Main Sequence Reconstruction and Linear Regression

Jon Heidema
Astro Statistics - Dr. Nihan Pol
2025 Spring Semester
Texas Tech University

Abstract:

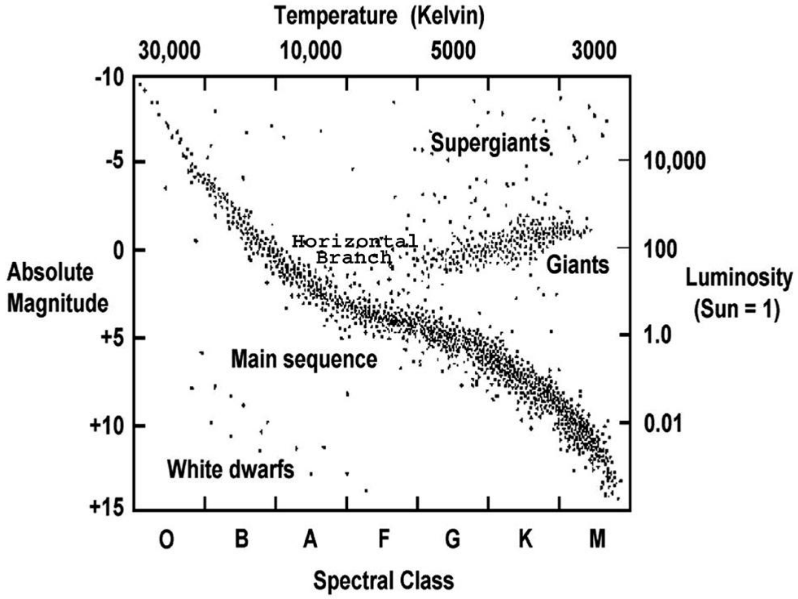
Using a Gaia dataset of stars, a H-R diagram was graphed and the main sequence data points isolated and a linear regression construction and analyzed. The data values of surface temperatures and luminosities of 600000 different stars were imported from a csv file and converted into an array. A log(L) vs log(T) [H-R diagram] was constructed using these values and the data was cleaned and analyzed. Mean-shift clustering was attempted with optimized bandwidth, giving 6 clusters as an ideal. A K-means proved more successful clustering and was used to reconstruct the Main Sequence branch and the data taken as a new array. This required prior knowledge of the structure of the main sequence branch and a manual selection from 10 clusters. This resulted in a dataset of as many values from just the main sequence as possible, while excluding those from the horizontal branch. The number of dwarf stars was insignificant in analysis, but with more clusters, those could be excluded as well. Using the data from the 8 selected clusters, a linear regression was constructed and analyzed. This was done for both all the data, and 'representatives' of each cluster (individual points [mean of cluster]). Using chi-squared and Hubber loss functions, best fit lines were constructed and compared. For the full data: squared loss (y = 6.25x - 23.3) (max log likelihood; -5506715), Hubber loss (y = 6.64x - 24.8) (max log likelihood; -1598871), with Hubber loss being better. For the point data: squared loss (y = 6.25x - 23.3) (max log likelihood; -29.1), Hubber loss (y = 6.77x - 25.4) (max log likelihood; -13.4), with Hubber loss being better.

Introduction:

Background: In astrophysics, there are many different parts of the cosmos that can be seen and measured with telescopes and other instruments. Some of the most common objects to measure and study are stars, black holes, nebulae, galaxies, etc. This paper focuses on stars. The temperature and luminosity of a star's core cannot be directly measured, so it is conventional to use surface temperature (T_{eff}) and absolute magnitude (luminosity (L)).

A standard way to represent this data is by comparing the $log_{10}(T_{eff})$ and $log_{10}(L)$. This is called a Hertzsprung-Russell Diagram. In this, the temperatures can be separated into Spectral Classes, labeled O, B, A, F, K, G, M, with O being the hottest and M the coolest. These also correspond to the color of the star. The luminosity can be represented in terms of factors of the sun's luminosity. It is convention to have the hottest stars on the left and the brightest at the top.

Figure 1: Hertzsprung-Russell Diagram [1]



The H-R Diagram is particularly useful because it helps convey stellar evolution, and how stars develop and evolve over time. [1]

There are some common classifications of stars: Main Sequence, Giants, Supergiants, White Dwarfs. The Main Sequence stars are the most abundant type and are where stars spend the majority of their lifespan. This section is going to be the focus of the paper.

The objective of this paper is to analyze the Main Sequence stars by isolating them from the Horizontal Branch stars and best fitting them.

Motivation: I am personally very interested in astrophysics and the physics of stars and their evolution. I was particularly interested in stellar evolution and how different starting masses led to different types of stars and lifetimes. The fact that there is a definite correlation between temperature and luminosity in the main sequence (see H-R diagram) sparked my curiosity about the mechanics behind star formation and development. I decided I would enjoy attempting to find a best fit line for the main sequence stars using just statistics, seeing as how we did a similar thing in my astrophysics class but using physics instead.

In order to find a best fit, I would have to isolate as many of the main sequence stars from the horizontal branch ones (giants) and possible supergiants and dwarfs if there were too many in my chosen dataset. This type of process and analysis leads well to clustering, which is another part of statistics I am quite fascinated by. I would use clustering to take as much data from just the main sequence as I could, while leaving out the horizontal branch giants and dwarves if there were too many. I would then fit different linear regressions and compare them. The process of selecting the clusters with only the main sequence and choosing a linear best fit is done because I have prior knowledge of what the main sequence branch looks like.

Methods:

Kaggle has a Gaia DR3 dataset of 600000+ stars [2]. The dataset included numerous statistics of the stars, including: rotation, velocity, apparent magnitude, color index, temperature, distance, radius, luminosity, mass, etc. The important data for this paper is temperature and luminosity. Importing the data and reading in just the temperature and luminosity data allows for the creation of a simple dataset. The direct and unfiltered dataset (Raw Data) was initially plotted, taking the $log_{10}(T_{eff})$ and $log_{10}(L)$ to compare to a standard H-R diagram and see if initial data does indeed appear to be an H-R diagram.

Plotting the raw data reveals a recognizable, albeit messy, H-R diagram. However there do appear to be strange vertical "artifacts" closer to the higher temperatures. This is not a coding or graphing error and is just what the data is. Many temperature values did not have corresponding luminosity values, so the data was parsed through and any temperatures with 'null' luminosity values were discarded. This was then graphed again and showed no discernable change, reinforcing that the vertical artifacts were just the raw data values. Lastly, a random sample of 60000 values was taken to create a smaller dataset, useful later on when attempting means - shift clustering.

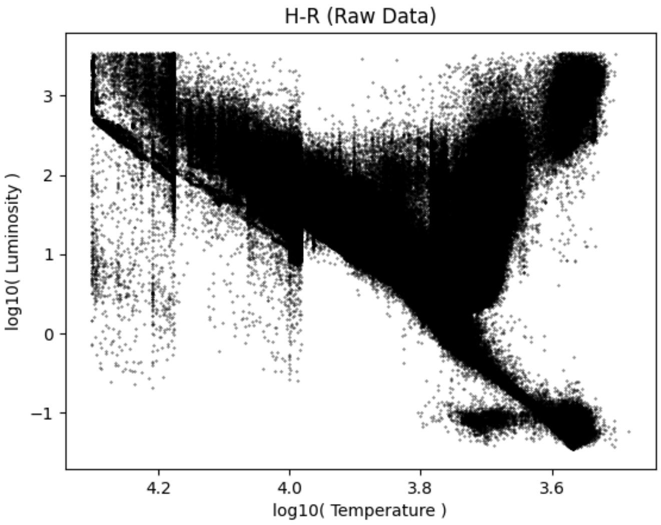


Figure 2: Raw Data - from Gaia DR3 [2]

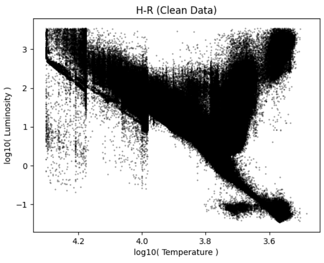


Figure 3: Clean Data - got rid of temperature data with null luminosity values

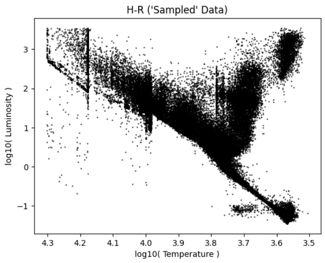
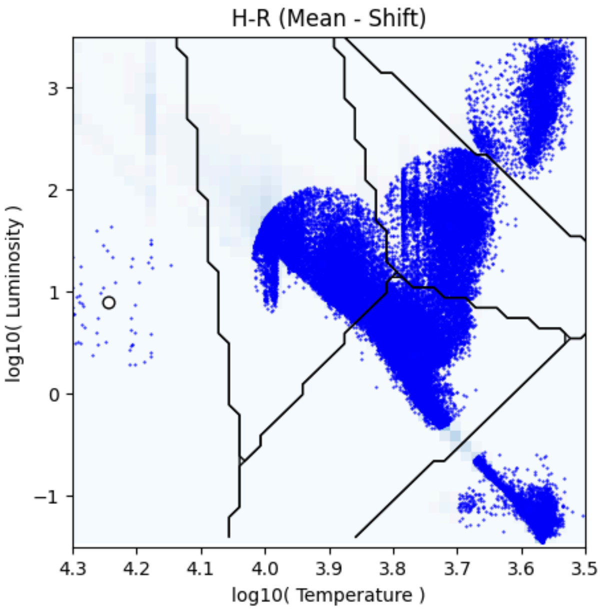


Figure 4: Sampled Data - randomly for smaller dataset (10x smaller)

After getting the data imported and cleaned, the next step was to use clustering to isolate the main sequence branch values. With clustering, only the clusters with the main sequence values would be used and put into a new array to act as the new dataset. The first attempt was to use mean-shift clustering, having the code optimize bandwidth and the optimal number of clusters for the data. However, optimizing the bandwidth takes a very long time, and is not able to be completed with the entire dataset, so the sampled data was used (10x fewer data points).

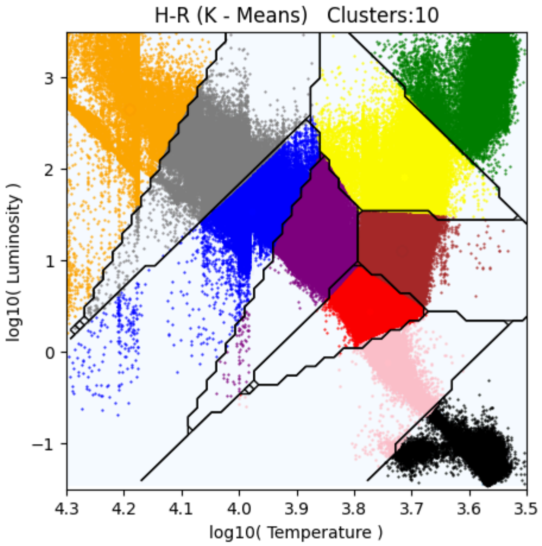
Figure 5: H-R Diagram 6 Mean-Shift Clusters



The optimal bandwidth was calculated: 0.7
The estimated number of clusters calculated: 6

While the bandwidth and number of clusters was successfully taken, and it appeared that the values in the clusters containing only the main sequence branch, when graphed, many data points seem to have disappeared. This is most likely a coding issue and not a mean-shift issue, but K-means was working better and allowed more flexibility. The clustering was also not particularly important in its optimization because the main sequence clusters were taken manually anyways.

For K-means clustering, theoretically using more clusters would provide more freedom in the ability to select which ones were truly main sequence values. However, because the data lacked labels and was already quite "messy" and it was already an estimation with the main purpose to be to exclude as many horizontal branch values, 10 clusters seemed appropriate and satisfactory to reconstruct a 'mock' main sequence. Using the prior information of what values were most likely to be on or near the main sequence [1], the clusters in the top right are most likely from the horizontal branch and excluded.



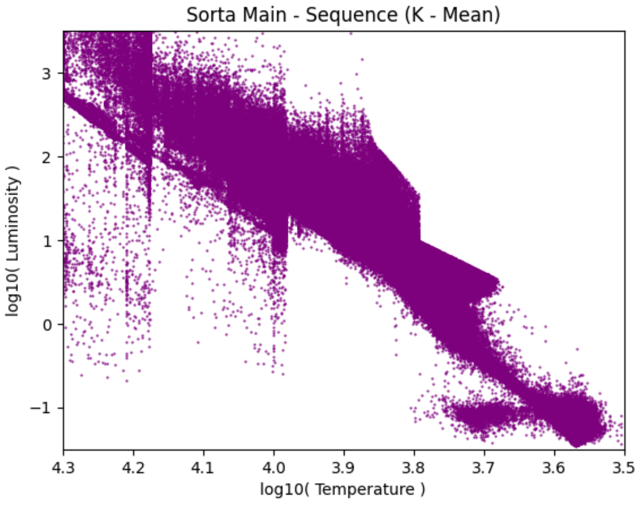


Figure 6: H-R Diagram 10 K-Means Clusters

Figure 7: Mock Main Sequence 7 Clusters

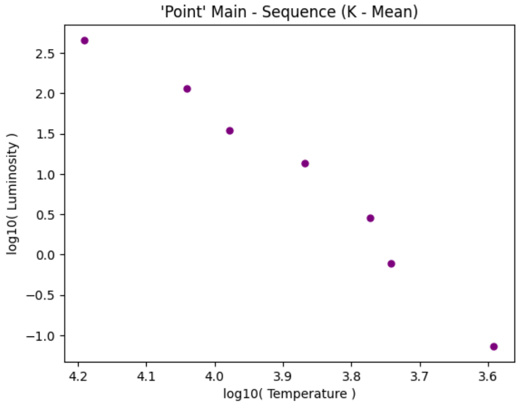


Figure 8: Mock Main Sequence 7 Points

The clusters can also be represented as individual points. These are the mean values of each cluster and can be plotted similarly. By this process, the more clusters used and manually selected, the more individual points would also be generated. With enough clusters, a more descriptive graph of just points (cluster mean) could be constructed. However, because the data is unlabeled, this process would be manual and tedious and inefficient and more prior dependent with a sufficient number of clusters and cluster exclusion.

To find the best fit, linear regressions were used and compared. Linear regression was chosen because when looking at the main sequence of just the points (means), it seems it would fit a linear fit decently well. Non-linear regressions could be used, but for this project only linear regression is considered. The regressions were constructed using both squared loss and Hubber loss. Hubber loss seemed more ideal because there were a lot of outliers in the data, especially towards the higher temperatures of the full main sequence data set. These were both done for both the full main sequence data and the cluster point data.

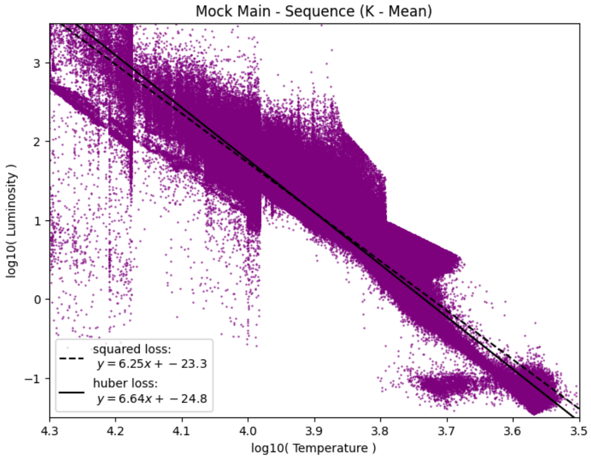


Figure 9: Mock Main Sequence Full Data
Squared Loss and Hubber Loss Linear Regressions

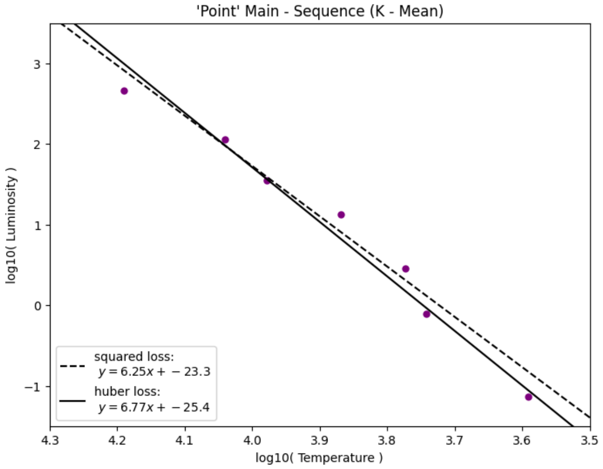


Figure 10: Mock Main Sequence 7 Points
Squared Loss and Hubber Loss Linear Regressions

Full main sequence: Squared Loss (y = 6.25x - 23.3) with maximum log likelihood: -5506715
Hubber Loss (y = 6.64x - 24.8) with maximum log likelihood: -1598871
Point main sequence: Squared Loss (y = 6.25x - 23.3) with maximum log likelihood: -29.1
Hubber Loss (y = 6.77x - 25.4) with maximum log likelihood: -13.4

Comparing the squared loss and Hubber loss likelihoods reveals that in both cases, the Hubber loss regression is better, which is to be expected due to the number of outliers.

References

Dataset - Gaia Dataset Stellar Classification
https://www.kaggle.com/datasets/realkiller69/gaia-data-set-for-stellar-classificationdr3

[1] - Hertzsprung-Russell Diagram
https://chandra.harvard.edu/edu/formal/variable_stars/bg_info.html

Code - Dr. Nihan Pol
Lectures and Homeworks - 2025 Spring Semester Astro Statistics

Appendix

Jupyter Notebook (code)