Revised manuscript of 'Fast and Scalable Gaussian Process Modeling with Applications to Astronomical Time Series'

Daniel Foreman-Mackey, Eric Agol, Sivaram Ambikasaran, Ruth Angus

I appreciate the authors for handling almost all of the major and minor issues that I posed last time. The revised manuscript has become crystal-clear in terms of statistical modeling. This time, I have only one major issue and a few minor issues. Once the authors handle these, I believe there would be no more statistical issues. The authors should use their own judgment in addressing (or ignoring) the points mentioned below.

Major issue:

1. MCMC convergence: Reporting the effective sample size does not guarantee the convergence of the Markov chains. It is hard to say that the Markov chains (at least) nearly converged to the stationary distribution with such small effective sample sizes in Sections 6.4 and 6.5 because the effective sample size is way smaller than the total number of iterations[1].

   If the parameter surface is unimodal, the authors may want to try thinning the Markov chain to increase the effective sample size, i.e., by running a long Markov chain with length n and then by taking every kth posterior sample (iteration) to have n/k posterior samples. The auto-correlation will decrease as k increases, which leads to a larger effective sample size.

   However, this thinning may not be helpful for increasing the effective sample size when the parameter surface is highly multi-modal[2], e.g., in Section 6.4 and potentially in Section 6.5 (considering that it is difficult to get such small effective sample size in a unimodal case). In this case, the authors need a multimodal sampler, such as parallel tempering (Kelly et al., 2014), that helps Markov chains jump between modes frequently. For reference, there are various choices for a multimodal MCMC sampler other than parallel tempering.

   If the authors think that using a multi-modal sampler is beyond the scope and decide to leave these two examples with such small effective sample sizes, then I

[1] Let us consider a unimodal case. For n posterior samples, the effective sample size is defined as $n/(1+2\sum_{j=1}^{\infty}\rho_j)$ and the auto-correlation function plot is just displaying ρ_j's for $j = 0, 1, 2, \ldots$, where $\rho_j = Cov(\theta^{(t)}, \theta^{(t+j)})/Var(\theta^{(t)})$. Roughly speaking, when $\rho_j = 2^{-j}$, i.e., the auto-correlation function geometrically decreases, then the effective sample size is $n/3$

[2] The problem of multimodality is that both auto-correlation function and effective sample size can be misleading because both measures indicate evidence of good convergence when a Markov chain never move between modes. The authors, however, can justify using these measures if each Markov chain has visited all of the (at least known) modes.

recommend the authors present `celerite` in a fair way by clearly specifying two limitations; (i) the current inference in Sections 6.4 and 6.5 may not reflect on the targeted stationary distribution, i.e., may not be based on the converged MCMC, and (ii) `celerite` is designed to improve the scalability but not the convergence rate of the MCMC especially for multimodal cases and thus a multimodal MCMC sampler, e.g., parallel tempering used in Kelly et al. (2014), is required for a statistically appropriate inference in Sections 6.4 and (potentially) 6.5. Additionally, I think reporting the (average) acceptance rate of the posterior sample is another easy-to-report criterion for the MCMC convergence considering that it may be quite burdensome for the authors to display some figures related to monitoring the MCMC convergence.

Minor issues:

1. The last two sentences in the first paragraph of Section 2: In the first paragraph of Section 2, the authors seem to provide an overview of Gaussian processes from modeling to estimation. As for estimation, the authors introduce MLE as a point estimate and Bayesian posterior sampling as an uncertainty quantification. It does not appear natural to me because most people (I believe) do not use Bayesian method to quantify the uncertainty of MLEs. For a more meaningful overview, it may be better to mention both approaches; frequentist's MLE and its asymptotic or bootstrapping-based uncertainty quantification and Bayesian posterior inference such as an MAP (maximum a posteriori) estimate and simultaneous uncertainty quantification via posterior distribution. Or, since the authors adopt a Bayesian approach for the rest of the manuscript, it may be also reasonable to introduce an MAP estimate (instead of MLE) and simultaneous uncertainty quantification via posterior distribution.

2. "This model is often called an Ornstein-Uhlenbeck process (...)" just after Equation (13) on page 6: Would it be reasonable to mention that the Ornstein-Uhlenbeck process is also called a damped random walk process because astrophysicists may be more familiar with the latter?

3. Equation (14) on page 6: The subscript k of k is confusing. Please consider using a different subscript, e.g., i or l.

4. "(...) we can derive a higher performance algorithm by restricting our method to positive definite matrices" in the second paragraph of Section 5: Here I am concerned about three things. (i) What is the meaning of restricting a method to matrices? (ii) Do the authors mean that the better performance is for a class of *celerite* models with semi-separable and positive-definite covariance matrices? Isn't a covariance matrix almost always a positive-definite matrix in practice? Considering that a covariance matrix is positive-semi-definite by definition, did Ambikasaran (2015) focus on a case with a semi-separable and positive-semi-definite covariance matrix? I do not understand why the positive-definiteness is

the key to the newly argued improvement since it is almost always the case (even in the previous manuscript). (iii) In addition, I feel that it is better to briefly mention here what the better performance means; I had to read all the technical details until the end of Section 5.1 to learn the improvement.

5. "(...) only formally valid under specific sets of assumptions, we cannot recommend their use in general." at the end of Section 5.4: What do the authors mean by 'formally valid under specific sets of assumptions'? I am curious because there is no reference cited. Do the authors happen to intend that physical motivation is more important than statistical motivation in model selection, i.e., some *celerite* models whose AICs or BICs are not optimal can be more meaningful in modeling physical phenomena?

6. "separable uniform priors" on the first line on page 17: I think the authors mean independent uniform priors. Am I correct? Can the authors use the words, 'separable' and 'independent', alternatively in the text to make it clear?

7. "The top left panel of Figure 4 shows the conditional mean and standard deviation of the MAP model over-plotted on the simulated data (...)" in the middle of the third paragraph in Section 6.1: The top left panel of Figure 4 does not exhibit the conditional mean and standard deviation, I mean, numbers. Maybe these are denoted by the blue contours, considering the caption.

8. "We initialize 32 walkers by (...)" in the fourth paragraph of Section 6.1: What are the walkers? Do the walkers mean parallel Markov chains (i.e., 32 chains are independently run)?

9. "effective model" repeatedly used in Section 6: What do the authors mean by 'effective model'? I have no idea about when statisticians call a model an effective model. Is this a jargon in Astrophysics? Then, when and in what sense do astrophysicists consider a model as an effective model?

10. Is the reported effective sample size (e.g., 2900 independent samples in Section 6.3) the average effective sample size that is averaged over the effective sample sizes of the sampled parameters? It is not clear in the text.

11. The right panel in Figure 7: It may be more informative if the authors display the uncertainty of the reported rotation period, i.e., ± 0.15, in some ways, e.g., two dashed vertical lines.

12. "This model requires about 10 CPU minutes to run the MCMC to convergence." in the eighth paragraph of Section 6.4: How do the authors know that the MCMC converged after 10 CPU minutes? I cannot see any evidence. Do the authors mean that it took 10 CPU minutes in total to run the 20,000 iterations?

13. The panels for the marginal distributions in Figure 9: It may be more informative if the authors display the published values of the asteroseismic parameters and their uncertainties in the marginal posterior distributions.

14. "(error bar)" in the caption of Figure 9: Isn't it clearer to say '(blue error bar)'?

15. "(...) because, in this case, do not set the deterministic (...)" in the first paragraph of Section 6.5: Maybe the authors intend "(...) in this case, we do not set (...)."

16. Table 6: It will be even more convincing to exhibit 'evaluations' and '$N_{\rm eff}$' of `direct`.

17. "(...) and the parameters a_j, b_j, c_j, and d_j can be easily computed analytically." in the second paragraph of Section 7: If it is easy to compute, then what about specifying the closed-form equations of the parameters in a footnote, e.g., $a_j = ...$, $b_j = ...$, etc., clearly connecting the CARMA and *celerite*?