Peer Community In Mathematical & Computational Biology

An empirical study on the impact of neglecting dependencies in the observed or the hidden layer of a H(S)MM model on decoding performances

Nathalie Peyrard (iD) *based on peer reviews by* **Sandra Plancade** *and 1 anonymous reviewer*

The article by Bez et al [1] addresses an important issue for statisticians and ecological modellers: the impact of modelling choices when considering state-space models to represent time series with hidden regimes.

The authors present an empirical study of the impact of model misspecification for models in the HMM and HSMM family. The misspecification can be at the level of the hidden chain (Markovian or semi-Markovian assumption) or at the level of the observed chain (AR0 or AR1 assumption).

The study uses data on the movements of fishing vessels. Vessels can exert pressure on fish stocks when they are fishing, and the aim is to identify the periods during which fishing vessels are fishing or not fishing, based on GPS tracking data. Two sets of data are available, from two vessels with contrasting fishing behaviour. The empirical study combines experiments on the two real datasets and on data simulated from models whose parameters are estimated on the real datasets. In both cases, the actual sequence of activities is available. The impact of a model misspecification is mainly evaluated on the restored hidden chain (decoding task), which is very relevant since in many applications we are more interested in the quality of decoding than in the accuracy of parameters estimation. Results on parameter estimation are also presented and metrics are developed to

help interpret the results. The study is conducted in a rigorous manner and extensive experiments are carried out, making the results robust.

The main conclusion of the study is that choosing the wrong AR model at the observed sequence level has more impact than choosing the wrong model at the hidden chain level.

The article ends with an interesting discussion of this finding, in particular the impact of resolution on the quality of the decoding results. As the authors point out in this discussion, the results of this study are not limited to the application of GPS data to the activities of fishing vessels Beyond ecology, H(S)MMs are also widely used epidemiology, seismology, speech recognition, human activity recognition ... The conclusion of this study will therefore be useful in a wide range of applications. It is a warning that should encourage modellers to design their hidden Markov models carefully or to interpret their results cautiously.

References:

[1] Nicolas Bez, Pierre Gloaguen, Marie-Pierre Etienne, Rocio Joo, Sophie Lanco, Etienne Rivot, Emily Walker, Mathieu Woillez, Stéphanie Mahévas (2024) Proper account of auto-correlations improves decoding performances of state-space (semi) Markov models. HAL, ver.3 peer-reviewed and recommended by PCI Math Comp Biol `https://hal.science/hal-04547315v3`

Reviews

Evaluation round #2

DOI or URL of the preprint: `https://hal.science/hal-04547315v2`
Version of the preprint: 2

Authors' reply, 14 February 2025

Download author's reply
Download tracked changes file

Decision by Nathalie Peyrard ⓘ**, posted 17 January 2025, validated 17 January 2025**

see attached file **Download recommender's annotations**

Reviewed by Sandra Plancade, 07 January 2025

Download the review

Reviewed by anonymous reviewer 1, 12 December 2024

Download the review

Evaluation round #1

DOI or URL of the preprint: `https://hal.science/hal-04547315`
Version of the preprint: 1

Authors' reply, 13 November 2024

Decision by Nathalie Peyrard 🔗**, posted 03 September 2024, validated 04 September 2024**

This preprint merits a 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?

Reviewed by Sandra Plancade, 18 July 2024

Overall Comment

This manuscript analyses the impact of model (mis-)specification on decoding accuracy in hidden markov and semi-markov settings, combining supervised experimental data and simulations. The subject is of high interest, and the analyses are conducted rigorously and thoroughly. The choice to restrict yourselves to two settings i.e. two real data sets used to determine realistic parameters for simulations, represents a good compromise to limit the complexity of the results while allowing variations in the setting characteristics. The results are both clear and comprehensive (with a remarkable job to include a large number of characteristics on the same graph while remaining understandable), the main finding is the strong impact of model mispecification on the observed layer on decoding accuracy, versus the weak impact of mispecification in the hidden layer.

The discussion represents a true added value to the results. The scope and limits, as well as the summary of the main results, are precise and accurate but the most valuable part is the reflexion on various issues regarding (mis-)modeling, notably based on literature. Even if the present version seems sufficient for a manuscript whose main goal is to provide novel analyses, it would be interesting to provide more details on these aspects.

Major comments

1) In the discussion section devoted to AR assumptions, you recommend to conduct exploratory analyses of the observation process data to define appropriate auto-correlation hypotheses. If these analyses are straightforward or at least already developped in literature, you should briefly present them and mention alternative models. If not, this issue should be discussed as potential perpectives for further work, if possible evocating leads and difficulties.

Moreover, these recommended prior analyses were not conducted on the real data presented in the manuscript. Indeed, the AR0 and AR1 model were chosen a priori, and only an a posteriori analysis on the validity of AR hypotheses is displayed (and concludes to a violation of the AR1 and even the general AR assumption). This appears as a contradiction, and as a minimum, this positionning should be clarified.

2) On real data, accuracy is less good with HSMM-AR1 than expected in a mixture model analysis in setting 2 (Figure 9). This observation is mentioned on l.429-430, but without explanation. The most surprising is that this occures for setting/vessel 2 only, while model mispecification is greater with setting/vessel 1. Do you have an explanation or hypotheses?

3) I am not sure that the title properly reflects the content of the manuscript. First of all, it seems to me that the terms "decoding", model (mis-)specification, and potentially "(semi-)Markov" should appear in the title. Besides, in the auto-regressive models, the term "long-term correlations" would allude to higher auto-regression levels than the AR1 actually implemented in the manuscript; moreover the AR level is not "properly accounted for" as demonstrated on Figure 14, since the real AR level is much higher than 1.

I take the liberty to make a suggestion but in an absolutely non-prescriptive way :
"Analysis of model specification in state-space models: account of correlations in the observations improves decoding performances".

Minor comments

1) l.158 (and several other times): It seems that you use the term "mixed model" for "mixture model".

2) l.160: $\widehat{S}$ could be written explicitly, and the notation s' could be properly defined (even if the latter is intuitive).

3) l.161: I don't understand what you mean by "empirical speed frequency distributions"; indeed, as far as I undestand, the "true" distribution parameters i.e. the ones used for simulations are actually the empirical ones. Could you clarify this point? Similarly, is there an empirical counterpart for π_s in the formula of d_V?

4) l.239: The definition of accuracy is unprecised. If I'm not wrong, S is never defined, does it correspond to a given S_t or to the chain $(S_t)_{t \geq 1}$? As the $(\widehat{S}_t)_{t\geq 1}$ are not mutually independent, formula for the chain $(S_t)_{t \geq 1}$ can not be directly deduced from single S_t formula. Moreover, are

the results displayed e.g. in Figure 9 computed with this theoretical accuracy formula or with an empirical counterpart?

I recommend to write the exact quantity (either as a sum or an integral) of the computed accuracy.

5) l.337: "nearly all parameters are impacted by this bi-layers discrepancy... ". As far as I understand, Figure 12 shows that a non-small proportion of parameters are weakly impacted, notably in setting 1.

6) Figures 14 and 15: provide the definition of the notions (delayed correction plots and coefficient of partial auto-correlation).

7) Considering mixture model as a basis of comparison is very relevant. Compared to the usual distance between PDF (Hellinger, etc), this discrepancy measure has an interpretation in terms of decoding. (This comment does not call for any modification!).

Typos

1) l.165-166: $S_\k+1\$ should be $S_\t+1\$

2) l.188: I think "by variable ($d_\AR_S\$)" should be "by state ($d_\AR_S\$)"

3) Figure 11 : the x-axis label should be $d_\AR\$ as mentioned on l.322.

Questionnaire

Title and abstract

Does the title clearly reflect the content of the article? No (cf major comment 3)

Does the abstract present the main findings of the study? Yes

Introduction

Are the research questions/hypotheses/predictions clearly presented? Yes

Does the introduction build on relevant research in the field? Yes

Materials and methods

Are the methods and analyses sufficiently detailed to allow replication by other researchers? Yes

Are the methods and statistical analyses appropriate and well described? Yes

Results

In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? Non relevant

Are the results described and interpreted correctly? Yes

Discussion

Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? Yes

Are the conclusions adequately supported by the results (without overstating the implications of the findings)? Yes

Reviewed by anonymous reviewer 1, 06 August 2024

Download the review