligned} \tag{A6}$$

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

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

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,

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

Appendix A.2. Proof of Theorem 2

Proof. (i) Fix a value $a \in A$, set two parts $L_i, L_j \in \mathcal{L}$, define $a_1 = N_n(L_i,a), b_1 = N_n(L_i)$ and $a_2 = N_n(L_j,a), b_2 = N_n(L_j)$.
$\sum_{s=1,2} a_s = N_n(L_{ij},a)$ and $\sum_{s=1,2} b_s = N_n(L_{ij})$. Thus, $a_1\ln(\frac{a_1}{b_1}) + a_2\ln(\frac{a_2}{b_2}) \geq \sum_{s=1,2} a_s \ln\left(\frac{\sum_{s=1,2} a_s}{\sum_{s=1,2} b_s}\right)$, with equality $\Leftrightarrow \frac{a_1}{b_1} = \frac{a_2}{b_2}$. This means that

$$\sum_{a\in A} N_n(L_i,a)\ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) + N_n(L_j,a)\ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right) \geq \sum_{a\in A} N_n(L_{ij},a)\ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right)$$

with equality $\Leftrightarrow \frac{N_n(L_i,a)}{N_n(L_i)} = \frac{N_n(L_j,a)}{N_n(L_j)}$. Thereby, (i) is proved.
(iii) $d_{\mathcal{L}}(i,k) \leq d_{\mathcal{L}}(i,j) + d_{\mathcal{L}}(j,k)$ if, and only if,

$$\begin{aligned} 0 \leq \sum_{s=i,k}\sum_{a\in A} N_n(L_j,a)\Big(\ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right) - \ln\left(\frac{N_n(L_{sj},a)}{N_n(L_{sj})}\right)\Big) + \\ \sum_{a\in A} N_n(L_i,a)\Big(\ln\left(\frac{N_n(L_{ik},a)}{N_n(L_{ik})}\right) - \ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right)\Big) + \\ \sum_{a\in A} N_n(L_k,a)\Big(\ln\left(\frac{N_n(L_{ik},a)}{N_n(L_{ik})}\right) - \ln\left(\frac{N_n(L_{kj},a)}{N_n(L_{kj})}\right)\Big), \end{aligned}$$

and the right side is equivalent to

$$\begin{aligned} \sum_{s=i,k} N_n(L_j)\sum_{a\in A}\frac{N_n(L_j,a)}{N_n(L_j)}\ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\Big/\frac{N_n(L_{sj},a)}{N_n(L_{sj})}\right) + \\ \sum_{s=i,k}\sum_{a\in A}\frac{N_n(L_s,a)N_n(L_{ik})}{N_n(L_{ik},a)}\frac{N_n(L_{ik},a)}{N_n(L_{ik})}\ln\left(\frac{N_n(L_{ik},a)}{N_n(L_{ik})}\Big/\frac{N_n(L_{sj},a)}{N_n(L_{sj})}\right), \end{aligned}$$

which is greater than

$$N_n(L_j)\sum_{s=i,k} D\left(\frac{N_n(L_j,.)}{N_n(L_j)}\middle\|\frac{N_n(L_{sj},.)}{N_n(L_{sj})}\right)+\frac{1}{n}\sum_{s=i,k} D\left(\frac{N_n(L_{ik},.)}{N_n(L_{ik})}\middle\|\frac{N_n(L_{sj},.)}{N_n(L_{sj})}\right)\geq 0.$$

Thus, (iii) is proved. □

Appendix A.3. Proof of Theorem 3

Proof. The first part of the proof will be devoted to show that the maximum BIC just can be attained on a good partition.

Consider $\mathcal{L}^b = \{L_1^b, L_2^b, \ldots, L^b_{|\mathcal{L}^b|}\}$ a partition of $\mathcal{S}$, which is not a good partition. This means that at least some part does not verify the Definition 3(i). Suppose, without loss of generality, that $\mathcal{L}^b$ has just one part that is not good, L_1^b. Suppose also that $BIC(\mathcal{L}^b, x_1^n) > BIC(\mathcal{L}, x_1^n), \forall\, \mathcal{L} \in \mathcal{P}$.

Let $\mathcal{L}^{b^*} = \{S_{11}, S_{12}, \ldots, S_{1k_1}, L_2^b, \ldots, L^b_{|\mathcal{L}^b|}\}$, where $\cup_{i=1}^{k_1} S_{1i} = L_1^b$ and each S_{1i} verifies the Definition 3(i). In addition, impose $\forall i, j \in \{1, \ldots, k_1\}, i \neq j$, $P(.|S_{1i}) \neq P(.|S_{1j})$. By Corollary 1 and Remark 3, we obtain $BIC(\mathcal{L}^b, x_1^n) < BIC(\mathcal{L}^{b^*}, x_1^n)$, eventually almost surely as $n \to \infty$. This is absurd coming from the supposition of the existence of $\mathcal{L}^b$ such that the BIC attains the maximum on $\mathcal{P}$.

The second part of the proof is dedicated to identifying the minimal good partition in the space of good partitions. Define the set of good partitions, $\mathcal{P}' = \{\mathcal{L} : \mathcal{L}$ is a good partition of $\mathcal{S}\}$.

Let an arbitrary good partition $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\} \in \mathcal{P}'$, by Corollary 1, $K_1 = |\mathcal{L}|$ and taking an appropriate $T \subseteq \{1, \ldots, |\mathcal{L}|\}$, $BIC(\mathcal{L}, x_1^n) \leq BIC(\mathcal{L}^T, x_1^n)$, eventually almost surely as $n \to \infty$.

Note that $|\mathcal{P}| < \infty$ because $|\mathcal{S}| < \infty$, and define $\mathcal{L}^* = \text{argmax}_{\mathcal{L} \in \mathcal{P}'} BIC(\mathcal{L}^T, x_1^n)$ when $n \to \infty$. By construction, $\mathcal{L}^* \in \mathcal{P}'$. If $\mathcal{L}^* = \{L_1^*, \ldots, L^*_{|\mathcal{L}^*|}\}$ is not minimal, then at least there are i and j such that $BIC(\mathcal{L}^*, x_1^n) < BIC(\mathcal{L}^{*ij}, x_1^n)$ and it is impossible because $\mathcal{L}^*$ was defined as being the maximum argument of the BIC criterion in the class $\mathcal{P}'$, and by construction $\mathcal{L}^{*ij} \in \mathcal{P}'$ (see Notation 1). As a consequence, $\mathcal{L}^*$ is the minimal good partition. □

Appendix B. Auxiliary Results

Proposition A1. *Allow the process $P(\cdot|L)$ on A, where $L \in \mathcal{L}$ and $\mathcal{L}$ is a good partition of $\mathcal{S}$. For arbitrary $\delta > 0$, there exists $\alpha > 0$ (depending on P), such that, eventually almost surely as $n \to \infty$,*

$$\left|\frac{N_n(L,a)}{N_n(L)} - P(a|L)\right| \leq \sqrt{\frac{\delta \ln(n)}{N_n(L)}},$$

with $M < \alpha \ln(n)$.

Proof. From Corollary 2 [6], we have that, for any $\epsilon > 0$, there exists $\alpha > 0$ (depending on P), such that eventually almost surely as $n \to \infty$,

$$\left|\frac{N_n(s,a)}{N_n(s)} - P(a|s)\right| \leq \sqrt{\frac{\epsilon \ln(N_n(s))}{N_n(s)}}, \tag{A7}$$

with $M < \alpha \ln(n)$. Letting $\delta > 0$ and $\epsilon = \frac{\delta}{|A|^{2M}}$ in Equation (A7), we obtain

$$\begin{aligned}\frac{N_n(s,a)}{N_n(s)} - P(a|s) &\leq \sqrt{\frac{\delta \ln(N_n(s))}{|A|^{2M} N_n(s)}},\\ N_n(s,a) - N_n(s)P(a|s) &\leq \sqrt{\frac{\delta \ln(N_n(s))}{|A|^{2M}} N_n(s)}.\end{aligned}$$

Because $\mathcal{L}$ is a good partition of $\mathcal{S}$, $s \in L$ and $L \in \mathcal{L}$, we obtain

$$\sum_{s\in L} N_n(s,a) - P(a|L)\sum_{s\in L} N_n(s) \quad \leq \quad \sum_{s\in L}\sqrt{\frac{\delta \ln(N_n(s))}{|A|^{2M}} N_n(s)}.$$

Following the equations in (1), we have

$$N_n(L,a) - P(a|L)N_n(L) \quad \leq \quad \frac{\sqrt{\delta \ln(n)}}{|A|^M}\sum_{s\in L}\sqrt{N_n(s)}.$$

Then,

$$\begin{aligned}
\frac{N_n(L,a)}{N_n(L)} - P(a|L) \quad &\leq \quad \frac{\sqrt{\delta \ln(n)}}{|A|^M N_n(L)}|L|\sqrt{\max_{s\in L}(N_n(s))} \\
&\leq \quad \frac{\sqrt{\delta \ln(n)}}{|A|^M N_n(L)}|A|^M\sqrt{\sum_{s\in L} N_n(s)} \\
&= \quad \frac{\sqrt{\delta \ln(n)}}{N_n(L)}\sqrt{N_n(L)} \\
&= \quad \sqrt{\frac{\delta \ln(n)}{N_n(L)}}.
\end{aligned}$$

□

References

1. Buhlmann, P.; Wyner, A. Variable length Markov chains. *Ann. Stat.* **1999,** *27*, 480–513.
2. Rissanen, J. A universal data compression system. *IEEE Trans. Inf. Theory* **1983**, *29*, 656–664.
3. Weinberger, M.; Rissanen, J.; Feder, M. A universal finite memory source. *IEEE Trans. Inf. Theory* **1995**, *41*, 643–652.
4. Csiszár, I.; Talata, Z. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inf. Theory* **2006**, *52*, 1007–1016.
5. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
6. Csiszár, I. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inf. Theory* **2002**, *48*, 1616–1628.
7. Csiszár, I.; Shields, P.C. The consistency of the BIC Markov order estimator. *Ann. Stat.* **2000**, *28*, 1601–1619.
8. Jääskinen, V.; Xiong, J.; Corander, J.; Koski, T. Sparse markov chains for sequence data. *Scand. J. Stat.* **2014**, *41*, 639–655.
9. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing;* MIT Press: Cambridge, MA, USA, 1999; Volume 999.
10. García, J.E.; González-López, V.A. Minimal Markov Models. In Proceedings of the Fourth Workshop on Information Theoretic Methods in Science and Engineering, Helsinki, Finland, 7–10 August 2011; Volume 1, pp. 25–28.
11. Farcomeni, A. Hidden Markov Partition Models. *Stat. Probab. Lett.* **2011**, *81*, 1766–1770.
12. García, J.E.; Fernández, M. Copula based model correction for bivariate Bernoulli financial series. In Proceedings of the 11th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2013), Rhodes, Greece, 21–27 September 2013; AIP Publishing: Melville, NY, USA, 2013; Volume 1558, pp. 1487–1490.
13. Fernández, M.; García Jesús, E.; González-López, V.A. Multivariate Markov chain predictions adjusted with copula models. In *New Trends in Stochastic Modeling and Data Analysis*; ISAST: Athens, Greece, 2015.
14. García, J.E.; González-López, V.A.; Hirsh, I.D. Copula-Based Prediction of Economic Movements. In Proceedings of the 13th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2015), Rhodes, Greece, 23–29 September, 2015; AIP Publishing: Melville, NY, USA, 2015; Volume 1738, p. 140005.

15. García, J.E.; González-López, V.A. Detecting regime changes in Markov models. In Proceedings of The Sixth Workshop on Information Theoretic Methods in Science and Engineering, Tokyo, Japan, 26–29 August 2013.
16. Gusfield, D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology;* Cambridge University Press: Cambridge, UK, 1997.
17. MSNBC.com Anonymous Web Data Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data (accessed on 5 April 2017).
18. Index of / jg/MSNBC. Available online: http://www.ime.unicamp.br/~jg/MSNBC/ (accessed on 5 April 2017).
19. Galves, A.; Galves, C.; García, J.; Garcia, N.L.; Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *Ann. Appl. Stat.* **2012**, *6*, 186–209.