MODEL SELECTION FOR MULTIVARIATE STOCHASTIC PROCESSES

Jesús E. García[1], *Verónica A. González-López*[2] *and M. L. L. Viola*[3]

[1 2] Department of statistics, University of Campinas,
Rua Sergio Buarque de Holanda,651 Cidade Universitária CEP 13083-859 Campinas, SP, Brazil.
[3] Department of statistics, Universidade Federal de São Carlos,
Via Washington Luís, km 235 - Bairro Monjolinho CEP 13.565-905 - São Carlos, SP, Brazil.

ABSTRACT

We address the problem of model selection for a multivariate source with finite alphabet. Families of Markov models and model selection algorithms are generalized for the multivariate case. For Markovian sources our model selection procedures are consistent in the sense that, eventually, as the collected data grows, the sources Markov model will be retrieved exactly and it will be described with a minimal number of parameters.

1. INTRODUCTION

Multivariate Markov chains are used for modeling stochastic processes arising on many areas as for example linguistics, biology and neuroscience. There are diverse models families from which to choose a model for a given data set. For example Markov chains of order m, variable length Markov chains (VLMC) see for example (5), (6), (2) or partition Markov models see (4). On each family, the selection of a specific Markov model gives information about the dependence structure for the dataset.

A recurrent problem is to model multiple streams of finite memory data with distributions that are suspected to be dependent or similar or equal. In the case of independent sources, the interest is to find the differences and similarities between the distribution of the sources. In the dependent case we want to find de dependence structure for the multivariate source. In this paper we propose a class of Markov models for each of that cases (dependent or independent sources), that generalize the partition Markov models for multivariate sources. We show procedures to, given a dataset, select a model in our class of models, that approximate the joint law of the source. The procedure are consistent in the sense that if the law of the source is Markovian, eventually, as the collected data grow, the source's Markov model will be retrieved exactly. This work extend and generalize previous results about minimal Markov models and context tree models as in (4), (6), (2), (1) and (3). In section 2 we revisit the family of partition Markov models. In section 3 we address the problem of simultaneously modeling multiple data sources. Finally in section 4 we show a procedure to estimate the internal structure of dependence between the coordinates of a multivariate stationary source.

2. MARKOV CHAIN WITH PARTITION $\mathcal{L}$

Let (X_t) be a discrete time, finite memory Markov chain on a finite alphabet A. Denote the string $a_m a_{m+1} \dots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$. Let M be the maximum memory for the process, and $S = A^M$.

For each $a \in A$ and $s \in \mathcal{S}$,

$$P(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s);$$

Definition 2.1. *Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. We will say that $s, r \in \mathcal{S}$ are equivalent (denoted by $s \sim_p r$) if $P(a|s) = P(a|r)$ $\forall a \in A$.*
For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.

Remark 2.1. *The equivalence relationship defines a partition of $\mathcal{S}$. The parts of this partition are the equivalence classes. The classes are the subsets of $\mathcal{S}$ with the same transition probabilities i.e. $s, r \in \mathcal{S}$ belongs to different classes if and only if they have different transition probabilities.*

Remark 2.2. *We can think that each element of $\mathcal{S}$ on the same equivalence class activates the same random mechanism to choose the next element in the Markov chain.*

We can define now the a Markov chain with partition $\mathcal{L}$.

Definition 2.2. *let (X_t) be a discrete time, order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$. We will say that (X_t) is a Markov chain with partition $\mathcal{L}$ if this partition is the one defined by the equivalence relationship $\sim_p$ introduced by definition 2.1.*

$$\text{Let } \mathcal{L} = \{L_1, L_2, \dots, L_K\} \text{ be the partition of } (X_t)$$

$$P(a|L_i) = P(a|s), \text{ for any } s \in L_i$$

Remark 2.3. *The set of parameters for a Markov chain over the alphabet A with partition $\mathcal{L}$ can be denoted by,*

$$\{P(a|L) : a \in A, L \in \mathcal{L}\}.$$

If we know the equivalence relationship for a given Markov chain, then we need $(|A| - 1)$ transition probabilities for each class to specify the model. Then the number of parameters for the model is $|\mathcal{L}|(|A| - 1)$.

2.1. Partition Markov model selection

Let x_1^n be a sample of the process $\big(X_t\big)$, $s \in \mathcal{S}$, $a \in A$ 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,

$$N_n(s,a) = \big|\{t : M < t \leq n, x_{t-M}^{t-1} = s, x_t = a\}\big|, \quad (1)$$

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

$$N_n(s) = \big|\{t : M < t \leq n, x_{t-M}^{t-1} = s\}\big|. \quad (2)$$

To simplify the notation we will omit the $_n$ on N_n.

2.2. A distance in S

Definition 2.3. *We define the distance d in S,*

$$\begin{aligned} d(s,r) &= \frac{1}{\ln(n)} \sum_{a \in A} \left\{ N(s,a) \ln\left(\frac{N(s,a)}{N(s)}\right) \right. \\ &+ N(r,a) \ln\left(\frac{N(r,a)}{N(r)}\right) \\ &- \left. (N(s,a) + N(r,a)) \ln\left(\frac{N(s,a)+N(r,a)}{N(s)+N(r)}\right)\right) \end{aligned}$$

for any $s, r \in S$.

Proposition 2.1. *For any $s, r \in S$,*

i. $d(s,r) \geq 0$ *with equality if and only if* $\frac{N(s,a)}{N(s)} = \frac{N(r,a)}{N(r)}$ $\forall a \in A$,

ii. $d(s,r) = d(r,s)$,

Remark 2.4. *d can be generalized to subsets (see (4)).*

Theorem 2.1. *(Consistence in the case of a Markov source) Let (X_t) be a discrete time, order M Markov chain on a finite alphabet A. Let x_1^n be a sample of the process, then for n large enough, for each $s, r \in S$, $d(r,s) < \frac{(|A|-1)}{2}$ iff s and r belong to the same class.*

Algorithm 2.1. *(Partition selection algorithm)*
Input*: $d(s,r) \forall s, r \in S$;* **Output***: $\hat{\mathcal{L}}_n$.*
$B = S$
$\hat{\mathcal{L}}_n = \emptyset$
while $B \neq \emptyset$

select $s \in B$

define $L_s = \{s\}$

$B = B \setminus \{s\}$

for each $r \in B, r \neq s$

if $d(s,r) < \frac{(|A|-1)}{2}$
$L_s = L_s \cup \{r\}$
$B = B \setminus \{r\}$

$\hat{\mathcal{L}}_n = \hat{\mathcal{L}}_n \cup \{L_s\}$

Return*: $\hat{\mathcal{L}}_n = \{L_1, L_2, \ldots, L_K\}$*

If the source is Markovian, for n large enough, the algorithm returns the partition for the source.

Corollary 2.1. *Under the assumptions of Theorem 2.1, $\hat{\mathcal{L}}_n$, given by the algorithm 2.1 converges almost surely eventually to $\mathcal{L}^*$, where $\mathcal{L}^*$ is the partition of $\mathcal{S}$ defined by the equivalence relationship.*

3. GENERALIZED PARTITION MARKOV MODELS FOR MULTIPLE INDEPENDENT FINITE MEMORY SOURCES

In this section we extend the family of models for multiple independent sources of data. We also extend our algorithm. As in (4), the procedure is consistent and tight, for Markovian sources, eventually, as the data grow, the source's Markov model will be retrieved exactly and described with the minimal number of parameters.

We will consider a dataset which consist of K sequences of size n_k, for $k = 1, ..., K$.

3.1. Model family

Let (X_t^k) for $k = 1, ..., K$ be the K independent finite memory stochastic processes, all of them stationary and ergodic. For each process (X_t^k) let S_k and d_k be the state space and order of the respective Markov model.

Definition 3.1. $\mathcal{S} = \{(s,k) : s \in S_k, k = 1, 2, ..., K.\}$

For each $a \in A$ and $(s,k) \in \mathcal{S}$,

$$P_k(a|s) = \text{Prob}(X_t^k = a | X_{t-M}^{k,t-1} = s);$$

The models in our family are indexed by the partition defined in the following equivalence relation.

Definition 3.2. *We will say that $(s,i), (r,j) \in \mathcal{S}$ are equivalent (denoted by $(s,i) \sim_{P,K} (r,j)$) if $P_i(a|s) = P_j(a|r)$ $\forall a \in A$. For any $(s,i) \in \mathcal{S}$, the equivalence class of (s,i) is given by $[(s,i)] = \{(r,j) \in \mathcal{S} | (r,j) \sim_{P,K} (s,i)\}$.*

We can define now the a set of Markov chain with partition $\mathcal{L}$.

Definition 3.3. *let X be a set of K independent Markov chains on A and let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$. We will say that X is a set of Markov chains with partition $\mathcal{L}$ if this partition is the one defined by the equivalence relationship $\sim_{P,K}$ introduced by definition 3.2.*

Remark 3.1. *The parameters for a set of independent Markov chains over the alphabet A with partition $\mathcal{L}$ is,*

$$\{P(a|L) : a \in A,\ L \in \mathcal{L}\}.,$$

where $P(a|L) = P_i(a|s)$ for any $(i,s) \in L$.
The number of parameters for the model is $|\mathcal{L}|(|A|-1)$.

3.2. A distance between sequences

Definition 3.4. *For any $(s,i),(r,j) \in \mathcal{S}$, we define the distance $d_K((s,i),(r,j))$ in S as*

$$\begin{aligned} d_K((s,i),(r,j)) &= \frac{1}{\ln(n)} \sum_{a\in A} \left\{ N_i(s,a) \ln\left(\frac{N_i(s,a)}{N_i(s)}\right) \right. \\ &+ N_j(r,a) \ln\left(\frac{N_j(r,a)}{N_j(r)}\right) \\ &- (N_i(s,a)+N_j(r,a)) \times \\ &\left. \times \ln\left(\frac{N_i(s,a)+N_i(r,a)}{N_i(s)+N_j(r)}\right) \right\}, \end{aligned}$$

where $N_i(s)$ and $N_i(s,a)$ are the number of times that the sequences s and sa respectively appear in the sample i.

Proposition 3.1. *$d_K(.,.)$ have the following properties,*

i. $d_K((s,i),(r,j)) \geq 0$ *with equality if and only if* $\frac{N_i(s,a)}{N_i(s)} = \frac{N_j(r,a)}{N_j(r)} \ \forall a \in A$,

ii. $d_K((s,i),(r,j)) = d_K((r,j),(s,i)))$,

To simplify the notation and without loss of generality we will suppose that all the sequences have the same size n.

Theorem 3.1. *(Consistence in the case of Markov sources) Let X be a set of independent Markov chain of finite order, $(x_1^{i,n})_{i=1}^K$ a size n sample of each process. For each $(s,i),(r,j) \in S$ for n large enough, $d_K((s,i),(r,j)) < \frac{|A|-1}{2}$ iff (s,i) and (r,j) belong to the same class.*

The same algorithm 2.1 can be used (with $d_K(.,.)$) to estimate the partition for the set of chains.

4. MULTIVARIATE SOURCES

In this section we will consider the case in which we have a multivariate source with dependent coordinates.

To simplify the notation, we will assume that the partition Markov model is known. Our objective is to obtain for each part a partition of the set of coordinates on independent sets. The same procedure can be used to find subsets of the coordinates that are conditionally independent.

Let (X_t) be a Markov chain on $A = B^l$ **with partition** $\mathcal{L}$. For $U = \{u_1,...u_k\} \subset \{1,2,...,l\}$ and $a = (a_1,...,a_l) \in A$, define:

i) $a^u = (a_{u_1},...,a_{u_k})$,

ii) for any $L \in \mathcal{L}$,

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

iii) for $s \in \mathcal{S}$

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

iv) for $L \in \mathcal{L}$

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

Example

Consider $B = \{0,1,2\}$ with dimension $l = 2$, the alphabet will be $A = B^2 = \{0,1,2\}^2$. For $L \in \mathcal{L}$, we need to specify $P(a|L)$, this means $(|A|-1) = 8$ parameters for each L. If for a fixed L the first coordinate is independent from the second then $P(a|L) = P(a_1|L)P(a_2|L) \ \forall a \in A$ and the number of parameter will be $(|B|-1)+(|B|-1) = 4$ for this L.

In general, for $A = B^l$, fix $L \in \mathcal{L}$ and a partition $\mathcal{I}_L$ of $\{1,2,...,l\}$ in independent coordinates, we have that

$$P(a|L) = \prod_{C\in\mathcal{I}_L} P(a^C|L) \ \forall a \in A$$

and the number of parameters needed for the part L will be

$$\sum_{C\in\mathcal{I}} (|B|^{|C|} - 1)$$

4.1. Conditional dependence structure

Definition 4.1. *For each $L \in \mathcal{L}$, define $\mathcal{I}_L$ as de maximal partition of $\{1,2,...l\}$ such that*

$$P(a|L) = \prod_{C\in\mathcal{I}_L} P(a^C|L) \ \forall a \in A.$$

We will say that $\mathcal{I}_{\mathcal{L}} = \{\mathcal{I}_L\}_{L\in\mathcal{L}}$ is the structure of conditional dependence for the process.

4.2. Estimating the conditional dependence structure

Our procedure to estimate $\mathcal{I}_{\mathcal{L}}$ is based on the Bayesian information criterion (BIC).

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

The maxima for $\prod_{L\in\mathcal{L},a\in A} \prod_{C\in\mathcal{I}_L} P(a^C|L)^{N_n^{\mathcal{L}}(L,a)}$ is

$$\text{ML}(\mathcal{L},\mathcal{I}_{\mathcal{L}},x_1^n) = \prod_{L\in\mathcal{L},a\in A} \prod_{C\in\mathcal{I}_L} \left(\frac{N_n^{\mathcal{L}}(L,a^C)}{N_n^{\mathcal{L}}(L)}\right)^{N_n^{\mathcal{L}}(L,a)},$$

and the BIC criterion for ou class of models,

$$\begin{aligned} BIC(\mathcal{L},\mathcal{I}_{\mathcal{L}},x_1^n) &= \ln\left(\text{ML}(\mathcal{L},\mathcal{I}_{\mathcal{L}},x_1^n)\right) \\ &- \sum_{L\in\mathcal{L}} \sum_{C\in\mathcal{I}_L} (|A|^{|C|}-1)\frac{\ln(n)}{2}. \end{aligned}$$

For a Markovian source the BIC model selection methodology is consistent.

4.3. Consistence

Theorem 4.1. *Let (X_t) be a Markov chain of order M over a finite alphabet A, with partition $\mathcal{L}^*$ and structure of conditional dependence $\mathcal{I}_{\mathcal{L}^*}$. Define,*

$$\mathcal{I}_{\mathcal{L}_n} = \arg\max_{\mathcal{I}\in\mathcal{D}} \{BIC(\mathcal{L}_n,\mathcal{I},x_1^n)\},$$

Where $\mathcal{D}$ is the set of all possible structures of dependences for A and $\mathcal{L}_n$, then, eventually almost surely as $n \to \infty$,

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

The next Theorem shows that is not necessary to search for the maxima on $\mathcal{D}$.

Consider any collection of partitions of $\{1, 2, ...l\}$,

$$\mathcal{D} = \{D_L\}_{L \in \mathcal{L}}.$$

Fix $L_0 \in \mathcal{L}$ and $U, V \in D_{L_0}, U \neq V$. Define $\mathcal{D}^{L_0,U,V}$ as the collection of partitions containing the same partitions than $\mathcal{D}$ except D_{L_0} is substituted by

$$D_{L_0} \setminus \{\{U\}, \{V\}\} \cup \{U \cup V\}.$$

Theorem 4.2. *Let (X_t) be a Markov chain over $A = B^l$ with partition $\mathcal{L}$, then,*

$$P(a^{U \cup V}|L_0) = P(a^U|L_0)P(a^V|L_0) \; \forall a \in A$$

if, and only if, eventually almost surely as $n \to \infty$,

$$BIC(\mathcal{L}, \mathcal{D}^{L_0,U,V}, x_1^n) < BIC(\mathcal{L}, \mathcal{D}, x_1^n).$$

5. CONCLUSION

In this paper we study two generalizations of previous results about minimal Markov models to the multivariate case. First, we consider the case in which we have multiple independent sources. We model all the sources simultaneously and the model selection algorithm returns not only the set of equivalent states for each source, it also identify all the states in all sources which can be considered equivalents between them. In this way, even strings activating the same random mechanism on different sources are identified and classified. The second generalization correspond to a stationary source with a multivariate alphabet. In this case we first choose a partition Markov model and then, for the transition probabilities of each part, we identify the maximal partition of the set of coordinates such that the different parts are independent. A similar procedure and algorithm can be used to find subsets of coordinates which are conditionally independent.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support for this research provided by CNPqs projects 485999/2007-2 and 476501/2009-1 and USP project "Mathematics, computation, language and the brain".

7. REFERENCES

[1] BUHLMANN P. and WYNER A. (1999). Variable length Markov chains. Ann. Statist. **27** 480–513.

[2] 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.

[3] GALVES, A., GALVES, C., GARCIA, J. E., GARCIA, N. L. and LEONARDI, F. (2012). Context tree selection and linguistic rhythm retrieval from written texts. Annals of Applied Statistics, **6** 1, 186 (2012).

[4] GARCIA, J. and GONZALEZ-LOPEZ, V. (20010)Minimal Markov Models, arXiv:1002.0729v1.

[5] RISSANEN J. (1983). A universal data compression system, IEEE Trans. Inform. Theory **29**(5) 656 – 664.

[6] WEINBERGER, M., RISSANEN, J. and FEDER, M. (1995). A universal finite memory source, IEEE Trans. Inform. Theory **41**(3) 643 – 652.