A Classification of Stars Using GMM

Caleb Caulk, Tyler Moll, Kevin Jacob

Introduction

Star clustering is a well-discussed problem in astronomy. Identifying and understanding different star clusters can contribute to a better understanding of the universe, and how it was formed. Historically, this task has been completed manually, but with the advent of software and various clustering algorithms, it is clear to see how the task of stellar clustering can be made exponentially quicker, and more accurate. In this project, we will use GMM and expectation maximization to accomplish this task.

Clustering stars based purely on their spatial proximity to each other is a classic example of stellar clustering and illustrates the difficulty of related problems. There are two primary methods for modern stellar distance measurement, relative to Earth. These methods are as follows: measurement by parallax, and measurement by calculated brightness. Measurement by parallax involves checking the relative movement of a star as the Earth moves in orbit. Measurement by brightness involves comparing the actual brightness of a star to the perceived brightness of a star. By measuring the drop in brightness, scientists are able to tell roughly how far away a star is.

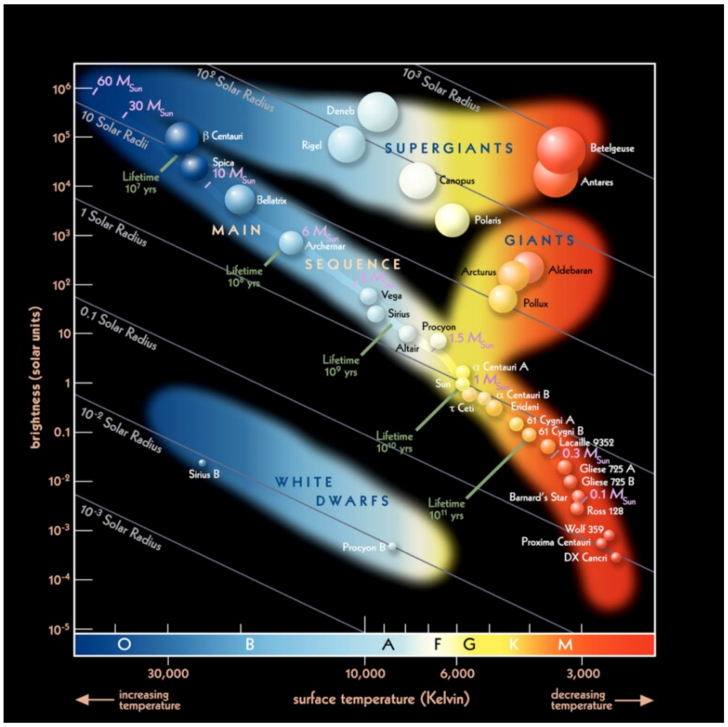
Various environmental factors can cause exaggerated error in stellar distance measures. Included are factors such as gravitational warping of light and space debris that cause errors in the measurements of brightness, color, and angle relative to Earth, which will eventually cause error in the spatial clustering of stars. Due to this concern and many others, researchers have examined various clustering algorithms and techniques that are able to group stars based on many factors.

Even a seemingly trivial example of clustering stars based on their physical proximity can be proven to be extremely difficult. Through the above example it can be easily understood that similar problems such as clustering stars to predict their type, size, or behavior will involve similar, and equally difficult problems.

This project will attempt to apply GMM and the expectation maximization algorithm to the problem of classifying star types in order to prove the feasibility and utility of GMM in the configuration that we have implemented. We will utilize several factors, such as color index, and luminosity to accomplish this.

Data

We will be using the HYG version 3 star data found in this repository. Generally, stars can be categorized into the following categories, using criteria such as surface temperature (Kelvin) and brightness (solar units). Primary categories include “main sequence stars”, “supergiants”, “giants” and “white dwarfs". White dwarfs are the remaining core of a low mass main sequence star that has lost its outer layers in a planetary nebula.



The database set linked above has 25 fields, each listed below with a general description, and a note if the field is likely to be important to most users. **All descriptions come from the original repository as the source.**

1.) **id**: The database primary key

2.) **hip**: The star's ID in the Hipparcos catalog, if known (Catalog of measurements of the star)

3.) **hd**: The star's ID in the Henry Draper catalog, if known (Another Catalog)

4.) **hr**: The star's ID in the Harvard Revised catalog, which is the same as its number in the Yale Bright Star Catalog (Another Catalog)

5.) **gl**: The star's ID in the third edition of the Gliese Catalog of Nearby Stars (Another Catalog)

6.) **bf**: The Bayer / Flamsteed designation, primarily from the Fifth Edition of the Yale Bright Star Catalog. This is a combination of the two designations. The Flamsteed number, if present, is given first; then a three-letter abbreviation for the Bayer Greek letter; the Bayer superscript number, if present; and finally, the three-letter constellation abbreviation. Thus Alpha Andromedae has the field value "21Alp And", and Kappa1 Sculptoris (no Flamsteed number) has "Kap1Scl" (Another Catalog)

7.) **ra, dec**: The star's right ascension and declination, for epoch and equinox 2000.0 (location in space relative to Earth)

8.) **proper**: A common name for the star, such as "Barnard's Star" or "Sirius". These are taken from the International Astronomical Union (Common name of the star)

9.) **dist**: The star's distance in parsecs, the most common unit in astrometry. To convert parsecs to light years, multiply by 3.262. A value >= 100000 indicates missing or dubious (e.g., negative) parallax data in Hipparcos. (Distance in parsec from Sun I think and over 100,000 is bad data so don't use)

10.) **pmra, pmdec**: The star's proper motion in right ascension and declination, in milliarcseconds per year. (how much the star moves)

11.) **rv**: The star's radial velocity in km/sec, where known (How fast the star is moving usually towards (negative) or away from Earth)

12.) **mag**: The star's apparent visual magnitude. (How bright the star looks)

13.) **absmag**: The star's absolute visual magnitude (its apparent magnitude from a distance of 10 parsecs) (How bright the star looks from 10 parsec)

14.) **spect**: The star's spectral type, if known (O, B, A, F, G, K, and M based on surface temp)

15.) **ci**: The star's color index (blue magnitude - visual magnitude), where known (Number to help determine Spectral type)

16.) **x,y,z**: The Cartesian coordinates of the star, in a system based on the equatorial coordinates as seen from Earth. +X is in the direction of the vernal equinox (at epoch 2000), +Z towards the north celestial pole, and +Y in the direction of R.A. 6 hours, declination 0 degrees (More position data)

17.) **vx,vy,vz**: The Cartesian velocity components of the star, in the same coordinate system described immediately above. They are determined from the proper motion and the radial velocity (when known). The velocity unit is parsecs per year; these are small values (around 1 millionth of a parsec per year), but they enormously simplify calculations using parsecs as base units for celestial mapping (Velocity data that will show where the stars are moving)
18.) **rarad, decrad, pmrarad, pmdecrad**: The positions in radians, and proper motions in radians per year (More motion and position data in a different format)
19.) **bayer**: The Bayer designation as a distinct value (A way to name the stars based on constellations)
20.) **flam**: The Flamsteed number as a distinct value (Another way to name/ number bright stars)
21.) **con**: The standard constellation abbreviation (Constellation)
22.) **comp, comp_primary, base**: Identifies a star in a multiple star system. comp = ID of companion star, comp_primary = ID of primary star for this component, and base = catalog ID or name for this multi-star system. Currently only used for Gliese stars (More naming stuff I think)
23.) **lum**: Star's luminosity as a multiple of Solar luminosity (Luminosity)
24.) **var**: Star's standard variable star designation, when known (Unique identifier to variable stars where a variable star is a star whose brightness changes overtime)
25.) **var_min, var_max**: Star's approximate magnitude range, for variables. This value is based on the Hp magnitudes for the range in the original Hipparcos catalog, adjusted to the V magnitude scale to match the "mag" field (Not super sure)

The two fields that will be primarily used for analysis and clustering are **color index** (field 15) and **luminosity** (field 23) fields.

Real-World/ Exploratory

There are many papers discussing similar problems to the ones explored in this project. A few of the most relevant ones are listed below:

1.) Stellar Cluster Detection using GMM with Deep Variational Autoencoder [2]
2.) Application of the Gaussian mixture model in pulsar astronomy - pulsar classification and candidates ranking for the Fermi 2FGL catalog [1]

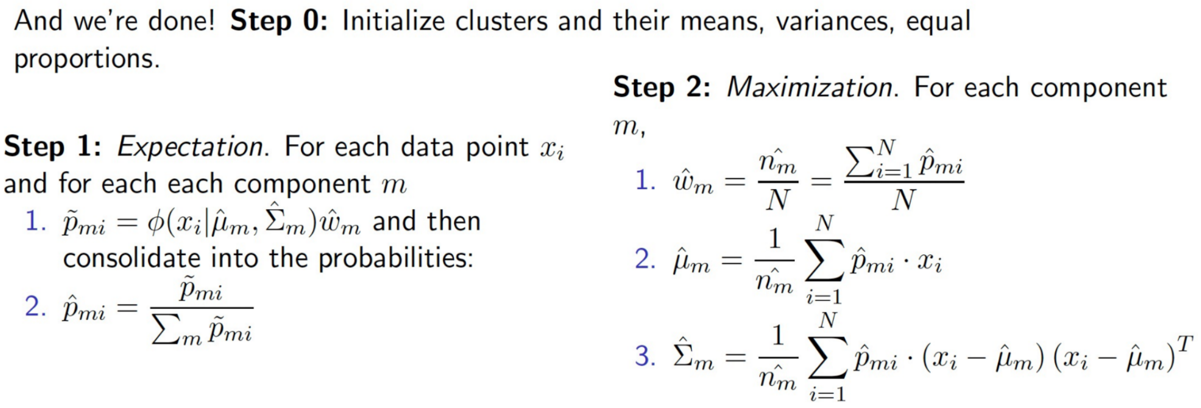
In the first paper, the three researchers discuss new techniques to detect stellar density using a GMM based classification with “Deep Variational Autoencoder” for star cluster detection. In the paper, the researchers exclusively use available image data, and forgo commonly used data repositories for their classification. Using the data provided in these images such as pixel position and pixel intensity, the researchers were able to accurately convert distances between stars, and in turn provide a basis for their technique to accurately cluster a star group. In the concluding section of the paper, it is listed that the new technique allows for stellar classification significantly faster than other approaches, while remaining fairly accurate.

The second paper reintroduces GMM and its application in data modeling and classification in the context of pulsars and the 2FGL catalog point sources. Specifically, the paper first aims to find an empirical definition of millisecond pulsars (MSPs) from a catalog of known pulsars. And second, the paper tries to describe the 2FGL catalog point-source distribution, to generate the likelihood that a particular source is a pulsar, and to produce a candidates list of pulsars for later confirmation. In this second application, the researchers use three variables to categorize their sources: Variability Index (VI), significance curve (Sc), and integral gamma-ray flux (between 1 and 100 GeV) for error correction.

These two papers, among many others, exhibit numerous similarities to the processes and goals of this project. Our project, however, is the only source that attempts to exclusively use GMM and expectation maximization to classify stars into the four main categories based on *color index* and *luminosity*.

Methods

We clustered the data using a GMM with luminosity and color index as our parameters. We iterated between 50 and 100 times until the various groups converged. We had two different versions. Our first version used starting values for the mean and covariance taken from points selected from each of the Hertzsprung-Russell groupings: the white dwarfs, main sequence, red giants, and supergiant star classifications. Our second version set the means randomly for all four clusters and set the covariance as the covariance of the entire data set. We used a normal distribution for the expectation maximization algorithm to represent the star groups as we believe the shapes present in the data would group well. The GMM ran using the methods and equations discussed in class that are shown below



And we're done! **Step 0:** Initialize clusters and their means, variances, equal proportions.

Step 1: *Expectation*. For each data point x_i and for each each component m

1. $\tilde{p}_{mi} = \phi(x_i | \hat{\mu}_m, \hat{\Sigma}_m)\hat{w}_m$ and then consolidate into the probabilities:
2. $\hat{p}_{mi} = \dfrac{\tilde{p}_{mi}}{\sum_m \tilde{p}_{mi}}$

Step 2: *Maximization*. For each component m,

1. $\hat{w}_m = \dfrac{\hat{n_m}}{N} = \dfrac{\sum_{i=1}^{N} \hat{p}_{mi}}{N}$
2. $\hat{\mu}_m = \dfrac{1}{\hat{n_m}} \displaystyle\sum_{i=1}^{N} \hat{p}_{mi} \cdot x_i$
3. $\hat{\Sigma}_m = \dfrac{1}{\hat{n_m}} \displaystyle\sum_{i=1}^{N} \hat{p}_{mi} \cdot (x_i - \hat{\mu}_m)(x_i - \hat{\mu}_m)^T$

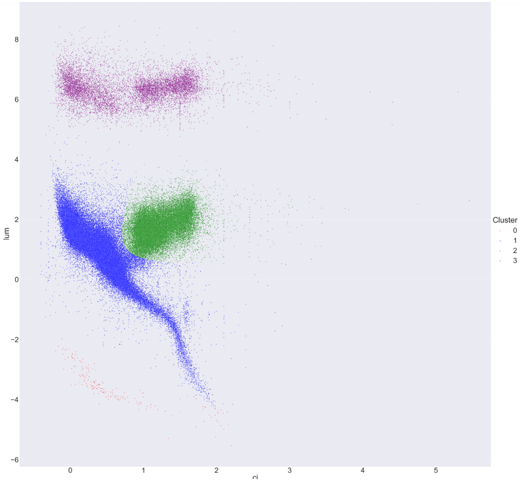
From class lecture notes February 17th, 2023

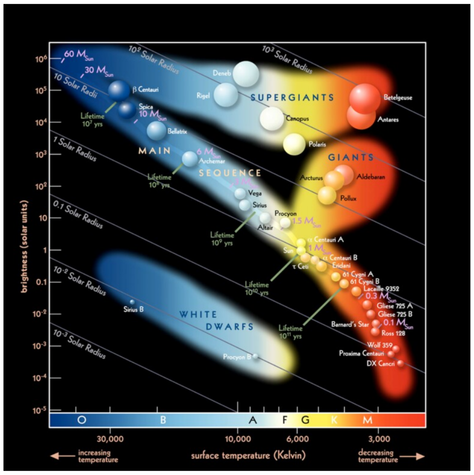
Results

Since GMM groups stars based on their likelihood to be in that cluster based on color index and luminosity, our model classifies stars differently as compared to the Hertzsprung-Russell diagram which is based purely on observed patterns and shapes of the data.

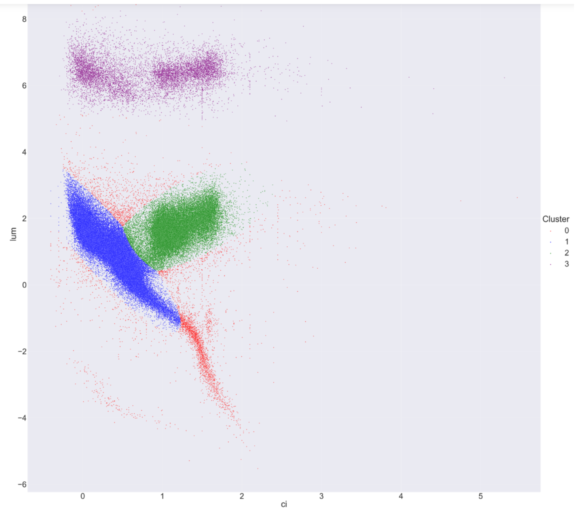
Generally, with preset starting points, and fairly low iterations, our model clustered the stars similarly to Hertzsprung-Russell. Though this was the only way our model was able to produce similar results to Hertzsprung-Russell as with an increased number of iterations, and different starting points, our results varied wildly.

With the use of predefined starting points and after about 5 iterations, the algorithm tends to cluster the stars similar to their official stellar classification. The white dwarf group is hard to see, but is grouped into red in the bottom left of the graph. The two graphs below show a very similar distribution of stars, where the graph on the left was generated by our project code.



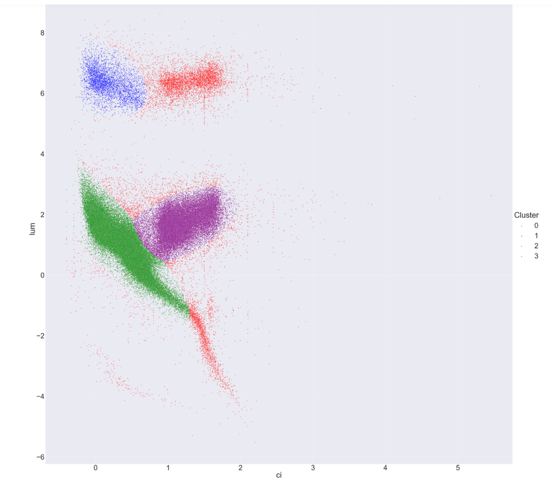


The following shows the results of GMM with predefined mean and covariance, and roughly 100 iterations.



This configuration of GMM is able to group the supergiants and red giants fairly well but struggles with distinguishing the main sequence and white dwarfs.

The following shows the results of GMM with random starting means and initial covariance matrices of the entire dataset.



GMM with random starting points varies significantly from official classification, and categorizes red giants into several different categories, which shouldn't have any significance.

Conclusion

Generally, the approach we took to categorize star types based on the Hertzsprung-Russell star chart was mildly effective, but due to differences in how GMM and Hertzsprung-Russell categorize stars, the results varied. Hertzsprung-Russell is a well-known star categorization diagram that is based on patterns visible in the chart to humans. Our algorithm classifies stars based on color index and luminosity similarity, so the natural pattern groupings of the Hertzsprung- Russell diagram wasn't quite followed as the GMM was clustering more on closeness to the factors instead of a general pattern. The best results were obtained with a few iterations of GMM and predefined means and covariances. Some reasons for the clustering not doing as well as we anticipated could be because this stellar classification might not follow a normal distribution, not enough data on certain star groups such as white dwarfs, and distinguishing between groups close together, like the main sequence and red giants, can be difficult. While our specific model needs more tuning to provide interesting results, this paper proved that GMM is a promising potential candidate to improve stellar categorization. Future work could include expanding GMM to account for more than two variables, or by mixing GMM classification with other machine learning classification methods, as was done in the several papers mentioned earlier. Overall there is room for improvement of our model but at a low number of iterations, our model showed promise.

Citations

1.) Lee, K. J, et al. “Application of the Gaussian Mixture Model in Pulsar Astronomy - Pulsar Classification and Candidates Ranking for the Fermi 2FGL Catalog.” *Academic.oup.com*, https://academic.oup.com/mnras/article/424/4/2832/1058592.

2.) Karmakar, Arnab, et al. “Stellar Cluster Detection Using GMM with Deep Variational ... - Arxiv.” *Stellar Cluster Detection Using GMM with Deep Variational Autoencoder*, https://arxiv.org/pdf/1809.01434.pdf.

3.) Zamani, Majid. 2023. CSCI 4022 Class Lecture Notes: GMMs the Expectation Maximization Algorithm. February 17