Response to Reviewers Comments on "A Preferential Attachment Model for the Stellar Initial Mass Function" [EJS1807]

Cisewski-Kehe, Weller, Schafer

Thank the reviewers for their thoughtful comments. Below we respond to the questions and comments. Additions to the manuscript appear in **bold** font and sentences and statements removed will have a ~~strikethrough~~.

Reviewer 1 Comments and Responses

1 Summary

The authors develop a preferential attachment model for estimating the stellar IMF. They use Approximate Bayesian Computation to derive posteriors for the model. Use of this methodology (PA + ABC) within astronomy has the potential to improve our understanding of the stellar IMF. Further, the application of this methodology to challenging data sets, such as those studied here, will improve statisticians understanding of the methods and help generate further methodological advances.

I have some questions/suggestions mostly regarding the Preferential Attachment model and ABC.

Response: Thank you for your careful reading of our work, and your helpful comments and questions. We believe your feedback has helped us especially to improve the presentation and clarity of the manuscript.

2 Preferential Attachment Model

- **Page 8: More references to statistics papers studying / developing PA models would be helpful. Is PA related to models such as the Chinese Restaurant Process? What sort of computational methods ABC or otherwise have been used for other PA models? Could those methods be used here, as competitors to ABC?**
  Response: We have added a paragraph at the start of Section 3 which lists some of the commonly used methods for performing inference with PA models. The term "preferential attachment" is fairly broad and refers to a wide range of processes, including the Chinese restaurant process (which itself encompasses many generalizations).

- **Page 8: In Equation 3.1 $\pi_t = \min(1, \alpha)$ but two sentences later the range of α is restricted to be [0, 1).**
  Response: This is an excellent point. We have removed the references to π_t and updated the text surrounding the noted equation.

- **Page 8: The π_{kt} are the probability that some mass joins existing star k at time t. But the 3.1 definition of π_{kt}, looks like it could be greater than 1. In particular it has units of some mass. Would**
$$\frac{\pi_{kt}}{\pi_t + \sum_j \pi_{jt}}$$

make mores sense for the probability of joining k?
Response: Yes, that is correct. We have updated the noted equation so the definition of π_{kt} has the precise definition instead of a proportionality.

- **Page 8 "The generating process is complete when the total mass of formed stars reaches M_{tot}" The total mass of stars will never equal exactly M_{tot} because the mass additions are from a continuous exp(λ) distribution, correct? Do we stop when the total mass first exceeds M_{tot}?**
Response: Yes, we stop when the total mass first exceeds M_{tot}. We have updated the text to clarify this.

- **Figure 2 and 3: Is it reassuring that the PA model can reproduce Kroupa (2001) and Chabrier (2003a,b), given that neither of these models is necessarily correct?**
Response: Yes, we consider it a positive feature of our model that it can capture the form of the Kroupa (2001) and Chabrier (2003a,b) models as special cases. In practice, both these model forms have been used for real star clusters so we believe this feature will help the astronomy community to accept our proposed model.

- **Page 15: The phenomenon of "all the mass that has to be distributed to the already existing stars (rather than forming a new star) tends to be assigned to the same, most massive star" sounds physically unrealistic. Wouldn't there be some upper limit, such as the star collapsing on itself and becoming a black hole? The PA model proposed seems related to clustering ideas such as Dirichlet process mixture models / Chinese restaurant processes. It has been noted in these models that "it is well known the DPMs favor introducing new components at a log rate as the sample size increases, and tend to produce some large clusters along with many small clusters." (see "Reducing over-clustering via the powered Chinese restaurant process" by Lu, Li, Dunson on arXiv) Is this the same/similar phenomenon happening here?**
Response: We agree that there are limits on how large the largest star could be, and that some parameter combinations will yield unphysical models. The point being made here is that as the parameters γ and α are varied, the model is able to produce a wide range of different behaviors, including the extremes effect described in the above quote. We believe that this model is more flexible than the classic DPM/Chinese restaurant models because of these additional parameters.

- **Page 25: "A goal of the proposed model and algorithm is to begin making a statistical connection between the observed stellar MF and the formation mechanism of the cluster, not that the proposed model shape is superior to the standard IMF models." The proposed model is superior in that we may not know that Chabrier or Kroupa IMF shape is correct but the PA model is flexible and includes both these models as (approximate) submodels. So rather than fitting both Chabrier and Kroupa and doing model selection, we just fit PA and interpret the posteriors. Right?**
Response: Yes, that is correct. Rather than attempting model selection between two model forms that may both be incorrect, our more general model and the resulting posteriors can be used for interpretation.

3 Analysis / Comparison of ABC algorithm

- **Section 4: What is the alternative to using ABC in Section 4? Why are these alternatives (such as MCMC or variational Bayes) impractical for this problem? It seems to me that evaluating the likelihood with the measurement errors 3.5 would require convolving the likelihood with normals of different variance and then renormalizing separately for each observation (star). The would require n 1-d integrals to evaluate the likelihood once. Is this true? Is this why one cannot use MCMC?**
  Response: Yes, it is difficult to conceive of how a likelihood could be written in this situation. Because of how the cluster evolves over time, the individual observed masses cannot be thought of as "iid draws" from some distribution parameterized by the unknown parameters. The observational effects, including measurement errors, only further complicate the form of the likelihood. While variational Bayes is an intriguing approach, it relies upon having a family of distributions that one believes the posterior has a sufficient approximation to. It seems more difficult to construct such a family in applications such as this with complex physical parameters and models.

- **Figure 7 a) Degeneracies in posteriors often cause problems for convergence of MCMC algorithms such as Gibbs or Metropolis. Would there be any reason to worry about convergence for sequential ABC? Is it possible to reparameterize to make the posteriors less dependent across these parameters?**
  Response: While it is generally a good idea to think about possible degeneracies, in our proposed model, degeneracy does not seem to be an issue. In the different empirical settings we considered, and to the extent it was possible to evaluate this, we did not notice any convergence issues for the sequential ABC algorithm. Unfortunately we are not aware of a way to reparametrize our model to reduce the dependency between the parameters. We had investigated different forms of the model previously, with an expanded parametrization, but it had seemed to result in a model that was not identifiable. With the current model form, we are able to get posteriors that have high probability around the input parameter values.

- **Section 4: How close is the ABC approximation to the actual posterior? Are there any ABC convergence diagnostics available?**
  Response: In Section 4, the form of the true posterior is unknown (since the likelihood function is unknown) so we are not able to assess how close the ABC approximation is to the true posterior. This is the common situation for ABC work since we are generally dealing with settings without a clearly specified likelihood function. There are different approaches for assess convergence. One idea is outlined in Ishida et al (2015, *Astronomy and Computing*) set a convergence threshold based on the number of draws that is needed for the algorithm to accept the requested N particles. For example, if their convergence threshold is set to 0.20, then the algorithm would stop once it takes at least $N/0.20$ draws to get the N required particles. In the Ishida et al. (2015) example in Section 4, they used a convergence threshold of 0.01. While we did not use this approach for assessing convergence, if we had used it, for example, for our Bate astrophysical simulation model example in Section 5, our convergence threshold would have been set at slightly less than 0.00021 since in the last step 4,785,511 draws were needed to achieve our desired 1,000 particles.

4 Other Issues

- **Section 3.1.1. I am trying to understand the relevance of this section to the rest of the work. "the power law model is a prevalent assumption in this application" If this assumption is (approximately) correct does the conclusion "the power law fit degrades quickly for γ outside (0.5, 1.5)" imply that priors on γ could/should put most mass in**

this range? More generally, connecting this section more strongly with the rest of the work would be helpful.
Response: [[**TODO:** Jessi will add comment]]

- **Page 4: "Focusing on the upper part." the "upper part" is large m? maybe "upper tail". does M_{min} define the left boundary of the upper part? if so, then why would "c" be chosen to make f_M a valid pdf? wouldn't f_M integrate to less than 1?**
Response: Yes, we mean the "upper tail" here. The text has been updated to clarify the purpose of the example provided in this paragraph. In particular, we explain that we are only providing an illustration of the Salpeter (1955) model, which is included to define the form of the power law model. Rather than using M_{min}, we specify $0.5M_{\odot}$ as the start of the upper tail. Since the Salpeter (1955) model only consider this section of the mass range, c is specified so we have a proper probability density (i.e., so that it integrate to 1). We reference the later definition of the Kroupa (2001) for the specification of the broken power law model, where the constants analogous to c are chosen so the model is continuous and integrates to 1.

- **Figure 3: I would suggest limiting the x-axes to the support of the posteriors, rather than the support of the priors. The current scaling may be masking differences in the distributions, especially for the a), the λ^{-1} parameter. I don't think the y-axis need to be the same for all densities. These changes will make comparison across plots more difficult, but I think they are worthwhile because the more important comparison is between the Kroupa and Chabrier model for a particular parameter. Similar comment for Figure 6, especially 6 a).**
Response: [[**Jessi:** I can respond]]

- **Is the completeness function equivalent to some form of probabilistic truncation? (I am using the survival analysis definition of truncation) If so, is there research on this within the survival analysis literature? If so, could the authors provide some citations?**
Response: While one can think of the completeness function as a form of probabilistic truncation, it does not seem to have a good analogy in the survival analysis setting. The closest related concept that we are familiar with may be the *survival function*, which specifies the probability of survival beyond some specified time (i.e. one minus the CDF for that value). For the completeness function, the idea is different enough that it may lead to confusion if we attempt to make a connection. The completeness function does not specify the probability of survival beyond a particular time (or mass, in the IMF setting), but, rather the "probability of survival" for that specific mass. It is possible that this notion of a completeness function could be an interesting object of study for those in the survival analysis community.

- **Is the completeness function 3.4 applied to the data before the measurement error 3.5. If so, why? (I don't see a clear reason for either ordering, perhaps the instrument somehow determines this.)**
Response: The completeness function is applied before the measurement error is added. Our reasoning for this ordering is because the intrinsic brightness of the star would determine our ability to detect it. However, an argument could be made to apply the measurement error before the completeness function since our ability to detect the star also depends on the instrument.

- **In Equation 3.6, is $f_M(m|\theta)$ the stellar initial mass function after accounting for measurement error (3.5) and completeness (3.4)? Shouldn't there be a proportionality, rather**

than equality due to need for normalization? Is the lhs of 3.6 the mass function (MF) referred to in the subsequent section.
Response: Equation 3.6 is referring to any generic IMF model form, $f_M(m \mid \theta)$, and is only meant to illustrate how to incorporate the effect of aging. A statement has been added below Equation 3.6 to clarify this, and the equality was turned into a proportionality (though one can also assume that the $f_M(m \mid \theta)$ includes the normalization, but we agree that it is clearer if we make it a proportionality). In subsequent sections, the form of the mass function is the result of the simulation plus any observational effects included (which depends on the simulation study...some include the observational effects and some do not). [[**Jessi:** Chad, do you think the above question requires additional changes in the text? I wonder if we should add an explicit statement clarifying how we write the IMF model?]]

- **Page 18: "4.2.1. Simulated data with observational effects" So the simulation in 4.2, before 4.2.1, does not have these effects? This is not explicitly stated and so is rather confusing. Perhaps have 4.2.1 be the simulation without effects and 4.2.2 (currently 4.2.1) the simulation with effects. At the beginning of 4.2 you could let readers know there will be two simulations.**
  Response: Thank you for this suggestion. We have added text at the beginning of 4.2 to note that the first part of the study does not include observational effects, and the next sub-section does include observational effects.

- **Effective application of ABC requires a good distance function ρ, tolerances ϵ, kernel, etc. Were there findings, perhaps qualitative, about how to choice these quantities that could be useful for other ABC practitioners? Perhaps the authors could summarize these findings in the conclusions, e.g. a few sentences of the form "In agreement with other studies, selection of a proper distance function was the most challenging... " or "In contrast to so and so..."**
  Response: [[**Jessi:** I can respond]]

- **Link / references to code and / or data to reproduce results?**
  Response: [[**Jessi:** I can respond]]

Reviewer 2 Comments and Responses

The manuscript "A Preferential Attachment Model for the Stellar Initial Mass Function" introduces a generative simulation model (depending on three parameters $(\alpha, \gamma, \lambda)$ for the stellar initial mass function science. ABC is used to obtain approximate posterior samples of these three parameters on simulations that incorporate complicated observational effects which are difficult to include without the ABC technology. In general I like this manuscript and think EJS is a good fit for publication. There appears to be two main contributions. The first is the introduction of the preferential attachment model (apparently a Yule-Simon stochastic model) as a more direct approach – compared with parametric stellar initial mass function – for modeling the distribution of star masses in a cluster. It seems as though this could be a useful tool for astronomers to have in their modeling tool box. I would also predict the generative procedure is general enough for astronomers to be able to add tweaks to the stochastic procedure which could incorporate other physical effects. The second contribution is the exposition of how ABC can be used with this generative model, along with complicated observational effects. As for revisions, I think the paper needs to be tighten up a bit. The writing can be somewhat wordy at times and there seems to be a couple superfluous sections

that could probably be dropped. Also, there are 17 figures, some with multiple panels and a lot of white space. I would encourage the authors to work on filtering out the diagrams which do not significantly contribute to the main points of the paper or help the reader understand the exposition. For example, Figure 1 is mentioned in one sentence on page 5 and doesn't seem to give the reader much more than isn't already written in the proceeding text. Another example comes in Figure 3 where the x-axis scale is chosen so wide that it's hard to discern multiple densities. Why have all that white space? Some of this boils down to reader preference, and of course the authors can't please everyone, but I think some editing iteration could really improve the paper. I've included other comments below (in no particular order). Again, many correspond to personal preferences which I think would improve the paper. I'll leave it up to the editors to evaluate what of these comments are important for publication.

Response: We thank Reviewer 2 for the helpful and thoughtful comments. We have made a number of updates to the manuscript, including removing and moving sections and figures to decrease the length and (hopefully) improve the readability of the manuscript. The specific updates are noted below.

Further comments

1. **Only after I read the Yule-Simon simulation procedure did I actually get a clear picture what the stellar initial mass function is: a continuous density describing the histogram of the list of star masses in a cluster. The exposition preceding somehow never quite gets to the point that an IMF isn't necessarily a physical model describing how the star masses are formed in the dynamic evolution of the cluster, but rather a non-physical summary of such a list of star masses. Also that the PA model avoids specification of a parametric IMF using a simple non-physical simulation procedure to create such a list. Perhaps this should be emphasized someplace in the first couple paragraphs of the paper.**
   Response: Statements have been added to the first and third paragraphs of the introduction to address these suggestions. The first paragraph has the following addition: "The IMF can be thought of as a continuous density describing the distribution of star masses that initially form in a stellar cluster." And the third paragraph has the following statement added: "[...] and avoids specification of a parametric model form by using on a new simulation model."

2. **In (3.1) why write $\pi_t = min(1, \alpha)$ then restrict $\alpha \in [0, 1)$ in the next sentence? Why write π_t depending on t? Why even have a new symbol for what amounts to α anyway?**
   Response: This is an excellent point. We have removed the references to π_t and updated the text surrounding the noted equation.

3. **Many of the estimated posterior densities which summarize the ABC output look to have small scale local fluctuations. I'm guessing these bumps and wiggles are just artifacts from the finite number of posterior samples, but I think it runs the risk of suggesting to the reader that ABC is less accurate than it actually is. It would really strengthen the authors main points if the plots didn't have as much visual finite sample fluctuation. Perhaps step-line histograms with wide enough bins would give a sufficient visual description of the posterior samples and also suggest to the reader that the actual samples may be very accurate but not necessarily the histogram derived from it.**

   On a related note, I wonder if Rao-Blackwellisation can be used here for reducing the sample variability in the posterior density estimates from ABC. The final posterior

samples $(\theta_1, \gamma_1, \alpha_1), \ldots, (\theta_n, \gamma_n, \alpha_n)$ carry along with them simulated auxiliary variables used in the process of generating a data sample that can be used to, effectively, make a better kernel smoother estimate of the marginal density. In particular, let N_i denote the number of stars generated by the particular data generating process associated with the ABC sample $(\theta_i, \gamma_i, \alpha_i)$. Notice that $N_i \sim \textbf{Poi}(M_{tot}/\lambda_i)$, or at least if your using M_{tot} to denote the upper limit for the mass of the system. Now instead of plotting a histogram or kernel density estimate based on the approximate posterior samples $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$, the rao-blackwell density estimate would be

$$\lambda^{-1} \to P(\lambda^{-1} \mid \textbf{data}) \approx \frac{1}{n} \sum_{i=1}^{n} P(\lambda^{-1} \mid N_i)$$

where $P(\lambda^{-1} \mid N_i)$ is easy to compute using the Poisson likelihood and the prior π. It is not entirely clear to me that this will be easy for the marginal posterior density of the other parameters. If it is easy and it makes the plots in this manuscript less variable, it might be worth while adding it to the paper. However, I think it would be certainty sufficient to instead simply clean up the plots by replacing the kernel density estimates with an appropriately binned histogram.

Response: Thank you for this well thought-out suggestion. We have decided to leave the posterior plots with the noted wiggles. Since the posteriors are based on a particle approximation, these sorts of wiggles are expected. Additional smoothing of the KDE would remove the wiggles, but we feel that it would be a misleading representation of the posteriors to suggest they are smoother than they actually are. We also have left the posteriors as KDEs since histograms would also appear variable and are known to be worse approximations to true densities than a KDE (e.g. see Wasserman 2006, *All of nonparametric statistics*).

4. **Section 3.1.1 seems like a bit of an afterthought. Not really sure I follow what I should get out of it. Is this probing the flexibility of the PA model or the ability of the data to constrain γ. If the main point of this section is something of direct relevance to astronomers, then I would suggest being a bit more clear and to the point what the authors are trying to get accross. Otherwise, I would suggest dropping it altogether.**
   Response: [[**TODO:** discuss]]

5. **In the beginning of Section 3.2 the authors write: "*The PA model describes the formation of a star cluster at initial formation. However, we are not generally able to observe the star cluster after initial formation due to observational uncertainties, measurement uncertainties, and aging and dynamical evolution of the cluster.*"**

   This statement confuses me a bit. It seems to suggest that we only observe the initial formation of star clusters. I would expect the opposite, i.e. that the star clusters we observe are a mix of old and new clusters that have developed over different time ranges. Perhaps the authors are getting at a selection effect where the old star clusters are more dim and are effectively censored due to the sensitivity of our instruments. Either way, this paragraph could use some cleaning up.
   Response: [[**Jessi:** I can respond]]

6. **In display eqn (3.4) the letter 'm' appearing on the left hand side should be italic.**
   Response: Thank you for pointing this out; we have fixed this.

7. **I think Section 4.1, which describes the ABC sampling algorithm, can go into an appendix. In fact, I would remove Section 4 (titled "Methods") altogether, moving Section 4.1 to an appendix and Section 4.2 to the next section on simulations. This will allow the authors to have all the simulations in one place and can sharpen their main points and conclusions.**

   Regarding the three simulations presented in the paper (currently given in Sections 4.2, 4.2.1 and 5), do the authors really need a preliminary simulation study, given before 4.2.1, which doesn't include observational effects? Why not just include two sets of simulations: (1) using the PA model to generate the data which includes observation effects; (2) using the astrophysical simulation to generate the data (also including observational effects). Reducing the length of the exposition, number of plots to examine and focusing on the main conclusions seems like it would greatly improve the manuscript.
   Response: [[**TODO:** Jessi will clean this up.]]