Review of JABES=D-18-00082

Synopsis and main points

Motivated by ant interactions this article attempts to tackle the issue of fitting a hidden Markov model to a large amount of temporal point-process data. Since no model is correct, the model does not fit the data and parameter estimates are unrealistic. A reformulation and a penalisation term seem to fix this.

The article is well written in terms of the English, and sometimes the probability/statistics is clear, but sometimes it is unclear or simply wrong. Most importantly, the fundamental reformulation between equations (16) and (34) is just plain wrong.

Also, given mistakes and lack of clarity in the formulation of the original (unaltered, unpenalised) model and algorithm, I was left with some uncertainty about the correctness any implementation. This uncertainty was then doubled when I saw the results in equation (15). There is no apriori reason why $\hat{P}$ should have this special form (the independent beta priors for the rows would lead to a very different matrix, albeit with the same symmetry as observed in (15)) - the fact that opposite terms agree to four decimal places rings alarm bells and suggests that it's possible that a correct fitting of the model would not lead to the biologically unrealistic estimates.

Having discovered such errors, I stopped reading after section 3.2.

Equations (16) to (34)

P8 'Rather than modelling the transition probabilities, P, directly, as we did in Section 2.1, we model these probabilities as a function of state switching rates. Because the resolution of the data provide a second-by-second account, we consider a time discretization of a continuous-time Markov chain such that state switching may only occur ... at times $t = 1, \ldots, T$. We choose not to model at a sub-second resolution (higher resolutions provide no additional benefit or interpretability) and discretization would need to occur for convariates to be included.'

Roughly speaking, the authors wish to use a continuous-time formulation because there is so much data. However, they then wish to time-discretize the continous-time formulation to transitions after one second because higher resolution doesn't help. This would appear to completely nullify any benefit of using a continuous-time formulation. Moreover the claim that covariates could not be included in a continuous-time formulation is false; e.g. $\gamma_{ij} = \gamma_{ij}^{(0)} e^{\beta'_{ij} x}$.

Next, in equations (16)-(19), there is an attempt at a formulation (via the probability density of the transition time) that only allows transitions to happen at integer multiples of one second. Firstly, this really does not fit with the point of continuous-time modelling. Secondly it is not even self consistent: consider the probability density of a transition after 2 seconds: this is either $\gamma_i e^{-2\gamma_i}$ (continuous-time, at 2 seconds) or $(1 - \gamma_i e^{-\gamma_i})\gamma_i e^{-\gamma_i}$ (using the authors' transition matrix). Thirdly, if you are going to have a discrete transition matrix then the standard formulation of transition probabilities is there for a reason - it is the simplest to write down, and all other forms must be equivalent to it. Finally, a density is not a probability! The P_{ij} in the article satisfies, for small δt,

$$\begin{aligned} P_{ij} \times \delta t \approx \mathbb{P}(&\exists \text{ transition between time 1 and } 1 + \delta t \mid X_0 = i) \\ &\times \mathbb{P}(\text{transition is to state } j \mid \exists \text{ transition from state } i). \end{aligned}$$

This is why the formulated transition 'probability' can exceed 1. There is no justification for the above. If the authors really wanted to mimic a continuous-time specification via a rate matrix then the correct formulation would be (for $i \neq j$)

$$\begin{aligned} P_{ij} = \frac{\gamma_{i,j}}{\gamma_i} e^{-\gamma_i} = \mathbb{P}(&\exists \text{ transition between time 0 and } 1 \mid X_0 = i) \\ &\times \mathbb{P}(\text{transition is to state } j \mid \exists \text{ transition from state } i). \end{aligned}$$

It would be fine to obtain inferences on the rate parameters here. Though why this would be better than using the standard formulation would need to be explained.

Equation (9)

It took me a while to make sense of most of this and I was still unable to make sense of one part. I am also unable to work out all the details of the MCMC algorithm from it - the steps should be succinctly summarised. The equation claims to give the joint distribution of all the information of interest, given the data. My remaining issue is with the right hand side,

but let's first consider the left. I think the left hand side should read

$$[\{X_t\}_{t=0}^T, \{\lambda_L, \tilde{\lambda}_H\}_{t=0}^T, P, \{N_{Lt}, N_{Ht}\}_{t=0}^T \mid \{N_t\}_{t=1}^T] \text{ not } [\{X\}, \{\lambda_L, \tilde{\lambda}_H\}, P, \{N_{Lt}, N_{Ht}\} \mid N_t]$$

If the authors really do not want to keep the set indices then that is acceptable as long as it is stated somewhere so that it is clear exactly what each set is, but surely it is conditioned on all the data, not just the point at some (what?) time t?

Secondly, the right hand side. The first term in the product is $[N_t|N_{Lt}, N_{Ht}]$. Firstly, surely this also depends on X_t? Secondly, it is a *deterministic* function of N_{Lt}, N_{Ht} and X_t, yet [stuff] is being used to indicate 'the distribution of' (this should also be explicitly stated before the notation is used); so what does this mean?

The third term in the equation is $[N_{Ht}|X_t = H, \lambda_L, \tilde{\lambda}_H]$. This does not seem right and it's because this term appears for *all* t yet what the term means is not even defined when $X_T = L$. What is needed is a term which is as specified when $X_t = H$ and 1 when $X_t = L$.

Minor points

- Abstract: chamber-level analysis (hyphen needed).
- 'overfit' is the first person singular. In several places it is used as third person past perfect or present perfect e.g. 'state process that is overfit', 'predictions were overfit'. This should be 'is overfitted' and 'were overfitted.'
- P6L6. Why have an initial distribution of $(0.5, 0.5)$? Starting from the stationary distribution of X_t would be more usual, unless you have particular knowledge (in which case, it would be helpful if you could share this with the reader).
- After equation (1), θ_ℓ does nto contain prior probabilities, it contains the parameters for the Dirichlet prior on the probabilities of row ℓ.
- Equation (23), the distribution of X_t depends on X_{t-1} as well as P.
- P11 'The hyperparameters detailed above' - it would be more helpful to reference the specific equations where you define the prior distributions (e.g. (4) and (5) for the λs; this already uses all 4 of the hyperparameters; where the priors were defined for the γs was not obvious).

- Having read section 3.4, I am still non entirely sure why the particular prior hyperparameters in (13) were chosen. Also in (13) the number 120000 is used, whereas in the wording below 12000 is used.