Minimal Markov Models [1]

García, J.[2] González-López, V.[3]

WITMSE 2011- University of Helsinki
Finland

[1]This work was partially supported by CNPq's projects grant numbers 485999/2007-2 and 476501/2009-1.

[2]IMECC-Unicamp. Brazil
[3]IMECC-Unicamp. Brazil

Introduction

- We introduce a new class $\mathcal{M}$ of finite order Markov chain models.

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 $\mathcal{M}$ with the minimal number of parameters.

Notation

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

- A the finite alphabet;

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;

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;

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 $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.

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 $\mathcal{S}$.

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.

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

- $A = \{0, 1\}$, $M = 2$, $\mathcal{S}(= A^M) = \{00, 01, 10, 11\}$, $P(0|00) = P(0|01) = 0.4$ $P(0|10) = P(0|11) = 0.2$;

Example

- $A = \{0, 1\}$, $M = 2$, $\mathcal{S}(= A^M) = \{00, 01, 10, 11\}$, $P(0|00) = P(0|01) = 0.4$ $P(0|10) = P(0|11) = 0.2$;
- $P(1|s) = 1 - P(0|s)$ $\forall s \in \mathcal{S}$;

Example

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

Example

- $A = \{0, 1\}$, $M = 2$, $\mathcal{S}(= A^M) = \{00, 01, 10, 11\}$, $P(0|00) = P(0|01) = 0.4$ $P(0|10) = P(0|11) = 0.2$;
- $P(1|s) = 1 - P(0|s)$ $\forall s \in \mathcal{S}$;
- 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 $\mathcal{S}$

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 $\mathcal{S}$
- We (now) are thinking in the NML estimator....

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.

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.

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

Starting the algorithm from the good partition defined by the VLMC

We can use any good partition to start the algorithm, in particular we can use the partition defined by a fitted VLMC model.

Example

Consider 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:

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

The procedure will be:

1. select a VLMC model
2. use the context tree corresponding to that model as the initial partition for the algorithm.

We can suspect, that this compound algorithm will perform better when the model chosen by the VLMC procedure is small. We will simulate two Markov models, model 1 will have a small context tree and model 2 will have a full context tree.

Model 1

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

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

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

The VLMC corresponding to this partition is

$$\{0\}, \{1\}, \{12\}, \{102\}, \{202\}, \{22\}, \{002\}.$$

Sample size	correct choices (from S)	correct choices (from $\hat{VLMC}$)
5000	361	835
10000	813	997
15000	908	999
20000	962	1000
25000	976	1000
30000	984	1000

Table: Number of times the MMM algorithm found the minimal good partition in 1000 samples, for model 1, when the initial partitions are $S = \{0, 1, 2\}^3$ and $\hat{VLMC}$

Model 2

It is a Markov chain of order $M = 3$ on the alphabet $A = \{0, 1, 2\}$ with minimal good partition,

$$\begin{aligned} L_1 &= \{000, 001, 002, 010, 011, 012, 020, 021, 220, 022, 221, 222\}, \\ L_2 &= \{100, 101, 102, 110, 111, 112, 120, 121, 122\}, \\ L_3 &= \{210, 211, 212, 200\}, \\ L_4 &= \{201\}, \\ L_5 &= \{202\}, \end{aligned} \tag{1}$$

and transition probabilities,

$$\begin{aligned} P(0|L_1) = 0.2, \quad P(1|L_1) = 0.3, \\ P(0|L_2) = 0.4, \quad P(1|L_2) = 0.3, \\ P(0|L_3) = 0.4, \quad P(1|L_3) = 0.1, \\ P(0|L_4) = 0.1, \quad P(1|L_4) = 0.4, \\ P(0|L_5) = 0.3, \quad P(1|L_5) = 0.5. \end{aligned} \tag{2}$$

The VLMC model corresponding to model 2's partition is the full order 3 Markov chain ($\mathcal{T} = S = \{0,1,2\}^3$).

Sample size	correct choices (from S)	correct choices (from $\hat{VLMC}$)
5000	296	9
10000	801	705
15000	937	933
20000	967	972
25000	974	975
30000	985	982

Table: Number of times that our algorithm found the minimal good partition in 1000 samples, for model 2,when the initial partitions are $S = \{0,1,2\}^3$ and $\hat{VLMC}$

As we can see, even for the extreme model 2, the strategy of selecting first a VLMC model an then a minimal Markov model is not so bad when the sample size is not too small.

Navegation Patterns on a Web Site (MSNBC)

Dataset

- Around one million users sessions recorded in 24 hours on the msnbc web site.
- the web pages in the site are divided in 17 categories:

frontpage	news	tech	local	opinion	on-air
misc	weather	msn-news	health	living	business
msn-sports	sports	summary	bbs	travel	

small sample of the dataset:

frontpage	tech	tech	frontpage	END			
weather	weather	weather	misc	local	weather	weather	weather
on-air	msn-news	msn-news	msn-news	msn-news	misc	msn-news	END
news	END						
msn-sports	sports	msn-sports	END				
frontpage	frontpage	frontpage	END				
news	business	tech	local	business	business	END	
frontpage	END						
local	END						
frontpage	tech	tech	END				
frontpage	frontpage	business	frontpage	END			
sports	sports	sports	sports	sports	sports	END	

Objective, to predict the category of the next page a user will choose.

Supposition, the different sessions are independent.

- We applied our algorithm with $M = 3$
- on the data set there are 3917 different sequences of size 3
- the selected Markov chain has a partition with 229 parts
- there are 73 parts that concentrate 80% of the probability

To predict the next place the user will go, we need to check in which of the 229 parts his/her path fall and use the corresponding probabilities.

- We also applied the CTM algorithm to fit a VLMC with $M = 3$
- the resulting tree has 437 contexts
- there are 97 contexts that concentrate 80% of the probability

Table: A =Alphabet of 20 amino acids + stop codon

Alanine (A)	Arginine (R)
Asparagine (N)	Aspartic acid (D)
Cysteine (C)	Glutamine (Q)
Glutamic acid (E)	Glycine (G)
Histidine (H)	Isoleucine (I)
Leucine (L)	Lysine (K)
Methionine (M)	Phenylanine (F)
Proline (P)	Serine (S)
Threonine (T)	Tryptophan (W)
Tyrosine (Y)	Valine (V)
Stop (4)	

The Data 1

- 3 complete genome sequences of EBV (Epstein Barr virus);

The Data 1

- 3 complete genome sequences of EBV (Epstein Barr virus);
- EBV is an agent associated with infectious mononucleosis, Burkitt lymphoma and nasopharyngeal carcinoma;

The Data 1

- 3 complete genome sequences of EBV (Epstein Barr virus);
- EBV is an agent associated with infectious mononucleosis, Burkitt lymphoma and nasopharyngeal carcinoma;
- i sequence B95-8-type I (57274 amino acids), from a North American case of infectious mononucleosis;
  ii sequence GD1-type I (57219 amino acids), from a Chinese case of nasopharyngeal carcinoma;
  iii sequence AG876-type II (57588 amino acids), from a West African case of Burkitt lymphoma.

Table: Minimal good partition ($M = 1$ and initial good partition=A)

part	B95-8 - type I	GD1 - type I	AG876 - type II
L_1	{4, E, K, M, Q, W}	{4, E, K, M, Q, **T**, W}	{4, E, K, M, Q, **T**, W}
L_2	{A, G, P, **R** }	{A, G, P}	{A, G, P}
L_3	{C, **D**, S, **T**, **V**}	{C, **R**, S}	{C, **D**, **R**, S}
L_4	{F, H, I, L, N, Y}	{**D**, F, H, I, L, N, **V**, Y}	{F, H, I, L, N, **V**, Y}

The differences...

i. B95-8 - type I vs GD1-type I: position of Threonine (T), Valine (V), Arginine (R) and Aspartic acid (D)

ii. B95-8 - type I vs AG876 - type II: position of the Threonine (T), Valine (V) and Arginine (R)

iii. GD1 - type I vs AG876 - type II: position of Aspartic acid (D)

- Data 2: 15 human sequences of Burkitt lymphoma/leukemia with the same diagnosis

Table: Minimal good partition ($M = 1$ and initial good partition=A)

part	Amino acids
L_1	{4, C, K, Q, W }
L_2	{A, D, E, G, M, R}
L_3	{F, I, L, N, V, Y}
L_4	{H, P, S, T}

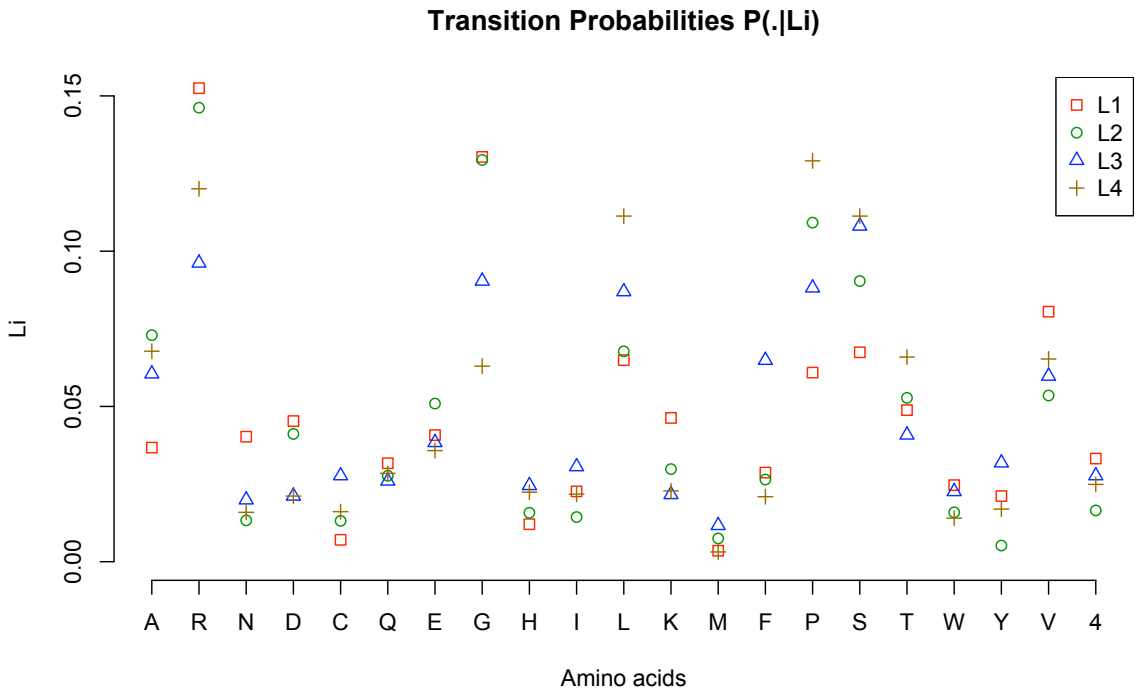

Figure: The figure shows the transition probabilities $\hat{P}(\cdot|L_i), i = 1, 2, 3, 4$.

References


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:0902.3619v2.


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


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

Tank you!
e-mail for contact:
veronica@ime.unicamp.br, jg@ime.unicamp.br

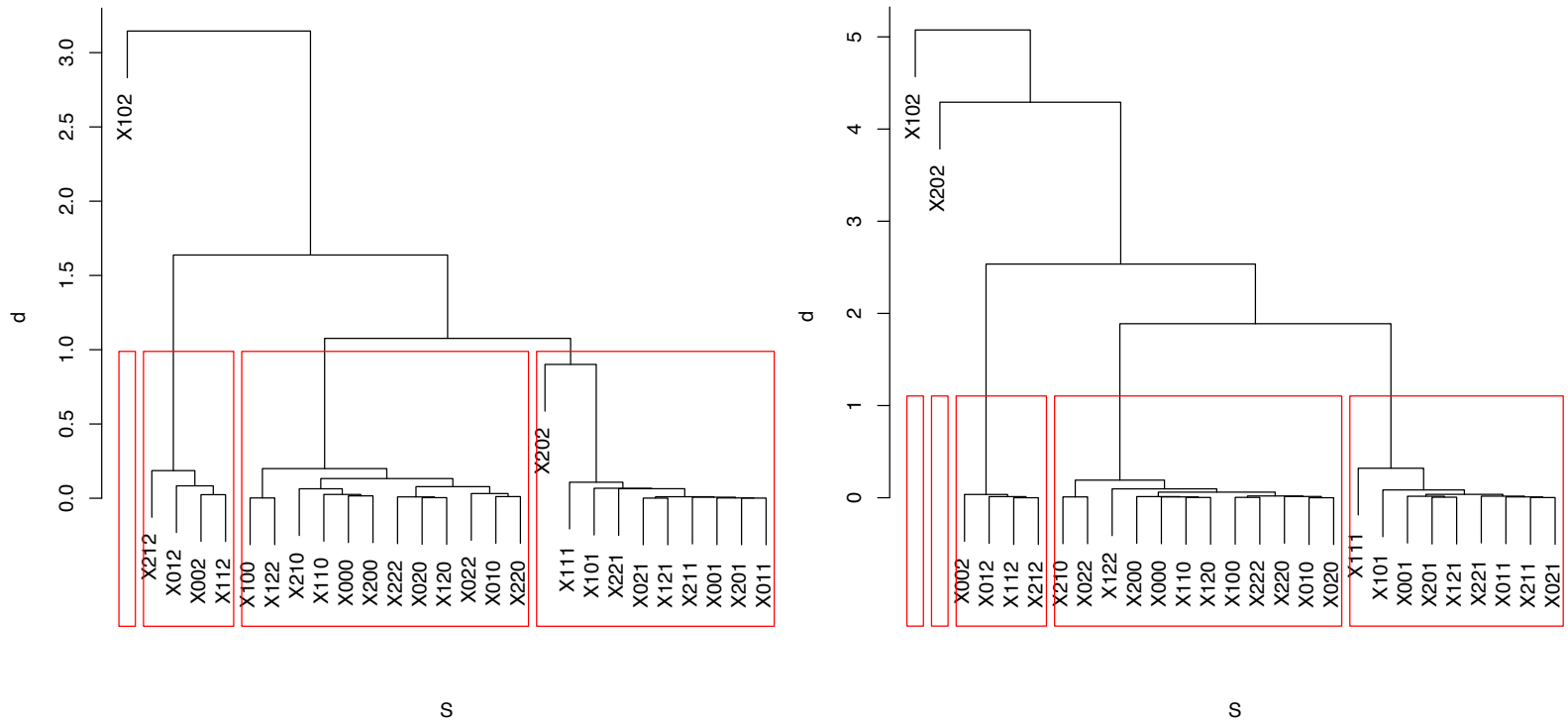

Figure: dendrograms for model 1 estimated for sample sizes of 5000 (left picture) and 9000 (right picture).

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 postfix 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) 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 $L_4 = \{111\}$.