Full Partition Markov Models

Verónica Andrea González-López (`veronica@ime.unicamp.br`)
Universidade Estadual de Campinas

Abstract. In this paper, we analyze the model proposed in [1] in which is considered a set of p independent samples of discrete time Markov chains, over a finite alphabet A and with finite order o. The model is obtained identifying the states on the state space A^o where two or more samples share the same transition probabilities (see also [2]). This identification establishes a partition on $\{1, ..., p\} \times A^o$ the set of samples and the state space. We show that by means of the Bayesian Information Criterion (BIC) the partition can be estimated eventually almost surely. Also in [1] is given a notion of divergence, derived from the BIC, which serves to identify the proximity/discrepancy between elements of $\{1, ..., p\} \times A^o$ (see also [3]). In the present article, we also prove that this notion is a metric in the space where the model is built and that it is statistically consistent to determine proximity/discrepancy between the elements of the space $\{1, ..., p\} \times A^o$. We apply the notions discussed here for the construction of a parsimonious model that represents the common stochastic structure of 153 complete genomic Zika sequences, coming from tropical and subtropical regions.

References

[1] García Jesús E., Londoño S.L.M., (2019, July). Optimal Model for a Set of Markov Processes, in AIP Conference Proceedings. (Vol. 2116: 130002). DOI: 10.1063/1.5114125

[2] García Jesús E., González-López V. A., Consistent Estimation of Partition Markov Models, Entropy 2017; 19(4): 160. DOI: 10.3390/e19040160

[3] García Jesús E., Gholizadeh R., González-López V.A., A BIC - based consistent metric between Markovian processes, Applied Stochastic Models in Business and Industry 2018; 34(6): 868-878. DOI: 10.1002/asmb.2346

Joint work with: M. T. A. Cordeiro, Jesús E. García and S. L. M. Londoño.