Model selection for multivariate Markov chains and language classification [1]

García, Jesús E.[2]
González-López, V. A.[3]
Viola, M. L. L. [4]

August 10, 2012

[1]This work was partially supported by CNPq's projects grant numbers 485999/2007-2 and 476501/2009-1. Capes doctoral grant for Viola, M. L. L.
[2]IMECC-UNICAMP e-mail: jg@ime.unicamp.br
[3]IMECC-UNICAMP e-mail: veronica@ime.unicamp.br

Introduction

- We introduce a new class $\mathcal{M}$ of finite order Markov chain models.
- Address the following model selection problem
  - given a sample generated by a stationary process (source);
  - find a Markov model in $\mathcal{M}$ for the sample with the minimal number of parameters.

Notation

Let (X_t) be a discrete time order M Markov chain

- A the finite alphabet;
- $x_1^n = x_1 x_2 \dots x_n$ a realization of the process;
- $\mathcal{S} = A^M$ the state space;
- $P(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s),\ a \in A,\ s \in \mathcal{S}$ the transition probabilities.

Equivalence relationship on $\mathcal{S}$

Definition

for $s, r \in \mathcal{S}$; $s \sim_p r \iff P(a|s) = P(a|r)\ \forall a \in A$.

- For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.
- The classes defined by $\sim_p$ are the subsets of $\mathcal{S}$ with the same transition probabilities.

Equivalence relationship on $\mathcal{S}$

- The equivalence relation defines a partition $\mathcal{L}$ of $\mathcal{S}$.
- We have $(|A|-1)$ transition probabilities for each "part" (element of $\mathcal{L}$), obtaining a model with $(|A|-1)|\mathcal{L}|$ parameters.
- The elements of $\mathcal{S}$ on the same equivalence class activate the same random mechanism to choose the next element in the Markov chain.

Markov model with partition $\mathcal{L}$

Definition

let (X_t) be a discrete time, order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ 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 $\sim_p$.

Example

- $A = \{0, 1\}$, $M = 2$, $\quad \mathcal{S}(= A^M) = \{00, 01, 10, 11\}$, $P(0|00) = P(0|01) = 0.4 \;\; P(0|10) = P(0|11) = 0.2$;
- $P(1|s) = 1 - P(0|s) \;\; \forall s \in \mathcal{S}$;
- the partition for this Markov chain is $\mathcal{L} = \{\{00, 01\}, \{10, 11\}\}$ with parts $L_1 = \{00, 01\}$ and $L_2 = \{10, 11\}$;
- the parameters of the Markov chain with partition $\mathcal{L}$ are $P(0|L_1) = 0.4$ and $P(0|L_2) = 0.2$.

Model selection problem

Given a sample generated by a finite memory stationary process, how to choose a partition defining a good Markov model for the source?
We use the BIC criterion

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

where $\mathcal{P}$ is the set of partitions of $\mathcal{S}$

BIC derivation

Let (X_t) be a Markov chain with partition $\mathcal{L}$.
For any $L \in \mathcal{L}$ and $s \in L$, define

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

Denote for $s \in \mathcal{S}$ and $a \in A$,

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

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

Denote for $L \in \mathcal{L}$ and $a \in A$,

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

BIC derivation 2

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

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

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

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

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

Consistence

Theorem

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

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

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

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

- **it is not necessary to look all the set $\mathcal{P}$ in order to find $\mathcal{L}_n$!**

Good partitions of $\mathcal{S}$

Definition

(1) Given and arbitrary partition $\mathcal{L}$ of $\mathcal{S}$. The part $L \in \mathcal{L}$ is a good part if $\forall s, s' \in L$, $P(a|s) = P(a|s')$ $\forall a \in A$.
(2) A partition $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of $\mathcal{S}$ is a good partition of $\mathcal{S}$ if each $L \in \mathcal{L}$ is a good part.

If (1) is true for L we can write

$$P(a|L) = P(a|s) \quad \text{for any } \ s \in L;$$

$$P(L) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s).$$

Theorem

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 $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$ and suppose that exist i and j, $i \neq j$ such that L_i and L_j 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 $\mathcal{L}^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_K\}$, and $L_{ij} = L_i \cup L_j$.

How to find $\arg\max_{\mathcal{L}\in\mathcal{P}}\{BIC(\mathcal{L}, x_1^n)\}$

The idea...
Consider a good partition $\mathcal{L}$

- if there is i, j such that $BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n)$ then $\mathcal{L} = \mathcal{L}^{ij}$, else end.
- (again...)if there is i, j such that $BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n)$ then $\mathcal{L} = \mathcal{L}^{ij}$, else end.
- etc

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

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

Corollary

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

Algorithm

Input: $\mathcal{L}$ ...**Output**: $\hat{\mathcal{L}}_n$.

$i \leftarrow 0,\ j \leftarrow 1$

while $i < K - 1$

$i \leftarrow i + 1$

while $j < K$

$j \leftarrow j + 1$

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

while $d < \frac{(|A|-1)}{2}$

$L_i \leftarrow L_i \cup L_j$

for each l **in** $\{j, ..., K-1\}$ **do** $L_l \leftarrow L_{l+1}$

$K \leftarrow K - 1$

$\mathcal{L} \leftarrow \{L_1, L_2, \ldots, L_K\}$

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

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

Multivariate case

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

- $a^U = (a_{u_1}, ..., a_{u_k})$,
- for any $L \in \mathcal{L}$,

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

- for $s \in \mathcal{S}$

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

- for $L \in \mathcal{L}$

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

Example

- $B = \{0, 1, 2\}$, $l = 2$, $A = B^2 = \{0, 1, 2\}^2$,
- for each $L \in \mathcal{L}$, we need to specify $P(a|L)$ this give us $(|A| - 1) = 8$ parameters,
- if the first coordenate is independent from the second then $P(a|L) = P(a_1|L)P(a_2|L)\ \forall a \in A$ and the numer of parameter will be $(|B| - 1) + (|B| - 1) = 4$,

In general, for $A = B^l$ fixed L and a partition $\mathcal{I}$ of $\{1, 2, ..., l\}$ in independent coordenates, $P(a|L) = \prod_{C \in \mathcal{I}} P(a^C|L)\ \forall a \in A$ and the number of parameters nedded will be $\sum_{C \in \mathcal{I}} (|B|^{|C|} - 1)$

Conditional dependence structure

Definition

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.

BIC derivation

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

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

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

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

Consistence

Theorem

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 $n \to \infty$,

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

For $\mathcal{D} = \{D_L\}_{L \in \mathcal{L}}$ any colection of partitions of $\{1, 2, ...l\}$.
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 colection of partitions containing the same partitions than $\mathcal{D}$ exept D_L is substituted by
$D_L \setminus \{\{U\}, \{V\}\} \cup \{U \cup V\}$

Theorem

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 $n \to \infty$,

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

Applications

Buhlmann P. and Wyner A. (1999). Variable length Markov chains. Ann. Statist. **27** 480–513.

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.

Cuesta, A., Fraiman, R., Galves, A., Garcia, J., Svarc, M. (2007). Identifying rhythmic classes of languages using their sonority: a Kolmogorov-Smirnov approach. *Journal of Applied Statistics* 34: 749-761.

Frota, S., Galves C., Vigário M., Gonzalez-Lopez, V. and Abaurre B. (2012). The phonology of rhythm from Classical to Modern Portuguese. *Journal of Historical Linguistics* 2 (2): 173-207.

Galves, A., Galves, C., Garcia, J., Garcia, N. L. and Leonardi, F. (2012). Context tree selection and linguistic rhythm retrieval from written texts. *Annals of Applied Statistics* 6(1): 186-209.

García, J. and González-López, V. A. (2010). Minimal Markov Models. arXiv:1002.0729v1.

Garcia, J., Gut U., Galves, A., 2002. Vocale - A Semi-Automatic Annotation Tool for Prosodic Research. In Speech Prosody 2002, International Conference.

García, Jesús E., González-López, V. A., Viola, M. L. L. (2012). Robust model selection and the statistical classification of languages. *In American Institute of Physics Conference Series* 1490: 160-170.

García, Jesús E., González-López, V. A., Viola, M. L. L. (2012). Robust model selection for stochastic processes. In fifth Workshop on information theoretic methods in science and engineering (p. 79).

García, Jesús E., González-López, V. A., Viola, M. L. L. (2012). Model selection for multivariate stochastic processes. In fifth Workshop on information theoretic methods in science and engineering (p. 68).

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