JAS1101H Winter 2020
Summary of Peer Feedback on Progress Reports

March 8, 2020

Group: Liu, Ngo, and Starkman

Below is the verbatim peer feedback on your progress report.

1. **In 4-6 sentences, describe the group's project.**

   - It is important to understand whether Intermediate-Mass Black Holes (IMBHs) dwell at the centre of globular clusters, as this would have significant implications for stellar populations and galac- tic dynamics. This group plans to estimate whether globular clusters as a population show a statistically significant preference for housing an IMBH. This will be accomplished by looking at the clusters' stacked velocity and stellar density distributions, as measured by the GAIA survey. Cluster membership will be determined using data cuts in parallax and proper motion, along with algorithms like HDBSCAN, followed by Bayesian parameter estimation to determine whether an IMBH is likely to exist in the clusters.
   - The group project by the authors above, hypothesise the possible existence of IMBHs and intend to show if the presence of IMBH in the centre of the GC population is statistically relevant, thus supporting the claim. The IMBHs are said to be important for the GC dynamics, as such the authors estimate 1velocity/proper motion distribution relative to scale radius or X/Y component, in order to test the relationship between the mass of black hole (M_{BH}) and velocity dispersion ($\sigma - v$) (so are IMBH present). However, as mentioned in the report, data filtering and processing are necessary in order to minimize the number of outliers (get cluster members) as well as remove error noise from the population. The author's use HDBSCAN in order to classify the data from GCs population, as it was noted in the report to be a superior technique to the previously discussed ones in the literature. Finally, the authors plan to use Bayesian statistics in order to model the posterior of black hole mass, given the dynamical model (Jeans model) chosen by the group and fitting to the data distribution. The significance of the Bayesian procedure is that it allows to test the proposed hypothesis (using Kullback-Divergence of different posteriors) and estimate the significance of M_{BH} to $\sigma - v$ relationship.
   - This project aims to investigate the presence of an Intermediate Mass Blackhole (IMBH) in Globular Cluster (GC) centres. This will be statistically tested by fitting GC dynamical models including IMBHs to GC data through Bayesian Parameter Estimation. Currently, they have done the analysis on the GC 47 Tucanae using data from the Gaia catalogue and they wish to extend the analysis to ˜150 Milky Way GCs in Gaia. Further, they want to explore whether a black hole mass - velocity dispersion correlation holds for IMBH. Thus, this study will shed light on the role that IMBHs play in globular clusters, and help us better understand the co-evolution of black holes with astronomical systems on different scales.
   - You want to do a population study of GCs in the Milky Way to see if you can place any constraints on the existence of IMBHs in their cores. You plan to do this by rescaling various properties of the GCs in order to consider them as a whole. It turns out that this can be hard for various reasons: Gaia data completeness, and contamination resulting from imperfect isolating of the stars in a GC from the

background. You intend to estimate the posterior distribution of parameters for a model that includes a central IMBH using nested sampling.

- The group has set the ambitious goal of trying to provide statistically significant evidence for the existence of IMBH's at the center of globular clusters. They are using 6D phase space and photometric data from Gaia. In an effort to develop a robust pipeline, they focus their EDA on a single GC entitled 47 Tuc. Their analysis reveals missing data towards the center of the globular cluster resulting from instrumental limitations. They discuss a method for determining cluster membership called, "Vasiliev's cluster membership threshold" which they argue is potentially insufficient. They provide an overview of HDBSCAN clustering and an argument for adopting it as opposed to the Vasiliev criterion. Next, they suggest a model for Globular cluster dynamics based on 'Plummer Spheres' and provide a prediction for how such a dynamical model might be affected by the presence of an IMBH. They conclude with discussions of model limitations and plans for moving forward.
- The purpose of the group project is to check whether the presence of IMBH is statistically significant and whether the black hole mass-velocity dispersion correlation holds for IMBH. Data of around 150 GCs within the Milky Way and obtained from the Vasiliev catalogues is used; the data however requires filtering and processing. The team discussed the methods for clustering the GCs and how they plan to expand the analysis to other GCs.

2. **What parts of the progress report were confusing and could be clarified?**

- I don't fully understand how a statistically significant measurement of the presence of IMBHs can be determined when the authors admit that they are missing data from the centre of GC's. It just seems impossible from my initial reading of the report. Clarification on how their goals could still be achieved based on the admitted limitations would be helpful.
- Why is "dynamic modeling" an EDA? I may have misunderstood but this is more like a literature review to support the model choice. I think it would be a great addition if you can substantiate how the model is appropriate given your data (i.e. what assumptions are required and whether your data support these assumptions). Same applies for Section 3.4.
- I personally thought that the sections on background vs cluster membership were rather confusing. A lot of conflicting ideas have been presented and while I understand that it is not a straightforward problem, it would have been better to cut it down to the essentials with possible ideas at the end.
- I felt that the flow for the report was confusing, maybe it needed method section where the different
- I think the introduction was a bit confusing as it lacked astronomy background/context. For example, the paper assumes that the reader knows about IMBHs or SMBHs and the differences. I also first understood that "Due to mass segregation IMBHs quickly sink to the center of their GC" but later read that "there is no unequivocal observational evidence for the existence of IMBHs", this got me a bit confused. In addition, the concept of black hole mass - velocity dispersion correlation is introduced without proper explanation of a) what is this relationship b) why is it important to look at c) how can this "help us better understand the co-evolution of black holes with astronomical systems on different scale".
- Section 2.1 about the further processing of the data was pretty confusing, and I felt like some of your processing techniques and models could have been explained in a bit more detail. Section 3.1 comparing the uncut Gaia data to Vasiliev's cut data was a bit confusing, and I was unsure of which one was better. Hopefully in the final report, you will be using a custom filter, as hinted in the end of this section, but try to make the section explaining and comparing the filtered data a bit more clear.
- So are you moving forward using cluster membership as identified by HDBSCAN, or using Vasiliev's cuts? It's not immediately clear from Figure 5 that HDBSCAN is any more or less ideal.
- The HDBSCAN part was a bit confusing, it was hard to compare the work by Vasiliev and the new method. Also, the order of some sections was a bit confusing and why figure 4 was present?

- Is it hugely problematic that the GCs are missing data in their centers? Following your dis- cussion in the introduction, it sounded like the effects of kinematic heating from IMBHs are most prominent towards the cores.

3. **What parts of the progress report were the most useful for your understanding?**

- I liked the introduction, it was easy to comprehend the problem and understand the challenges as well as some methods which will follow. Although I was surprised by the use of MCMC, which I think they could talk in their intro a bit.
- The figures showing the data histograms was very helpful in understanding the role of data cuts. Figure 2, showing the two population of GC stars and background stars clearly separated in velocity was particularly neat. The explanation of DBSCAN and HDBSCAN was very clear and well written.
- Super clear introduction, nice! This project sounds awesome.
- I found the sections on Dynamical Modelling and Bayesian Parameter Estimation with Toy Model Example particularly useful for my understanding. They perform Bayesian fitting on a toy model that mimics GC velocity dispersion using nested sampling which I thought was a good first step to understand and visualise the analysis.
- Section 3.1 was a very nice read — it very well outlined potential challenges of data filtering and processing. It demonstrates that the team is not blinding applying filters that were previously used but rather critically assessing its impact and developing alternative filters. I'd be quite interested to know whether the filtering being developed vs Vasiliev's one would provide different results. Not to mention, in general, your plots are "pretty cool"! The future work also helps to bind the story together.
- They gave a good overview of the data, processing and analysis techniques they are using. There is a lot about this report that is very good. It is well formatted with some nice plots. The most helpful part for me was the exposition in the introduction, limitations, and data sections.
- Your nested sampling implementation looks good! Looking forward to seeing the final result and that constraint on M_{BH}! Good luck!
- Your discussion of the assumptions you make and their limitations in your first attempts is really great. I think Figure 6 is awesome as well.
- I also liked the intro, but I feel kinematic heating could be explained a bit more. Additionally, you mention 2-body interaction, should it be N-body perhaps?

4. **Suggest at least 3 improvements that could be made to the report.**

- There is an inconsistent use of commas for introductory statements. Ex. "In the following we provide a more detailed description of the data...." should have a comma after following. I found many instances of the same thing, but they tended to be bunched together and not ubiquitous. This is usually the result of different people working on different sections each having different writing styles.
- Definitely an issue with figures, they are hard to read (axis) and it was hard to understand the density without the presence of colour bars.
- Was Figure 4 relevant, since no mention in the text also small type Fig. X on page 4.
- 1. It is minor, but there is no informative project title; a title is generally important as it is the essence of the paper and, if attractive, helps to catch the reader's attention. Furthermore, additional astro context – as mentioned above – will help readers from all backgrounds. 2. The plots can be improved with a) a title to provide the reader an overall understanding of what they're looking at (for example in Figure 1, a title would have easily allowed the reader to know which one is cut vs uncut, without having to read through the generally longer captions) b) more informative labels (e.g. defining the X[normed] and say what X is, including the units if available). 3. Overall, I think that the group should further expand on the EDA – i.e what is the structure of the data, how does the data allow the use of models mentioned,

what are the limitations of the data, any measurement error with the data, are the results for the single GC expandable across all other GCs in the dataset? Maybe expand Section 4 to provide more on data issues (e.g. what's the percentage of missing data, what are the reasons for the measurement errors, etc.), your treatment and their impact on your model choice.

- I liked the Baysian part, but could you include a link to the code used or reference for the sampler (code name/GitHub).
- I would also like to see more equations for the models (dynamic, Plummer, etc.), instead of just pure text. This way the reader can understand the various scaling relationships present.
- Is rescaling the GCs safe? I.e., do we expect whatever IMBH physics we're interested in to be independent of the physical extent of the GC and its proper motion?
- In general, I think the level of required knowledge about globular clusters and dynamics is higher than my own. While I believe that it is an appropriate level for those familiar with GC's and IMBH's, those of us who are only partly familiar (or not familiar at all) will have a hard time digesting this report. An example of this might be Figure 6, the axes of the plot are not explained in the text or the caption, but would otherwise be obvious to someone familiar with GC Dynamics.
- Comparison of HDBSCAN and Vasiliev method would be nice, maybe in a form of CMF or KDE.
- Explain the priors used in your Bayesian analysis. How do you even describe a prior on the probability of an IMBH in a population of globular clusters?
- Explain more about how you plan on performing a population-level inference on the presence of IMBHs in GCs. How do you plan on stacking and rescaling all the star data from different GCs so it can be analyzed as a single population?
- Have people not been searching for the existence of IMBHs for a while? It seems to me that a better approach to solve this problem is to get images of a nearby GC and look closely at its dynamics. I just have a hard time seeing how GAIA data could constrain their existence given that it does not resolve many GC centres and the authors state that realistic dynamics of GCs are not being taken into account. Seems like an incredibly ambitious goal on the two month timescale, but this could be a limitation of my knowledge.
- The first few sections mention a lot of concepts like Probabilistic Gaussian mixture model, Spatially-dependent priors, 2D Gaussian model, Eigenvalues of covariance matrix, etc. While I know these terms at least a little, it is hard to grasp all of them in the context of the problem. I am not so convinced by the HDBSCAN results. How are the major and minor populations differentiated and why are the others strictly outliers? Gaussian proper motion distributions and spatial distributions seems to be not enough to do so considering typical uncertainties.
- Explain how you will calculate the Likelihoods on the cluster mass profiles, and how you will incorporate uncertainty into these calculations.
- All the sections are good to be stand alone which is great; however, I found the flow a little hard to follow. Highlighting only important parts of Data Preprocessing and Background/Cluster Membership followed by Dynamical Modelling and Bayesian analysis would be better for the Report. Issues and potential avenues could be mentioned later.
- I don't think all the Figures are properly referenced and used. For e.g. Figure 4 which talks about the King Profile Fit hasn't been referenced in the writeup.