Clustering and Principle Component Analysis on Gaia Data.

Sarah Nicole Straw

17th March 2025 — Word Count: 1027

1 Part 1: Reading and filtering the data

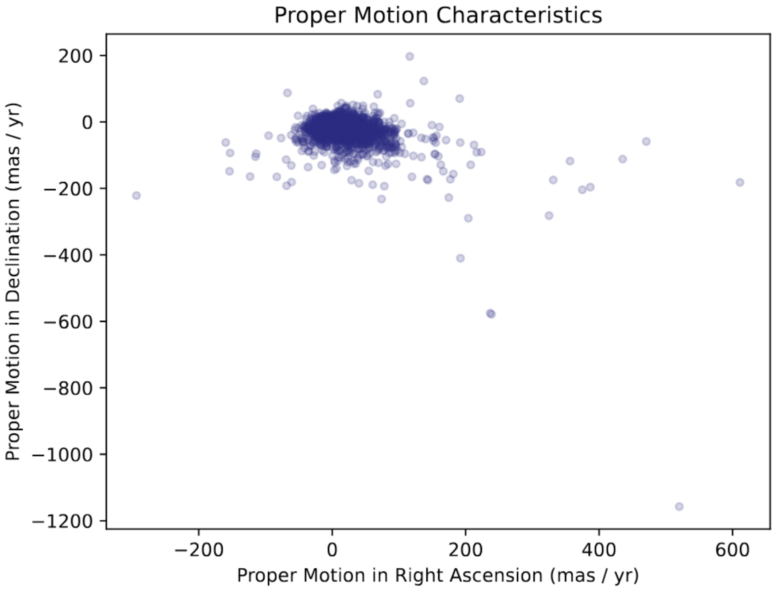


Figure 1: Scatter plot of proper motion in right ascension by proper motion in declination for the full data set, showing a core cluster with outliers spread around it.

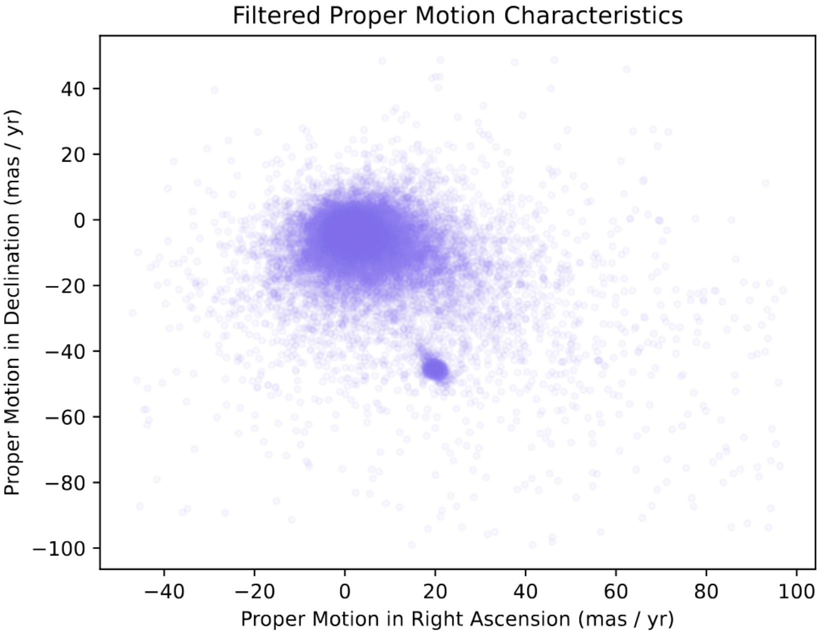


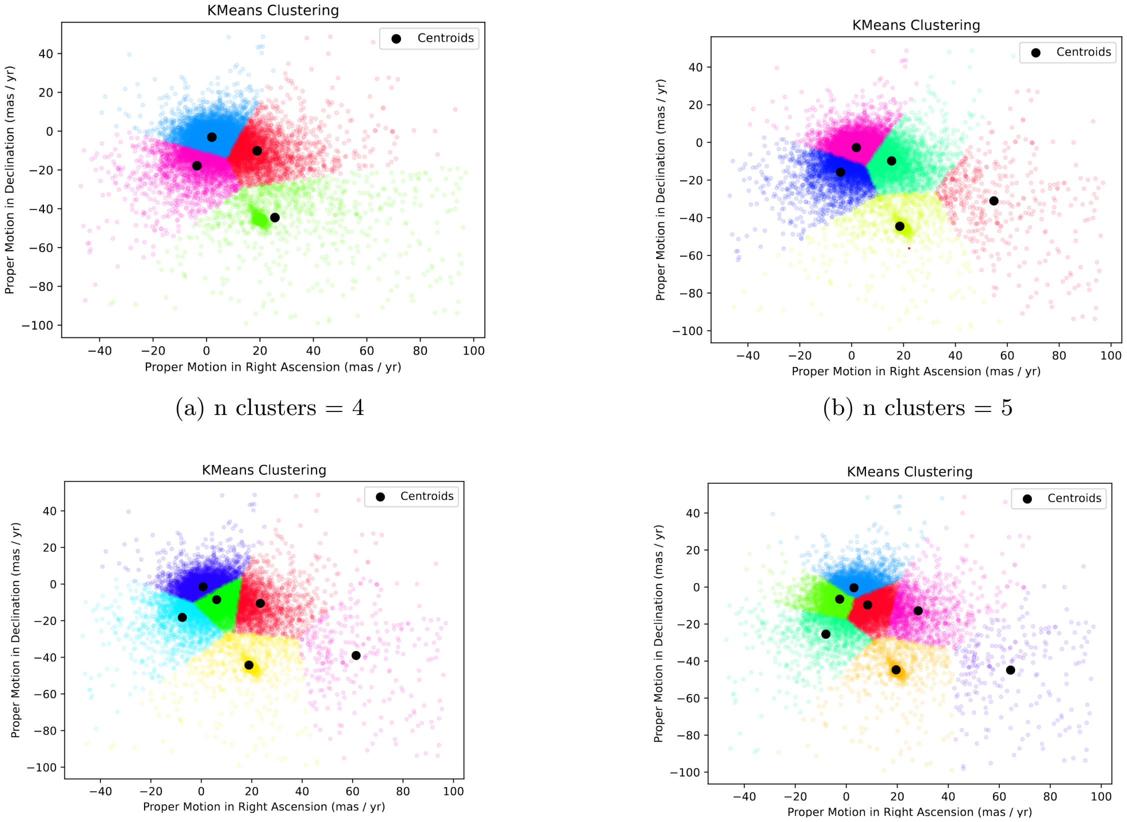
Figure 2: Scatter plot of proper motion in right ascension by proper motion in declination for a filtered data frame.

In figure 2 the data frame is filtered by discarding points with relative error on the parallax > 20% and any extreme values of proper motion declination and ascension. Each data point has a transparency of 95% to allow the smaller cluster to be seen clearly.

2 Part 2: Standardizing

Due to the data frame being so large and it's features having inhomogeneous units, the data is scaled using StandardScaler from scikit-learn. This is beneficial in this case to make sure no larger unit features dominate over the others, and all features contribute equally.

3 Part 3: Clustering



(a) n clusters = 4

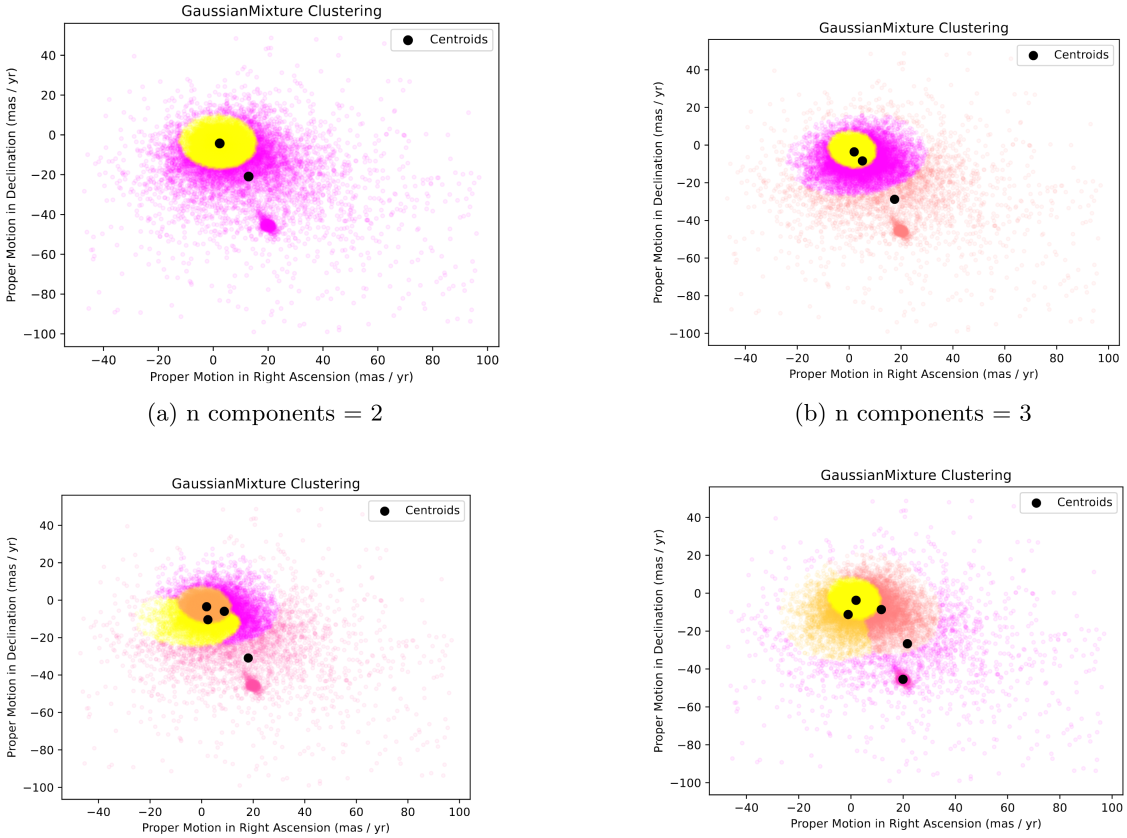
(b) n clusters = 5

(c) n clusters = 6

(d) n clusters = 7

Figure 3: K-means clustering with varying n clusters value.

K-means clustering fails to disentangle the small cluster of interest from the rest of the stars, even increasing n clusters to 7 it's still grouping the cluster with a large portion of nearby stars, proving it to be an unsuitable clustering technique for this exercise. This would be more suitable for well separated clusters of similar sizes.



(a) n components = 2

(b) n components = 3

(c) n components = 4

(d) n components = 5

Figure 4: Gaussian mixtures clustering with varying n components value.

Both techniques fail to separate the smaller cluster from the rest when the number of clusters is set to 2. This is likely due to background stars populating the data, they're spread fairly uniformly and overlap with the clusters causing the k-means and gaussian mixtures to wrongly recognize them as apart of them. This means two clusters cannot accurately account for the distribution of stars in the data. However gaussian mixtures clustering successfully recognises the smaller cluster and disentangles it when n components is set to 5, indicating that due to the clusters population being a small fraction of the total stars it requires a high number of clusters to be sought out for it to be individually recognised.

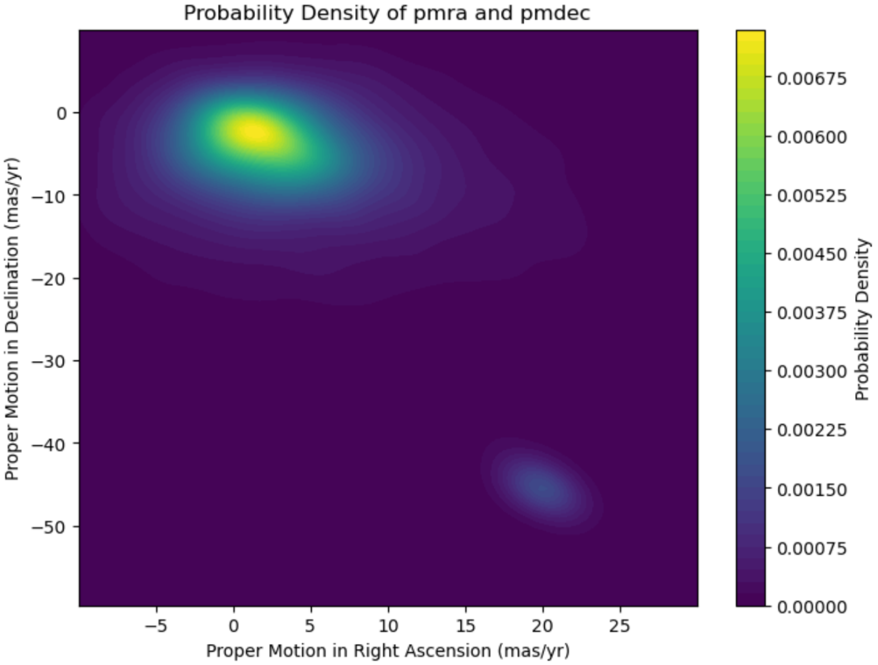


Figure 5: Probability distribution evaluated by Gaussian Kernel Density Estimation.

A gaussian kernel density estimation is used to plot the estimated probability distribution of the stars with a colour map showing more probable regions. This shows the smaller cluster of interest has an incredibly low probability density, nearly blending with the background stars. This visualizes why modelling 2 clusters fails, as the cluster of interest is drowned out by the larger cluster and surrounding stars. The resulting cluster data showed a population of 1033 stars, accounting for 0.91% of the total

dataset.

4 Part 4: Principal component analysis

Next the cluster data is converted to a numpy array and it's covariance matrix C is found, from this the diagonalisation problem is solved:

$$C = VDV^T \tag{1}$$

Matrix V contains 12 columns for the 12 eigenvectors of C, also known as the principle components. Each eigenvector has 12 components, being the weight of each original feature contributing to the particular principle component. D is a diagonal matrix containing the eigenvalues, each representing the variance from each principle component. PCA maximizes variance along the principal components, so the principle components with the smallest variance can be dropped and the data dimensionality can be reduced while keeping most of the information.

Principal Component	**Variance**	**Variance Ratio**
PC1	3.34	27.77%
PC2	2.28	18.99%

Table 1: Variance and Variance Ratio for PC1 and PC2

Total variance explained by the top two principle components is 46.75%.

5 Part 5: Projection and factor analysis

To assess how well the top 2 principle components represent the data, it's projected onto the principle components. Tjis means for each data point, its dot product with the principal component vectors $PC1$ and $PC2$ is calculated and plotted. This shows how much each data point aligns with each principal component.

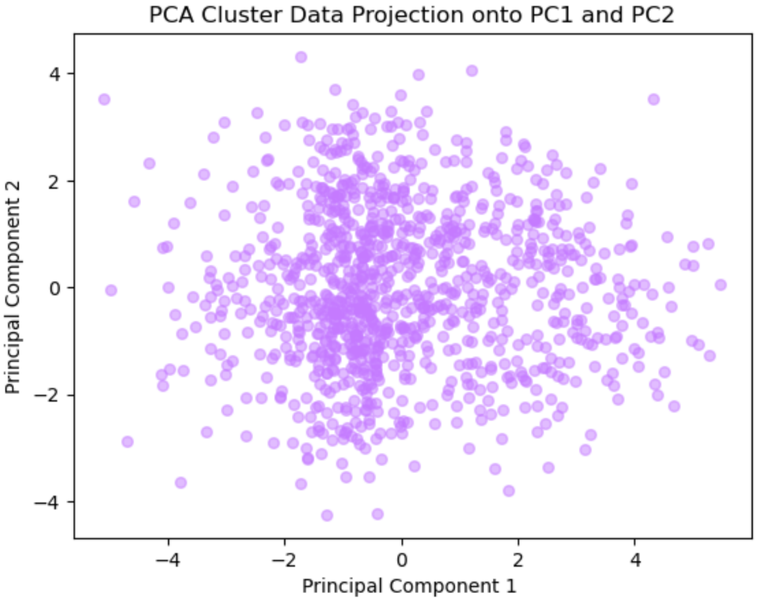


Figure 6: Scatter plot of principle component 1 against principle component 2

The plot shows no strong trends due to the small variance difference between $PC1$ and $PC2$. However, a clump spread along $PC1 = -1$ shows for a lot of the data the projection along $PC1$ is similar while the projection along $PC2$ is more varied. Interestingly, $PC2$ captures more variance across the data despite having lower variance than $PC1$. This could be due to correlated features or $PC1$ having a smaller range, which is investigate further using histograms.

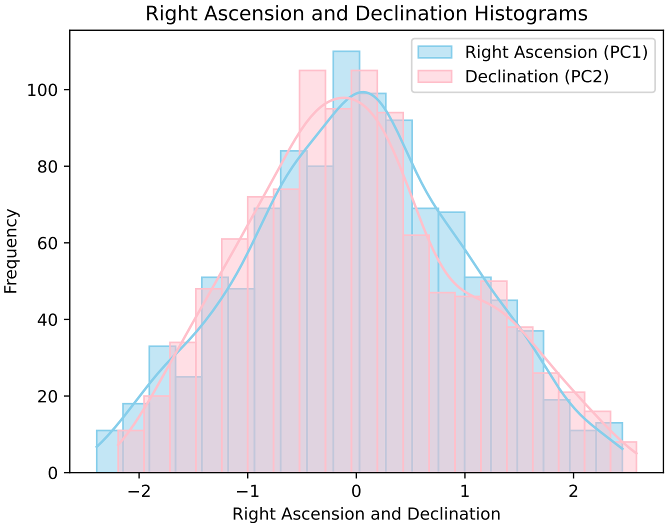


Figure 7: Overlayed Histogram of features for the two Principle Components of largest variance.

The features for $PC1$ and $PC2$ are right ascension and declination respectively. Their histogram shows no significant difference in variance, the range for them both is fairly small due to the data given being from a very small section of the sky. Since right ascension doesn't bunch at -1 the clump in figure 6 could then be due to the data aligning with $PC2$ across a larger range than $PC1$.

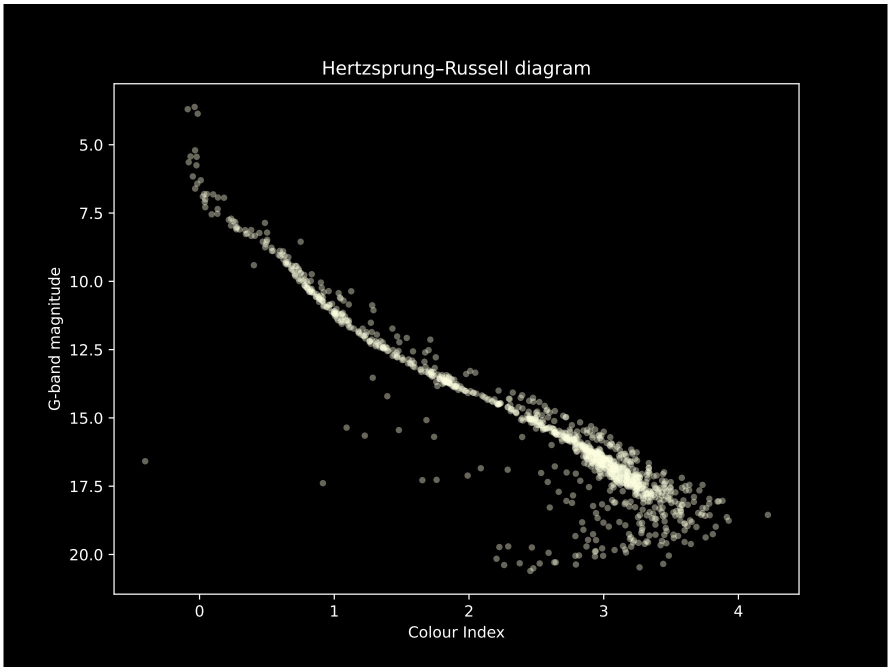


Figure 8: Hertzsprung–Russell Diagram obtained by scatter plot of cluster data G-band magnitude by colour index.

By grouping the difference in blue and red photometric band magnitude into colour index, plotting against G-band magnitude (a measure of the star's brightness in the Gaia G band) a Hertzsprung–Russell Diagram can be shown. Comparing the curve with the Hertzsprung–Russell Diagram (Figure 9) shows the cluster is in the 'main sequence' region, with it's stars at varying stages in their life cycles.

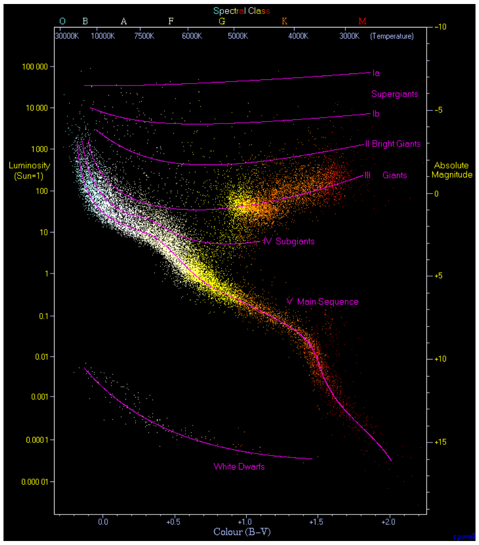


Figure 9: Hertzsprung–Russell Diagram