Dissimilarity between Markovian Processes applied to Industrial Processes

ICNAAM 2016

Rhodes, Greece

Jesús E. García; V. A. González-López and F. H. Kubo de Andrade

September 24, 2016

Abstract

- In this paper we introduce a methodology to measure the discrepancy between two Markovian stochastic processes.
- The discrepancy is based on a consistent measure derived from a generalization of the Bayesian Information Criterion.
- We apply this concept to analyzing the similarity between two parallel lines of distillation of sugar cane, for the production of fuel.

The dataset

- After the fermentation, the product is heated in the same batch and immediately it is introduced in two different columns, in order to extract the hydrated alcohol.
- For each column, there are **5 variables collected** in a period of 1 month. The **sample size is** $n = 44643$.
- The variables are
  1. alcoholic contents $M_t^i(1)$,
  2. fill level $M_t^i(2)$,
  3. entrance temperature $M_t^i(3)$,
  4. exit temperature $M_t^i(4)$,
  5. vapor pressure $M_t^i(5)$.

Notation and definitions

- X_t is a multivariate discrete time order $o < \infty$ Markov chain on a finite alphabet A,
- $\mathcal{S} = A^o$ the state space.
- $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ a partition of $\mathcal{S}$.
- $a_m^n = a_m a_{m+1} \ldots a_n$ where $a_i \in A$, $m \leq i \leq n$.

Notation and definitions

- For each $a \in A$, $r \in \mathcal{S}$ and $L \in \mathcal{L}$,
  - $P(a|r) = \text{Prob}(X_t = a | X_{t-o}^{t-1} = r)$,
  - $P(L) = \sum_{s \in L} \text{Prob}(X_{t-o}^{t-1} = s)$,
  - $P(L, a) = \sum_{s \in L} \text{Prob}(X_{t-o}^{t-1} = s, X_t = a)$ and
  - $P(a|L) = \frac{P(L,a)}{P(L)}$ with $P(L) > 0$.

Definition

Let X_t be a discrete time order o Markov chain on a finite alphabet A. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$, where for each $i = 1, \ldots, K$, $s, r \in L_i$ if and only if $P(a|s) = P(a|r)\ \forall a \in A$. Then X_t is a Markov chain with partition $\mathcal{L}$.

Model selection

- The model, fitted following definition 1 is called Partition Markov Model (PMM).
- Given a sample x_1^n of the process X_t, $L \in \mathcal{L}$, $a \in A$ and $s \in A^o$,
  - $N_n(s) = \sum_{i=1}^{n-(o-1)} 1_{\{x_i^{i+o-1}=s\}}$
  - $N_n(s,a) = \sum_{i=1}^{n-o} 1_{\{x_i^{i+o-1}=s, x_{i+o}=a\}}$
  - $N_n^{\mathcal{L}}(L) = \sum_{s \in L} N_n(s)$
  - $N_n^{\mathcal{L}}(L,a) = \sum_{s \in L} N_n(s,a)$

Model selection

- To choose a model in the PMM family in a consistent way we can apply the algorithm introduced in García & González-López (2011) [1].
- The procedure works maximizing the Bayesian Information Criterion (BIC_α), associated to the sample x_1^n over the space of all the partitions of $\mathcal{S}$.
- The procedure is consistent for any positive and finite value α, and the expression of the criterion is

$$\mathrm{BIC}_\alpha(x_1^n, \mathcal{L}) = \sum_{a \in A, L \in \mathcal{L}} N_n^{\mathcal{L}}(L, a) \ln\left(\frac{N_n^{\mathcal{L}}(L, a)}{N_n^{\mathcal{L}}(L)}\right) - \frac{(|A| - 1)|\mathcal{L}|}{\alpha} \ln(n). \tag{1}$$

Model selection

- Also, for each $a \in A$ and $s \in \mathcal{S}$, the estimator of $P(a|s)$ is defined by $\hat{P}(a|s) = \frac{\sum_{r \in \hat{L}} N(r,a)}{\sum_{r \in \hat{L}} N(r)}$, where $\hat{L}$ is a part of $\hat{\mathcal{L}}_n$ such that $s \in \hat{L}$.
- This criterion BIC_α not only allows the estimation of the minimal partition, it also allows to compare processes laws, see García & González-López [3].

Dissimilarity

- Given the stochastic processes X_t^1 and X_t^2 of order $o < \infty$ with independent, size n, samples $(x_t^1)_1^n$ and $(x_t^2)_1^n$,
- X_t^1 and X_t^2 have the same distribution if, and only if, for n large enough,
  $\mathrm{BIC}_\alpha\left((x_t^1)_1^n, (x_t^2)_1^n, \mathcal{L}\right) > \sum_{i=1,2} \mathrm{BIC}_\alpha((x_t^i)_1^n, \mathcal{L}^i)$
-

$$\mathrm{BIC}_\alpha\left((x_t^1)_1^n, (x_t^2)_1^n, \mathcal{L}\right) = \sum_{a\in A, L\in\mathcal{L}} N_{2n}^{\mathcal{L}}(L,a) \ln\left(\frac{N_{2n}^{\mathcal{L}}(L,a)}{N_{2n}^{\mathcal{L}}(L)}\right) \quad (2)$$

$$- \frac{(|A|-1)}{\alpha}|\mathcal{L}| \ln(2n). \quad (3)$$

Dissimilarity

- Moreover, García & González-López [3] defines

$$d_\alpha^n\left((x_t^1)_1^n,(x_t^2)_1^n\right)=\alpha\frac{\sum_{a\in A}(B(\mathcal{L}^1,n,a)+B(\mathcal{L}^2,n,a)-B(\mathcal{L},2n,a))}{(|A|-1)\left\{(|\mathcal{L}^1|+|\mathcal{L}^2|)\ln(n)-|\mathcal{L}|\ln(2n)\right\}} \quad (4)$$

  with $B(\mathcal{L},n,a)=\sum_{L\in\mathcal{L}}N_n^{\mathcal{L}}(L,a)\ln\left(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)}\right)$.

- They show that X_t^1 and X_t^2 have the same distribution if, and only if, $d_\alpha^n<1$ for n large enough.
- In practice, for finite samples, the problem of choosing α remains and is used in general $\alpha=2$, see Schwarz (1978) [4].

The methodology that we use in this work is based on the following theorem.

Theorem

Given the stochastic process X_t^i *of order* $o < \infty$ *with sample* $(x_t^i)_1^n$ *of size* n, $i = 1, 2$ *and* $(x_t^1)_1^n \perp (x_t^2)_1^n$ *and* $\alpha > 0$,

$$X_t^1 =^d X_t^2 \text{ if, and only if } \lim_{n\to\infty} d_\alpha^n\left((x_t^1)_1^n, (x_t^2)_1^n\right) = 0 \text{ almost surely.} \tag{5}$$

We will use d_α^n as a measure of dissimilarity between the processes. This new result sets us free from the restriction $d_\alpha^n < 1$, showing that for all value of α the measure d_α^n should be closer to 0 when the compared processes follow the same law, if n is large enough.

Data and Results

- $X_t = (X(1)_t, ..., X(k)_t)$, where $X(j)_t \in B$ and it is the state of the j-source at time t for $j = 1, ..., k$, $X_t \in A$, where $A = B^k$ and B is the finite alphabet for the one-dimensional marginal processes (considered equal by simplicity).
- In the case treated here, for each time t and each column $i = 1, 2$ define $X_t^i(j) = 1$ if the value of $M_t^i > \text{median}(M_t^i, t \geq 1)$ and $X_t^i(j) = 0$ otherwise, for $j = 1, \ldots, k = 5$. Then, $B = \{0, 1\}$ and $|A| = 2^5$.

Data and Results

The samples from each column are splitted into two equally sized subsamples $(X_t^i)_{t=1}^{\lfloor \frac{n}{2} \rfloor}$ $(X_t^i)_{t=\lfloor \frac{n}{2} \rfloor+1}^{n}$, $i = 1, 2$, the dissimilarity measure d_α is applied to the different combinations of these samples.

Table: Predictive ability of a k-variate PMM in the two portions of the two columns $i = 1, 2$. The order used was $o = 2 \approx \lfloor \log_{|A|}(22320) \rfloor$.

Portion of the Sample	Proportion of Success-PMM
$(X_t^1)_{t=1}^{\frac{n}{2}}$	0.79293
$(X_t^1)_{t=\frac{n}{2}+1}^{n}$	0.80646
$(X_t^2)_{t=1}^{\frac{n}{2}}$	0.73282
$(X_t^2)_{t=\frac{n}{2}+1}^{n}$	0.75204

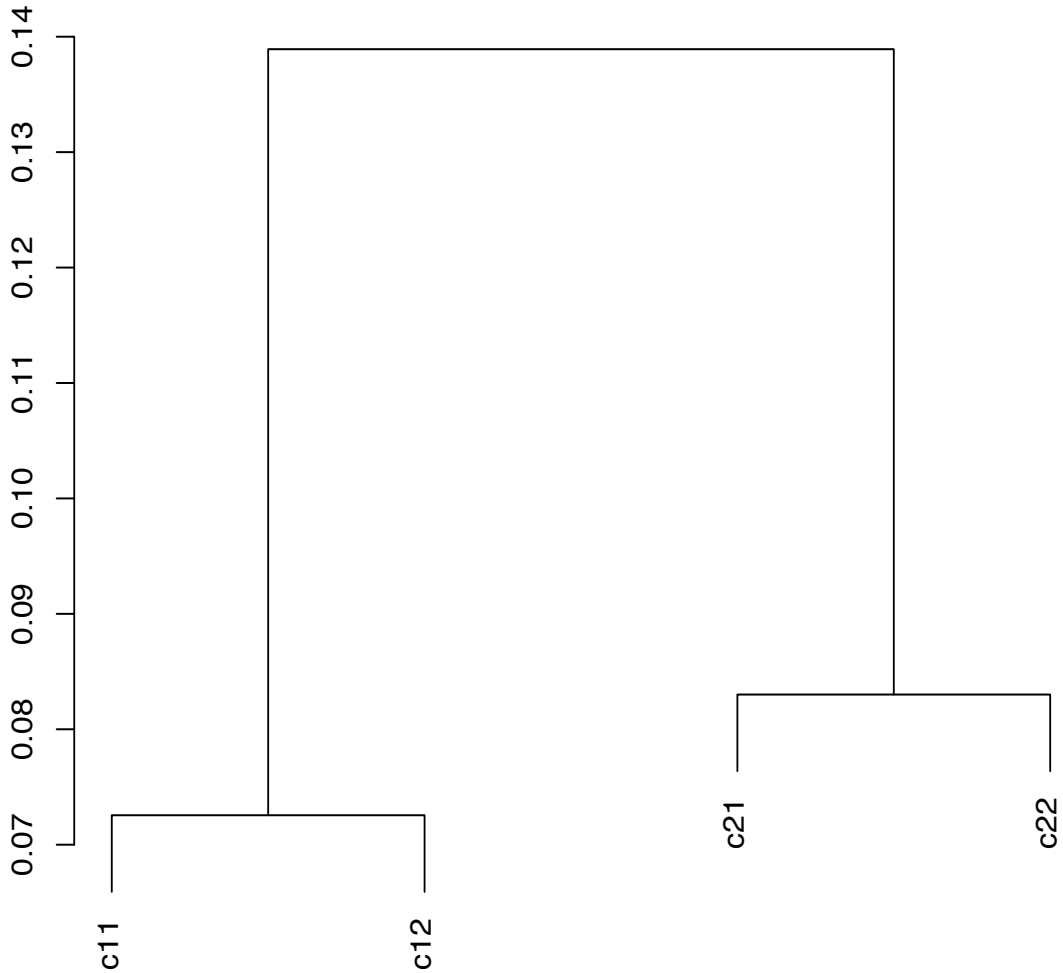
Data and Results

The distance between subsamples belonging to the same column is smaller than between subsamples belonging to different columns. This indicates the hypothesis of different laws for the columns.

Table: d_α, $\alpha = 2$, applied to the different combinations of subsamples: $(X_t^i)_{t=1}^{\frac{n}{2}}$ $(X_t^i)_{t=\frac{n}{2}+1}^{n}$, $i = 1, 2$.

d_α	$(X_t^1)_{t=1}^{\frac{n}{2}}$	$(X_t^1)_{t=\frac{n}{2}+1}^{n}$	$(X_t^2)_{t=1}^{\frac{n}{2}}$	$(X_t^2)_{t=\frac{n}{2}+1}^{n}$
$(X_t^1)_{t=1}^{\frac{n}{2}}$	0	0.07255	0.17333	0.20236
$(X_t^1)_{t=\frac{n}{2}+1}^{n}$	0.07255	0	0.17743	0.15812
$(X_t^2)_{t=1}^{\frac{n}{2}}$	0.17333	0.17743	0	0.08301
$(X_t^2)_{t=\frac{n}{2}+1}^{n}$	0.20236	0.15812	0.08301	0

Dendrogram of the values of d_α^n



- The BIC_α criterion is consistent for any $0 < \alpha < \infty$,
- There is not a clear way to choose a specific α value for a (finite) sample.
- Schwarz (1978) [4] used $\alpha = 2$, this value was derived from the specific assumptions of a discrete uniform prior distribution in the dimensions of the spaces, that can be considered to construct the model.
- We sidestep this problem using the dissimilarity measure d_α which in the application shows a proximity between the results coming from the two parts of the same column. And a difference between the columns, as illustrated by the dendrogram.

Concluding Remarks

- Is proposed a dissimilarity measure between Markovian processes based on a generalization of the Bayesian Information Criterion.
- A parameter α is introduced and we show that two processes are identical if $d_\alpha < 1$, for n large enough.
- Moreover, it is not necessary to know the best value for α because the condition $d_\alpha < 1$ is quite unnecessary, since theorem 2 implies that the measure d_α converges to zero, when n go to infinite, for processes with the same law.
- We use the measure to verify the performance of two columns of fuel alcohol production and we conclude that there is a difference between the production lines. An open question is whether this difference is significant or not.

Acknowledgments

The authors would like to thank Raiza Balbino, from I. Systems, Brazil, for making available the data analyzed in this study


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, Detecting regime changes in Markov models. *SMTDA2014 Book* (to appear).


G. Schwarz, Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461-464, 1978.