# Lab 1: Gaia, RR Lyrae stars, and Galactic Dust
# Final Lab Report Due Monday March 4, 2024

Astro 128 / 256 (UC Berkeley)

Feb 6, 2024

In this lab, we'll explore the sample of RR Lyrae stars curated by Gaia. The final goal is to build a dust map of the Milky Way. We'll make use of the fact that RR Lyrae stars are "standardizable candles" – that is, they follow a period-luminosity trend (the Leavitt Law) that can be used to infer intrinsic luminosity using measurements of brightness and periodicity. This, in turn, can be used to infer distances.

As in Lab 0, the catalogs we'll need can all be found on the Gaia archive. To see which catalogs are available, 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 catalogs from DR3 and eDR3.

## Technical Components

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

## Preamble

*Note: you likely already did some or all of Problem 0 for Lab 0.*

(a) For this lab, you will have to write a lot of ADQL queries. If you run into difficulties, you can consult the Gaia ADQL cookbook. This should contain everything you need for the lab. We particularly recommend reading through the "Query timeouts" section.

This tutorial is for your own edification, so spend as much or as little time as you need on it.

(b) Make yourself an account on the Gaia Archive by going to the *Gaia* archive and clicking "sign in" and "register new user" in the upper right-hand corner. You don't need an account to query the Gaia catalogs, but having an account will allow you to save your previous queries and to upload and query your own tables.

(c) Download Astroquery. (Already done for you if you are working on Datahub). The astroquery.utils.tap.core.TapPlus utility will allow you to combine ADQL queries with Python code.

*Note: For Lab 0, some people used a different package, `astroquery.gaia.Gaia`. Feel free to use whatever packages you like to submit queries, but be aware that this one seems to time out faster than the `TapPlus` package.*

(d) As in Lab 0, you should develop a system for saving the results of web queries (i.e., ADQL queries, light curves, anything else that requires an internet connection) locally. This is a safeguard in case, say, the *Gaia* archive crashes and hour before the lab is due. After Lab 0, you know this is not just a hypothetical! That is, once you're run a cell or query once, you should be able to run it again without a web connection. You'll also want to implement a system for overwriting the cache when you improve your queries.

# Problem 1

1. The `gaiadr3.vari_rrlyrae` catalog contains the results of RR Lyrae classification done by the Gaia data processing and analysis consortium for more than $10^5$ RR Lyrae stars. 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.

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. This is most easily accomplished with a Python utility for fetching web urls. `Urllib` is one good option. Plot one of the light curves.

3. Estimate the period and mean G-band magnitude of the 100 light curves you downloaded above. A simple estimate of the period can be obtained using a Lomb-Scargle 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.

   When estimating mean magnitudes, note that magnitudes are a logarithmic quantity, so simply taking the mean of the measured value is *not* correct.

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. Suppose we want to predict how bright an RR Lyrae star will be at some time in the future. To do this, we need to know a functional form, $m_G(t)$, which we can evaluate at some future time $t$. Stellar atmospheres are complicated, so predicting a closed-form expression for $m_G(t)$ from first principles is hard. However, since we expect the fluctuations to be periodic, we can use some results from Fourier analysis to get a good estimate of $m_G(t)$.

Any periodic, smooth 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 ($a_k$ and $b_k$ are arrays that are $K$ terms long), and $\omega = 2\pi/P$ is the angular frequency. This equations is exact in the limit of $K \to \infty$. The reason it is useful for us is that for typical perioidic functions, it is reasonably accurate even for small K (say, $K = 5$ or 10).

In the most general case, determining all the $2K + 2$ free parameters needed to represent $f(t)$ is a complex nonlinear optimization problem. However, if $\omega$ is already known (say, from a periodogram), the problem becomes much simpler.

Show that if $\omega$ is known, then the problem of determining the remaining $2K + 1$ free parameters can be re-cast as a linear algebra problem; i.e., $y = \mathbf{X}\beta$. Here $y$ is an array of the measured fluxes, $\mathbf{X}$ is a matrix that can be constructed from known quantities, and $\beta$ is an array of unknowns. What are $\mathbf{X}$ and $\beta$? (write out the terms).

6. For the star with Gaia DR3 id = 5817567360327589632 (which should be the 7th entry in your table), 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. 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 $\chi^2/N_{\rm data}$ 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.

8. Using a value of $K$ chosen based on (7), predict the expected 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.

**Required checkpoint for Tuesday, Feb 13. Submit to bCourses (a) the plot produced in part (6), and (b) the plot produced in part (8). Submit a pdf, named Firstname_Lastname.pdf**

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.

10. All the RR Lyrae you have looked at so far 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. We'll also want an estimate of the distance that uses our knowledge of Galactic structure to place a prior on the distance inferred from parallaxes. A catalog of distances obtained in this way is described in Bailer-Jones et al. 2020 and can be queried through the *Gaia* archive (`external.gaiaedr3_distance`). You'll want the geometric (not photo-geometric) distances.

    Write another ADQL query that gets these distance estimates and uncertainties for the targets obtained in your query in part (12). The relevant columns are `r_med_geo`, `r_hi_geo`, and `r_lo_geo`. Explain how these are calculated. Besides the prior, what's the important difference between these distance and what you'd get by just inverting the parallaxes?

14. Plot the distribution of targets you obtained in Galactic coordinates. Verify that your ADQL query has removed stars in the Galactic disk.

    Compare the distance estimate from the Bailer-Jones catalog to the naive distance estimate from $d = 1/\varpi$ (where $\varpi$ is the parallax in appropriate units). Explain the trend you see.

    Plot the EDR3 vs Bailer-Jones parallaxes for this sample, showing uncertainties as error bars. How different are the DR3 parallaxes on average? Are the DR3 and Bailer-Jones parallaxes generally consistent within their respective uncertainties? Why or why not?

15. Plot period vs. absolute G-band magnitude for all stars returned by your query. Use the Bailer-Jones distances. You can use the symmetrized errors., i.e, (`r_hi_geo - r_low_geo`)/2.

16. 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?

17. 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.

18. For the same sample, try plotting the period-absolute magnitude relation based on *Gaia* DR3 parallaxes instead of Bailer-Jones parallaxes. Geometric distance estimates based on EDR3 can be found in the `external.gaiaedr3_distance` catalog in the *Gaia* archive. Note that these are indexed based on EDR3 source ids, which are identical to the DR3 ids.

19. Fit a line to the period vs. absolute magnitude relation in the G band, using DR3 data. Assume a model $M_G = a \times \log[P/\text{day}] + b$, where $a$ and $b$ are free parameters. Allow for intrinsic scatter in your relation. That is, fit for a positive constant $\sigma_{\text{scatter}}$, such that if there were no measurement uncertainties, values of $M_G$ at a given $P$ would follow a Gaussian distribution with variance $\sigma^2_{\text{scatter}}$. Do this in several steps:

    (i) Code up your own Metropolis-Hastings (M-H) MCMC sampler using a Gaussian proposal distribution. To verify that your sampler works, use it to draw 10,000 samples from a one-dimensional Gaussian distribution, $p(x) = 1/\sqrt{2\pi\sigma^2} \exp\left[-(\mu - x)^2 / (2\sigma^2)\right]$, with $\mu = 1$ and $\sigma = 0.1$. Compare a normalized histogram of these samples to the analytic $p(x)$. Choose a step size such that the acceptance fraction is of order 0.5. To qualitatively demonstrate convergence, plot $x$ and $\ln P$ versus step number.

    Once you are satisfied that your sampler works, use it to constrain the values of $a$, $b$, and $\sigma_{\text{scatter}}$ in the RR Lyrae period–luminosity relation. You will need to write down a likelihood function, $p\left(\vec{d}|a, b, \sigma_{\text{scatter}}\right)$, where $\vec{d}$ is the data, and a prior, $p(a, b, \sigma_{\text{scatter}})$. State clearly what priors you are assuming. Use Bayes' theorem to calculate the posterior probability distribution, $p\left(a, b, \sigma_{\text{scatter}}|\vec{d}\right)$, 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.

    Use the corner package to visualize constraints on the posterior for the fit.

    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 $\mu = a \log(P/\text{day}) + b$, and to follow a normal distribution with mean $\mu$ and variance $\sigma^2_{\text{scatter}} + \sigma_i^2$, where $\sigma_i$ are the measurement uncertainties.

    On a single corner plot, compare the posterior constraints you obtained in parts (i), (ii), and (iii). 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 (i) have provided some insight into what is happening under the hood.

**Required checkpoint for Friday, Feb 23. Submit to bCourses (a) the plot produced in part (19.i), where you are sampling from a Gaussian distribution and (b) the corner plot produced at the end of part (19.iii), where you compare posterior constraints for the**

**period-luminosity relation from three sampling methods. Submit a pdf, named Firstname_Lastname.pdf, and bring your plots to class for discussion.**

20. 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.

21. Repeat step (19), 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 (iii).

22. 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?

23. 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?

24. Following a similar procedure to the one in step (19), 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.

25. 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.

26. 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

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

27. Compare your calculated $A_G$ values to the "G_absorption" value provided in the RR Lyrae catalog.

28. 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.

    You will notice that the distribution of RR Lyrae stars in the catalog is not at all uniform. Why is this?

    **Required checkpoint for Thursday, Feb 29. Submit to bCourses the cleaned reddening map produced in part (28). Submit a pdf, named Firstname_Lastname.pdf.**

29. 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, $E\,(B-V)$, 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?

30. 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?