Recommender’s review and decision for preprint entitled: Proper account of long-term correlations in the observations improves state-space models' performances

Decision : This preprint merits revision

Review: The authors propose a study of the impact of model mis-specification for models on the family of HMM and HSMM. Mis-specification can be at the level of the hidden chain (Markov versus semi-Markov) or at the level of the observed chain (AR0 versus AR1). The study is made in the context of data from fishery vessel movements. The impact of model mis-specification is assessed on the restored hidden chain (decoding task), which I find very relevant since in many applications we are more interested by decoding quality rather than by precise parameters estimation. The main conclusion of the study is that choosing the wrong AR model at the observed sequence level has more impact that choosing the wrong model on the hidden chain.

This work addresses a very interesting topic for statisticians and ecological modelers. As underlined by the two reviewers it is very clearing presented.

However, they have made comments that require answers from the authors and clarification in the manuscript.

I also have some remarks and questions listed below.

One main conclusion is that ‘imposing a Markov structure while the state process is semi-Markov does not impair the state decoding performance’. Actually, this result is obtained for a particular semi-Markov model, with Negative Binomial sojourn time distribution. I understand the reasons for restricting the analysis to this distribution but then, the conclusion cannot be so general. Maybe with another sojourn time distribution, the impact on decoding would be more important.

Figure 1. There are some approximations in the legend text. The legend starts with ‘Directed acyclic graph for HMM and HSMM, …’. I agree with the term in the case of HMM. But not in the HSMM case. In a DAG representation of a Bayesian network, the nodes of the graph are the model variables and in the classical HSMM representation variables are jump time, sojourn duration or date of jump, and observations indexed on calendar time. Since the value of each sojourn time is not known in advance it is not possible to draw the arrows from hidden states towards observations. Still about the legend, the arrows in a DAG can enable to recover conditional independencies. But an absence of arrow does not indicate that the two variables are conditionally independent (it would depend on which variable is the conditionning). This can be seen in the case of a V-structure.

HSMM description (p 6). The definition of a HSMM is too superficial. It should be more concrete/explicit, like for the HMM model. The random variables involved are not described. Also a standard reference like one of these two should be added:

Shun-Zheng Yu. *Hidden Semi-Markov Models Theory*, Algorithms and Applications. Elsevier, 2016

or

Barbu, V-S. et N. Limnios. 2008. *Semi-Markov Chains and Hidden Semi-Markov Models towards Applications.* Springer.

Line 152 : a ‘conditionally to the state sequence’ is missing

Line 158: do you mean ‘by ignoring **temporal dependencies** in the hidden layer’? There are hidden variables in the mixture model, even if they are independent.

Line 218: ‘For each case’ instead of ‘For each model’?

Page 10: About the definition of the loss. Since MRA and MSA are probabilities, with values between 0 and 1, wouldn’t it be easier to interpret the loss if it was defined as the difference instead of the relative difference?