JCGS-18-220.R1: Efficient Parameter Sampling for Markov Jump Processes

1 Summary

I thank the authors for the detailed response to my previous comments (made in `review.pdf`) in the previous round of reviews. I also wish to apologise for my rather convoluted (and clearly unsuccessful) attempt at explaining how the ideas in the present manuscript relate to previous work on non-centred parametrisations (NCPs) as discussed in Papaspiliopoulos et al. (2007) and in particular the second method from Roberts et al. (2004, Section 3.6). I still believe that making this connection more clear would greatly benefit the reader and I have outlined the main points in the *General Comments* section below. However, I am happy to recommend publication in the present journal (assuming the general comments given below are addressed) because the authors have convinced me that the present work is sufficiently distinct from Roberts et al. (2004, Section 3.6) in two ways:

1. The present work assumes a finite (and sufficiently small) state space so that the states can be integrated out analytically via standard forward-backward recursions for hidden Markov models. The main proposed algorithm (Algorithm 4) therefore induces a Markov chain on the (marginal) space of the parameters and the jump times. I expect that this marginalisation can lead to substantially improved mixing in many models relative to the algorithm from Roberts et al. (2004, Section 3.6) which induces a Markov chain on the (larger) space that also includes the states and which may not mix well outside the class of Ornstein–Uhlenbeck type models considered in that work.

2. The present work includes a geometric ergodicity proof which, in particular, holds without the need for further updates of the path conditional on the parameters (in contrast, updates of the path conditional on the parameters seem to be needed in Papaspiliopoulos et al. (2007); Roberts et al. (2004)).

2 General comments

1. The main point in my previous review was that Algorithm 4 (rather than Algorithm 3) can be seen as a version of the non-centred parameter update from the second method outlined in Roberts et al. (2004, Section 3.6) in which the states are integrated out analytically rather than sampled (I have formalised this in Point 2 below). While, as stated in the *Summary* above, this marginalisation can dramatically improve mixing (and other properties) of the chain, the "symmetrisation" idea at the heart of current work in Algorithm 4 addresses exactly the same issue: circumventing correlation between parameters and latent variables under the prior, as

the NCPs in Papaspiliopoulos et al. (2007); Roberts et al. (2004). More specifically, note that a NCP immediately implies such a "symmetrised" Metropolis–Hastings (MH) update. Thus, I believe that explaining this link between NCPs and "symmetrisation" would ultimately strengthen Section 5 (as well as the motivation at the end of Section 4) by relating it more clearly to the existing body of research on NCPs.

For instance, it may then be possible to draw upon results from the literature on NCPs in order to give guidance to the user under which conditions the non-centred/symmetrised parametrisation (e.g. Algorithm 4) outperforms the standard centred/non-symmetrised parametrisation (e.g. Algorithm 2); or, to argue whether a mixture of centred and non-centred updates could be sensible (Yu and Meng, 2011).

2. As mentioned above, Algorithm 4 can be viewed as a version of the second non-centred parameter update Roberts et al. (2004, Section 3.6). The only difference is that Algorithm 4 integrates the states out analytically which is practical here since the state space $\mathcal{S}$ is finite and sufficiently small. To show this connection, the steps below outline the second non-centred parameter update from Roberts et al. (2004, Section 3.6) when applied to Markov jump processes (MJPs) (with all undefined notation is as in Algorithm 4, e.g. in particular, $\Omega = \Omega(\theta, \vartheta)$).

   Given an MJP path (S, T) and parameters θ, perform the following steps

   (a) Sample $\vartheta \sim q(\,\cdot\,|\theta)$.

   (b) Sample $U = U_{1:|U|} \sim \text{PoissProc}(\Omega, A_{S(t)}(\theta))$ (the latter being a Poisson process on $[0, t_{end})$) as in Alg. 4.

   (c) Sample a vector of $|U|$ independent and identically distributed (IID) standard uniform random variables $\widetilde{R} = \widetilde{R}_{1:|U|} \in [0,1]^{|U|}$.

   (d) Transform $S = S_{1:|S|} \to \widetilde{S} = \widetilde{S}_{1:|S|} \in [0,1]^{|S|}$ by setting $\widetilde{S}_i := A_{S_{i-1}S_i}(\theta)/A_{S_{i-1}}(\theta)$.

   (e) Set $W = T \cup U$ as well as $\widetilde{V} = \widetilde{S} \cup \widetilde{R}$ and sort the elements of W in ascending order; also order the elements of $\widetilde{V}$ accordingly (i.e. by applying the same permutation to the indices of its components).

   (f) Accept (S', T', ϑ) with probability

$$1 \wedge \frac{P(X|W,\widetilde{V},\vartheta,\theta)P(\vartheta)q(\theta|\vartheta)}{P(X|W,\widetilde{V},\theta,\vartheta)P(\theta)q(\vartheta|\theta)} \underbrace{\frac{P(W,\widetilde{V}|\vartheta,\theta)}{P(W,\widetilde{V}|\theta,\vartheta)}}_{=1},$$

   where

$$P(X|W,\widetilde{V},\theta,\vartheta) = P(X|S,T,\theta)$$

   is equal to the likelihood of the MJP iff $(W, \widetilde{V})$ and (S, T) are satisfy the following recursion.

   Set $j = 1$. For $i = 1, \ldots, |W|$ (and with obvious modifications for $i = 1$ and $i = |W|$), if $\widetilde{V}_i \leq A_{S_{j-1}}(\theta)/\Omega$ then set $T_j = W_i$ and $S_j = F_i^{-1}(1 - \widetilde{V}_i/\Omega)$, where F_i denotes the CDF associated with the discrete

distribution on $\mathcal{S}\backslash S_{j-1}$ with probabilities $\{A_{S_{j-1},k}(\theta)/A_{S_{j-1}}(\theta)\}_{k\in\mathcal{S}\backslash S_{j-1}}$. The quantities S' and T' are determined in the same way but with θ replaced by ϑ.

Some things to note.

- This update is "symmetrised" (because the prior on $P(W,\widetilde{V}|\theta,\vartheta) = \text{PoissProc}(\Omega(\theta,\vartheta))(W)\text{Uniform}(0,1)^{\otimes|W|}(\widetilde{V})$ is symmetric in (θ,ϑ)). Though, as pointed out regarding Algorithm 4 by the authors in their rebuttal to my previous review, it does not strictly speaking constitute an NCP (because the above-mentioned prior still depends on the parameter(s)). However, this construction can be formally turned into a NCP by further extending the space by drawing $U \sim \text{PoissProc}(\bar{\Omega}, A_{S(t)}(\theta))$ in Step b, where $\bar{\Omega} \coloneqq \sup_{\theta,\vartheta}\Omega(\theta,\vartheta)$ but then removing/thinning the ith component from $(W,\widetilde{V})$ if $\widetilde{V}_i \geq (\bar{\Omega}-\Omega)/\bar{\Omega}$ and rescaling the remaining components of $\widetilde{V}$ by the factor $\bar{\Omega}/(\bar{\Omega}-\Omega)$ ahead of Step f. Note that while this formally allows us to view the construction as an NCP, it leads to computationally exactly the same algorithm as above (because the additional points need not be sampled in the first place).
- From the parameter update outlined above, we immediately obtain Algorithm 4 if we integrate $\widetilde{V}$ out analytically, hence demonstrating the abovementioned connection between the work on NCPs and the "symmetrisation" idea in the present manuscript.
- I hope that this makes the connection between the work on NCPs and the "symmetrisation" idea which I was asserting more clear. However, I do stress that this algorithm likely performs very poorly in most MJPs and would not actually suggest using this update for such models. That is, marginalising out the states rather than sampling them is likely to be crucial for the MJP models considered in the present manuscript.

3. I do not see how Algorithm 3 is an NCP as mentioned on Page 10 except in the trivial case that $\Omega \geq \sup_\theta \max_i A_i(\theta)$. As far as I can tell, it is a standard Metropolis-within-Gibbs update for the parameters which exploits the fact that the states can be integrated out analytically. It is possible that I am missing something here but if not, I do not think Section 4 needs to contain references to NCPs.

3 Specific comments

1. Throughout the document: The notation $S(t)$ is still used in many places in the manuscript to refer to two different quantities which is confusing:
   (a) sometimes $S(t) = (s_0, S, T)$ denotes the entire MJP path over the interval $[0, t_{end})$;
   (b) at other times, $S(t)$ denotes the value of the MJP path at some specific time $t \in [0, t_{end})$.

I think it would help the reader to find a different notation in Case (a). Maybe writing $S([0, t_{end}))$ instead of $S(t)$ when referring to the entire path would be more clear.

2. Algorithm 1, Step 1: (S, T) -> (s_0, S, T) (?)

3. P. 7, l. 44: is there a redundant closing parenthesis here?

4. P. 10, l. 52: do -> to

5. P. 27, l. 39: distribution -> distributions

6. Bibliography: there are still some typos in the bibliography:

   (a) some entries have a redundant period right after the article title

   (b) inconsistent use of abbreviated/non-abbreviated journal names

References

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, 22(1):59–73.

Roberts, G. O., Papaspiliopoulos, O., and Dellaportas, P. (2004). Bayesian inference for non-Gaussian Ornstein–Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):369–393.

Yu, Y. and Meng, X.-L. (2011). To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570.