Draft version October 7, 2022
Typeset using LaTeX default style in AASTeX631

Astron 128 - Lab 1: An Extinction Map of the Milky Way

Shrihan Agarwal[1]

[1] University of California, Berkeley, Department of Astronomy, Berkeley, CA 94720

1. INTRODUCTION

Magnitudes are the most important quantity for astronomy, giving us crucial information for multiple science objectives. However, dust and gas in the way impact these objectives by scattering light off from stars. This is especially prevalent in the galactic bulge, where this "extinction" becomes so strong that accurate estimates of absolute magnitudes become extremely difficult. In order to solve this, we build extinction maps, which estimate the degree of extinction throughout the galaxy. In this lab, we generate 2D extinction maps of the Milky Way, using RR Lyrae as an absolute magnitude basis.

In this lab, we query GAIA DR3 and WISE for various categories of RR Lyrae. We fit their periodic variations, for their mean magnitudes, their instrinsic luminosities, their period-luminosity relation, and eventually develop an extinction map for the Milky Way.

We begin by drawing multiple RR Lyrae and fitting their respective periodic curves with Lomb-Scargle Periodograms and Fourier series components in section 2. We use mean magnitudes found from this in order to generate period-luminosity relations, fitting them with MCMC in section 3. Finally, we use these period-luminosity relations in order to calculate the extinction and thus generate a dustmap, in section 4.

2. ANALYSIS OF RR LYRAE

2.1. *Data Retrieval*

We query the *gaiaedr3.vari_rrlyrae* catalog and select the top 100 stars for which a period 'pf' has been calculated by GAIA, and the number of clean epochs of data for each target is greater than 40. See Appendix 4 for the exact query. These were cached using dump_to_file = True in the GAIA request. The photometry for these sources was obtained by specifying GAIA Source IDs and retrieving combined epoch photometry. We used these light curves for the majority of the remaining RRLyrae analysis in the proceeding section. A new query was created to select the top 3 RRc type Lyrae, as well as the top 3 RRab Lyrae. These are also provided in 4. This query selected for at least 80 good quality observations, as well as a mean magnitude brighter than 15.

2.2. *Period Determination With Lomb-Scargle Periodograms*

We use Lomb-Scargle Periodograms in order to estimate the frequency of the periodic oscillations of RR Lyrae in the catalog. These periodograms use a Fourier series decomposition of the original waveform in order to highlight the best fitting frequency, as highlighted by it's Fourier series coefficient, or the 'power'. In Figure 1, we show an example of the power spectrum of the periodic brightening for an RRab variable.

The frequency space was searched between 0.25/day and 2/day, following (Clementini et al. 2016), and the frequency with maximum power was chosen. This range had to be altered to encompass RRc Lyrae as well. We found accurate fitting results when using a frequency range from 0.01/day to 4/day for RRc Lyrae specifically, covering a very large range of periodic oscillation. We depict the power spectrum and estimated results for RRab Lyrae in Figure 1 and Table 1.

Corresponding author: Shrihan Agarwal
shrihan@berkeley.edu

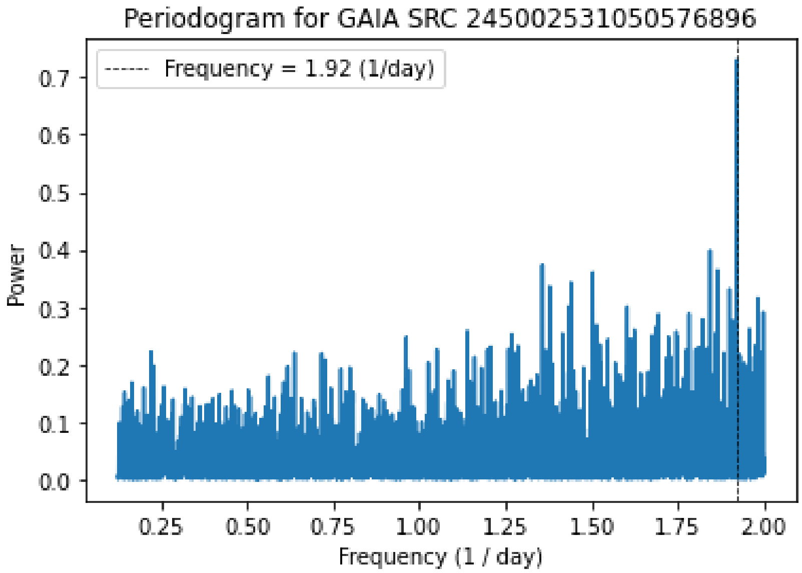


Figure 1: Periodogram for randomly selected GAIA source. The maximum power corresponds to the best-fit frequency as specified by Lomb-Scargle. This particular star had a frequency of 1.92 oscillations per day.

	Source ID	Estimated Period (d)	GAIA Period (d)
0	1934784777174637056	0.710674	0.710655
1	1934869366550952448	0.535673	0.535698
2	1935104150941800448	0.803901	0.803898
3	1935230693559394816	0.521132	0.481906
4	1935418400810447872	0.556609	0.556610

Table 1: Source ID for 5 RR Lyrae, along with the estimated period using Lomb-Scargle, compared to GAIA's period in the *gaiadr3.vari_rrlyrae* catalog.

2.3. *Fitting Fourier Series to Lightcurves*

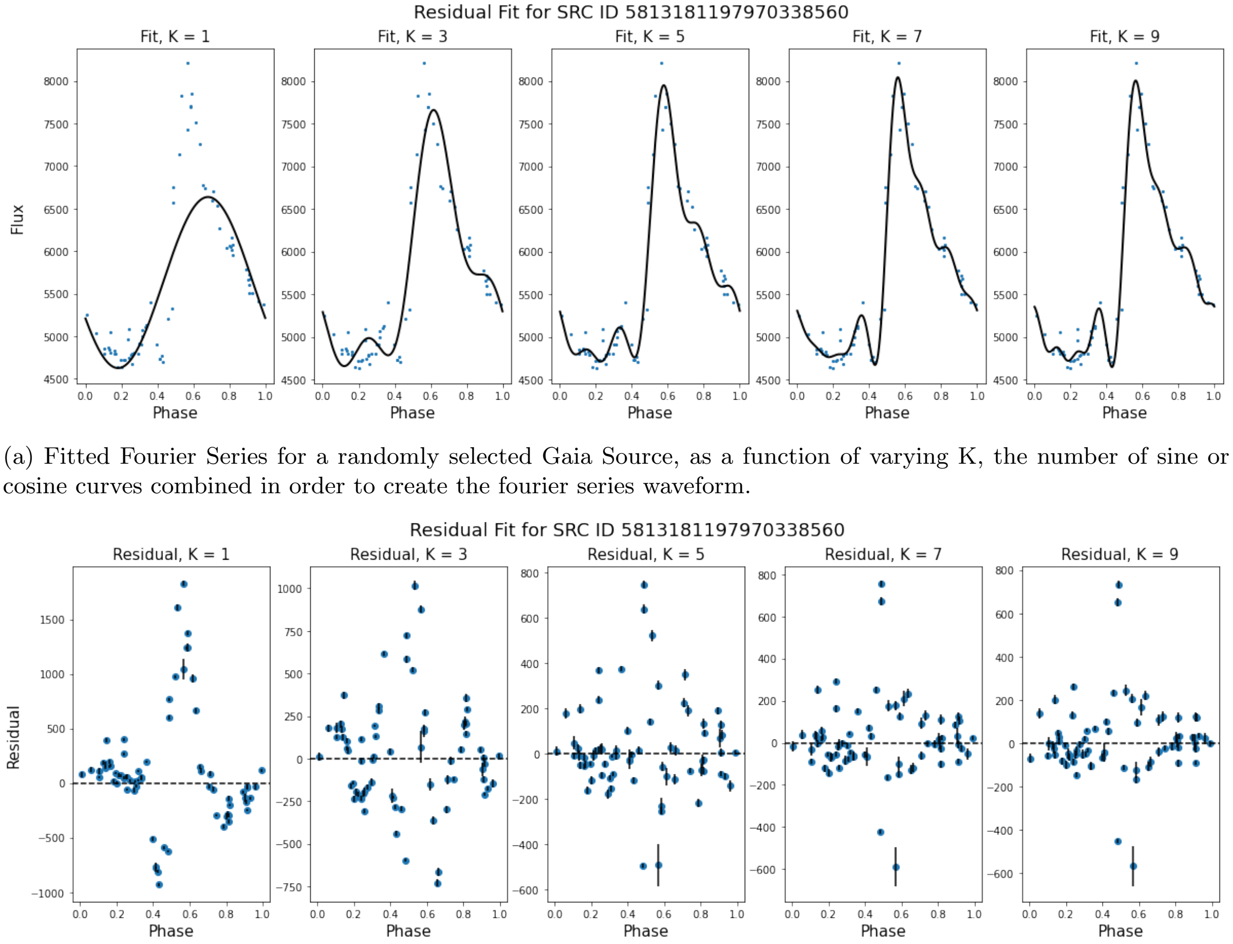
Though Lyrae do not follow a strict sinusoidal motion, we can fit for them using a linear combination of sine waves with a Fourier series that accurately approximates the waveform. To find the best fit Fourier series coefficients, we solve for β in the following equation. Here, K is a free parameter corresponding to the number of sin or cosine coefficients being fit.

$$y = X\beta,$$

where y_n is the range of magnitudes, β_{2K+1} is a set of coefficients to be solved for, and $X_{n\times(2K+1)}$ is a matrix with the following structure:

$$X = \begin{bmatrix} 1 & \sin(\omega t_1) & \sin(2\omega t_1) & \dots & \cos(\omega t_1) & \cos(2\omega t_1) & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \dots \\ 1 & \sin(\omega t_n) & \sin(2\omega t_n) & \dots & \cos(\omega t_n) & \cos(2\omega t_n) & \dots \end{bmatrix}$$

We solve for β using *numpy.linalg.lstsq*. We present the results of our fits to an example of a GAIA source in Figure 2. Here, we phase fold the light curve according to the period as determined in Section 2.2, and use the folded lightcurve in order to visually see the results of coefficient determination. By changing values of the hyperparameter K, we notice that we can tend to overfit the data to noise, creating a spurious fit that does not obey the true distribution



(a) Fitted Fourier Series for a randomly selected Gaia Source, as a function of varying K, the number of sine or cosine curves combined in order to create the fourier series waveform.

(b) Residuals of the above Fourier Series for a randomly selected Gaia Source, as a function of varying K, the number of sine or cosine curves combined in order to create the fourier series waveform.

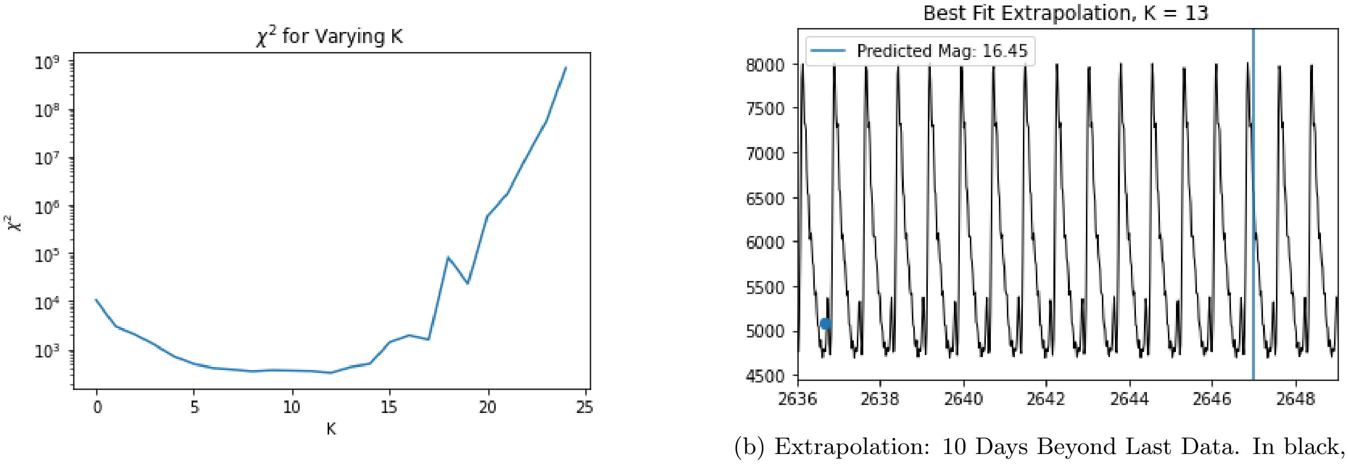
Figure 2: In the figures above we depict the phase-folded lightcurve (blue scatter) alongside the best fit Fourier series combination (solid black fit) for the particular value of K chosen. We also depict uncertainties in flux retrieved from GAIA. We note that increasing the number of coefficients leads to better fitted data, but may lead to overfitting - fitting noise as a unique signal.

of data. This is a problem.

In order to prevent overfitting, we adopt a cross-validation approach. We fit 80% of the data points for the example source, and calculate χ^2 on the remaining 20% of the dataset. This ensures that our model does not pick outliers as surious noise and fit to the dataset. We calculate this for K from 1 to 20, and compare χ^2/dof for each of the cases. On fitting on the majority of our 100 RR Lyrae, we find an optimal K of 5 for the majority of the stars. We fit all the Lyrae with a K of 5. We argue that this retains the same results as a careful analysis determining all K. This is confirmed in 2.4.

We also fit 3 RRc Lyrae using the same procedure as a proof of concept. This procedure holds generally for RRc Lyrae as well. We present a fit for 3 RRc Lyrae and compare to 3 RRab Lyrae in Figure 4. We find that the RRc Lyrae tend to have a more sinusoidal magnitude variation, and additionally have a higher frequency of oscillation. However, these claims are hindered by our limited sample size of 3 variables. Additionally, we see no clear Blazhko effect in our sample, or deviations from the typical phase (Netzel et al. 2018).

2.4. *Calculating Mean Magnitudes*



(a) Chi-Squared as a function of varying K. We find K = 13 to be the optimal hyperparameter for this star.

(b) Extrapolation: 10 Days Beyond Last Data. In black, we present the best fit model with K = 13. The blue data point is the last GAIA measurement.

Figure 3: For a randomly selected source, we present chi-squared as a function of K. This particular star (SRC ID: 6061867986376466688) had interesting structure in its variablity, suggesting the model to recommend a K = 13 as the best chi-squared. In general for a larger sample, K = 5 seemed to perform best.

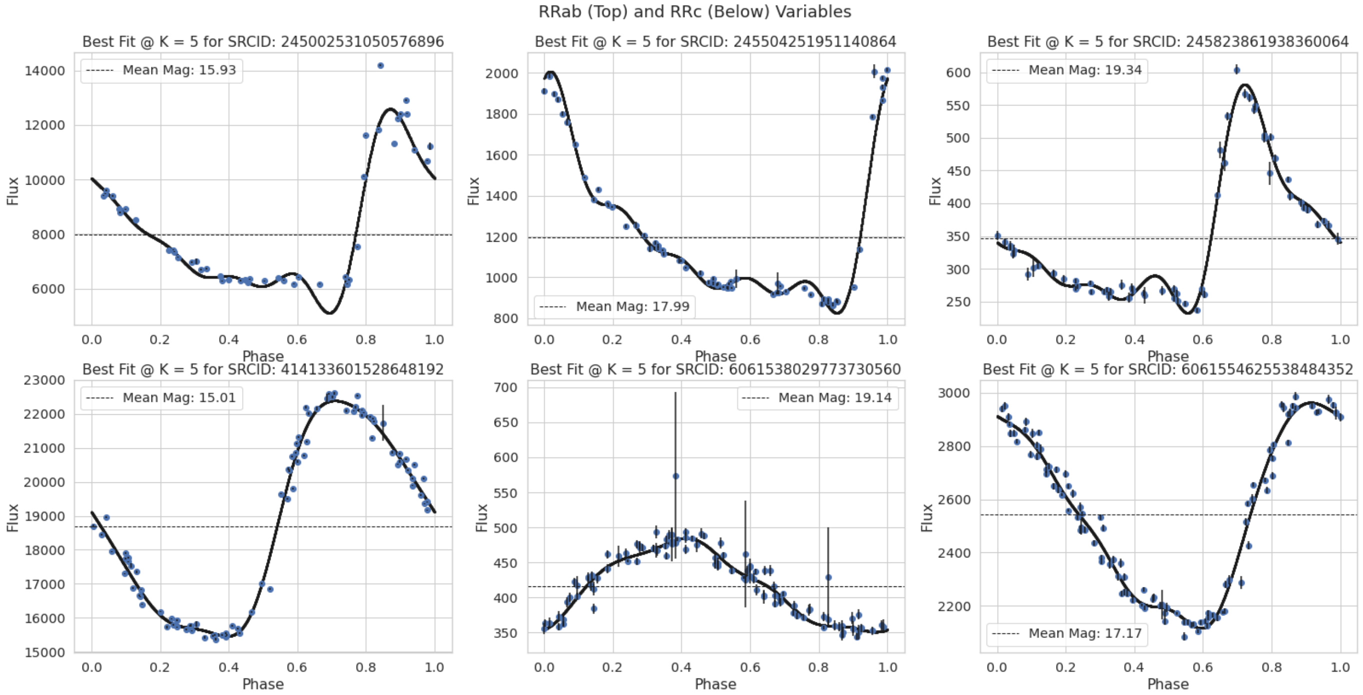


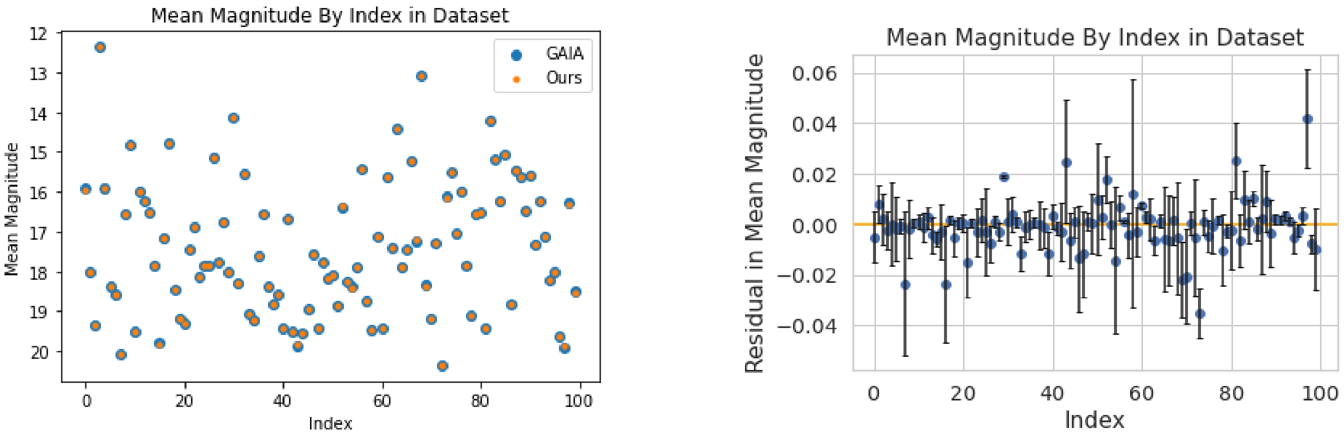
Figure 4: Fit comparisons for 3 RRab and RRc lightcurves. The RRab lightcurves are on top, with a more sawtooth-like waveform. The RRc lightcurves have a more sinusoidal waveform.

Mean magnitudes for RR Lyrae are necessary to determine to RR Lyrae period-luminosity relation for our dustmap. We determine the mean magnitude by averaging the flux of an RR Lyrae over its phase, and converting it to magnitude with a zero-point offset. Since GAIA provided both magnitude and flux in it's lightcurve, we use these to calculate the zero-point for GAIA G.

$$m_s = -2.5\log(f_s) + ZP$$
$$ZP = \text{mean}_i(m_i + 2.5\log(f_i))$$

	Source ID	Estimated G Magnitude	GAIA G Magnitude
0	1934784777174637056	19.289039	19.289589
1	1934869366550952448	17.465653	17.450500
2	1935104150941800448	16.878619	16.879194
3	1935230693559394816	18.123521	18.120331
4	1935418400810447872	17.859488	17.860941

Table 2: Comparison of GAIA estimated mean magnitudes to the Estimated GAIA mean magnitude. This allows us to create a realtion between period and luminosity for RR Lyrae. We find our results are in remarkable agreement, within GAIA uncertainties.



(a) All 100 GAIA magnitudes to our best estimate.

(b) Residuals: Estimated Magnitude - GAIA Magnitude

Figure 5: For all 100 RRab Lyrae in our dataset, we compare the best fit mean magnitudes and find that all are in strong agreement with our predictions. In blue, we show GAIA magnitudes, and in orange, our estimates. The right figure shows a residual of the same, indicating all expected amount of residuals are within uncertainties.

where i indicates the index of a datapoint in the GAIA lightcurves, and s is the index of the star in the dataset. f_s is the mean flux of the RR Lyrae over it's phase, according to its Fourier series model (since the data may be undersampled). We present our results on calculation of RR Lyrae mean magnitudes, and compare them to the current GAIA estimates in Table 2. Additionally, we show results with uncertainties in Figures 5. With this, we complete our analysis to find GAIA magnitudes for RR Lyrae. We can being to analyze the period-luminosity relationship between them.

3. PERIOD-LUMINOSITY RELATIONSHIP OF RR LYRAE

3.1. *Geometric vs. Parallax Distance Estimation*

In this section, we utilize Bailer-Jones et al. (2018) geometric distances. Geometric distances with good quality use a galactic prior, astrometry, and parallax information to give an uncertain but accurate estimate of stellar distances in the Milky Way. It solves the issue of the non-linear transformation of uncertainty on conversion from a parallax uncertainty to a distance uncertainty, especially for distant stars with parallaxes consistent with 0. On validation of our data, we find that the Bailer-Jones distances are largely consistent with with GAIA parallaxes for Lyrae with distances within 4kpc.

3.2. *Data Retrieval*

We combine the external catalog developed by Bailer-Jones et al. (2018) with our existing set of GAIA RR Lyrae. We select on the top 600 Lyrae, with low extinction, at high latitudes (—b— ¿ 30). This will allow us to get an estimate of the intrinsic luminosity. We also fit to WISE W2 magnitudes by cross-matching GAIA and WISE stars, and finally, to GAIA BP_RP color,which is independent of distance estimates. The AQDL queries is provided in 4.

We clean the data in the period-luminosity diagram by performing a variety of quality cuts in the data, according to equations C1 and C2 in (Lindegren et al. 2018). For remaining outliers, we strip magnitudes above 2 in GAIA and WISE, as well as color-magnitudes above 1.2 in GAIA BP-RP. Though crude, the outliers lied significantly outside of the reasonable period-luminosity and range and so could be cut off conveniently. As a test, the data were fitted with and without outliers, and the results were found to be negligibly different.

3.3. *Monte-Carlo Markov-Chain Fit*

The period luminosity relation between RR Lyrae is a beautiful quirk of stellar physics that allows us to predict the absolute magnitude of a RR Lyrae independent of the distance to it. We fit a line between the period $\ln(P)$ and the absolute magnitude or color of the RR Lyrae using *np.polyfit* in order to gauge the prior range of slopes and magnitudes to be searched. Although *np.polyfit* provides us with a strong fit, it can only give us the covariance in the result, and not a true posterior. While that may have been sufficient for our case, we wanted to fit an additional variance parameter $\ln\sigma$ that describes the variance in the fit, and obtain posteriors for our data.

Monte-Carlo Markov-Chain (MCMC) is a process of numerical estimation of a posterior given a prior and data as evidence. This data can be used to evaluate the likelihood function, and retrace the set of degenerate parameters that could create the result in question. We take the *maximum a postieri*, or MAP estimate, which maximizes the posterior rather than the likelihood function (MLE).

While generally not an ideal method to search for the minimum loss, which is better performed by informed processes such as variations of gradient descent, it is one of the few methods that can give us an accurate posterior set of solutions by faithfully evaluating the evidence based on the priors.

Although this was described in lecture in detail, the MCMC process works by having a number of chains (walkers) to explore an N-dimensional parameter space. Initialized by the prior, their movements are determined by a simple Markov Chain, where at each step, the walker proposes a direction, and either rejects or accepts that proposal with a probability that is higher for a solution with lower loss. If accepted, the walker moves in the direction in question with a step size (hyperparameter). As such, the walker tends to move toward the minima while exploring the space around the minima. The walker will tend to spend more time in regions of lower loss, and less in regions of higher loss by the nature of randomly drawn sampling. By taking a histogram of the positions which each walker has traced, we can obtain an N-dimensional posterior of solutions, each with their respective log-Likelihood. The final posterior is a weighted sample of these solutions.

I coded up my own example of this MCMC process, after which I used `pymc3`, a Python package designed for parallelized MCMC evaluation. I present the best fit results from the MCMC fitting procedure for three period-luminosity relationships: (i) GAIA G-band magnitudes based on Bailer-Jones et al. (2018) distances, (ii) WISE W2 magnitudes on the same, (iii) GAIA BP-RP color.

We initialize the chains with priors as determined by our least square linear fits earlier, taking a uniform range of ±1.0 around the fit value for both slope and intercept. For $\log\sigma$, we set a uniform, but wide range from -10 to 10. We run 5 chains with 10000 draws and 1000 tuning steps - overkill for a linear fit, but gives us a well-informed posterior sample. We present the trace and cornerplots for these in the Appendix, Section 4.

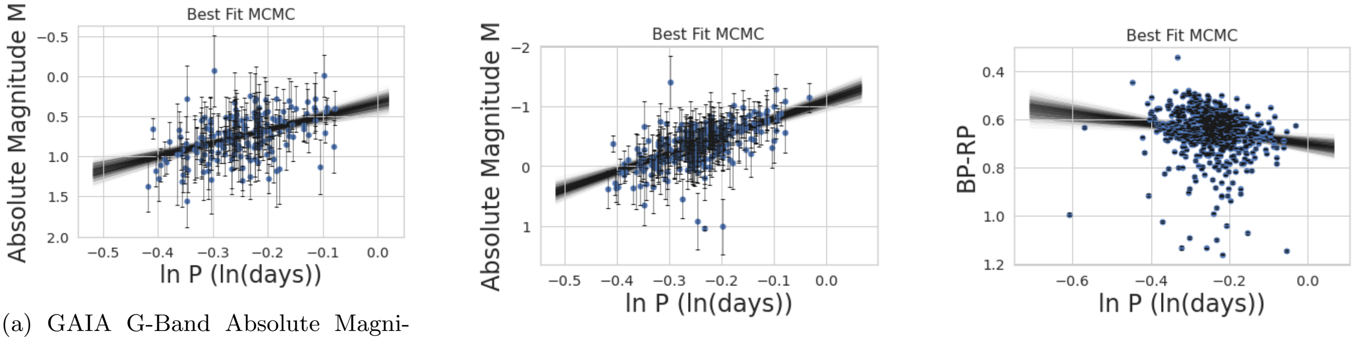
The distribution of possible linear fits for is shown in Figure 6, for all the 3 cases. The final fitted values and uncertainties are presented in Table 3.

3.4. *Discussion of Period-Luminosity Fits*

The Wise W2 band has an effective wavelength of about 4.6 μm, whereas the GAIA G-band has a wavelength range of 330nm to 1050nm. Based on Beaton et al. 2018, for a wavelength of 4.6 μm, the period-luminosity slope was -2.3, whereas our fit produces a value of -2.95 ± 0.25 (WISE). For GAIA G, the period luminosity slope is around -1.45. Our predictions match at -1.652 ± 0.220. These comparisons are on the basis of Figure 8 in Beaton et al. 2018.

	Gaia G Mean	Gaia G SD (+-)	WISE W2 Mean	Wise W2 SD (+-)	BP-RP Mean	BP-RP SD (+-)
m	-1.650000	0.219000	-2.956000	0.251000	0.211000	0.055000
b	0.339000	0.056000	-1.095000	0.063000	0.701000	0.014000
logsig	-2.735000	0.414000	-1.661000	0.070000	-2.339000	0.028000

Table 3: Best-fit / MAP Results for `pymc3` fit to data.



(a) GAIA G-Band Absolute Magnitudes

(b) WISE W2 Absolute Magnitudes

(c) GAIA BP-RP

Figure 6: We present the distribution of best-fit linear solutions as evaluated by the MCMC process (black). The exact best-fit MAP parameters are provided in Table 3

.

4. GENERATION OF A MILKY WAY DUSTMAP

The data were retrieved similarly to before, however ignoring areas with extinction. We maintain earlier quality cuts on photometric SNR and BP/RP flux. We take all RRab Lyrae in GAIA DR3 (87523 targets), and determine the difference between their intrinsic BP-RP color inferred from earlier period-luminosity relations.

$$A_G = 2E(G_{BP} - G_{RP}) = 2((G_{BP} - G_{RP})_{obs} - (G_{BP} - G_{RP})_{int})$$

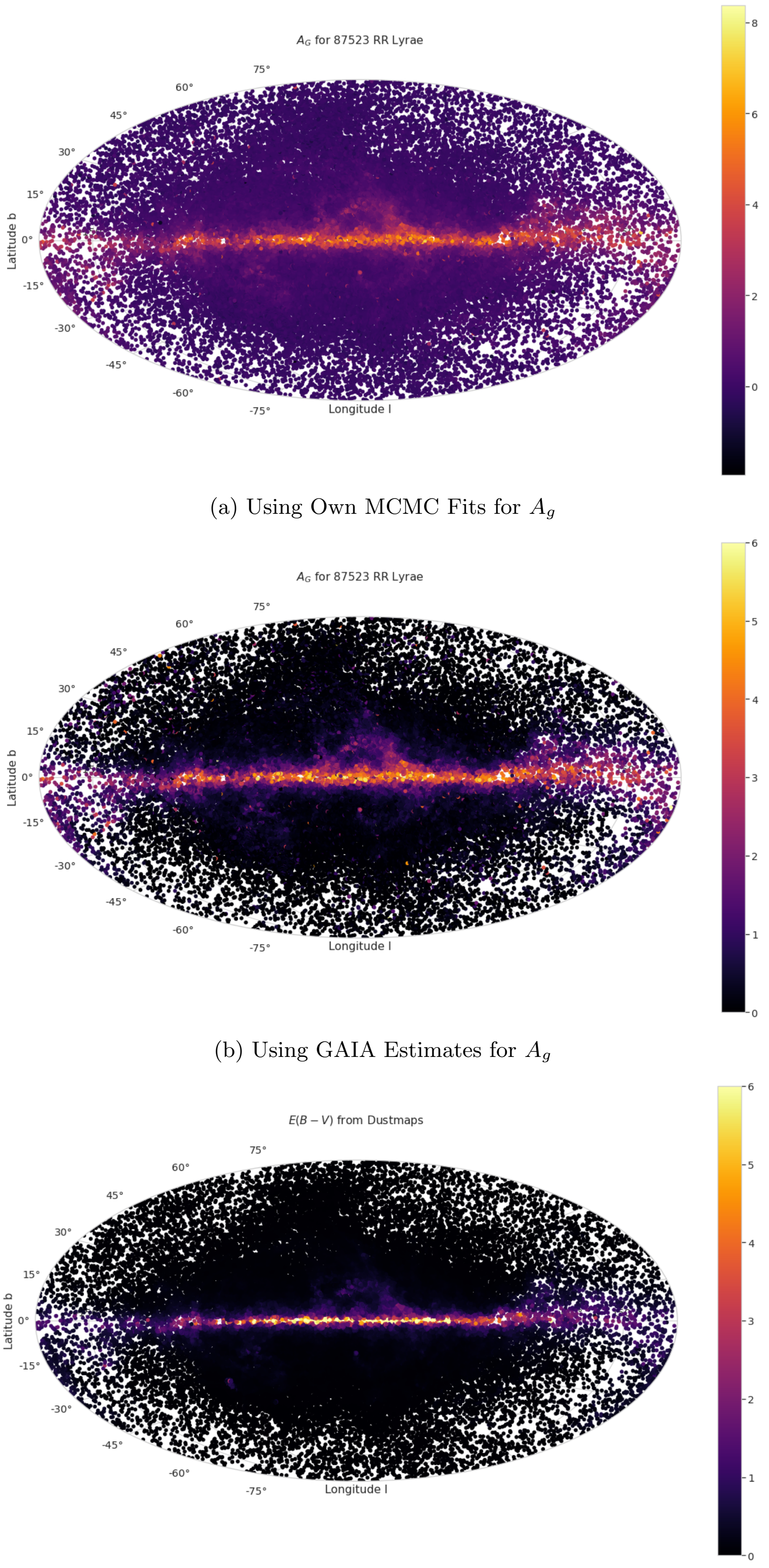
With this intrinsic color difference, we can estimate the A_g for all RR Lyrae in the catalog, and overplot them on an Aitoff-projected map of the sky. We compare our results to those of GAIA's g_absorption in the Figure 7 (next page). Additionally, we generate the SFD '98 dustmap from the Python `dustmaps` package.

On comparison, we find significant agreement between all three sets of parameters. Although there may be small systematic shifts and likely scaling errors as seen in the diagrams, the diagrams agree with the conventional wisdom that the dust near the Galactic bulge and milky way disk causes high extinction in the region, while outside the disk we see lower extinction. A caveat to these extinction maps is that they are presented in 2D, which is not necessarily the most accurate use, depending on the science case. A science case focused on a target nearer to us, in front of the galactic bulge, will have a lower extinction that is made out to be by this 2D dustmap.

We also find a variation in the density of the RR Lyrae. Towards the galactic bulge region with still lower extinction, there are a multitude of RR Lyrae. However, within the bulge and within far out in the disk, there are few RR Lyrae. This is likely just a result of a combination of a variation in stellar density in the galaxy, as well as extinction in the Bulge causing far fewer RR Lyrae to be visible by GAIA, which experiences high extinction, being an optical telescope.

We have covered the science, data processing, and analysis for all portions of the lab. This concludes our analysis for Lab 1.

APPENDIX



(a) Using Own MCMC Fits for A_g

(b) Using GAIA Estimates for A_g

(c) Using Python `dustmaps` E(B-V) estimate.

Figure 7: For the 3 cases, besides differences in scaling, all dustmaps seem to follow an identical structure for Milky Way extinction. The colorbar indicates the extinction A_g. In the case of the `dustmaps` map, it indicates the E(B-V) extinction

.

A. AQDL QUERIES

- Extracting Top 100 RRab Lyrae:

```
    SELECT TOP 100 * \
FROM gaiadr3.vari_rrlyrae \
WHERE pf IS NOT NULL \
AND num_clean_epochs_g > 40;
```

- Extracting Top 100 Lyrae for RRc Selection:

```
    SELECT TOP 100 * \
FROM gaiadr3.vari_rrlyrae \
WHERE num_clean_epochs_g > 80
AND int_average_g > 15;
```

- Extracting Top 100 Lyrae for RRab Selection:

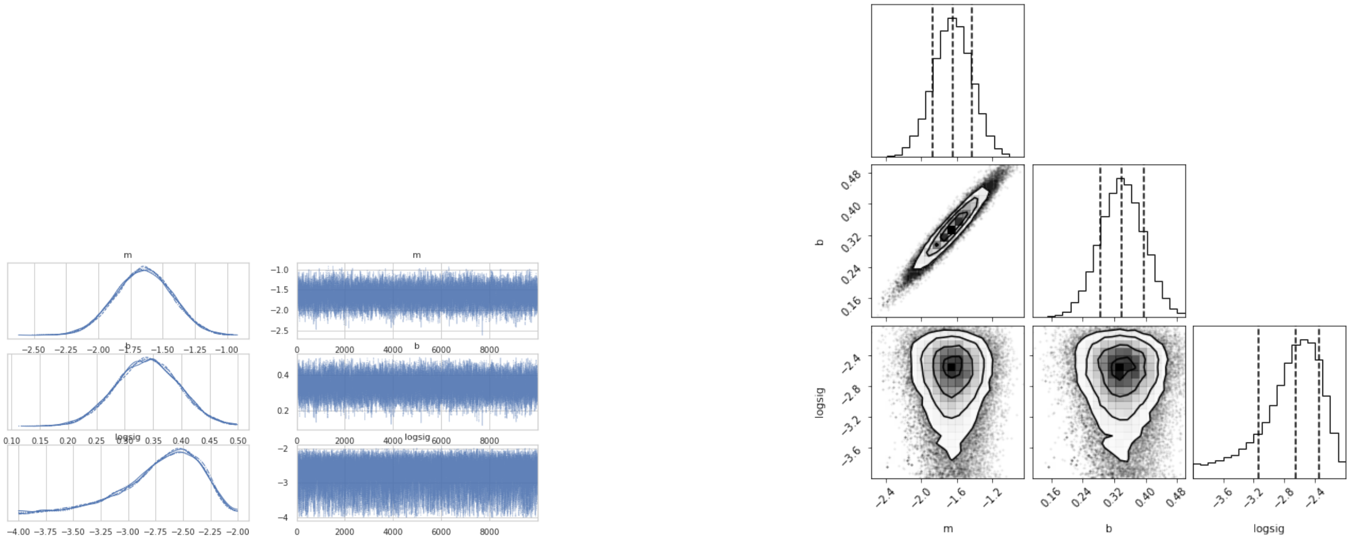
```
    SELECT TOP 100 * \
FROM gaiadr3.vari_rrlyrae \
WHERE num_clean_epochs_g > 40;
```

- Extracting Top Lyrae with Quality Cuts and Low Extinction:

```
    SELECT TOP 600 (1/parallax) as dist,phot_g_mean_mag+5*log10(parallax)-10 as mg, * \
FROM gaiadr3.vari_rrlyrae AS rrly \
JOIN gaiadr3.gaia_source AS dr3 ON
    rrly.source_id = dr3.source_id \
JOIN external.gaiaedr3_distance AS ext ON
    rrly.source_id = ext.source_id
WHERE parallax_error < 0.2 \
AND parallax > 0 \
AND ABS(b) > 30 \
AND (1 / parallax) < 4
AND SQRT(astrometric_chi2_al / (astrometric_n_good_obs_al - 5)) <
    1.2 * GREATEST(1, EXP(-0.2 * (phot_g_mean_mag - 19.5))) \
AND 1 + 0.015 * POWER(bp_rp, 2) < phot_bp_rp_excess_factor \
AND phot_bp_rp_excess_factor < 1.3 + 0.06 * POWER(bp_rp, 2);
```

- Extracting Top Lyrae With Low Extinction Matched with WISE W2:

```
    SELECT TOP 600 (1/parallax) as dist,phot_g_mean_mag+5*log10(parallax)-10 as mg, * \
FROM gaiadr3.vari_rrlyrae AS rrly \
JOIN gaiadr3.gaia_source AS dr3 ON
    rrly.source_id = dr3.source_id \
JOIN external.gaiaedr3_distance AS ext ON
    rrly.source_id = ext.source_id \
JOIN gaiadr3.allwise_best_neighbour AS wbn ON
    rrly.source_id = wbn.source_id \
JOIN gaiadr1.allwise_original_valid AS allw USING (allwise_oid) \
```



(a) Traceplot

(b) Cornerplot

Figure 8: Traceplots and Cornerplots for GAIA G-Band fit.

.

```
WHERE parallax_error < 0.2 \
AND parallax > 0 \
AND ABS(b) > 30 \
AND (1 / parallax) < 4 \
AND SQRT(astrometric_chi2_al / (astrometric_n_good_obs_al - 5)) <
    1.2 * GREATEST(1, EXP(-0.2 * (phot_g_mean_mag - 19.5))) \
AND 1 + 0.015 * POWER(bp_rp, 2) < phot_bp_rp_excess_factor \
AND phot_bp_rp_excess_factor < 1.3 + 0.06 * POWER(bp_rp, 2);
```

- Extracting All RR Lyrae from GAIA catalog that pass Quality Cuts:

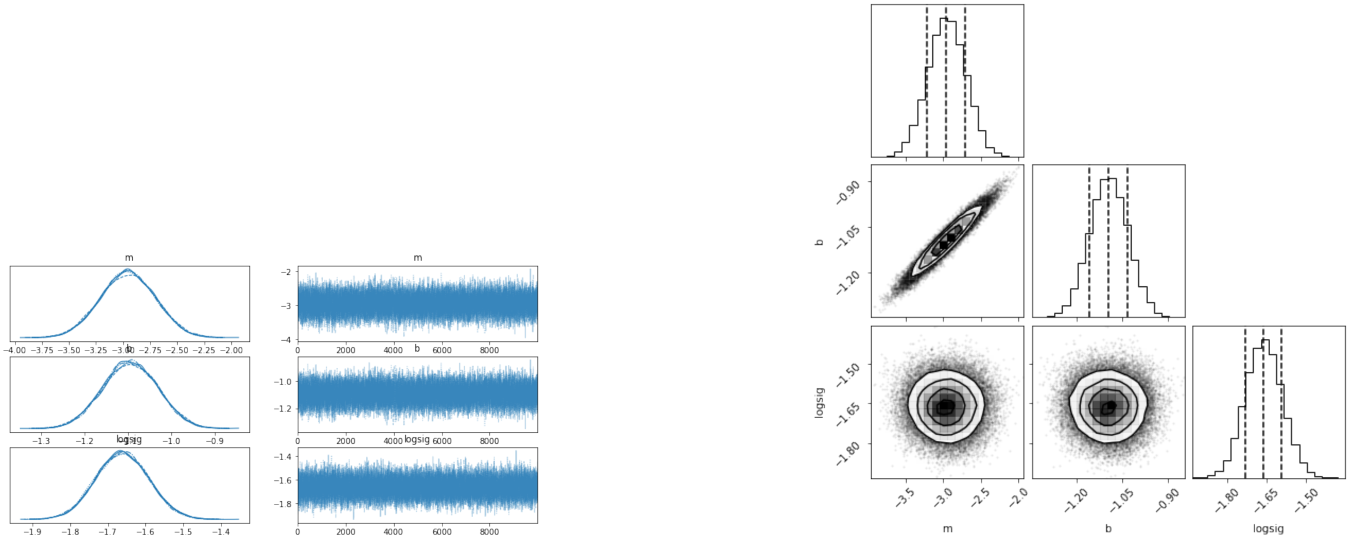
```
    SELECT TOP 300000 bp_rp, pf, g_absorption, best_classification, l, b \
FROM gaiadr3.vari_rrlyrae AS rrly \
JOIN gaiadr3.gaia_source AS dr3 ON
    rrly.source_id = dr3.source_id \
LEFT JOIN external.gaiaedr3_distance AS ext ON
    rrly.source_id = ext.source_id \
WHERE parallax > 0 \
AND SQRT(astrometric_chi2_al / (astrometric_n_good_obs_al - 5)) <
    1.2 * GREATEST(1, EXP(-0.2 * (phot_g_mean_mag - 19.5))) \
AND 1 + 0.015 * POWER(bp_rp, 2) < phot_bp_rp_excess_factor \
AND phot_bp_rp_excess_factor < 1.3 + 0.06 * POWER(bp_rp, 2) \
AND pf IS NOT NULL;
```

B. TRACEPLOTS AND POSTERIORS

REFERENCES

Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G., & Andrae, R. 2018, AJ, 156, 58, doi: 10.3847/1538-3881/aacb21

Clementini, G., Ripepi, V., Leccia, S., et al. 2016, A&A, 595, A133, doi: 10.1051/0004-6361/201629583

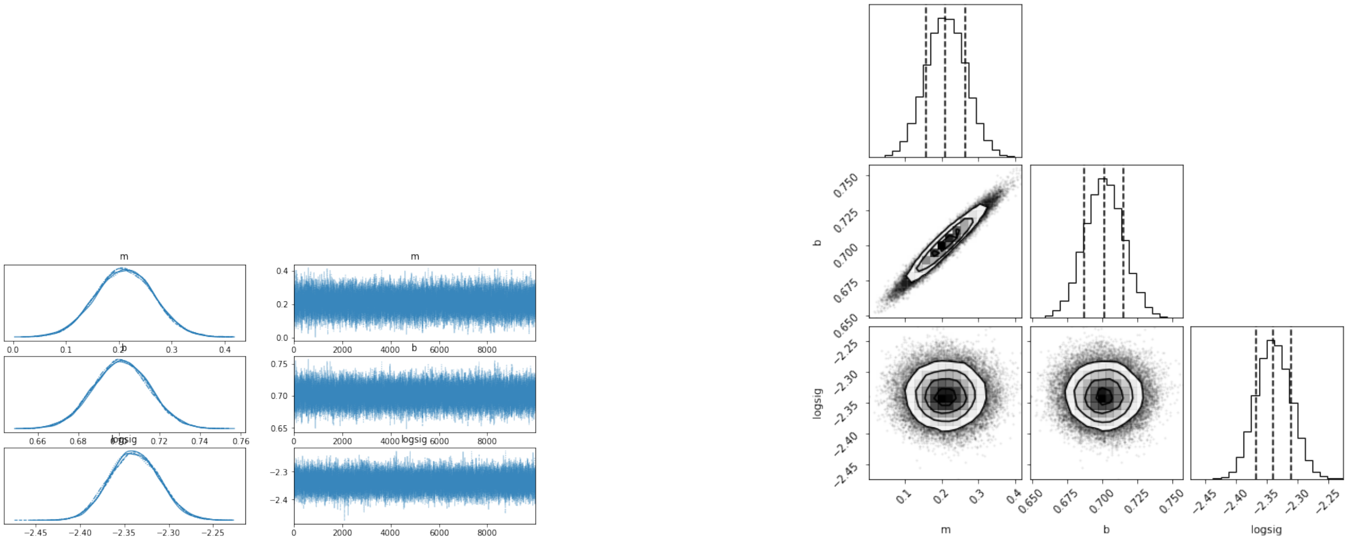


(a) Traceplot

(b) Cornerplot

Figure 9: Traceplots and Cornerplots for WISE W2 fit.

.



(a) Traceplot

(b) Cornerplot

Figure 10: Traceplots and Cornerplots for GAIA BP-RP color fit.

.

Netzel, H., Smolec, R., Soszynski, I., & Udalski, A. 2018, VizieR Online Data Catalog, J/MNRAS/480/1229