Interaction partition. Structure of interacting coordinates for a multivariate stochastic process

Jesús E. García[1] and V. A. González-López[2]

[1] University of Campinas, Brazil
(e-mail: jg@ime.unicamp.br)
[2] University of Campinas, Brazil
(e-mail: veronica@ime.unicamp.br)

Abstract. In this paper we show a methodology to infere the dependence structure between the coordinates of a k-variate Markovian source. The methodology is based on the Bayesian information criterion (BIC). It is consistent in the sense that if the source is Markovian and the dataset is large enough, the true partition will be retrieved. Consider a set of k sources. For each $t \in \{0, 1, 2, ...\}$ each source produces a letter in the alphabet $A = \{0, 1\}$. The sources interact between them depending on the past states of the set of k sources. Our methodology, obtains a partition of the past such that two possible pasts are in the same part of the partition if and only if, the set of interacting coordinates given any of this two pasts, is the same. We also obtain, for each possible past, the set of sources which interact between them. We can imagine a very simplified model of interacting neurons. Consider k neurons. Discretize time in intervals of a fixed size. For each time interval we will say that the value of a particular neuron is 1 if there was at least one spike from that neuron in that interval of time. In this setting, two neurons interact mean that, when one of them fire, the probabilities of the other one to fire, changes. This interaction can be in any of two kinds, it can increase or decrease the probability of firing for the other neuron.
Keywords: Multivariate Markov chain, Dependence structure, Partition Markov model.

1 Introduction

Estimation in multiple interacting processes is a difficult task. Even if the multivariate process is Markovian. The number of parameters grow exponentialy not only with the dimmension of the alphabet but also with the length of the memory. To mitigate this problem we will use the family of partition Markov models (PMM) (see [3] and [4]) which are a generalization of the variable length Markov chain models (VLMC) family (see [5], [7], [1] and [2]). The PMM family is more economic that the VLMC family in the number of paremeters required to describe a given Markov process.

Note that the interaction between the neurons does not necessarily have the same dependence from the past that the joint probability of the whole set of k neurons. Also the structures of dependence from the past for the marginal probabilities are not necesarily the same as the dependence from the past for the interaction between them. It is possible to have different marginals distributions with the same set of neurons interacting. The family of partition Markov models

2 Notation

Let X_t be the state of the set of k sources at time t. $X_t = (X(1)_t, ..., X(k)_t)$, where $X(i)_t$ is the state of the source number i at time t. $X(i)_t \in \{0,1\}$ and $X_t \in A = \{0,1\}^k$. We will assume that X_t is an order M Markov chain, $M < \infty$. Denote the string $a_m a_{m+1} \ldots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$. x_1^n will be a size n realization of X_t. For each $s \in S = A^M$, $a \in A$, $b \in \{0,1\}$ and $1 \leq i \leq k$.

$$N(s) = \sum_{i=M+1}^{n-1} 1_{\{x_{i-M}^{i-1}=s\}},$$

$$N(s,a) = \sum_{i=M+1}^{n-1} 1_{\{x_{i-M}^{i-1}=s, x_i=a\}},$$

we denote the conditional joint probability of the process by,

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

and the conditional marginal probability of the source i by,

$$P_i(b|s) = \text{Prob}(X(i)_t = b | X_{t-M}^{t-1} = s).$$

3 Equivalences

Definition 1. (Equivalence relationship based on the joint distribution) For each $s, r \in S$, $s \sim r$ if $P(a|s) = P(a|r)$ $\forall a \in A$.

Definition 2. (Equivalence relationship based on the marginal distributions) For each $i \in \{1, 2, ...k\}$ and $s, r \in S$,
$s \sim_i r$ if $P_i(b|s) = P_i(b|r)\ \forall b \in \{0, 1\}$.

Remark 1 *For each $s, r \in S$, $s \sim r \Rightarrow s \sim_i r\ \forall i \in \{1, 2, ...k\}$.*

Proposition 1. *If all the sources are independent all the time then, for each $s, r \in S$,*

$$s \sim r \iff s \sim_i r \ \ \forall i \in \{1, 2, ...k\},$$

Example 1. Bi-variate case, $k = 2$, $A = \{0, 1\}^2$ and $S = A^M$. In this case, for any $s, r \in S, s \sim r$ if and only if $P_1(0|s) = P_1(0|r), P_2(0|s) = P_2(0|r)$ and $P((0, 0)|s) = P((0, 0)|r)$. Given $s \in S$, the two neurons interact if and only if $P_1(0|s) \neq P((0, 0)|s)/P_2(0|s)$ that is, if and only if $1 \neq \frac{P((0,0)|s)}{P_1(0|s)P_2(0|s)}$. If the two neurons are always independent then $P((0, 0)|s) = P_1(0|s)P_2(0|s)$ and we have that if $s \sim_1 r$ and $s \sim_2 r$ then $s \sim r$. Suppose $M = 2$ and the following set of conditional probabilities:

s	$P_1(0\|s)$	$P_2(0\|s)$	$P((0,0)\|s)$	$\sim$	$\mathcal{I}$	$\mathcal{F}$
$(0,0),(0,0)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,0),(0,1)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,1),(0,0)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,1),(0,1)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,0),(1,0)$	0.1	0.1	0.02	L_2	M_2	F_2
$(0,0),(1,1)$	0.1	0.1	0.02	L_2	M_2	F_2
$(0,1),(1,0)$	0.1	0.1	0.02	L_2	M_2	F_2
$(0,1),(1,1)$	0.1	0.1	0.02	L_2	M_2	F_1
$(1,0),(0,0)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,0),(0,1)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,1),(0,0)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,1),(0,1)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,0),(1,0)$	0.2	0.2	0.02	L_4	M_2	F_3
$(1,0),(1,1)$	0.2	0.2	0.02	L_4	M_2	F_3
$(1,1),(1,0)$	0.2	0.2	0.02	L_4	M_2	F_3
$(1,1),(1,1)$	0.2	0.2	0.02	L_4	M_2	F_3

In this example, for each marginal, $\sim_i$ have two classes. The partition $\mathcal{L}$ corresponding to $\sim$ have four parts, the fifth column of the table indicates the part to which each $s \in S$ belongs. In the parts L_2 and L_4 the two sources are interacting while in the parts L_1 and L_3 they are independent. The partition $\mathcal{M} = \{L_1 \cup L_3, L_2 \cup L_4\}$ indicates when the two neurons interact. Note that the partition $\mathcal{M}$ indicate when the neurons interact but not how. The partition $\mathcal{F} = \{L_1 \cup L_3, L_2, L_4\}$ indicates exactly when the two neurons interact and wich kind of interaction they have.

In the next section we will see how to estimate the partition $\mathcal{L}$ corresponding to $\sim$ in a way that is consistent (for more detail see [3]).

4 Partition Markov Models

In this section we give a summarized introduction to the PMM, see [3] and [4] for a more complete explanation. Let (X_t) be a discrete time, Markov chain with memory M on a finite alphabet A, with state space $S = A^M$.

Definition 3. 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 $\sim$ introduced by definition 1.

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. The total number of parameters for the model is $|\mathcal{L}|(|A| - 1)$. To choose a model in the family in a consistent way (see [3]), we can use the following distance in $S_{/\sim_n}$, where $s \sim_n r \iff \frac{N(s,a)}{N(s)} = \frac{N(r,a)}{N(r)} \ \forall a \in A$, n is the size of the dataset.

Definition 4. Let n be the size of the dataset. For any $s, r \in S$, $N(\{s,r\}, a) = N(s,a) + N(r,a)$,

$$\begin{aligned} d_n(s,r) = \frac{2}{(|A|-1)\ln(n)} \sum_{a \in A} \Bigg\{ & N(s,a) \ln\left(\frac{N(s,a)}{N(s)}\right) \\ & + N(r,a) \ln\left(\frac{N(r,a)}{N(r)}\right) \\ & - (N(\{s,r\},a) \ln\left(\frac{N(\{s,r\},a)}{N(s)+N(r)}\right)\Bigg\}, \end{aligned}$$

d_n can be generalized to sub sets of S and it have the very nice property of being equivalent to the BIC criterion to decide if $s \sim r$ for any $s, r \in S$ (see [3]).

Theorem 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 $s, r \in S$, $d(r, s) < 1$ iff s and r belong to the same class.*

Algorithm 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 $d(s, r) < 1$
      $L_s = L_s \cup \{r\}$
      $B = B \setminus \{r\}$
  $\hat{\mathcal{L}}_n = \hat{\mathcal{L}}_n \cup \{L_s\}$

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 true partition for the source.

Corollary 1. *Under the assumptions of Theorem 3, $\hat{\mathcal{L}}_n$, given by the algorithm 1 converges almost surely eventually to $\mathcal{L}^*$, where $\mathcal{L}^*$ is the partition of $\mathcal{S}$ defined by the equivalence relationship.*

5 Estimation of the Interaction structure on a multivariate PMM

For simplicity, we will assume that a PMM has been already estimated and we have a partition $\mathcal{L}$ corresponding to $\sim$. Our objective is to obtain for each part of the estimated partition a new partition of the set of coordinates on independent sets. After that, we will put together all the parts of $\mathcal{L}$ with the same partition of the coordinate space.

Let (X_t) be a Markov chain on $A = \{0,1\}^k$, **with partition** $\mathcal{L}$. For $U = \{u_1, ...u_l\} \subset \{1, 2, ..., k\}$ and $a = (a_1, ..., a_k) \in A$, define:

i) $a^u = (a_{u_1}, ..., a_{u_l})$,
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 $s \in \mathcal{S}$

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

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

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

v) for $i \in \{1, 2, ..., k\}$, $b \in \{0, 1\}$ and $c \in \{0, 1\}^l$

$$N_n(s, i(b)) = \sum_{t=M}^{n} 1_{\{x_{t-M}^{t-1} = s, x_t^i = b\}},$$
$$N_n(s, u(c)) = \sum_{t=M}^{n} 1_{\{x_{t-M}^{t-1} = s, x_t^{u_j} = c_j,\ 1 \leq j \leq l\}}.$$

In general, for $A = \{0,1\}^k$, fix $L \in \mathcal{L}$ and a partition $\mathcal{I}_L$ of $\{1, 2, ..., k\}$ 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

$$\sum_{C \in \mathcal{I}_L} (2^{|C|} - 1).$$

5.1 Dependence structure

Viola in [6] defines the dependence structure between the coordenates for Context Tree models and shows that this dependence structure can be estimated using the Bayesian information criterion (BIC) in the following way. First is fitted a Context Tree Model and then, for each context, the BIC criterion is used on the transition probabilities corresponding to that context to find a partition of the coordinates on dependent sets. The results in [6] are valid for any family of Markovian models as they only depend on the individual transition probabilities and not on the model structure.

In this paper we will simultaneously estimate the partition of our PMM and the interaction structure using the BIC criterion.

Definition 5. Let (X_t) For each $L \in \mathcal{L}$, define $\mathcal{D}_L$ as de biggest partition of $\{1, 2, ...k\}$ such that

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

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

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

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

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

and the BIC criterion for ou class of models,

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

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

Theorem 2 *Let (X_t) be a Markov chain of order M over a finite alphabet A, with partition $\mathcal{L}^*$ and structure of conditional dependence $\mathcal{D}^*_{\mathcal{L}}$. Define,*

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

Where $\mathbf{D}$ *is the set of all possible structures of dependences for* A *and* $\mathcal{L}_n$, $\mathcal{L}_n$ *obtained using algorithm 1, then, eventually almost surely as* $n \to \infty$,

$$\mathcal{D}^*_{\mathcal{L}} = \mathcal{D}_{\mathcal{L}n}.$$

6 Simultaneous estimation of the partition and the interaction structure

We will introduce the following measure of dependence between pairs of coordinates conditioned to a past $t \in S$,

Definition 6. For any $t \in S$, and $i, j \in \{1, 2, ...k\}$

$$d_t^n(i,j) = \frac{2}{\ln(n)} \sum_{b \in B} \left\{ N(t, i(b)) \ln \left(\frac{N(t, i(b))}{N(t)} \right) + N(t, j(b)) \ln \left(\frac{N(t, j(b))}{N(t)} \right) \right\} - \sum_{c \in B^2} \left\{ (N(t, \{i,j\}(c))) \ln \left(\frac{N(t, \{i,j\}(c))}{N(t)} \right) \right\},$$

This measure of dependence can be also used with parts of the partition defining a PMM, substituing t by L.

The next Theorem shows that this distance between coordinates can be used to find the structure of interactions for a given past $t \in S$ in a consistent way.

Theorem 3 *(Consistence in the case of a Markov source) For* n *large enough, for* $t \in S$, *and* $i, j \in \{1, 2, ...k\}$, $d_t^n(i,j) < 1$ *iff* i *and* j *are dependent.*

Using the distances in definition 4 and definition 6, we can define the following algorithm to estimate $\mathcal{D}_{\mathcal{L}}$.

Algorithm 2 *(Coordinate partition selection algorithm)*
Input*: for a fixed* $t \in S$, $d_t(i,j) \forall 1 \le i, j \le k$.
Output*:* $\hat{D}_t^n$;
$B = \{1, 2, ..., k\}$
$\hat{D}_t^n = \emptyset$
while $B \neq \emptyset$

select $i \in B$
define $D_i = \{i\}$
$B = B \setminus \{i\}$
for each $j \in B, j \neq i$
  if $d_t^n(i,j) < 1$
    $D_i = D_i \cup \{j\}$
    $B = B \setminus \{j\}$
$\hat{D}_t^n = \hat{D}_t^n \cup \{D_i\}$

Return: $\hat{D}_t^n$

The following modification of the distance 4, which will account with the possible, change on the degree of freedom caused by the dependence structure, will be used to find a PMM and dependence structure for a multivariate dataset.

Definition 7. Let n be the size of the dataset. For any $s, r \in S$,

$$\begin{aligned} &d'_n(s,r) = \\ &= \frac{2}{M(s,r)\ln(n)} \sum_{a \in A} \left\{ \sum_{C \in \hat{D}_s^n} N(s,a) \ln\left(\frac{N(s,a^C)}{N(s)}\right) \right. \\ &+ \sum_{C \in \hat{D}_r^n} N(r,a) \ln\left(\frac{N(r,a^C)}{N(r)}\right) \\ &- \left. \sum_{C \in \hat{D}_{\{s,r\}}^n} (N(\{s,r\},a)) \ln\left(\frac{N(s,a^C) + N(r,a^C)}{N(s) + N(r)}\right) \right\}, \end{aligned}$$

where $M(s,r) = \sum_{C \in \hat{D}_s^n}(|B|^{|D|} - 1) + \sum_{C \in \hat{D}_r^n}(|B|^{|D|} - 1) - \sum_{C \in \hat{D}_{r,t}^n}(|B|^{|D|} - 1)$.

Now, substituting the input $d_n(s,r)$ by $d'_n(s,r) \forall s, r \in S$ in algorithm 1. We will obtain simultaneously the PMM and dependence structure estimators, where the estimator of the dependence structure for part $\hat{L}$ will be the partition obtained using $d_{\hat{L}}(i,j),\ i,j \in \{1,2,...k\}$.

6.1 Interaction structure of the coordinates

Consider now

$$I = \cup_{\{L\in\mathcal{L}\}}\mathcal{D}_L.$$

I will contain each kind of partition of the coordinates appearing in $\mathcal{D}_{\mathcal{L}}$. Now we want to put together all the parts of $\mathcal{L}$ with the same partitioning of the coordinates. For each $K \in I$, let be

$$M_K = \cup_{\{L\in\mathcal{L}:\mathcal{D}_L=K\}}L \quad \text{and} \quad \mathcal{M} = \{M_K\}_{K\in I}.$$

$\mathcal{M}$ is a partition of S such that two sequences $s, r \in S$ are in the same part of $\mathcal{M}$ if and only if, given each of this two pasts, the set of sources interacting is the same. The partition $\mathcal{M}$ tell us for each possible past, which of the different sources interact.

7 Conclusion

Fix $M_K \in \mathcal{M}$. Conditioned to the part M_K, each marginal $X^C,\ C \in K$, being independent from the others coordinates, can be analyzed by itself, which in general require less data than the simultaneous analysis of all k coordinates. The same for $\mathcal{D}_{\mathcal{L}}$, once it is estimated, we can identify the specific kind of iteration using standard statistical methods on each interacting set. For example fixed a part $L \in \mathcal{L}$ if $\mathcal{D}_L = \{C_1, ...C_{m_L}\}$, we only need to work with the marginals $X^{C_i},\ i \in \{1, ..., m_L\}$ to determine the kind of dependence between the coordinates of A.

8 ACKNOWLEDGMENTS

The authors gratefully acknowledge the support for this research provided by USP project “Mathematics, computation, language and the brain” and FAPESP’s projects (a)“Portuguese in time and space: linguistic contact, grammars in competition and parametric change” 2012/06078-9 and (b)“Research, Innovation and Dissemination Center for Neuromathematics - NeuroMat” 2013/07699-0.

References

1. Csiszár, I. and Talata, Z., "Context tree estimation for not necessarily finite memory processes, via BIC and MDL", *IEEE Trans. Inform. Theory* **52**, 1007–1016, 2006.
2. Galves, A., Galves, C., Garcia, J. E., Garcia, N. L. and Leonardi, F., "Context tree selection and linguistic rhythm retrieval from written texts", *Annals of Applied Statistics*, **6** 1, 186 – 209, 2012.
3. Garcia, J. and Gonzalez-Lopez, V. A., "Minimal Markov Models". arXiv preprint *arXiv:1002.0729*, 2010.
4. Garcia, J. E., Gonzalez-Lopez, V. A. and Viola, M. L. L., "Robust model selection and the statistical classification of languages", *AIP Conference Proceedings*, vol. 1490, p.160, 2012.
5. Rissanen J., "A universal data compression system", *IEEE Trans. Inform. Theory* **29**(5), 656 – 664, 1983.
6. Viola, M., L. (2011). Tópicos em seleção de modelos markovianos. PhD Tesis. `http://www.bibliotecadigital.unicamp.br/document/?code=000844289`
7. Weinberger, M., Rissanen, J. and Feder, M. (1995). A universal finite memory source, IEEE Trans. Inform. Theory **41**(3) 643 – 652.