Final Lab Report: Determining Distance to Andromeda Galaxy Using Cepheid Variables

Aaryan Thusoo

April 24th 2024

1. Introduction

In 1610, Galileo Galilei's discoveries challenged our view of the universe by revealing that Earth and our Solar System belong to a vast collection of stars within the Milky Way galaxy. He observed that stars, rather than being randomly scattered, clustered within this galaxy. For centuries, astronomers believed all visible celestial bodies were part of the Milky Way until the early 20th century when the nature of the Andromeda Galaxy, then known as M31, sparked debate.

Cepheid variables are stars that pulse radially, causing their brightness to fluctuate in a regular pattern due to expansions and contractions. This pulsation affects their luminosity, which increases as they expand and decreases as they contract, following a predictable Period-Luminosity relationship. This relationship allows astronomers to use Cepheids as 'standard candles' to measure cosmic distances. By determining a Cepheid's pulsation period, its true luminosity can be calculated, and from its apparent magnitude, the distance to the star can be accurately estimated.

Edwin Hubble's application of this method in 1924 to several Cepheids in M31 was revolutionary. His measurements confirmed that M31 was not part of the Milky Way but a separate galaxy, significantly expanding our understanding of the universe. This marked a new era in astronomy, reminiscent of Galileo's transformative discoveries.

This report aims to replicate Hubble's method using modern data from the Cepheid variable star M31-V1 in the Andromeda Galaxy. Data from the American Association of Variable Star Observers (AAVSO 2023) provides the apparent magnitude (m), which reflects perceived brightness and varies with distance, and absolute magnitude (M), which represents brightness at a fixed distance of 10 parsecs. The relationship between a Cepheid's oscillation period and its luminosity can be expressed using Equation 1 below. Then using the difference between M and m, the distance to the star can be calculated using Equation 2.

$$M = -\left[2.43\left(\log_{10}\left(P\right) - 1.0\right)\right] - 4.05 \tag{1}$$

$$d = 10^{(m-M+5)/5} \tag{2}$$

The course has introduced several noise reduction techniques that will be employed to clarify the data. First following with a Gaussian convolution to smooth the data, then a high pass FIR filter will be used to remove low frequency noise. This is followed by a power spectrum analysis to determine the period of M31-V1, using Equation 3 to correlate time and frequencies.

$$T = \frac{1}{f} \tag{3}$$

2. Analysis and Results

There is collected data ranging back to the early 1900's but we will focus on the last two decades of observations. First we spliced the data into several sections. In these sections, light curve data was taken on M31-V1 over periods of 1-2 years before conducted research stalls again. Thus from the original data, the best continuous cycle I could find is used.

As with many observations in Astronomy, setting a consistent set of times to conduct data collection is difficult. Factors such as weather and atmosphere can skew or completely prevent observations entirely. In the provided data set, not only is the time step not consistent, but also there are duplicate times which also need to be accounted for. Having a consistent time step is not possible so first a linear interpolation was conducted to keep the data as untouched as possible while also creating an equally spaced data set. Checking the time steps, most fall to a value of roughly 0.8 but some are much larger and so by instead of taking a mean value, the median time step value is used to get a better result from the interpolation. In the Appendix, Figure 4 shows the plot of the difference values to show the uneven distribution of points and why taking a median is a much stronger choice for an approximated time step. After interpolation the data set expanded from 214 points to 248 and so this did not drastically alter the data we started with.

2.1. Gaussian Convolution

Now with even time steps, the first step in analyzing the data from the Cepheid variable star involves using a Gaussian convolution to smooth out unwanted noise variations picked up during observation. This process involves applying a function, described by Equation 4,

with a specific parameter known as t_h—the half-time of the Gaussian curve, which we set to 3 days. This setting was chosen through trial and error to find a Gaussian which effectively reduces noise while preserving important as much of the cycle details.

$$g(t, t_h) = \frac{1}{\sqrt{\pi} t_h} e^{-\left(\frac{t}{t_h}\right)^2} \tag{4}$$

In Gaussian convolution, each point in the data is replaced by a weighted average of its neighboring points, with weights given by the Gaussian function. By choosing a smaller t_h, the weighting focuses more closely on each point and its immediate neighbors, rather than distant ones. This targeted smoothing helps in maintaining the essential features of the brightness fluctuations of the star, important for our analysis. Figure 1 shows the original and smoothed data, highlighting three distinct cycles of brightness changes, which are crucial for further investigation of the star's behavior.

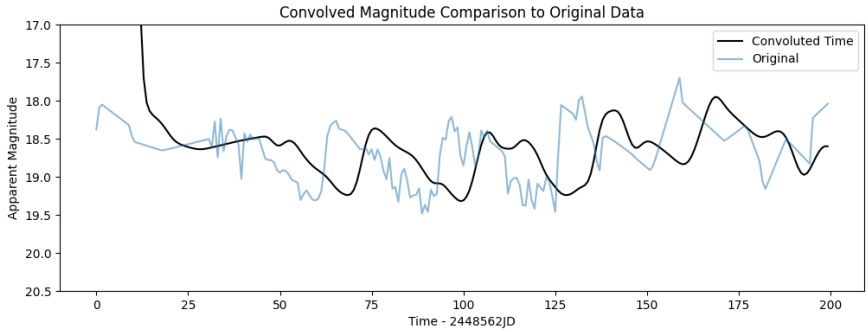

Fig. 1.—: The above plot shows the comparison of the Gaussian convolved data and the original data. There is noticeable reduction in the noise and the overall shape is kept untouched. The positioning of the data is shifted but this is not an issue as the important information is the cycle and not its initial positions.

Before moving on in the analysis, we require the mean apparent magnitude to compare with the absolute magnitude. Now that the majority of noise has been removed, the mean can be taken more accurately. The data for this step is limited in between day 40 and 125 since this is where the cycle of the light curve is most accurate. From this I find an average apparent magnitude of 18.85.

2.2. Window Filter

To focus on the three key peaks in the data while minimizing external influence, we've considered applying a window function, specifically evaluating options like the Boxcar, Hann, and Hamming windows. The selection of an appropriate window function is crucial, especially since we intend to perform a Fourier transform on this filtered data. Spectral leakage is a concern in our analysis—it occurs when the data within the window is not perfectly periodic, which can distort the frequency spectrum. While our data exhibits a consistent pattern, the edges don't perfectly align with this trend, making some degree of spectral leakage inevitable. Applying a window function can help to taper those ends off to help in the frequency domain.

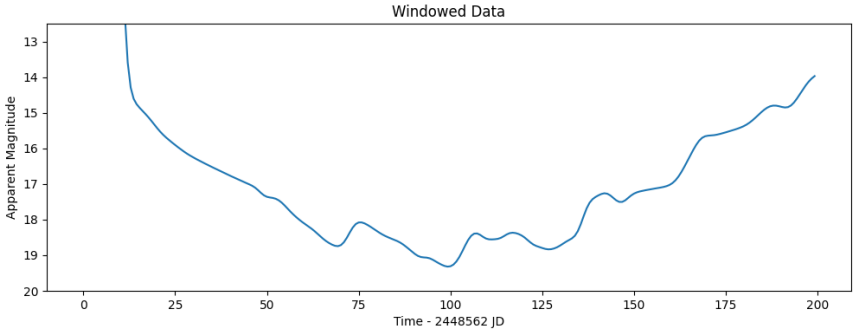

Fig. 2.—: A Hann window function is applied to the data to limit the influence of the outer points in future steps. This will help in lowering the leakage when taking the Fourier Transform. The magnitudes have been affected but this amount does not alter our power spectrum.

A Hann window is chosen for the windowing since the Boxcar Window while perfectly maintaining the magnitude values, does introduce a large amount of spectral leakage into our Fourier Transforms later. However, applying a simple Hann window over the entire length of the data would remove a massive portion of important data making the process pointless. To manage this the size of the window is made 3 times larger than the data array and is centered at the middle. This helps to smooth out the edges and also not alter the magnitudes enough to only change our power spectrum negligibly. Figure 2 shows the placement of the Hann window over our data, which highlights the interval used for analysis.

2.3. High Pass Filtering

Our data is set up now to manage leakage but we must account for one more process in our data which will affect our results. There is an intrinsic peak in our power spectrum at 0 cycles per day and this comes from the average value of the magnitude. This mean value acts as a constant in the light curve data and this forms this peak which we want to remove. To manage this we implement a high-pass FIR filter to provide a cutoff of frequencies below $0.02\,\mathrm{day}^{-1}$. To make this filter, we needed to start by making a low pass filter (Lyons 2004) Thus the first thing we do is determine the Nyquist frequency using the half of the sampling frequency. This is the highest frequency that can be shown in the filter. From here with the setup cutoff, we need the filter coefficients for this data and to do so we generate a sinc function to give the best discrete array representation. This is then limited and tapered off using a quick Hamming window to mitigate leakage later on. Next for normalization we divide all points by the total sum so that results are not amplified or attenuated. This work is all for a low pass filter and so to flip it to a high pass filter we need to apply spectral inversion where all except the middle coefficient have their signs flipped (Vilardell et al. 2010). Now using the filter we can attenuate the peak at the 0 frequency point to give a better view of the intended peak.

2.4. Power Spectrum:

While looking at the data we can assume the period of the cycle is in the range of 25-50 days. This is important as we want to look at the power spectrum to find the peak frequency. This frequency when plugged into Equation 3 gives us the period of our data but to get an accurate result we need to make sure the frequency resolution is adequate. By taking the Fast Fourier transform of the data set we can then determine the power spectrum which will return a peak frequency which correlates to the period of M31-V1. An issue that arose when taking the transform, only one peak came at 0 cycles per day. This result does not make sense and the reason is our time step was too small to provide a small enough frequency resolution and so we need to decide on a method for better viewing the data. Using the same equation as before and rearranging to solve for frequency, and taking the low and high end of our estimated period, we see we need a frequency resolution on the order of 10^{-2} cycles per day.

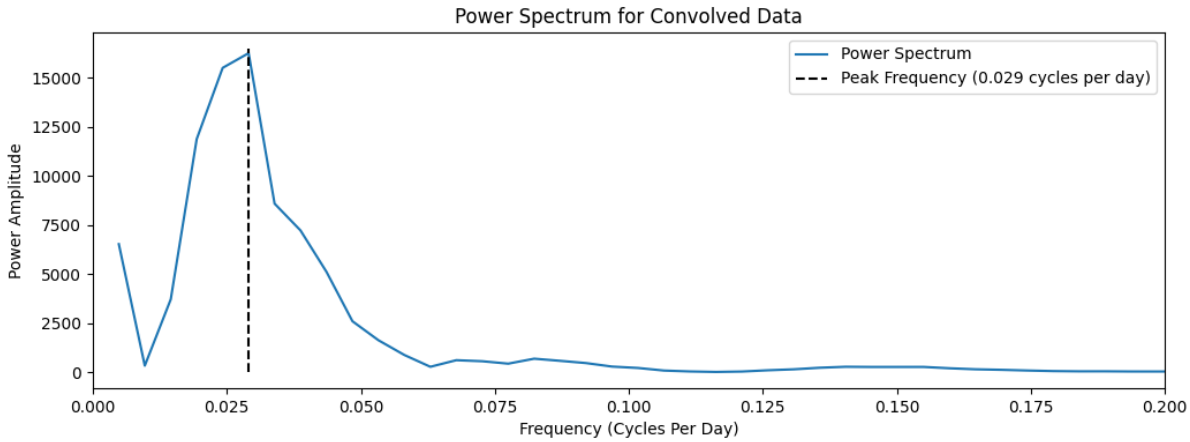

Fig. 3.—: Here the power spectrum has a major peak sitting at the 0.029 cycles per day mark. Taking the inverse of this value provides us with the period of M31-V1

The method chosen to increase the resolution of our Fourier transform is to instead pad the data with zeros. By applying a discrete array of zeros to the end of the data we make the length of observation artificially longer. The corresponding resolution of the frequency does not become smaller but rather increases the density of the frequency points. With this we have an easier time observing finer details in the power spectrum. When padding, it is best to extend the data to the nearest higher power of 2 and so the data is padded up to a data set of 256 points. Before padding the analysis of the peaks was too difficult and so after padding, the frequencies become visible and when checking the values we find the peak sits at $0.029\,\mathrm{day}^{-1}$. The final power spectrum plot is shown below in Figure 3.

2.5. Finding the Distance

The next step in determining the distance to the Andromeda galaxy involves analyzing the peak frequency of the observed light from the Cepheid variable star M31-V1. Using Equation 3, we calculate the period of pulsation associated with this peak frequency to be 34.4 days. This calculated period is crucial because it helps classify M31-V1 as a Type I Cepheid Variable, also known as a Classical Cepheid. This classification is important because Classical Cepheid's exhibit a well-defined Period-Luminosity relationship, allowing us to directly link the pulsation period to its absolute magnitude using Equation 1.

Once the absolute magnitude of M31-V1 is determined, we proceed to estimate its apparent magnitude. This is done by recording the peak and minimum brightness values over several pulsation cycles. From these values, we calculate the mean apparent magnitude

by averaging the maximum and minimum magnitudes, simplifying our approach to handling the variable brightness characteristic of Cepheid's. Finally, with both the absolute and apparent magnitudes known, we apply the distance modulus formula Equation 2. This formula relates these magnitudes through the logarithm of the distance to M31-V1, providing a direct measure of the distance to the Andromeda galaxy. The computed is provided in parsecs so to find the result in light years, we multiply by 3.262 giving the final answer of 2.26×10^6 lyr.

3. Discussion

The real distance to Andromeda as determined by scientists is 2.43×10^6 lyr $\pm$ 0.11×10^6 lyr (Vilardell et al. 2010). Our findings are close to this value, indicating that the methods we used are effective and comparable to advanced techniques used by modern astronomers. While our result is encouraging, it is important to consider the possible sources of the slight discrepancies, such as the tools we used, weather conditions during observations, or our data analysis methods. Investigating these factors could help improve the accuracy of future measurements and demonstrates the dynamic nature of scientific discovery, which evolves with new insights and technologies.

First was the interpolation which may have been a source of error. The linear assumption made was useful but there are portions where potentially some data was lost due to the uniform spacing set up. The spacing set left limited room for the frequency resolution and so if possible having a smaller time step could help get an even more accurate peak frequency value. Of course however we know adding more time steps removed more of the raw information from our data. Following from this, if the Hann window application distorted the periodic motion of the data we may have lost some information there too. A Box window like mentioned before would avoid any attenuation or flattening of data but the spectral leakage would become a cause for concern in the Fourier analysis. Potentially a low pass filter could be used on a Box car window to limit the spread outwards.

In this study, we attempted to use auto-correlation to analyze the periodicity of Cepheid Variable star magnitudes, which helps in identifying patterns by comparing the data to its shifted versions over time. However, this approach was less effective for us due to the limited amount of data, which only covered a few cycles. This made it hard to detect clear, statistically significant patterns in the auto-correlation function. The presence of noise in our observational data also made it difficult to identify clear periodic signals. To improve results in future studies, we could extend our observation period to include more cycles and apply methods to reduce noise.

In 1924, Hubble discovered the first galaxy outside our own, suggesting that the Universe might be much larger than previously thought. His initial calculations, though incorrect, estimated the distance to this galaxy at 900,000 light years, significantly less than what we know today. Despite the error, his findings motivated more astronomers to gather data and confirm the existence of other galaxies. This report follows a general approach to such analyses, but more sophisticated methods could yield better results. It demonstrates that the time series methods taught in this course are applicable not only to classroom topics but also to the field of astronomy.

4. Bibliography

REFERENCES

AAVSO. 2023, M31 V0619, https://www.aavso.org, accessed: April 24th 2024

Lyons. 2004, Understanding Digital Signal Processing (Bernard Goodwin)

Vilardell, F., Ribas, I., Jordi, C., Fitzpatrick, E. L., & Guinan, E. F. 2010, Astronomy and Astrophysics, 509, A70

This preprint was prepared with the AAS LaTeX macros v5.0.

5. Appendix

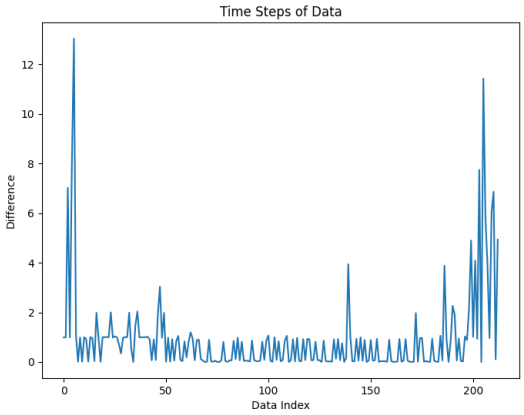

Fig. 4.—: This figure shows the time step values for each point in the data set. There are clear inconsistencies and so to choose a new time step to interpolate over, choosing the mean would be heavily affected by the few high values. Taking the median make for a much fairer and accurate time step.