

Optimal Model for a Set of Markov Processes

Jesús E. García and S. L. M. Londoño

September 14, 2018

Introduction

- We develop a model selection procedure for the joint modeling of multiple Markov processes over the same finite alphabet A .
- The methodology is based on the Bayesian Information Criterion (BIC) by Schwarz [5].
- The procedure consists on finding the elements on the state space sharing the same transition probabilities.
- This is done inside each process and between processes.
- We build a partition of the set of state spaces, such that states having the same transition probabilities are in the same part.
- In this family of models the number of parameters is minimized.
- The joint model developed here is a generalization of the partition Markov models by García and González-López [1], [2], [3] and [4].

Notation

- $\mathcal{F} = \{X_t^j\}_{j=1}^p$, a collection of p independent, discrete time, Markov chains on the same finite alphabet A .
- All the processes have the same memory length o .
- $\mathcal{S} = A^o$ is the state space of each Markov chain.
- $a_m a_{m+1} \dots a_n = a_m^n$, where $a_i \in A$, $m \leq i \leq n$.
- $P^j(a|s) = \text{Prob}(X_t^j = a | X_{t-o}^j = s)$, for each $j \in J = \{1, 2, \dots, p\}$, $a \in A$ and $s \in \mathcal{S}$.
- $M = J \times \mathcal{S}$ is the joint state space.

Theoretical Framework

Definition

The elements $(i, s), (j, r) \in M$, are equivalent (denoted by $(i, s) \sim (j, r)$) if $P^i(a|s) = P^j(a|r)$ for all $a \in A$.

Definition

\mathcal{F} has Markov partition $\mathcal{L} = \{L_1, L_2, \dots, L_k\}$ if \mathcal{L} is the partition of M defined by the equivalence relationship \sim .

Remark

Under the assumption of definition 2, we need a set of k transition probabilities to specify the model. We can think that (i, s) and (j, r) are in the same part if they share the same random mechanism to choose the next element in the sequence.



Definition

If \mathcal{F} has Markov partition $\mathcal{L} = \{L_1, L_2, \dots, L_k\}$, for any $L \in \mathcal{L}$, we will denote

$$P_L(a) = P^i(a|s)$$

for any $(i, s) \in L$.

Example

Consider the collection of order two Markov chains over the alphabet $A = \{0, 1\}$ and state space $\mathcal{S} = \{00, 01, 10, 11\}$,

s	$P^1(0 s)$	s	$P^2(0 s)$	s	$P^3(0 s)$
00	.2	00	.3	00	.25
01	.5	01	.5	01	.5
10	.2	10	.1	10	.15
11	.1	11	.1	11	.1

The number of parameters for the three models is twelve. While, we can observe that

- $P^1(.|01) = P^2(.|01) = P^3(.|01)$ and there is not other transition probability with the same value, this mean that the Markov partition contains the part $L_1 = \{(1, 01), (2, 01), (3, 01)\}$,
- $P^1(.|00) = P^1(.|10)$ and there is not other transition probability with the same value, this mean that the Markov partition contains the part $L_2 = \{(1, 00), (1, 10)\}$,
- $P^1(.|11) = P^2(.|10) = P^2(.|11) = P^3(.|11)$ and there is not other transition probability with the same value, we obtain the part $L_3 = \{(1, 11), (2, 10), (2, 11), (3, 11)\}$,
- finally, the rest of the transition probabilities do not have repetitions an they are unitary parts.

The joint model has six parameters with the following Markov partition and associated transition probabilities,

$L \in \mathcal{L}$	$P_L(0)$
$\{(1, 01), (2, 01), (3, 01)\}$.5
$\{(1, 00), (1, 10)\}$.2
$\{(1, 11), (2, 10), (2, 11), (3, 11)\}$.1
$\{(2, 00)\}$.3
$\{(3, 00)\}$.25
$\{(3, 10)\}$.15

Model selection procedure

Notation

- Let $\{x_1^{j n_j}\}_{j=1}^p$ be samples of the processes $\{X_t^j\}_{j=1}^p$,
- with sample sizes $\{n_j\}_{j=1}^p$ and $n_m = \min\{n_1, \dots, n_p\}$.
- $N(j, s)$ is the number of occurrences of $s \in \mathcal{S}$ in the sample $x_1^{j n_j}$,
- $N((j, s), a)$ is the number of occurrences of s followed by a in the sample $x_1^{j n_j}$,
- $N(L) = \sum_{(i,s) \in L} N(i, s)$, $L \in \mathcal{L}$ is the number of occurrences of elements in L and
- $N(L, a) = \sum_{(i,s) \in L} N((i, s), a)$ is the number of occurrences of elements in L followed by $a \in A$.

Likelihood of the sample

$$\left\{x_1^{jn_j}\right\}_{j=1}^p = \left\{\{x_1^1\}_1^{n_1}, \{x_1^2\}_1^{n_2}, \dots, \{x_1^p\}_1^{n_p}\right\}$$

is a sample of \mathcal{F}

Suppose that $\mathcal{L} = \{L_1, L_2, \dots, L_k\}$ is the Markov partition of \mathcal{F} .

The likelihood of the sample is

$$\text{Prob}\left(\left\{x_1^{jn_j}\right\}_{j=1}^p\right) = \prod_{j=1}^p \text{Prob}^j(x_1^{jn_j}) \quad (1)$$

$$= \prod_{j=1}^p \text{Prob}^j(x_1^{j^o}) \prod_{L \in \mathcal{L}} \prod_{a \in A} P_L(a)^{N(L,a)}. \quad (2)$$

The pseudo log-likelihood can be written as

$$\sum_{L \in \mathcal{L}} \sum_{a \in A} N(L, a) \ln(P_L(a)),$$

the maximum is

$$\sum_{L \in \mathcal{L}} \sum_{a \in A} N(L, a) \ln \left(\frac{N(L, a)}{N(L)} \right).$$

The estimator, based on the BIC, for the partition \mathcal{L} of M is

$$\hat{\mathcal{L}}_n = \operatorname{argmax}_{\mathcal{L}} BIC \left(\mathcal{L}, \left\{ x_1^{j n_j} \right\}_{j=1}^p \right),$$

where

$$BIC \left(\mathcal{L}, \left\{ x_1^{j n_j} \right\}_{j=1}^p \right) = \sum_{L \in \mathcal{L}} \sum_{a \in A} N(L, a) \ln \left(\frac{N(L, a)}{N(L)} \right) - \frac{|A| - 1}{2} |\mathcal{L}| \ln \left(\sum_{j=1}^p n_j \right).$$

Problem

$$\hat{\mathcal{L}}_n = \operatorname{argmax}_{\mathcal{L}} \operatorname{BIC} \left(\mathcal{L}, \left\{ x_1^{j n_j} \right\}_{j=1}^P \right),$$

- The set of partitions for M is too large,
- the maximization step is not practical,
- we build the optimal partition locally, using the next divergence notion.

Divergence

Definition

For $(i, s), (j, r) \in M$, set

$$\begin{aligned}d((i, s), (j, r)) &= N((i, s), a) \ln \left(\frac{N((i, s), a)}{N(i, s)} \right) \\ &+ N((j, r), a) \ln \left(\frac{N((j, r), a)}{N(j, r)} \right) \\ &- N(\{(i, s), (j, r)\}, a) \ln \left(\frac{N(\{(i, s), (j, r)\}, a)}{N(\{(i, s), (j, r)\})} \right),\end{aligned}$$

where $N(\{(i, s), (j, r)\}) = N(\{(i, s)\}) + N(\{(j, r)\})$ and
 $N(\{(i, s), (j, r)\}, a) = N(\{(i, s)\}, a) + N(\{(j, r)\}, a)$.

This divergence is related to the BIC criterion in the following way, the BIC criterion indicates that $(i, s), (j, r) \in M$ should be in the same part if and only if $d((i, s), (j, r)) < \frac{|A|-1}{2} \ln \left(\sum_{j=1}^p n_j \right)$.

Remark

The divergence is statistically consistent,

$$\lim_{n_m \rightarrow \infty} d((i, s), (j, r)) = 0 \iff P^i(\cdot|s) = P^j(\cdot|r),$$

$$\lim_{n_m \rightarrow \infty} d((i, s), (j, r)) = \infty \iff P^i(\cdot|s) \neq P^j(\cdot|r).$$

- d is symmetric and nonnegative.
- From the Remark, we can apply any clustering algorithm, using d as a dissimilarity measure,
- See for example the algorithm introduced in García & González-López [1].

Remark






Equivalent to d , we can compute

$$D((i, s), (j, r)) = \frac{\left(\frac{N((i, s), a)}{N(i, s)}\right)^{N((i, s), a)} \left(\frac{N((j, r), a)}{N(j, r)}\right)^{N((j, r), a)}}{\left(\frac{N(\{(i, s), (j, r)\}, a)}{N(\{(i, s), (j, r)\})}\right)^{N(\{(i, s), (j, r)\}, a)}}.$$

- To obtain D , $p|S|(p|S| + 1)/2$ elements of the matrix must be computed.
- Considering that the number of parameters have to be smaller than the size of the dataset,
- the maximum order for the processes should be smaller than $\log(n_m)$ and
- we should be able to calculate d with order n_m^2 float point operations.

Concluding Remarks

- As the model selection procedure is based on the BIC criterion, a balance between complexity and significance is natural to this kind of model selection methodology.
- This means, the procedure will have a tendency to produce more complex models (larger Markov partitions) for a larger data set, until the right model is attained.
- The traditional constant $\frac{1}{2}$ appearing on the BIC criterion, see Schwarz [5], is arbitrary and the criterion will still be consistent if changed for any positive and finite constant.

-  Jesús E. García and V. A. González-López, Minimal Markov Models In *Fourth Workshop on Information Theoretic Methods in Sciences and Engineering*, p. 25, 2011.
-  J. E. García and V. A. González-López, Minimal Markov Models. *arXiv preprint arXiv:1002.0729*, 2010.
-  Jesús E. García and V. A. González-López, Consistent Estimation of Partition Markov Models. *Entropy* **19** (4), 180. 2017.
-  Jesús E. García and V. A. González-López, Detecting regime changes in Markov models. *SMTDA2014 Book* .
-  G. Schwarz, Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461-464, 1978.