Lab 1: Gaia, RR Lyrae stars, and Galactic Dust

Astro 128 / 256 (UC Berkeley)

Assigned: Mon Sep. 15, 2025
Checkpoints: Wed Sep. 24, 2025; Wed Oct. 1, 2025; Wed Oct. 8, 2025
Final Write Up Due: 11:59 PM Sun Oct. 12, 2025

In this lab, we explore a sample of RR Lyrae stars observed by Gaia with the goal of mapping dust in the Milky Way. We make use of the fact that RR Lyrae stars are "standardizable candles"—they follow Leavitt's Law, allowing us to infer intrinsic luminosity (and distance) from periodicity.

All necessary catalogs can be found on the Gaia archive. Go to the *Gaia* archive, click "search", and click on the "Advanced (ADQL)" tab on the upper left. Available catalogs are listed on the left side of the page in nested drop-down menus. We will use both the DR3 and eDR3 catalogs.

We strongly encourage you to read through the linked material, especially regarding error propagation and MCMC fitting.

Technical Components

- Periodograms
- Fourier decomposition
- Databases; SQL
- Python web queries
- Linear and nonlinear optimization
- Markov Chain Monte Carlo (MCMC)
- Bayesian modeling
- Data visualization

Outline

In broad terms, these are the steps you will follow:

- Write ADL queries to download RR Lyrae stars from the Gaia archive.
- The Gaia archive provides time-averaged magnitude, periods, etc. of varying quality. Confirm the reliability of these values. Estimate the periods and mean magnitudes of RR Lyrae and compare them with reported Gaia values.
- Establish a period-relationship for the RR Lyrae. Query parallaxes to nearby RR Lyrae and explore the reliability of the reported distances.
- Determine the period-luminosity (or absolute magnitude) relationship using a Markov chain Monte Carlo (MCMC) fit. Write and test your own simple MCMC routine, and compare to results from `pymc`. (This can be time consuming).
- Finally, create a 2D map of dust in the Milky Way by applying your measured period-luminosity relationship to Gaia's full RR Lyrae catalog, then computing the "color excess" to determine line-of-sight dust extinction as a function of direction.

Preamble

(a) If you run into difficulties with ADQL queries, consult the Gaia ADQL cookbook. The *Query timeouts* section may be particularly helpful.

(b) From Lab 0, you should already have an account on the *Gaia* archive. If not, click "sign in" and "register new user" in the upper right-hand corner. This allows you to save queries and upload/query your own tables.

(c) Download Astroquery. (it is already installed on Datahub). The `astroquery.utils.tap.core.TapPlus` and `astroquery.gaia.Gaia`utilities allow you to combine ADQL queries with Python code. You may use either, but `Gaia` seems to time out faster than the `TapPlus`.

(d) As in Lab 0, cache query results to safeguard against archive crashes. Make sure you can overwrite the cache when you improve your queries.

Assignment

1. The `gaiadr3.vari_rrlyrae` catalog contains $> 10^5$ results for Gaia's data processing and analysis for RR Lyrae classification. Write an ADQL query to download the first 100 rows of the table for which a fundamental pulsation frequency ("pf") has been measured and more than 40 clean epochs were obtained in the G-band. Submit your query using Astroquery and display the first 10 rows of the thus-obtained catalog.
   To find the meaning of individual columns, you can search for them in the *Gaia* DR3 data model. To learn about the meaning of RR Lyrae–specific parameters and how they were derived, see Clementini et al. 2016 and Celmentini et al. 2022. Table 12 lists the relevant columns for the data your download from the Gaia archive.
2. The table you downloaded above contains the results of fitting the light curves of RR Lyrae stars, but it does not contain the raw light curves. The raw light curves can be accessed as described in the Datalink and light curves tutorial.
   Download the light curves (i.e., G-band magnitude vs. time, with magnitude uncertainties) for the 100 RR Lyrae stars in your table. `astroquery` is one option for fetching web urls within Python. Plot a light curve.
3. Estimate the period and mean G-band magnitude for the 100 downloaded light curves. A simple estimate of the period can be obtained using a Lomb-Scargle (L-S) periodogram. An implementation is available in `astropy.timeseries.LombScargle`. Plot the periodogram for one light curve, marking your estimate of the period on the plot.
   Magnitudes are a logarithmic quantity, so simply taking the mean of the measured value is *not* correct. Include error bars on the absolute magnitudes, which are a combination of the flux errors and the zero point errors. More information on error propagation can be found here and here. *Hint: Setting a plausible range of periods for RR Lyrae helps avoid aliasing in your L-S analysis.*
4. Compare the periods you computed from the 100 light curves to the values reported in the `vari_rrlyrae` catalog. Comment on your results.
5. If we want to predict how bright an RR Lyrae will be in the future, we need a functional form, $m_G(t)$, that can be evaluated at a future time t. Stellar atmospheres are complicated, so predicting a closed-form expression for $m_G(t)$ from fundamental physics is hard, but if the fluctuations are periodic, we can use Fourier analysis to estimate $m_G(t)$.
   Any periodic function $f(t)$ can be described by a sum of sines and cosines:
   $$f(t) = A_0 + \sum_{k=1}^{K} \left[a_k \sin(k\omega t) + b_k \cos(k\omega t)\right], \tag{1}$$
   where A_0, a_k, and b_k are to-be-determined constants and $\omega = 2\pi/P$ is the angular frequency. This equation is exact in the limit of $K \to \infty$, but for smooth, periodic functions, it can be accurate even for small K (say, $K = 5$ or 10).
   Show that if ω is known, then the problem of determining the remaining $2K+1$ free parameters can be re-cast as a linear algebra problem; i.e., $y = \mathbf{X}\beta$, where y is an array of measured fluxes, $\mathbf{X}$ is a matrix of known quantities, and β is an array of unknowns. What are $\mathbf{X}$ and β? (write out the terms)?
6. **Note: Past students have indicated this is a particularly challenging problem** For the star with Gaia DR3 id = 4659713442253931776 , determine the series representations for $K = 1, 3, 5, 7$, and 9. Plot the phased light curve, the series representation on a fine grid, and the residuals between the data and series representation for each K.
7. **Note: Past students have indicated this is a particularly challenging problem** How many terms is enough? Use cross-validation to find the optimal K. Designate 20% of the observed points as the cross-validation set. For K ranging from 1 to 25, calculate χ^2_r for both the training data (which

enters y) and the cross-validation data (which does not). Plot this quantity as a function of K. Discuss what an appropriate value of K for this data set might be. It maybe useful to use a log-scale on the y-axis.

8. Use the value of K from (7) to predict the magnitude of the star exactly 10 days after the last observed data point (in the units returned from the Gaia archive). Plot the light curve, showing the last few days of Gaia data points, the extrapolation over the next 12 days, and indicating your best estimate for the magnitude exactly 10 days after the last data point.
9. When you calculated the average magnitude in part (3), you probably used some sort of mean in flux space of the observed data points. This is not ideal because the observations are not necessarily evenly sampled in phase. A more accurate value can be obtained by calculating the mean magnitude based on the mean flux when averaging over one pulsation period. Now estimate the mean magnitude using your Fourier model for each of the 100 light curves. Once again, the averaging should be done in flux space, not magnitude space. Make a plot comparing your estimates of the mean G-band magnitude – both from this estimate and from part (3) – to the estimate in the Gaia catalog. Also plot the residuals, and comment on your results.

Required checkpoint 1, due Wed Sep. 24, 2025: (a) the plot produced in part (6), and (b) the plot produced in part (8). Submit this via gradescope.
It should be a pdf, named Firstname_Lastname_lab1_cp1.pdf

10. So far, all the RR Lyrae we have looked at are in the variability class "RRab". (This is a consequence of our requirement in part (1) that there be a measured "pf"). There are, however, other observationally-motivated classes of RR Lyrae, the most common of which is "RRc".
    Let's compare the light curve shapes of RR Lyrae in the these two classes. Following a similar procedure to the one in part (2), download light curves for the top 3 RR Lyrae in the vari_rrlyrae catalog that have best_classification = "RRc", mean G-band magnitudes brighter than 15, and more than 80 clean epochs in the G-band. Using a suitable number of terms, compute Fourier expansions for these three light curves. Plot the phased light curves and overplot their Fourier models.
    Repeat this for the top 3 RR Lyrae with best_classification = "RRab". Compare the 6 phased light curves and models, using some plotting scheme that makes it easy to see and compare their shapes. Comment on the difference in light curve shape between the two classes. Do some reading (published papers, Wikipedia, etc) about the difference between the classes.
11. Are the 6 light curves you plotted well-described by a single period, or is there evidence of intrinsic scatter? Read about deviations from simple periodicity in RR Lyrae in Netzel et al. 2018 and discuss your findings.
12. Now that you've looked at the light curves for a few objects, we'll assume for the rest of the problem that the periods reported in the Gaia catalog are reliable. We'll now use the data in the `vari_rrlyrae` table to infer the RR Lyrae period-luminosity relation in different bandpasses.
    First, we'll need to use Gaia distances to estimate the absolute magnitude of RR Lyraes. This will only work if there isn't a lot of dust between us and the RR Lyraes. Explain why this is.
    It turns out that most of the dust in the Milky Way is in the disk, at low Galactic latitude. Write an ADQL query to select RR Lyrae stars that (a) have accurately measured distances, with parallax errors of less than 20%, (b) are above or below the disk, with $|b| > 30$ degrees, where b is Galactic latitude, and (c) are relatively nearby, with distances less than 4 kpc. To do this, you'll need to join the `gaiadr3.vari_rrlyrae` and `gaiadr3.gaia_source` catalogs. You should find about 500 objects.
13. Plot the distribution of targets you obtained in Galactic coordinates. Verify that your ADQL query has removed stars in the Galactic disk.
14. Plot period vs. absolute G-band magnitude for all stars returned by your query. Include error bars on your plots.
15. You should see that the majority of the stars have similar absolute magnitude, but a non-negligible fraction of them scatter far off the median relation. This is mostly due to incorrectly measured parallaxes. Apply the quality cuts in Equations C1 and C2 of Lindegren et al. 2018 to the sample. Then plot the period-luminosity relation again. Has the scatter decreased?
16. Most of the "bad" objects should have been removed by the above cut, but a few will remain. You

can crudely remove these outliers by removing objects with absolute G magnitudes greater than some threshold of your choice. You can also try more sophisticated outlier rejection, if you wish. Plot the period-absolute magnitude relation, including error bars on absolute magnitude due to distance uncertainties, from the resulting cleaned sample.

17. **(this problem will be started as an in-class group coding activity. Results from your implementation should be lab report)**. We are going to use Markov chain Monte Carlo (MCMC) techniques to fit models to data. This technique is quite powerful and frequently used in cutting-edge research, in part because it can yield reliable uncertainty estimates. However, it is also likely to be new to you and it is not intuitive when learning it for the first time.
To develop an understanding for how Markov chain Monte Carlo (MCMC) routines work, you're going to start by coding up your own Metropolis-Hastings (M-H) MCMC sampler using a Gaussian proposal distribution. To verify that your sampler works, pretend you have a single measurement of a number, $x = 1$, with error $\sigma = 0.1$. Draw 10,000 MCMC samples assuming a one-dimensional Gaussian likelihood, $p(\mu) = 1/\sqrt{2\pi\sigma^2}\exp\left[-(\mu - x)^2/(2\sigma^2)\right]$. Compare a normalized histogram of these samples to the analytic likelihood $p(\mu)$. Choose a step size such that the acceptance fraction is of order 0.5. To qualitatively demonstrate convergence, plot μ and $\ln P$ versus step number.
18. **Note: Past students have indicated this is a particularly challenging problem.** Fit a line to the period vs. absolute magnitude relation in the G band, using DR3 data. Assume a model $M_G = a \times \log[P/\mathrm{day}] + b$, where a and b are free parameters. Allow for intrinsic scatter in your relation. That is, fit for a positive constant $\sigma_{\mathrm{scatter}}$, such that if there were no measurement uncertainties, values of M_G at a given P would follow a Gaussian distribution with variance $\sigma^2_{\mathrm{scatter}}$. Do this in several steps:
    (i) **(optional extra credit – whether you do this or not, some of the terminology and definitions will be useful for (ii) and (iii))** Once you are satisfied that your sampler works, use it to constrain the values of a, b, and $\sigma_{\mathrm{scatter}}$ in the RR Lyrae period–luminosity relation. You will need to write down a likelihood function, $p\left(\vec{d}|a, b, \sigma_{\mathrm{scatter}}\right)$, where $\vec{d}$ is the data, and a prior, $p(a, b, \sigma_{\mathrm{scatter}})$. State clearly what priors you are assuming. Use Bayes' theorem to calculate the posterior probability distribution, $p\left(a, b, \sigma_{\mathrm{scatter}}|\vec{d}\right)$, which will include an unknown multiplicative factor $p\left(\vec{d}\right)$, the "evidence". Use your sampler to draw samples from the posterior distribution (you don't need to know the evidence to do this). You may need to modify it slightly to account for the fact that the function you are sampling from is now three-dimensional rather than one-dimensional. Once again, tune the step size so that the acceptance fraction is about 0.5, and make diagnostic plots to verify that the sampler has converged.
    Plot 50 random, independent samples from the posterior over the data. Does the spread between samples as a function of period seem consistent with what you'd expect given the data? Explain.
    (ii) Repeat your fit to the period-luminosity relation, using the same likelihood and priors as in (i). But instead of using your M-H sampler, use the no-U-turn Hamiltonian Monte Carlo sampler provided in pymc. Discuss briefly what this sampler is and what some of the advantages are of using it over Metropolis-Hastings. Add your likelihood function to your pymc model as a "potential" term.
    (iii) Repeat step (ii), but instead of enrolling your likelihood function into the pymc model explicitly, simply tell pymc your model: namely, that you expect values of M_G at a given P to have a mean value $\mu = a\log(P/\mathrm{day}) + b$, and to follow a normal distribution with mean μ and variance $\sigma^2_{\mathrm{scatter}} + \sigma_i^2$, where σ_i are the measurement uncertainties.
    On a single corner plot, compare the posterior constraints you obtained in parts (i) and (iii) [, and (i) if doing the extra credit]. If everything worked, they should be basically identical. When we do MCMC fits in the future, you will usually be able to skip right to (iii). We hope (i) and (ii) have provided some insight into what is happening under the hood.

Required checkpoint 2, due Wed Oct. 1, 2025: the corner plot produced at the end of part (18.iii), where you compare posterior constraints for the period-luminosity relation from two (or three, if doing extra credit) sampling methods. Submit this via gradescope. It should be a pdf, named Firstname_Lastname_lab1_cp2.pdf

19. **(optional for extra credit in AY128)** Many of the RR Lyrae identified by Gaia were also observed by the `WISE` survey, which observed the whole sky in the near-infrared. Cross-match your sample of clean RR-Lyrae stars with WISE. The catalogs `gaiadr3.allwise_best_neighbour` and `gaiadr3.allwise_neighborhood`, available on Gaia archive, will be useful.
20. **(optional for extra credit in AY128)** Repeat step (17), now using the WISE "W2" magnitude rather than the G band. You can simply use the magnitude reported in the Wise catalog; you don't need to average any light curves. You can skip directly to fitting method (iv).
21. **(optional for extra credit in AY128)** Comment on the differences between your inferred period-luminosity relations in the optical and in the near-infrared. In which band is the period-luminosity relationship steeper?
22. Compare your derived period-luminosity relations to results in the literature, which can be found in Beaton et al. 2018 or Klein & Bloom 2014. You can compare your G-band relation to their V-band relation. If there are systematic differences between your results and the literature, what might account for them?
23. Following a similar procedure to the one in step (17), derive a period-color relation for RR Lyrae stars in the Gaia bands. That is, a relation between log(period) and the Gaia $G_{\rm BP} - G_{\rm RP}$ color. You will need to calculate color uncertainties using the uncertainties in BP and RP flux. You don't need to use your own sampler, but can skip right to the pymc version.
24. Write an ADQL query to download the full Gaia RR Lyrae catalog (i.e., no longer excluding stars with imprecise parallaxes or low galactic latitude) and cross-match it with the Gaia source catalog.
25. Calculate the *color excess*, $E\left(G_{BP} - G_{RP}\right) = \left(G_{BP} - G_{RP}\right)_{\rm observed} - \left(G_{BP} - G_{RP}\right)_{\rm intrinsic}$, for all RR Lyraes in the catalog. From this, calculate A_G, the G-band extinction, for each star. You may assume

$$R_G \equiv \frac{A_G}{E\left(G_{BP} - G_{RP}\right)} = 2.0.$$

26. Compare your calculated A_G values to the "G_absorption" value provided in the RR Lyrae catalog.
27. Plot a 2-d map of $E\left(G_{BP} - G_{RP}\right)$ as a function of Galactic longitude and latitude, using an Aitoff projection. A simple way to do this is to plot each RR Lyrae as a point in a scatter plot with semi-transparent points, coloring each point by its A_G.
    When you first make the plot, you will find that some large-scale structure is clearly visible, but also that there are some points for which the reddening looks clearly wrong; i.e., the color excess in one point is very different from the adjacent points. Construct appropriate quality cuts to remove these objects. You may want to cut on photometric signal-to-noise and/or BP/RP excess. To ensure reasonable spatial coverage, you should not have any less than $\sim 60,000$ stars (and you may have more). Be sure to clearly explain your cuts and the effects they have on the resulting map.
    You will notice that the distribution of RR Lyrae stars in the catalog is not at all uniform. Why is this?
    Required checkpoint 3, due Wed Oct. 8, 2025: the cleaned reddening map produced in part (27). Submit this via gradescope.
    It should be a pdf, named Firstname_Lastname_lab1_cp3.pdf
28. The most widely used Galactic dust map is the "SFD" map produced by Schlegel, Finkbeiner, and Davis 1998. Incidentally, all three authors have strong ties to UC Berkeley.
    Compare the attenuation map you produced above to the SFD map. You can query SFD using the "dustmaps" Python package. You only need the SFD map; don't worry about installing the larger maps (some of which are several GB).
    Plot the SFD optical reddening map, $E\left(B - V\right)$, sampled at the same positions as your RR Lyrae map. Does the general structure of your map agree with SFD? What about the small-scale details?
29. Do you *expect* your map to look exactly like SFD? Hint: think about how the SFD map is constructed. What are the differences between what it measured and what your map measures?

Next Steps

Your full lab write up is due 11:59 PM Sun Oct. 12, 2025. It should be a full lab report including discussion of all plots and sections highlighted in magenta, named Firstname_Lastname_lab1.pdf.