Inspecting Discrepancies between Industrial Processes ICNAAM 2017

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

September 25, 2017

Introduction

- A distance between samples of Markovian processes is introduced.
- Given a sample of each process, the distance also allows to decide if the laws of the processes are the same.
- In the case where the processes are considered different, the distance allows to find the elements of the state space where the discrepancy in probabilities is manifested.
- The distance is applied to the inspection of processes that work in parallel. The objective is to use the distance to help make adjustments to the processes to make the laws of the two processes the more similar possible.

Introduction

- The case investigated in this article is the process of *bauxite digestion*, where the bauxite undergoes a reaction with a caustic liquor resulting in the alumina paste.
- In the application, the distance reveals the configurations of the variables that expose discrepancies between the processes. This information can be used to fine tuning the processes.

Notation

For (X_t) a discrete time, order $o < \infty$ Markov chain on a finite alphabet A,

- $\mathcal{S} = A^o$ is the state space,
- $a_m^n = a_m a_{m+1} \dots a_n$ where $a_l \in A$, $m \leq l \leq n$,
- for each $a \in A$ and $s \in \mathcal{S}$, $P_X(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s)$.
- $x_1^{n_x}$, a size n_x sample of process (X_t).
- $N_x(s)$ is the number of occurrences of s in the sample x_1^n.
- $N_x(s, a)$ is he number of occurrences of s followed by a in the sample x_1^n,

For (X_t) and (Y_t) two discrete time, order $o < \infty$ Markov chains on the same finite alphabet A, consider $x_1^{n_x}$, a size n_x sample of process (X_t) and $y_1^{n_y}$, a size n_y sample of process (Y_t).

- $N_{x,y}(s) = N_x(s) + N_y(s)$,
- $N_{x,y}(s,a) = N_x(s,a) + N_y(s,a)$,

Distance

Definition

For two such samples $x_1^{n_x}$, $y_1^{n_y}$ and $s \in \mathcal{S}$,

$$d_s(x_1^{n_x}, y_1^{n_y}) = \frac{1}{b(x_1^{n_x}, y_1^{n_y})} \sum_{a \in A} \left\{ N_x(s,a) \ln \left(\frac{N_x(s,a)}{N_x(s)} \right) \right.$$
$$\left. + N_y(s,a) \ln \left(\frac{N_y(s,a)}{N_y(s)} \right) - N_{x,y}(s,a) \ln \left(\frac{N_{x,y}(s,a)}{N_{x,y}(s)} \right) \right\}$$

where $b(x_1^{n_x}, y_1^{n_y}) = (|A| - 1) \ln(n_x + n_y)$

Properties

i. *For a fixed string $s \in \mathcal{S}$, the function $d_s(x_1^{n_x}, y_1^{n_y})$ is a distance.* Consider $x_1^{n_x}, y_1^{n_y}, z_1^{n_z}$ three samples of order o Markov Chains on the alphabeth A,

$$d_s(x_1^{n_x}, y_1^{n_y}) \geq 0 \text{ with equality } \Leftrightarrow \frac{N_x(s,a)}{N_x(s)} = \frac{N_y(s,a)}{N_y(s)} \ \forall a \in A,$$

$$d_s(x_1^{n_x}, y_1^{n_y}) = d_s(y_1^{n_y}, x_1^{n_x}),$$

$$d_s(x_1^{n_x}, y_1^{n_y}) \leq d_s(x_1^{n_x}, z_1^{n_z}) + d_s(z_1^{n_z}, y_1^{n_y}).$$

ii. *Local consistence.* If the stochastic laws of (X_t) and (Y_t) are the same in s, then $d_s(x_1^{n_x}, y_1^{n_y}) \underset{\min(n_x,n_y)\to\infty}{\longrightarrow} 0$.

Otherwise $d_s(x_1^{n_x}, y_1^{n_y}) \underset{\min(n_x,n_y)\to\infty}{\longrightarrow} \infty$.

Data and Results

- The case investigated in this article is the process of *bauxite digestion*,
- the bauxite pulp undergoes a reaction with a caustic liquor resulting in the alumina paste.
- We will analyze two digesters $i = 1, 2$ that work in parallel and of which we have the seven measures.
- Those seven variables were collected in the same period of time and operate in parallel.

- $M_{i,t}(1)$: temperature of the bauxite pulp and material that reacts with the caustic soda and the Bayer liquor to recover the alumina;
- $M_{i,t}(2)$: temperature of the caustic soda which is injected into the digester, with the bauxita pulp and the Bayer liquor;
- $M_{i,t}(3)$: temperature of the bauxite pulp at the exit of the digester, after undergoing reaction with the Bayer liquor and the caustic soda;
- $M_{i,t}(4)$: temperature of the bauxite pulp at the entrance of the digester, prior to undergoing reaction with the Bayer liquor and the caustic soda;
- $M_{i,t}(5)$: internal temperature of the bauxite digester, measured to control the reaction of the components;
- $M_{i,t}(6)$: caustic soda flow entering the digester which together with its temperature, interferes with the reaction and the internal temperature of the digester;
- $M_{i,t}(7)$: flow of the bauxite pulp, which enters the digester. It is the material that undergoes the reaction to recover the alumina.

- $X(j)_t \in \{0,1\}$ and it is the state of the j-source at time t for $j = 1, ..., k$,
- $X_t = (X(1)_t, ..., X(7)_t)$, $X_t \in A = \{0,1\}^k$.
- for each time t and each column $i = 1, 2$ we define: $X_{i,t}(j) = 1$ if the value of $M_{i,t}(j) > M_{i,t-1}(j)$ and $X_{i,t}(j) = 0$ otherwise, for $j = 1, \ldots, k = 7$.

For the analise of the ressults we will use,

- *dmean* : the average value of d, over all strings common to both processes,
- d^1 : the maximun value of d,
- s^1 : the string where the maximun is attained,
- d^2 : the second maximun value of d and
- s^2 : the string where the second maximun is attained,

Table: Marginal comparison $B = \{0, 1\}$. The order used was $o = 12 = \lfloor \log_{|B|}(11674) \rfloor - 1$.

Marginal	*dmean*	d_1	s_1	d_2	s_2
1	0.219	97.91801	111111111111	11.224	111111110111
2	0.081	15.77245	111111111111	2.784	111111111100
3	0.067	2.07549	111111111111	1.963	111111111110
4	0.150	366.02234	111111111111	1.651	111111111110
5	0.098	27.66434	111111111111	3.194	011111111111
6	0.474	36.59304	111111111111	34.040	111111111110
7	0.256	45.77704	111111111110	27.128	000000000001

The marginal comparisons show that the processes are different, since there are values of d greater than one in all the cases.

Table: Joint comparison between $X_{1,t}$ and $X_{2,t}$, $A = \{0,1\}^7$. The order used was $o = 2 = \lceil \log_{|A|}(11674) \rceil$.

$dmean$	d_1	s_1	$\hat{P}(\cdot \vert s_1)$
0.031	1.255	(1111111)(1111111)	$\hat{P}_1(1111111 \vert s_1) = 0.306$
			$\hat{P}_2(1111111 \vert s_1) = 0.504$
			$\hat{P}_1(a \vert s_1) < 0.1,\ a \neq 1111111$
			$\hat{P}_2(a \vert s_1) < 0.1,\ a \neq 1111111$

In tables 3-5 we list (by pairs of variables compared) the strings that show a value of $d > 1$.

Table: Strings for intervals of d-values, such that $d > 1$.

$d \in$	Marginal 1	Marginal 2
[1, 2)	010011111111, 001110111111, 111101000000, 110111111101 111110100000, 111111110011, 000000000101, 111101011111 111111000111, 000000010011, 000110111111, 000000001001 111111010000, 000000000111, 110000000000, 111111001111 111001111111, 111111010111, 111100111111, 000101111111 111000111111, 010111111111, 000100111111, 101011111111 111100011111, 111110011111, 111000000000, 111111001000	111110011111, 001111111111 111111110000, 110111111111 101111111111, 111101111111 000011111111, 111111111000
[2, 3)	000000001111, 000000011111, 000000000001, 111100000000 111111101000, 000001111111, 000000111111, 110011111111	111011111111, 011111111111 **111111111100**
[3, 4)	000000000011, 100111111111, 111110000000, 111111100111	
[4, 5)	111111111100, 111111000000, 000011111111, 111111111101	
[6, 7)	001111111111, 111111110000, 111111100000	
[7, 8)	111110111111, 111101111111, 111011111111	
[8, 9)	111111111000, 000111111111, 011111111111, 110111111111	
[9, 10)	101111111111, 111111011111	
[10, 11)	111111111011, 111111101111	
[11, 12)	**111111110111**	
[15, 16)		**111111111111**
[90, 100]	**111111111111**	

Table: Strings for intervals of d-values, such that $d > 1$.

$d \in$	Marginal 3	Marginal 5
$[1, 2)$	111111111000, **111111111110**	111111110010, 111000000000, 111111010011, 111110001000 000000000000, 111111101100, 111111110000, 110111111111 111111111100, 001111111111
$[2, 3)$	**111111111111**	111111000000
$[3, 4)$		111111100000, 111111111000, **011111111111**
$[27, 28)$		**111111111111**

Table: Strings for intervals of d-values, such that $d > 1$.

$d \in$	Marginal 6	Marginal 7
$[1, 2)$	100000000011, 000001111110, 111110000111 110000001111, 111111000111, 111000001111 111100000001, 000001111000, 000111111100 111111001111, 001111111110, 111110000011 111110001111, 111111110001, 000011111110 111100000011, 001111111100, 111100011111 100000001111, 111111100011, 111111000001 111110000001	000011000000, 000000110000, 011110000000 111001111111, 111111011100, 110001111111 011111011110, 001111100000, 001111000000 111111111010, 110000001111, 000001100000 011111100000, 000010000001, 110000011111 011000000001, 110000111111, 111111101110 111011111110, 100011111111, 111111100111 011111000000, 111111101111, 001000000001 000000100001, 111000111111, 000011110000 111111110111, 111000011111, 100000000011 100000001111, 111100001111, 110000000000 111100011111, 111111000111, 111110111110 000000000100, 111111111101, 000111110000 111111011110, 110111111110, 000001110000 111101111110, 101111111110, 100000000111 100000011111, 111101111111, 111111101100 011111111000, 110000000011, 111111011111 011111110000, 000111100000, 011111111111 000001111000
$[2, 3)$	100000000111, 000111111110, 011111111110 111000000111, 111100000111, 111111100001 100000000001, 111000000001, 110000000001	000011000000, 000000110000, 011110000000 000000001110, 000111111111, 111110000000 100111111110, 111100111111, 111110000111 001111111000, 111111111111, 111100000000 111111001111, 001111111100, 111110001111 000000011110, 000011111100, 100000111111 000000000110, 111111100001, 010000000001

Concluding Remarks

The marginals 4 and 1 show the values of d most extreme $d = 366.02234$ and $d = 97.91801$ respectively, but in each case it is produced by the same string 111111111111. That is, we can see that the consecutive growths in temperatures recorded by $M_t(1)$ and $M_t(4)$ ("*temperature of the bauxite pulp and material that reacts with the caustic soda and the Bayer liquor*" and "*temperature of the bauxite pulp at the entrance of the digester*") experience markedly different transition probabilities when compared between columns 1 and 2.

On the oder hand, the marginals 7, 1 and 6 show a greater volume of strings whose values of d are greater than 1, but without reaching the magnitudes of the previous cases. The *flow of the bauxite pulp*: $M_t(7)$, can be considered as a mayor contributor of discrepancies between the columns of production, with 98 strings showing a d value in $(1, 4)$. In this sense, if we observe the processes $M_t(7), M_t(1), M_t(6)$ in relation to the total of strings with values of d in $(1, 2)$, each of them shows 55, 28 and 22 strings respectively, which again places the *flow of the bauxite pulp* as the largest contributor, followed by *temperature of the bauxite pulp and material that reacts with the caustic soda and the Bayer liquor* - and - *caustic soda flow entering the digester*.

Bibliography


I. Csiszar and P. Shields. The consistency of the bic markov order estimator. *The Annals of Statistics* 28, (6), 1601-1619, 2000


A. Galves, C. Galves, J.E. Garcia, N.L. Garcia and F. Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *The Annals of Applied Statistics* 6(1), 186-209, 2012.


J.E. García, R. Gholizadeh and V.A. González-López. Linguistic Compositions Highly Volatile in Portuguese. *Submitted*.


J.E. García and M. Fernández. Copula based model correction for bivariate Bernoulli financial series. AIP vol. 1558, no. 1, pp. 1487-1490. AIP Publishing, 2013.


J. Garcia and V.A. Gonzalez-Lopez. Minimal markov models. arXiv preprint arXiv:1002.0729, 2010.


J.E. García and V.A. González-López. Minimal Markov Models. In *Fourth Workshop on Information Theoretic Methods in Science and Engineering. Helsinki.* v.1. p.25 - 28, 2011.


J.E. García, V.A. González-López and M.L.L. Viola. Robust model selection and the statistical classification of languages. In *American Institute of Physics Conference Series* (Vol. 1490, pp. 160-170), 2012.


García, J.E.; González-López, V.A. Consistent Estimation of Partition Markov Models. Entropy 2017, 19, 160.


J.E. García, V.A. González-López and M.L.L. Viola. Robust Model Selection for Stochastic Processes.