Markov Partition Models [1]

Jesús E. García Verónica A. González-López

April 11, 2011

[1] This work is partially supported by CNPq Edital Universal (485999/2007-2), project: "Padrões rítmicos, domínios prosódicos e modelagem probabilística em corpora do português" and CNPq Edital Universal (476501/2009-1), project: "Sistemas estocásticos com interação de alcance variável".

Introduction

- Based on the work of Jorma Rissanen, Buhlmann, Wyner, Csiszár, Talata and others for Variable length Markov chain models.
- We introduce a new class of finite order Markov chain models.
- Address the following model selection problem, given a sample generated by a stationary process, find the Markov model inside this class of models with the minimal number of parameters which is necessary to represent the source as a Markov chain of finite order.

Notation

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

- A the finite alphabet.
- $x_1^n = x_1, x_2, ..., x_n$ 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 class on this equivalence relationship are the subsets of $\mathcal{S}$ with the same transition probabilities.
- The elements of $\mathcal{S}$ on the same equivalence class activate the same random mechanism to choose the next element in the Markov chain.
- The equivalence relation defines a partition $\mathcal{L}$ of $\mathcal{S}$.
- We have one set of $(|A| - 1)$ transition probabilities for each class, obtaining a model with $(|A| - 1)|\mathcal{L}|$ parameters.

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 the equivalence relationship $\sim_p$.

Example

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

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 $\mathcal{S}$
Problem: $\mathcal{P}$ is huge!

Good partitions of $\mathcal{S}$

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

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

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

Definition

A partition $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of $\mathcal{S}$ is a good partition of $\mathcal{S}$ if for each $L \in \mathcal{L}$ and $s, s' \in L$, $P(a|s) = P(a|s') \; \forall a \in A$.

BIC derivation

If $\mathcal{L}$ is a good partition of $\mathcal{S}$,

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

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

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

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

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

Consistence

Theorem

Let $\{X_t, t = 0, 1, 2, \ldots\}$ *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* $\mathcal{S} = \mathcal{A}^{\mathcal{M}}$*. Define,*

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

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

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

Theorem

Theorem

Let $\{X_t, t = 0, 1, 2, \ldots\}$ *be a Markov chain with order* M *over a finite alphabet* A, $\mathcal{S} = A^M$ *the state space. If* $\mathcal{L}$ *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* $n \to \infty$,

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

if, and only if

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

where $\mathcal{L}^{ij}$ *denote the partition* $\mathcal{L}^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_K\}$, *and* $L_{ij} = L_i \cup L_j$.

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

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 $\mathcal{L}_{(1)} = \mathcal{L}^{ij}$, else end.
- if there is i, j such that $BIC(\mathcal{L}_{(1)}^{ij}, x_1^n) > BIC(\mathcal{L}_{(1)}, x_1^n)$ then $\mathcal{L}_{(2)} = \mathcal{L}_{(1)}^{ij}$, 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

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

Model 1

$M = 3$, $A = \{0, 1, 2\}$, with minimal good partition,

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

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

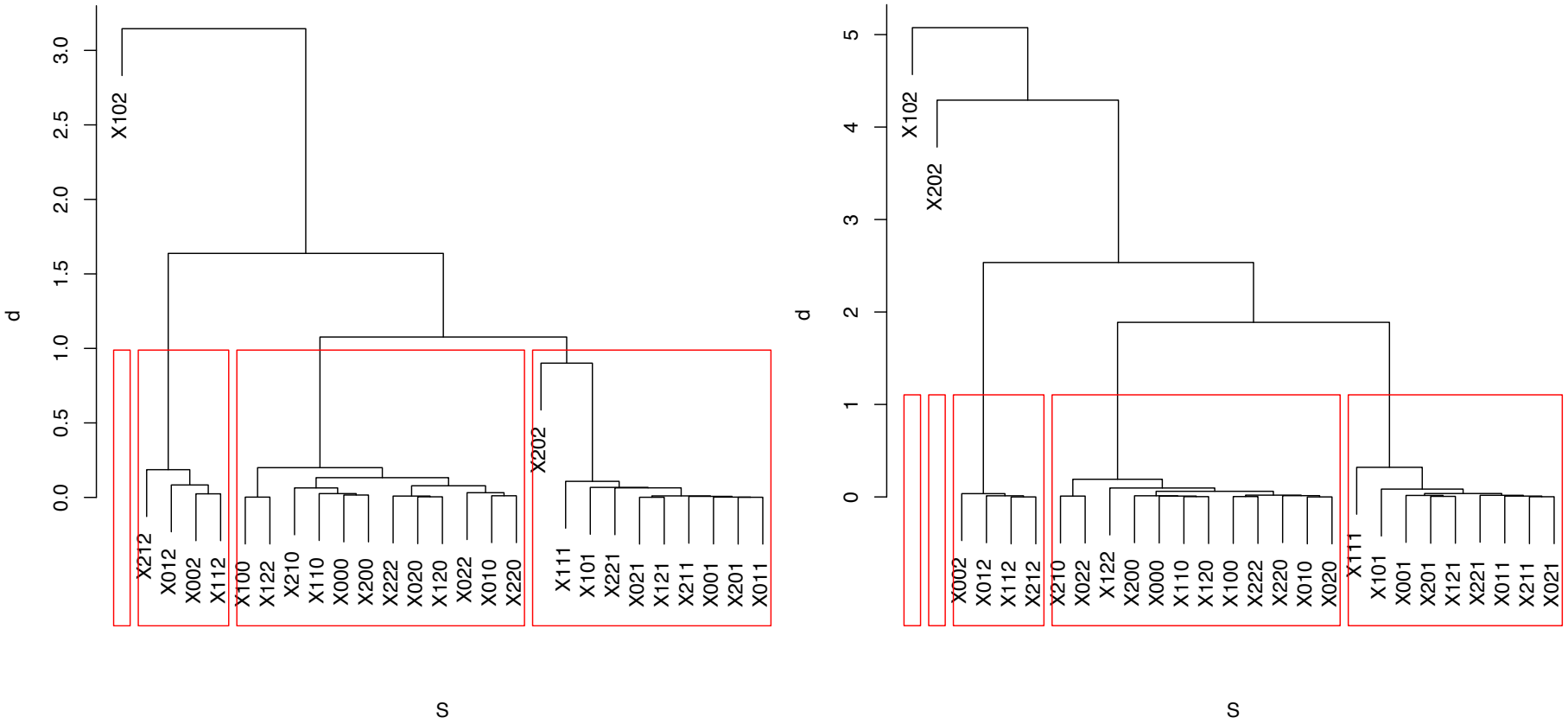

Figure: dendrograms for model 1 estimated for sample sizes of 5000 (left picture) and 9000 (right picture).

For each sample size of $5000, 10000, 15000$ and 20000 we simulated 1000 samples of this Markov chain.

Sample size	Number of correct choices
5000	361
10000	813
15000	908
20000	962

Table: Number of times that our algorithm found the true good partition in 1000 samples, for model 1, when the initial good partition is $S = \{0, 1, 2\}^3$

Model 2

$M = 4$, $A = \{0, 1\}$, with minimal good partition,

$$\begin{aligned} L_1 &= \{0000, 0111\}, \\ L_2 &= \{0001, 0011, 0101, 1001, 1011, 1101\}, \\ L_3 &= \{0010, 0100, 0110, 1010, 1100, 1110\}, \\ L_4 &= \{1000, 1111\}, \end{aligned}$$

and transition probabilities,

$$P(0|L_1) = 0.2, \ P(0|L_2) = 0.8, \ P(0|L_3) = 0.6, \ P(0|L_4) = 0.4$$

For each sample size of $5000, 10000, 15000$ and 20000 we simulated 1000 samples of this Markov chain.

Sample size	Number of correct choices
5000	686
10000	876
15000	926
20000	940

Table: Number of times that our algorithm found the true good partition in 1000 samples, for model 2, when the initial good partition is $S = \{0, 1, 2\}^3$

Good partitions and context trees

Let $\mathcal{T}$ be 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 $s \in \mathcal{T}$, $d(\mathcal{T}) = \max\left(l(s), s \in \mathcal{T}\right)$ where $l(s)$ denote the length of the string s, with $l(\emptyset) = 0$ if the string is the empty string.

Definition

$\mathcal{T}$ is a context tree for the process $\{X_t, t = 0, 1, 2, \ldots\}$ 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

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

Good partitions and context trees

The context tree for a Markov chain with finite depth M, define a good partition on the space $\mathcal{S} = A^M$,

Example

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:

$$\big\{\{000\}, \{100\}, \{010\}, \{110\}\big\}, \big\{\{001\}, \{101\}\big\}, \{011\}, \{111\}$$

where $L_1 = \{\{000\}, \{100\}, \{010\}, \{110\}\}$
$L_2 = \{\{001\}, \{101\}\}$, $L_3 = \{011\}$ and $L_4 = \{111\}$.

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

 Csiszár, I. and Shields, P. C. (2000).The consistency of the BIC Markov order estimator. Ann. Statist. **28** 1601–1619.

 Csiszár, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. IEEE Trans. Inform. Theory **48** 1616–1628.

 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.

 Galves, A., Galves, C., Garcia N. L. and Leonardi F. (2009). Context tree selection and linguistic rhythm retrieval from written texts. arXiv:0902.3619.

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