Combining multivariate Markov chains

Jesús E. García

In this paper we address the problem of modelling multivariate finite order Markov chains, when the dataset is not large enough to apply the usual methodology.

The curse of dimensionality in model selection for multivariate stochastic processes

- Consider the alphabet $B = \{1, 2, ..., |B|\}$
- For an **order 1 Markov Chain** on B the set of parameters is the set of transition probabilities, $\{p(s, b) : b, s \in B\}$ that is $|B|^2$ parameters, considering that $1 = \sum_b p(s, b)$, the number of parameters to estimate is $|B|(|B| - 1)$.
- For an **order** o **Markov chain** over the alphabet B, the set of transition probabilities is $\{p(s, b) : \ s \in B^o, \ b \in B\}$ corresponding to $|B|^o(|B| - 1)$ parameters.
- For an **order** o k**-variate Markov chain** over the alphabet $A = B^k$, the set of transition probabilities is $\{p(s, a) : \ s \in \{B^k\}^o, \ a \in B^k\}$ corresponding to $|B|^{ok}(|B|^k - 1)$ parameters.

The curse of dimensionality in model selection for multivariate stochastic processes

- For an **order** o k**-variate Markov chain** over the alphabet B^k, we need to fit $|B|^{ok}(|B|^k - 1)$ parameters
- The **number of parameters** needed for a multivariate Markov chain **grows exponentially with the process order and the dimension of the chain's alphabet**.
- **The size of the dataset needed to fit multivariate Markov** chain grows exponentially with the process order and the dimension of the chain's alphabet.
- Given a dataset, **the set of multivariate Markov chain models from where to choose a model for the dataset is limited by the curse of dimensionality**.

The curse of dimensionality in model selection for multivariate stochastic processes

- In general during model selection, we can't reduce the dimension of the alphabet.
- When the data set is not large enough for the number of parameters of the "true" model, the order of the fitted model will be smaller than the order of the "true" model from which the sample was produced.
- In this work we introduce a new strategy to estimate a model for a multivariate process. That allows the estimation of a model with greater order than the standard procedure.
- The family of models used for the model selection procedure is the family of *partition Markov models*.

Partition Markov models

Notation

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

- A the finite alphabet;
- $\mathcal{S} = A^o$ the state space;
- $P(a|s) = \text{Prob}(X_t = a | X_{t-o}^{t-1} = s),\ a \in A,\ s \in \mathcal{S}$ 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 chain with partition $\mathcal{L}$

Definition

let (X_t) be a discrete time, order o 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

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

Notation

Let x_1^n be a sample of the process $\left(X_t\right)$, $s \in \mathcal{S}$, $a \in A$ and $n > o$.

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

$$N_n(s) = \left|\left\{t : o < t \leq n, x_{t-o}^{t-1} = s\right\}\right|. \tag{2}$$

To simplify the notation we will omit the $_n$ on N_n.

A distance in S

Definition

We define the distance d in S,

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

for any $s, r \in S$.

A distance in S

Theorem

For any $s, r, t \in S$,

i. $d(r, s) \geq 0$ *with equality if and only if* $\frac{N(s,a)}{N(s)} = \frac{N(r,a)}{N(r)} \ \ \forall a \in A$,

ii. $d(r, s) = d(s, r)$,

iii. $d(r, t) \leq d(r, s) + d(s, t)$.

Remark

d can be generalized to subsets (see García, J. and González-López, V. A. (2010)).

Consistence in the case of a Markov source

Theorem

Let (X_t) be a discrete time, order o Markov chain on a finite alphabet A. Let x_1^n be a sample of the process and $0 < \alpha < \infty$, then for n large enough and for each $s, r \in S$, $d_n(r, s) < \alpha$ iff s and r belong to the same class.

Algorithm

Input: $d(s,r) \forall s \neq r \in S$; **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 $d(s,r) < \alpha$

$L_s = L_s \cup \{r\}$

$B = B \setminus \{r\}$

$\hat{\mathcal{L}}_n = \hat{\mathcal{L}}_n \cup \{L_s\}$

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

Notation for the k-variate process

Let X_t be the state of the k dimensional Markov process at time t,

- $A = B^k$.
- $X_t = (X(1)_t, ..., X(k)_t)$ will be the value of the k-dimensional source at time t.
- $X_t \in A$,
- $X(i)_t \in B$ is the value of the coordinate number i at time t.

k-variate copula

- $B = \{1, 2, ..., |B|\}$ a finite set.
- For $1 \leq i \leq k$, $X(i)$ is a random variable taking values on the set B.
- $(X(1), ..., X(k))$ is a random vector with joint probability function
  $p(b_1, ..., b_k) = P(X(1) = b_1, ..., X(k) = b_k),\ b_1, ...b_k \in B.$
- For $1 \leq i \leq k$, $X(i)$ have probability $p_i(b) = P(X(i) = b)$ and cumulative distribution $F_i(b) = P(X(i) \leq b)$,
- For $u = (u_1, ..., u_k)$, with $0 \leq u_i \leq 1$, for $i = 1, 2, ..., |B|$.
  The k-variate copula density is given by
  $c(u_1, \ldots, u_K) = \frac{p(b_1,...,b_k)}{p_1(b_1)...p_k(b_k)}$.
  where b_i is choosen such that $F(b_i - 1) \leq u_i \leq F(b_i)$ for $1 \leq i \leq k$. And $F(0) = 0$.

k-variate copula

The function $c(u_1, \ldots, u_K)$ satisfies the following characteristics

- It is a probability mass function, displayed in $[0,1]^K$.
- the univariate marginal distributions are $U(0,1)$ and
- the cumulative distribution C of c verifies
  $\text{Prob}(X(1) \leq b_1, \ldots, X(k) \leq b_k) =$
  $C\big(F_1(b_1), \ldots, F_k(b_k)\big), \quad$ for all $(b_1, \ldots, b_k) \in B^k$.

Example: 2-variate copula

Consider $k = 2,\ B = \{1, 2\}$, the copula density is

$$c(u,v) = \begin{cases} \frac{p(1,1)}{p_1(1)p_2(1)}, & \text{if} \quad (u,v) \in [0, F_1(1)[\times [0, F_2(1)[\\ \frac{p(1,2)}{p_1(1)p_2(2)}, & \text{if} \quad (u,v) \in [0, F_1(1)[\times [F_2(1), 1] \\ \frac{p(2,1)}{p_1(2)p_2(1)}, & \text{if} \quad (u,v) \in [F_1(1), 1] \times [0, F_2(1)[\\ \frac{p(2,2,)}{p_1(2)p_2(2)}, & \text{if} \quad (u,v) \in [F_1(1), 1] \times [1, F_2(2)] \\ 0 & \text{otherwise.} \end{cases}$$

Mixed Markov Partitions procedure

- We will assume that for $1 \leq i \leq k$, $X(i)_t$ is an order o_m Markov chain, with $o_m < \infty$.
- The marginal state space is B^{o_m}.
- For each $s \in B^{o_m}$ and $b \in B$, $P_i(b|s) = \text{Prob}(X(i)_t = b | X(i)_{t-o_m}^{t-1} = s)$.

Mixed Markov Partitions procedure, first part

On the first part of our procedure we fit a model for the multivariate process,

- Divide the dataset in two parts.
- Use the first half of the dataset to fit a PMM to the multivariate process X_t with a maximum order equal to o_c.
- Call $\mathcal{L}^{o_c}$, the partition of A^{o_c} corresponding to the fitted model.
- Extend the partition $\mathcal{L}^{o_c}$, to a partition of A^{o_m} denoted by $\mathcal{P}_c$ in the following way. If $\mathcal{L}^{o_c} = \{L_1^{o_c}, ..., L_{m_c}^{o_c}\}$, then

  $$\mathcal{P}_c = \{L_1^c, ..., L_{m_c}^c\}, \text{ with, } L_j^c = \cup_{s \in L_j^{o_c}} \{w.s : w \in A^{(o_m - o_c)}\}, \ 1 \leq j$$

  Where "." denotes the concatenation between strings.

Mixed Markov Partitions procedure, second part

On the second part of our procedure we fit a model for each marginal process, using the second half of the dataset.

- Divide the second half of the dataset into k independent subsets of equal length.
- Fit a PMM to each marginal process using the corresponding subset of the dataset.
- For $i = 1, 2, ..., k$ let $\mathcal{L}^i = \{L^i_1, ..., L^i_{m_i}\}$ be the partition of B^{o_m} corresponding to the model fitted to the marginal process $X(i)_t$.
- From the collection of partitions $\{\mathcal{L}^1, ..., \mathcal{L}^k\}$ define the following partition of A^{o_m}.

$$\mathcal{P}_m = \{L^1_{j_1} \times ... \times L^k_{j_k} : 1 \leq j_1 \leq m_1, ..., 1 \leq j_k \leq m_k\}.$$

Mixed Markov Partitions procedure, third part

On the third part of our procedure we build a new partition from the joint and marginal partitions fitted on the first and second part ofthe procedure.

- Build the partition $\mathcal{P}$ of A^{o_m} combining $\mathcal{P}_c$ and $\mathcal{P}_m$. $\mathcal{P}$ as the refinement of the partitions $\mathcal{P}_m$ and $\mathcal{P}_c$. Corresponding to the following equivalence relationship in A^{o_m},

$$s \sim r \text{ if } \exists L \in \mathcal{P}_m \text{ and } \exists L' \in \mathcal{P}_c \text{ such that } s, r \in L \cap L'.$$

- Two states s and r belong to the same part of $\mathcal{P}$ if and only if they belong to the same part of both $\mathcal{P}_m$ and $\mathcal{P}_c$.

Transition probabilities estimation using copula theory

Given $s \in A^{o_m}$ and $a \in A$, we will show how to compute $P(a|s)$.

- Let w be the size o_c suffix of s, that means $s = q.w$ for an appropriated string q.
- Consider the estimator $\hat{P}^c(a|w)$ for the joint probability $P^c(a|w)$ of the process of order o_c, obtained from the first step of our procedure.
-
  - For $1 \leq i \leq k$, let $s(i)$ be the sequence in B^{o_m}, that is the sequence consisting of the concatenation of elements of s in the coordinate i.
  - Denote by $\hat{P}_i(a(i)|s(i))$ the estimate of the marginal probability from the i process of order o_m, obtained from the second part of our procedure, where $a(i)$ is the i-coordinate of a.
  - Denote by $\hat{F}_i(a(i)|s(i))$ the corresponding marginal distribution function.

Transition probabilities estimation using copula theory

The two set of probabilities are combined in the following way,

- Define a k-dimensional copula distribution $\hat{C}\left((u_1, ...u_k)|w\right)$ from the joint probabilities $\hat{P}^c(a|s)$, following, for example the idea presented in [2]. There is more than one way of choosing copula distributions in the case of discrete random variables, see [13].
- Evaluate the copula distribution on the marginal distributions, as follows

$$\hat{P}(a|s) = \hat{C}\left((\hat{F}_1(a(1)|s(1)), ..., \hat{F}_k(a(k)|s(k)))|w\right).$$

- It is easy to check that for any $L \in \mathcal{P}$ if $s, r \in L$ then $\hat{P}(a|s) = \hat{P}(a|r)$, $\forall a \in A$.
- In the approach proposed in this paper the number of parameters to estimate is $(|A| - 1)|\mathcal{L}^{o_c}| + \sum_{i=1}^{k}(|B| - 1)|\mathcal{L}^i|$ which gives a notion of computational complexity of the procedure.

Conclusions

- In this paper, we show a procedure to fit an approximated PMM to a multivariate process, when the dataset is not large enough to apply the usual methodology.
- The procedure combines individual PMM from the marginal processes and a PMM model for the joint process, this last with an order smaller to the real one.
- We show how to combine the corresponding partitions on a unique partition and how to combine the probabilities of all the models to a unique set of transition probabilities, for the fitted model.
- This methodology can be modified to be used with others families of Markov models, such as the fixed and variable length Markov chains, for which there exists several model selection methods (see [14], [1] and [4]).

Acknowledgments

This article was produced as part of the activities of FAPESP Center for Neuromathematics (grant 2013/ 07699-0 , S.Paulo Research Foundation).
The authors acknowledge the support provided by USP project "Mathematics, computation, language and the brain and FAPESP project "Portuguese in time and space: linguistic contact, grammars in competition and parametric change."

Bibliography


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.


Fernández, M. and González-López, V. A., "A Bayesian approach for convex combination of two Gumbel-Barnett copulas" In American Institute of Physics Conference Series, vol. 1558, pp. 1491-1494. 2013.


Fernández, M. and González-López, V. A., "A copula model to analyze minimum admission scores " In American Institute of Physics Conference Series, vol. 1558, pp. 1479-1482. 2013.


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.


Garcia, Jesus E., and Fernandez, M., "Copula based model correction for bivariate Bernoulli financial series." In American Institute of Physics Conference Series, vol. 1558, pp. 1487-1490. 2013.


Garcia, J. and Gonzalez-Lopez, V. A., "Independence tests for continuous random variables based on the longest increasing subsequence". *Journal of Multivariate Analysis*, 127, 126-146. 2014.


Garcia, J. and Gonzalez-Lopez, V. A., "Minimal Markov Models", *arXiv:1002.0729v1*, 2010.


Garcia, J. and Gonzalez-Lopez, V. A., "Detecting Regime Changes In Markov Models", Proceedings of The Sixth Workshop on Information Theoretic Methods in Science and Engineering, 2013.


Garcia, J. and Gonzalez-Lopez, V. A., "Modeling of acoustic signal energies with a generalized Frank copula. A linguistic conjecture is reviewed ", *Communications in Statistics - Theory and Methods* **43** (10-12), 2034–2044, 2013.


Garcia, J. E., Gonzalez-Lopez, V. A., and Nelsen, R. B., "A new index to measure positive dependence in trivariate distributions". *Journal of Multivariate Analysis*, 115, 481-495. 2013.


Garcia, J. E., Gonzalez-Lopez, V. A. and Viola, M. L. L., "Robust model selection and the statistical classification of languages", *AIP Conference Proceedings* **1490**, 160 – 170. 2012.


Garcia, J. E., Gonzalez-Lopez, V. A. and Viola, M. L. L., "Robust Model Selection for Stochastic Processes", *Communications in Statistics-Theory and Methods* **43** (10-12), 2516-2526. 2014.