December 18, 2015

Ms. Wilma H. Trick
Max-Planck-Institut fuer Astronomie
Koenigstuhl 17
Heidelberg, Baden-Wuerttemberg 69117
Germany

Title: Action-based Dynamical Modelling for the Milky Way Disk

Dear Ms. Trick,

I have received the referee's report on your submission to The Astrophysical Journal, and append it below.

I hope that you will agree with my assessment that the report is constructive in tone, but you will note that the referee asks that a number of issues be addressed before publication in the ApJ.

Following the referee's report, I have also appended some comments regarding the statistical aspects of your manuscript. ApJ submissions with a statistical component are routinely previewed by Dr. Eric Feigelson, the member of the ApJ editorial board with a broad expertise in statistical matters in an astrophysical context. Perhaps these comments will be of use to you as you prepare your revision. If you have any questions about the comments, however, please direct them to me, not to Dr. Feigelson. Please consider the report carefully. When you resubmit, please include a detailed cover letter indicating point-by-point your responses to the referee's report, and also indicating any other changes you have made to the text. Reviewers find it helpful if the changes in the text of the manuscript are easily distinguishable from the rest of the text. Therefore we ask you to print changes in bold face; this highlighting can be removed easily after the review.

And two minor matters: Please adhere to the ApJ instructions to authors (see `http://aas.org/journals/authors/common_instruct#_Toc3.2`) regarding the abstract – "The abstract should be a single paragraph of not more than 250 words ..." Please also use the appropriate journal abbreviations in the reference list: these abbreviations for refereed journals are given on the ADS site `http://adsabs.harvard.edu/abs_doc/refereed.html`, and on the site `http://adsabs.harvard.edu/abs_doc/non_refereed.html` for non-refereed publications.

Click this link (it will work one time) to upload your revised manuscript: `https://apj.msubmit.net/cgi-bin/main.plex?el=A7Ew3QPh3A7DACX7F5A9ftdsyk8KtDBL1q74yfVAJnw0A`

Alternatively, you can also log into your account at the EJ Press web site, http://apj.msubmit.net. Please use your user's login name: wilmatrick. You can then ask for a new password via the Unknown/Forgotten Password link if you have forgotten your password.

The policy of The Astrophysical Journal is to view manuscripts as withdrawn if no revised version is received within six months after the most recent referee's report goes to the authors.

If you have any questions, please contact me.

With best wishes,

Butler Burton

————————-

Prof. W. Butler Burton
Associate Editor-in-Chief, The Astrophysical Journal
Professor Emeritus, Leiden University
National Radio Astronomy Observatory

+++
The text of the review is appended below.
===

1 Reviewer's Comments:

This paper details and tests a framework for constraining the Milky Way potential using action-based dynamical distribution functions. The work very much builds on previous work from two of the authors as well as the dynamics group in Oxford. However, for the first time the machinery and its assumptions are quite rigorously tested and the paper thoroughly discusses the limitations and downfalls of the methods.

I believe that this work should be published but that there are parts of the paper where the presentation could be clearer and where the discussion could be expanded or reduced. Additionally, I think that this paper will be a good reference paper for those who will fit dynamical models to the Gaia data, but there are a number of places where the paper could be made more 'usable' without much effort.

My main complaint with the paper, however, is that it doesn't tackle the 'real' problem for three reasons:

- 1. The selection volumes considered seem unrealistic. It is unclear how they relate to what one might use with the Gaia data.
  [Comment from HW: E.g. how to deal with complex sampling volumes has been demonstrated in Bovy et al 2013 and 2015; the point here is to make a generic exploration of search volume shapes.]

- 2. It is unclear how the error distributions relate to the anticipated errors from the Gaia data. $\implies$ *This is a good point and I included two paragraphs and a footnote in Section 3.4, that try to address this. The proper motion errors we considered were inspired by the range of proper motion errors achieved by ground-based surveys. I estimated also the Gaia errors for suitable giant tracer stars and Rene Andrae used the GUMS catalogue to estimate for us the magnitude limit required to ensure that the distance errors for all Gaia stars in a sample would be small enough for our method to work.*

- 3. The majority of the tests seem to be done with the isochrone potential. This is obviously done for speed reasons. A few cases are tackled with the

KKS-Pot and one case with the MW13-Pot. $\implies$ *This was indeed a caveat of our work. Four out of six aspects investigated in this paper were done with the `Iso-Pot`. To alleviate this, I rerun two of the more important aspects (on DF misjudgments and on SF misjudgements) with a more realistic Galaxy potential with disk, halo and bulge, the `DHB-Pot`. (All tests with the `MW13-Pot` were also replaced with the `DHB-Pot` to have no more than four different potentials in the paper. Because of the immense computational effort I could however not re-run the tests on the measurement uncertainties with the `DHB-Pot`. There we still use the `Iso-Pot`.) The new analyses using the `DHB-Pot` are presented in Figures [TO DO] and Tests [TO DO].*

It is then not clear how the precision of the potential parameter recovery for the isochrone relates to more realistic potentials (e.g. will the circular velocity always be so well constrained irrespective of the potential form?). $\implies$ *Now, that we re-run some of the tests with the more realistic `DHB-Pot`, we found that the $v_{circ}(R_\odot)$ parameter is indeed recovered to a very similar precision as for the `Iso-Pot`. Overall, the new results with the `DHB-Pot` are very similar to the `Iso-Pot`—for example for which differences in DF and misjudgment of the SF the recovery of the potential broke down—which is very encouraging for the transferability of the results, also considering the tests that we could not repeat with the `DHB-Pot`.*

The issue of computational speed is mentioned in the discussion but I think this issue is slightly glossed over. $\implies$ *I included a paragraph in the discussion section that discusses how computation speed limits the potential models that we are currently able to use in RoadMapping. I also included a measurement how long some of the analyses in the paper need on how many cores.* Using more realistic potentials in all cases would obviously be much better and would also test the 'Staeckel fudge' apparatus more fully. The improvements to this apparatus are highlighted in the paper but I don't think that it is shown that it has been fully tested. [Comment from Wilma: The Staeckel fudge was tested by other people excessively. I think we do not have to do that. Concerning the use of realistic potentials in all my tests: Just not possible. Sorry. But maybe we could do the following: I could have a potential with a fixed Hernquist halo+bulge and a Miyamoto-Nagai disk with scale length a=0.9kpc (or would that make the potential too unrealistic?). Then isochrone and this more realistic potential have the same free parameters. I could for a few selected test cases in 3.3 (incomplete data) and 3.5 (deviations from qDF) do the analysis with this more realistic potential and overplot these violins with my existing violin plots. Let's hope that we get the same trend...] I think that the heavy use of the isochrone potential should be flagged in the abstract. $\implies$ *As it is now only two of the tests that still use the `Iso-Pot` I hope it is okay if I do not mention the Iso-Pot specifically in the abstract.*

It would be good if some of these concerns were brought to the fore and highlighted in the paper. I have detailed places in the paper where things could be changed. There is also a short list of typos at the end.

2 "Abstract"

- 1. 4th sentence – explain what 'slightly wrong' means more quantitatively. *⟹ I had to delete this sentence because the abstract was too long. I tried however to give some qualitative statements for all the results I'm summarizing in the abstract.*

- 2. 5th sentence – Are the constraints of high precision on the potential or DF? – clarify *⟹ This sentence was also removed. In the original abstract however I meant that the constraints on both potential and DF parameters are precise.*

- 3. The tests referred to in the abstract have been performed independently but the lists suggest that you have shown that when all the listed conditions are satisfied the constraints are of high precision. This should be clarified. *⟹ This is a fair point. I also had to remove the sentence about precision, but I mentioned that we investigated "isolated test cases" for "idealized mock data" and added an "either ... or" between the results on "unbiased estimates". I hope that makes it clearer.*

- *⟹ **Additional changes:** The abstract was partly rewritten, to (a) meet the limit of 250 words, (b) focus more on the key results of the work, (c) make it clearer that we used idealized test cases.*

3 "1. Introduction"

- 1. Magorrian (2014) has provided a framework for constraining the potential without assuming a particular parametrized form for the DF. Whilst Magorrian's method is computationally intensive, it should be referenced in the introduction as it relates to the later discussion of choosing a particular DF parametrization. *⟹ Thanks for the suggestion. I included the reference in the paragraph on different modelling methods in the introduction. I also added a little paragraph in the discussion of different modelling methods.*

- 2. At the end of the introduction I think that you should refer people more strongly to the results section as much of section 2 is presenting a framework that appears elsewhere. *⟹ Most of the introduction on page 2 is a list of the different questions we try to investigate in the result section of this paper paper. But to make sure that the reader knows immediately what the key results are, I added a sentence pointing again to the corresponding sections at the very end of the introduction.*

4 "2. Dynamical Modelling"

4.1 Section 2.2 - Now: "2.2 Actions" & "2.3 Potential models"

- 1. A slightly fuller introduction of the actions is merited. Mention why the actions are introduced. What advantages do they present? Also state that

the action-angles are canonical – this is important later for transformation of the pdf. $\implies$ *I completely agree. I added a corresponding section.*

- 2. The third sentence isn't quite right – the most general are the triaxial Staeckel potentials of which the axisymmetric Staeckel potentials are special cases and all spherical potentials are special cases of these. THe isochrone potential is the most general potential in which the actions are not computed as a quadrature. $\implies$ *Thanks for making it clearer for me. I changed the text accordingly.*

- 3. The potential discussion could be put into a separate section. Also mention that the circular speed at the Sun is the same for all three potentials. $\implies$ *Done.*
  Re-iterate that the reason for using the isochrone and Staeckel potential is the ease with which the actions can be computed. $\implies$ *Done. I added a paragraph in Section 2.3 summarizing the advantages of the different potentials we choose to investigate in this paper.*

- $\implies$ ***Additional changes:*** *I removed the `MW13-Pot` completely from the paper and replaced it with the `DHB-Pot` everywhere, also in Table 1 and Figure 1. I wanted to have only four different potentials in the paper, and the `DHB-Pot` was much better suited to re-run some of the tests in short time than the `MW13-Pot`, for which the analysis is very slow due to the use of an exponential disk.*

4.2 Section 2.3 - Now: "2.4 Stellar distribution functions"

- 1. Guiding-center should not be in brackets as this is important. $\implies$ *Done.*

- 2. The final sentence of the left-hand column of Fig 3 should read $L << L_0$. $\implies$ *Thanks for pointing out that typo. I wrote $L < L_0$ instead.*

- 3. Top of page 4 – explain what X is in the text. $\implies$ *Done.*

- 4. Do you interpolate in log density? $\implies$ *Yes, I did. I now mention it in the paper as well.*

- 5. The footnote says 'should be chosen as' – add a forward reference to Fig. 4 $\implies$ *Done.*

- $\implies$ ***Additional changes:*** *I added a paragraph at the beginning of this section which discusses why the Jacobian = 1 (which I mentioned in Section 2.2 as requested by the referee) allows us to consider DF(x,v) and DF(J) equivalently as probability of a star to be at (x,v).*

4.3 Section 2.4 - Now: "Appendix B: Selection Functions"

- 1. An entire section dedicated to this topic seems unnecessary. See below. $\implies$ *I moved this section to the Appendix (because I also wanted to further motivate the "unrealistic" selection functions we use in this work).*

4.4 Section 2.5 - Now: "Appendix A: Mock data"

- 1. As this section only explains technical details rather than testing the apparatus I think that this section can be put in an appendix along with Fig. 2 and 3. Fig 2 and 3 are illustrative but I think that they are similar to BR13 Fig2 and 3 so do not need to appear in the main body. I don't think it is a 'test' so should be removed from Table 3. $\implies$ *I agree. I moved the mock data section and figures to the appendix and removed the parameters belonging to Figure 2 + 3 from the 'tests' Table 3.*

- 2. The discussion of selection on very errorneous x coordinates is interesting but surely this isn't the way the data will actually be handled? $\implies$ *We are not 100% sure that we understand the referee's main point. If one considers several percent of distance errors at the edge of the survey volume, the volume from which stars can get scattered outwards or inwards due to Poisson noise is larger than the survey volume. This is why we create perfect data in a large volume, before we first perturb it according to the measurement errors and then select the stars that would enter the actual survey volume and therefore data catalogue. For contiguous surveys like Gaia or Hipparcos we would of course not have a survey volume with a sharp edge and the survey selection function is more likely to be a magnitude cut. But for intrinsic standard candles like red clump stars a magnitude cut translates to a distance cut. We realise it's idealized and there won't be a sharp distance cut. But scatter in and out of survey volume will play a role and we explore an ideal case.* [TO DO: Move part of it to the above discussion of selection function.]

4.5 Section 2.6 - Now: "2.5 Data likelihood", "2.6 Likelihood normalisation" & "2.7 Measurement uncertainties"

- 1. The selection function can be briefly mentioned at the beginning of this section and stated that you assume here for simplicity it is a function of $\vec{x}$. $\implies$ *Done.*

- 2. The Jacobian from J,theta to x,v should be mentioned here. $\implies$ *I already put a little bit of the implications of the Jacobian for the DF at the beginning of Section 2.4 "Stellar distribution functions", but I added also a footnote here that explains why we can integrate DF(J) easily over (x,v) in the likelihood normalisation.*

- 3. pdf should be defined in a separate equation. $\implies$ *Done.*

- 4. Figure 4. – it wasn't clear to me that the 'truth' normalization used a high enough set of parameters. $N_x = 20$, $N_v = 56$ and $n_\sigma = 7$ only seem slightly larger than the values actually compared to. $\implies$ *Yes. I re-run this test with a higher accuracy $N_x = 32$, $N_v = 68$, $n_\sigma = 7$ for the "true" normalisation. I also used the `DHB-Pot` in this plot instead of the `KKS-Pot`, which was used in the original plot. The reason for the change is, that the `DHB-Pot` is used in more tests. I was also interested how this test depends on the kinematic temperature of the population. And as there seems to be*

6

a dependence, I added a new row of panels for the cool population in this plot. It turned out that for the cool population the accuracy condition was not satisfied for all volumes. So overplotted little black stars that show which volumes are actually used in tests of this paper, to assure the reader that for the actual tests the condition was satisfied.

- 5. Make the normalization discussion a separate section. $\implies$ *Done.*

- 6. The discussion of the likelihood normalization should reference and compare with McMillan and Binney (2013) as the discussion is very similar. $\implies$ *Thanks for the suggestion. I included a paragraph where I compare their work on the likelihood normalisation with ours.*

- 7. Is there any general advice on how to choose N_x, N_v and N_σ? The authors have shown it is OK for the mock datasets but do I have to redo the authors' exercise when I have a real dataset? $\implies$ *The short answer is unfortunately "yes", you would need to go through the full exercise. I think if you use a similar DF and potential than we did in Figure 2, you could use Figure 2 to look what accuracy you need for your number of stars. Or use our values as starting values and optimize later. But otherwise you need to check on a case-by-case basis if your accuracy was high enough. I added a corresponding sentence in the paper and explain (in a footnote) why it depends on both kinematic temperature and also the particular potential model what accuracy you need.*

- 8. Put error discussion in separate section. $\implies$ *Done.*

- 9. Reduce size of caption for Fig 5. More of the details could go in the text. $\implies$ *Done.*

- 10. Equation (15) is a novelty. It is troubling that the tests that use this approximation all seem to use the isochrone but the approximation is still necessary. Is that because it is computationally awkward to calculate this integral or just very slow? $\implies$ *In our simple test scenarios it should be indeed computationally possible to convolve the selection function with the homoscedastic uncertainties and not use the approximation in Equation (15) - independent of potential. However, in each realistic case (heteroscedastic errors, more complex selection functions, etc.) the proper treatment would be actually computationally way too expensive as the normalisation has to be calculated for each star separately. We will therefore never actually want to use the proper likelihood formula, and always use the approximation. Ignoring measurement uncertainties in the normalisation is by the way also done in similar studies, like McMillan & Binney (2013) and Das & Binney (2016). We are the first to test this approximation a bit. That we used the Iso-Pot instead of a more realistic potential was simply for its computational speed, because we wanted to investigate a large number of mock data sets at different accuracies to be able to do some statistics. If it would have been feasible, we would have done the test with a more realistic potential. We still think that it should not make a huge difference. In the tests that we run with both Iso-Pot and a more realistic DHB-Pot we always got pretty similar results - also quantitatively. I added a sentences in the section, to stress more the*

computational expense when NOT using the approximation and added a caveat in the discussion that we tested the approximation with the Iso-Pot only.

- 11. The penultimate sentence of this section contradicts the previous sentence without validation. Why is this? ⟹ *Thanks for pointing it out. That was indeed confusing. After running more tests for both the* `Iso-Pot` *and* `DHB-Pot` *I think the reason is, that the recovery of $v_{circ}(R_\odot)$ behaves in the* `Iso-Pot` *always as expected and similar to* `DHB-Pot`*. To not get biases in $v_{circ}(R_\odot)$ one needs more $N_{samples}$ for higher N_*, as expected. For the isochrone scale length b this trend was not so clear. I think that this is a specific property of the Iso-Pot. I rewrote those two sentences in the paper and specified a bit more closely what we observed for smaller N_* and different parameters.*

- ⟹ **Additional changes:** *In the section "Data likelihood" I added a paragraph that argues why we chose to use a Bayesian framework, to address the comments by Dr. Feigelson (see below).*

4.6 Section 2.7 - Now: "2.8 Fitting procedure"

- 1. I liked this section – it was well thought out and informative. ⟹ *Thanks. :-)*

- 2. Here a fixed sampling is used for the error samples. I think again you should reference McMillan & Binney (2013) as they discussed the numerical stability of this method. ⟹ *I referenced this paper, but not in this section, which is more about fitting and less about measurment errors. I mentioned it in "2.7 Measurement uncertainties" as a similar approach using MC sampling for convolution with the measurement errors.*

5 "3. Results"

- 1. It is stated that the breakdown of axisymmetry and steady state assumptions is not explored. I wonder as well about the impact of resonances, particularly when the data are very high quality. This cannot be explored in the current setup as the data are generated from an action-based DF but perhaps should be mentioned as a potential limitation of the approach. ⟹ *Thanks for this suggestion. I added a corresponding comment in this section and also mentioned that we are planning to further investigate non-axisymmetries by applying RoadMapping to a galaxy simulation in a future paper.*

- ⟹ **Additional changes:** *As I moved the Mock data section to the Appendix, I stress here explicitly that we use mock data for our tests.*

5.1 Section 3.1 - "Model parameter estimates in the limit of large data sets"

- 1. This seems a good sanity check but should it be published? Fig 6. seems sufficient to me to demonstrate that your code works. I don't think the

paper would miss this section. $\implies$ *I wanted to have one section in this paper that summarizes the (statistical) implications of using large data sets for RoadMapping separately and not just in the context of "Likelihood normalisation" as in Section 2.6. But as you are in principle right and it is mostly a sanity check, I edited the whole section to make it much shorter. I also removed the figure, that shows how the width of the pdf decreases with $1/\sqrt{N_*}$, completely from the paper and only mention the result in one sentence.*

5.2 Section 3.2

- 1. I understand that the selections used in Fig 9 are illustrative but the pink selection just doesn't seem realistic. I think Fig 8. is a sufficient demonstration of the difference between different selections. Fig. 9 doesn't add anything and is barely discussed in the text. Also, without observational uncertainties (which will be greater for the more distant boxes) the discussion seems superficial. I would consider removing this.
  [Comment by Wilma: Maybe I should make the point of this figure clearer. It does consider unrealistic selection functions, but I just wanted to figure out if there are regions within the Galaxy where stars are in general on more informative orbits. Or if we, for example, have a natural disadvantage or advantage due to the sun's position within the MW. I vote for keeping the figure.] TO DO: Hans-Walter said: we have chosen extreme cases to see the effects most dramatic, but we don't claim that all of them are realistic

5.3 Section 3.3

- 1. Isn't the reason for the cold population being more robust that it doesn't have as many stars at large distance as the hot population so it is less affected by the cuts? I suppose this not necessarily true for lines-of-sight in the plane.
  [Comment by Wilma: Good point. Any ideas?]

5.4 Section 3.4

- 1. It would be nice to state how the considered errors are related to the anticipated Gaia errors or other surveys. [Comment by HW: The proper motion errors which we choose cover very coarsely the proper motion error regions of ground based surveys. TO DO: Look it up in more detail. Rene helped me with the distance errors.

5.5 Section 3.5

- 1. I think this and section 3.6 are the most valuable in the paper as they really explore potential systematics. In my opinion, these are the key results.

- 2. Fig 15. – it would be interesting to see the difference between the fits and the truth. Do the fits break down in particular places?

5.6 Section 3.6

- 1. Fig 19 is difficult to interpret. Is it possible to display the difference? [Comment by Wilma: I don't think that it is difficult to interpret. Showing the difference would let me include much less information into the plot + I personally think reading residual plots is even more difficult. What I COULD do, is adding two panels that compare cuts through the density along z @R=8 and along R @z=0 for all models.]

- 2. The fact the density is not well recovered seems interesting as it points to possible biases in the surface density of the disc/dark matter measurements if one uses the wrong potential. It would be good to have the discreapncy quantized in the text.
[Comment by Wilma: How to quantize the discrepancy?]

- 3. I think Fig. 20 could be removed. As mentioned it doesn't make sense to compare the DF parameters between different potentials so I am not sure what Fig 20 is telling us.
[Comment by Wilma: I admit it is not a very important figure, but it makes the point in the text clearer that the qDF parameters in themselves are not that informative depending on which potential is used. I think that is easily forgotten and people should be reminded of this.]

5.7 Section 3.7

- 1. Should this section be moved to the discussion section?

6 Summary & Discussion

- 1. Perhaps add statements comparing the errors explored to those anticipated from Gaia.

- 2. The two approaches mentioned at the end of 'Gravitational potential beyond the...' are stated as formally similar but I think it is clear that one is better than the other. The true Staeckel approach limits you to potentials with the same foci. This is an obvious limitation and has been discussed before.

- 3. The definition of X in f(J,[X/H]) doesn't seem to make sense. [Comment by Wilma: Why not?]

- 4. The first section in future work is very interesting. Use of different DFs and potentials as explored in this paper is interesting but a true test of the apparatus on a more realistic galaxy would make the 'RoadMapping' tool much more attractive.

- 5. I think that the final two questions of the future work section are weak. Clearly the rotation curve is only describing the in-plane force not the force everywhere. Parametrizations will naturally convince you that the rotation curve is well measured but I think there is a lot more flexibility. Also, the advantage of using the approximate actions is that more realistic potentials can be considered. ⟹ *You are right. Thank you for making things clearer for me. I changed the discussion accordingly and we also thought of some more interesting science questions to conclude the paper with.*

6.1   Table 3

- 1. Can you add a summary column that summarizes the result? This would make the paper much more 'usable'. ⟹ *Done. Great idea.*

7   Typos

⟹ *Thanks for pointing them out.*

Abstract.   1. 3rd sentence – 'rules of thumb' for how data, model and machinery most affect ... and DF. ⟹ *Sentence was removed anyway.*

Introduction.   1. Start of 3rd para: 'to constrain' → 'constraining' ⟹ *Done.* 2. Start of penultimate para: 'to restrict' → 'restricting' ⟹ *Done.*

Table 1.   1. 'troughout' in caption. ⟹ *Done.*

Section 2.3.   1. Second sentence: 'about'→'on' ⟹ *Done.* 2. 'the circular orbit' to 'near-circular orbit'? ⟹ *Done.* 3. Top of page 4 two 'in's ⟹ *Done.*

Section 2.6   1. Top of right column page 7 – replace 'besides' with 'not only... but also' ⟹ *Done.*

Section 2.7   1. Need '(MCMC)' after MCMC ⟹ *Done.*

Fig 14.   1. 'pest'→'best' ⟹ *Done.*

Fig 15.   1. 'refereed'→'referred' ⟹ *Done.*

Section 3.5.   1. Second paragraph right column page 13 – 'sun'→'Sun' ⟹ *Done.*

Table 3.   'analyis'→'analysis' ⟹ *Done.*

8   Comment from Dr. Eric Feigelson

==
Below are comments on statistical aspects of the manuscript: ApJ submissions with a statistical component are previewed by Dr. Eric Feigelson, the member of the ApJ editorial board with a broad expertise in statistical matters in an astrophysical context.
++

An elaborate Bayesian inferential procedure is described in sec 2.6-2.7 for parameter estimation with results shown in Fig 6. But with uninformative uniform priors and a simply unimodal likelihood with a nearly multivariate normal distribution, this effort is unnecessary. The same result would be obtained with maximum likelihood estimation via the EM Algorithm (probably ¿¿100 iterations with convergence guaranteed by theorem) and parameter uncertainties estimated from the Fisher Information Matrix. The confluence of Bayesian and MLE procedures in this case should be presented. [Comment by Wilma: Jo said, we could ignore that... Can we really? What should I answer the referee?]