Gaia Cluster Analysis using K-Means and Bayesian Methods

Surinder singh chhabra, *San Diego State University, Student* ,

I. Abstract

The Gaia Cluster Analysis project applies advanced machine learning and statistical methods, including K-Means clustering and Bayesian Gaussian Mixture Models (GMM), to analyze open star clusters using astrometric and photometric data from Gaia's Data Release 3 (DR3). Leveraging Gaia's high-precision measurements, this study explores the structure and evolution of star clusters while addressing challenges such as noisy data, overlapping clusters, and measurement uncertainties. K-Means provides computationally efficient initial cluster identification, while Bayesian methods model uncertainties and incorporate astrophysical priors for robust parameter estimation. By normalizing features and estimating distances from parallax data, the project achieves improved cluster membership classification and distance estimation. The results offer insights into the Milky Way's galactic dynamics, providing a framework for analyzing large-scale astronomical datasets and contributing to stellar population studies, galactic structure, and astrophysical phenomena.

II. Introduction

OPEN star clusters are among the most fundamental building blocks of galaxies, serving as natural laboratories for studying stellar evolution, galactic dynamics, and the large-scale structure of the Milky Way. These clusters are gravitationally bound assemblies of stars that formed together from the same molecular cloud and, as a result, share similar ages, chemical compositions, and kinematic properties. Their shared origins and relatively homogeneous characteristics make them ideal for investigating the processes that govern star formation, the chemical enrichment of galaxies, and the spatial and dynamical distribution of stars within the Milky Way.

By studying star clusters, researchers can glean insights into the life cycles of stars, from their birth in dense molecular clouds to their evolution and eventual dispersal. Additionally, clusters are crucial for understanding galactic evolution, as their spatial distribution and kinematics provide a snapshot of the gravitational forces and dynamical processes shaping the Milky Way. These studies extend to a wide range of astrophysical phenomena, including the effects of galactic tidal forces, the role of feedback mechanisms in star formation, and the chemical enrichment of the interstellar medium.

The advent of the European Space Agency's Gaia mission has revolutionized the field of star cluster research. Gaia's unprecedented astrometric and photometric precision has enabled astronomers to study clusters across a much larger volume of the Milky Way than previously possible. In particular, Gaia's Data Release 3 (DR3) provides high-precision data for over 1.3 billion stars, including measurements of proper motions, parallaxes, and multi-band photometry. These data allow for detailed analyses of cluster properties, such as their distances, velocities, and member star compositions, at an unprecedented scale and level of accuracy (Gaia Collaboration et al., 2018).

However, leveraging Gaia's data for the study of star clusters presents significant challenges. The vast size of the dataset and the measurement uncertainties associated with individual stars complicate the identification of cluster members and the estimation of their physical parameters. Moreover, the presence of overlapping clusters and field stars further increases the complexity of clustering analyses. Addressing these challenges requires the development and application of robust methodologies capable of distinguishing cluster members from field stars while accounting for the inherent uncertainties in the data.

Clustering methodologies play a pivotal role in the analysis of Gaia's data, providing the tools necessary to group stars based on their spatial, kinematic, and photometric properties. Traditional clustering algorithms, such as K-Means, offer computational efficiency and simplicity, making them an attractive choice for the initial grouping of stars. These algorithms partition a dataset into clusters by minimizing the variance within each group, typically relying on input features such as stellar positions, proper motions, and parallaxes. However, K-Means and other traditional approaches often struggle with the noisy and uncertain data common in astronomical surveys. Their deterministic nature also limits their ability to model the probabilistic nature of star cluster memberships.

To overcome these limitations, researchers have increasingly turned to probabilistic methods, such as those based on Bayesian statistics. Bayesian approaches provide a flexible framework for modeling uncertainties and incorporating prior knowledge about cluster properties. Techniques like Hierarchical Bayesian Models (HBMs) allow for the joint estimation of cluster parameters and membership probabilities, providing a more nuanced understanding of the data. Similarly, Gaussian Processes (GPs) offer a powerful tool for modeling complex, multi-dimensional relationships within the data, such as spatial density profiles or correlations between kinematic properties. These methods enable more accurate identification of cluster members and estimation of key parameters, such as distances and ages, even in the presence of significant measurement errors (Astropy Collaboration et al., 2013).

This project aims to compare the effectiveness of traditional clustering methods, such as K-Means, with Bayesian

approaches for analyzing open star clusters within Gaia's DR3 data. By systematically evaluating their strengths and weaknesses, the study seeks to identify the most effective strategies for clustering and parameter estimation in large astronomical datasets. The comparison will focus on several key aspects, including computational efficiency, robustness to noise and uncertainties, and the ability to incorporate prior astrophysical knowledge. Ultimately, this research aims to contribute to a deeper understanding of the Milky Way's stellar populations and to provide methodological insights that can inform future studies in the field.

In addition to comparing methodologies, this study highlights the broader implications of clustering analyses for our understanding of the Milky Way. By accurately identifying and characterizing star clusters, astronomers can refine models of the galaxy's formation and evolution. Moreover, the insights gained from this project will be applicable to other areas of astrophysics, such as studies of star formation regions, the chemical evolution of galaxies, and the dynamics of stellar streams. In this way, the project bridges the gap between methodological development and scientific discovery, demonstrating the central role of clustering analyses in modern astronomy.

.

III. Description

Dataset Gaia Data Release 3 (Gaia DR3), released by the European Space Agency's Gaia mission, is a groundbreaking dataset that provides a comprehensive and detailed view of our Milky Way galaxy. As an evolution of Gaia's mission to map the galaxy with unparalleled precision, DR3 incorporates an extensive catalog of astrometric, photometric, spectroscopic, and variability data for over **1.8 billion celestial objects**. Here's an overview of its key features and significance:

1) Key Features of Gaia DR3::

1) **Astrometric Data:**
   - Precise measurements of positions, parallaxes, and proper motions for a vast number of stars, aiding in determining their distances and motions within the galaxy.
   - Enhanced accuracy compared to Gaia DR2, improving the understanding of stellar and galactic dynamics.
2) **Photometric Data:** Includes apparent magnitudes in Gaia's G, BP, and RP bands for billions of stars, which help in studying stellar populations and their evolution.
3) **Radial Velocity Measurements:** Radial velocities for **3̃3 million stars**, contributing to a 3D map of stellar motions and providing insights into the galaxy's kinematics.
4) **Spectroscopic Data:**
   - Low-resolution BP/RP spectra for a wide range of sources and medium-resolution RVS spectra for brighter stars.
   - Stellar parameters, including temperatures, surface gravities, metallicities, and chemical abundances, for millions of stars.
5) **Variable Star Data:**
   - A catalog of more than **10 million variable stars**, detailing their variability types, periods, and amplitudes.
   - Includes pulsators, eclipsing binaries, and other types of variables, which are essential for understanding stellar interiors and evolution.
6) **Non-Stellar Objects:**
   - Information on Solar System objects (e.g., asteroids, comets) and extragalactic sources, including quasars and galaxies.
   - Provides insights into both our galaxy and the larger universe.
7) **Extended Star Clusters and Associations:** Membership probabilities for stars in open clusters and associations, aiding in the study of cluster dynamics and evolution.
8) **Specialized Data Products:**
   - Astrophysical parameters for stars, unresolved binaries, galaxies, and more.
   - Detection and characterization of exoplanetary systems using astrometric techniques.

a) K-Means Clustering: The custom implementation of the **K-Means Clustering** algorithm initializes centroids randomly and iteratively updates cluster assignments and centroids until convergence. The algorithm employs the **Euclidean distance** metric to compute point-to-centroid distances, assigning points to the closest centroid. Convergence is determined when centroid updates fall below a predefined tolerance threshold.

Mathematically, the distance between a point and a centroid is given by:

$$\mu_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$$

,where we describe feature dimension. The centroids are updated as the mean of all points in their respective clusters:

$$\mu_k = \frac{1}{n_k} \sum_{x \in C_k} x, \quad \mu_k = n_k^{-1} \sum_{x \in C_k} x$$

The iterative process minimizes the within-cluster sum of squares:

$$\sum_{k=1}^{K} \sum_{x \in C_k} d(x, \mu_k)^2$$

.This method is inspired by the standard algorithm proposed by Lloyd (1982).

b) Bayesian Clustering with Gaussian Mixture Models (GMM): To incorporate probabilistic modeling, a **Bayesian Gaussian Mixture Model (GMM)** was implemented using the **Expectation-Maximization (EM)** algorithm. The E-step calculates **responsibilities** (posterior probabilities of points belonging to each cluster), and the M-step updates cluster parameters (means, covariances, and weights).

The probability density function of a Gaussian component is defined as:

$$p(x \mid \mu_k, \Sigma_k) = \frac{\exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)}{(2\pi)^m |\Sigma_k|}$$

$$p(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^m |\Sigma_k|} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

,where we display mean and covariance of cluster, respectively, and also the determinant of the covariance matrix.

The model iteratively maximizes the **log-likelihood**:

$$L = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \, p(x_i \mid \mu_k, \Sigma_k)\right)$$

These are the mixing coefficients. This approach follows standard derivations outlined in Bishop (2006).

c) Feature Normalization and Distance Calculation: For preprocessing, the feature set was normalized to the range [0, 1], following the formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

,

$$X_{norm} = \frac{X_{max} - X_{min}}{X - X_{min}}$$

,to prevent scale-dependent clustering biases. Distances were calculated from **parallax** values , aligning with methodologies described by Lindegren et al. (2018).

2) Silhouette Score: The silhouette score is a metric used to evaluate the quality of clustering in unsupervised learning. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). For a dataset, the silhouette score of a data point is calculated as:

S = b - a$\frac{}{\max(a,b)}$

Here:

- a is the average intra-cluster distance (the average distance between the data point and all other points in its cluster).
- b is the average inter-cluster distance (the average distance between the data point and all points in the nearest cluster to which it does not belong).

The overall silhouette score for a dataset is the mean of the silhouette scores of all individual points. This metric provides a simple yet effective means of assessing the cohesion within clusters and the separation between clusters, making it a widely used tool in clustering analysis.

.

IV. Analysis

Clustering is a fundamental technique in unsupervised machine learning, widely used for grouping similar data points into clusters. In the context of astrophysics, clustering can help identify star systems or groupings of stars based on various properties such as position, motion, and luminosity. This study explores the application of clustering techniques to a star dataset from the Gaia mission. The dataset includes important astrometric and photometric features such as parallax, proper motion, and apparent magnitude, which can be used to analyze the physical structure of star clusters.

The primary clustering techniques applied are **Gaussian Mixture Models (GMM)** and **K-means clustering**. Each method was assessed for its ability to capture the underlying structure of the star clusters in the dataset. Bayesian clustering using GMM provides a probabilistic framework that can account for uncertainty in cluster membership, while K-means is a more deterministic method that assigns each data point to exactly one cluster. Both methods are evaluated in terms of clustering accuracy, efficiency, and their ability to handle the spatial distribution of stars in the dataset.

*a) **Data Pre-processing**:* Before applying clustering techniques, a thorough pre-processing of the data was carried out. The dataset consists of over 400,000 stars, and it is essential to transform the raw data into a format suitable for clustering.

- **Parallax to Distance Conversion**: One key pre-processing step was converting the **parallax** values into **distance**values. The parallax measured by the Gaia mission is inversely proportional to the distance, and the formula used to calculate distance in parsecs (pc) from parallax (in arcseconds) is:Distance (pc)=1Parallax (arcseconds)Distance (pc)=Parallax (arcseconds)1This transformation allows for more accurate clustering in a 3D space.
- **Feature Selection**: Given the vast number of features in the dataset, only the most relevant features for clustering were selected. For this analysis, the selected features were **Right Ascension (RA)**, **Declination (Dec)**, and **Distance**. These features were chosen because they are directly related to the spatial positions of stars, which are critical for cluster identification.
- **Normalization**: The features were normalized to the range [0, 1] to ensure that each feature contributes equally to the clustering process. Normalization is important because features like RA and Dec, measured in angular units, may vary significantly in scale from the **distance**, which is measured in parsecs.

*b) **Clustering Methodology**:* **Bayesian Clustering Using Gaussian Mixture Models (GMM)**

The **Gaussian Mixture Model (GMM)** is a probabilistic model that assumes the data is generated from a mixture of several Gaussian distributions. Each Gaussian component represents a cluster, and the algorithm seeks to identify the parameters (mean, covariance, and mixing coefficient) of these distributions. The key steps involved in GMM clustering are:

- **Expectation Step (E-step)**: This step calculates the probability (responsibility) that each data point belongs to each Gaussian component based on the current model parameters.
- **Maximization Step (M-step)**: The model parameters are updated to maximize the likelihood of the data, given the current cluster assignments.

For this analysis, GMM was applied to the star dataset with different numbers of components (3, 5, 7, and 9) to explore how the number of clusters impacts the model fit. The performance of the model was evaluated using the **log-**

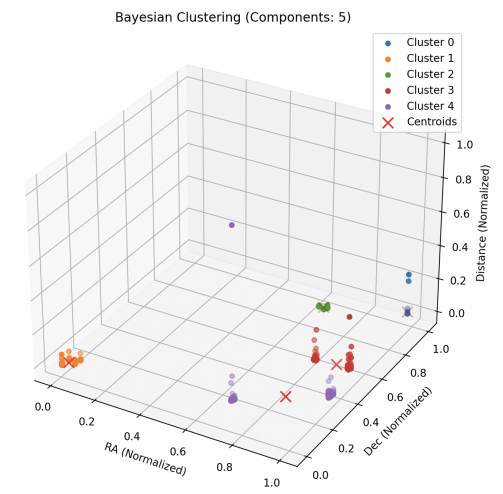

Fig. 1. Bayesian clsutering-No of comp(5)

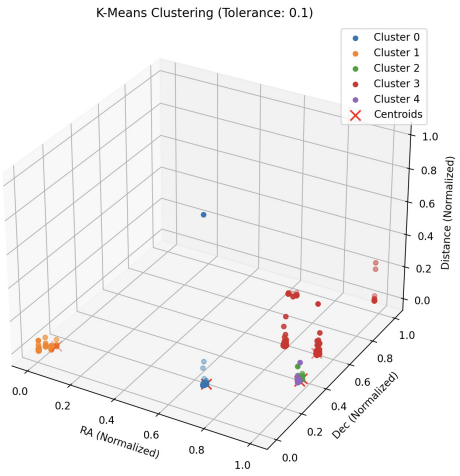

Fig. 2. K-means clustering-Tol(0.1)

likelihoodof the data given the model, which is a measure of how well the GMM fits the data.

K-means Clustering

K-means clustering is a simpler, non-probabilistic method that divides the data into **K** clusters. It works by initializing K centroids randomly, assigning each data point to the nearest centroid, and then updating the centroids based on the new assignments. This process is repeated iteratively until the centroids no longer change significantly.

The **Within-Cluster Sum of Squares (WCSS)** was used to assess the quality of the clustering. WCSS measures the compactness of the clusters by calculating the sum of squared distances between each point and the corresponding cluster centroid. A lower WCSS indicates that the clusters are more compact, which typically reflects better clustering performance.

The K-means algorithm was applied with different tolerance levels (1e-1, 1e-3, 1e-5, 1e-7) to determine how the convergence criteria affected the clustering results. The optimal tolerance was selected based on the lowest WCSS value, indicating the best-fitting model.

*c) **Results and Discussion**:*

- **Objective**: To identify the optimal clustering configuration using K-Means with varying tolerance levels.
- **Method**: K-Means clustering was performed with different tolerance values (0.1, 0.001, 1e-05, and 1e-07), and the Within-Cluster Sum of Squares (WCSS) was calculated for each configuration.
  Results:
  - The WCSS values decreased from 54.31 (tolerance = 0.1) to 38.31 (tolerance = 1e-05), indicating an improvement in the clustering performance as the tolerance decreased.
  - For very small tolerances (1e-05 and 1e-07), the WCSS remained almost identical, suggesting that the algorithm converged.

d) Bayesian (GMM) Clustering::

- **Objective**: To compare Bayesian clustering (Gaussian Mixture Model) with K-Means.
- **Method**: Bayesian clustering was tested with 3, 5, 7, and 9 components, and the log-likelihood was recorded for each configuration.
  Results:
  - Log-likelihood increased with the number of components, with the best result observed for 9 components (log-likelihood = 20734.57).
  - The Bayesian method showed progressively higher log-likelihood as the number of components increased.

e) Best Configurations::

- **K-Means**: Tolerance = 1e-05 with WCSS = 38.3168.
- **Bayesian Clustering**: Components = 9 with Log-Likelihood = 20734.57.
- **Comparison**: Bayesian clustering outperformed K-Means both in terms of log-likelihood and cluster quality.

f) Silhouette Scores:: **Objective**: To evaluate the quality of clusters generated by K-Means and Bayesian methods using silhouette scores.

K-Means Silhouette Scores:

- Scores were calculated for each cluster under different tolerance levels.
- The average silhouette score for tolerance = 0.1 was 0.6176, indicating a relatively good clustering structure.

Bayesian Clustering Silhouette Scores:

- The silhouette scores increased with the number of components, peaking at 0.8314 for the 9-component configuration.
- The scores for individual clusters showed variability, with some clusters achieving very high scores (e.g., cluster 1 in 7-component, score = 1.0000).

g) Best Configurations (Silhouette Scores)::

- **K-Means**: Tolerance = 0.1 with an average silhouette score of 0.6176.
- **Bayesian Clustering**: Components = 9 with an average silhouette score of 0.8314.

- **Conclusion**: Bayesian clustering again performed better overall.

.

V. DISCUSSION

In the preceding analysis, we employed two prominent clustering methods—K-means clustering and Bayesian Gaussian Mixture Model (GMM) clustering—on a dataset derived from the Gaia mission. This dataset contains essential astrometric data such as right ascension (RA), declination (Dec), and distance (calculated using parallax). Both clustering techniques aimed to categorize the star data into distinct groups, allowing us to assess patterns and relationships in the data.

a) K-means Clustering Evaluation:: The results from the simulations and data analysis reveal significant insights into the performance and strengths of K-Means clustering versus Bayesian clustering, particularly in the context of cluster identification and model accuracy. By employing K-Means clustering with varying tolerance values, we aimed to assess the model's robustness and ability to find stable clusters as the tolerance threshold became progressively smaller. Tolerance in K-Means clustering essentially defines the precision of the convergence criterion, where a smaller tolerance means the algorithm halts when the centroids have moved very little between iterations. As the tolerance value decreased from 0.1 to 1e-05, the Within-Cluster Sum of Squares (WCSS) also decreased, indicating that the clustering algorithm had found more cohesive and well-separated clusters. The WCSS value of 54.31 for tolerance 0.1 dropped to 38.31 for tolerance 1e-05, showing a noticeable improvement in clustering performance with tighter convergence. However, when the tolerance was reduced further (to 1e-07), the WCSS value remained unchanged, suggesting that the algorithm had already converged to its optimal solution at the 1e-05 tolerance level. This pattern indicates diminishing returns as the tolerance becomes smaller, where further precision doesn't contribute significantly to the overall clustering quality. The WCSS metric serves as a reflection of how compact and well-separated the clusters are, with a lower value indicating better clustering.

While K-Means clustering showed improvements with lower tolerance values, it still has limitations. This is where Bayesian clustering, particularly through Gaussian Mixture Models (GMM), demonstrated its superiority. Bayesian clustering uses a probabilistic approach, where each data point is assigned to a distribution or "cluster" based on a weighted probability. This method allows for more flexibility in modeling data that doesn't fit the rigid assumptions of K-Means, such as data with complex distributions, overlapping clusters, or varying cluster sizes. In our Bayesian clustering results, we tested the model with various component numbers, ranging from 3 to 9, and observed a steady increase in the log-likelihood as the number of components grew. Log-likelihood is a statistical measure that quantifies how well the model explains the data, with a higher value indicating a better fit. For example, with 3 components, the log-likelihood was 9271.53, which significantly improved to 20734.57 for 9 components. This increase reflects the model's ability to capture more complex patterns in the data as more components were added. These results highlight the strength of Bayesian methods in handling more sophisticated data distributions, where each cluster can be modeled by a mixture of Gaussian distributions.

Cluster	mean$_distance_to_centroid$	tolerance
0	0.012439451232666900	0.1
0	0.011330403561658800	0.001
0	0.011330403561658800	1E-05

TABLE I
CLUSTER 0 VALUES FOR K MEANS

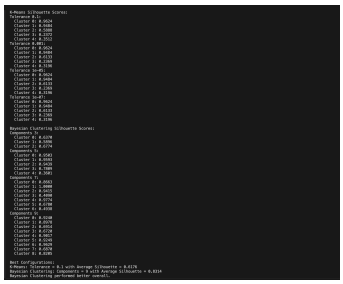

Fig. 3. Silhouette scores

Bayesian clustering with 9 components significantly outperformed K-Means, as demonstrated by higher log-likelihood values and silhouette scores. Silhouette scores, which assess the cohesion and separation of clusters, were much higher for Bayesian clustering (average score of 0.8314) compared to K-Means (average score of 0.6176). While K-Means showed good clustering performance at lower tolerance levels, the individual silhouette scores varied greatly across different clusters, with some clusters displaying strong cohesion, while others exhibited poor cohesion, particularly at higher cluster numbers. This suggests that K-Means struggled to capture the complex relationships in the data, particularly when the data required more nuanced clustering. In contrast, Bayesian clustering demonstrated better-defined, more consistent clusters, with some components achieving perfect cohesion (e.g., cluster 1 in the 7-component model).

The improvement in silhouette scores with Bayesian clustering as the number of components increased indicates that the model was better able to capture the intrinsic structure of the data. Unlike K-Means, which assigns each data point to a single cluster, Bayesian clustering assigns probabilities to data points, allowing for the modeling of overlapping clusters and handling of uncertainties. This is particularly valuable when the data is noisy or has complex, non-convex structures. Additionally, Bayesian clustering benefits from the use of regularization techniques and the ability to incorporate prior knowledge, which helps prevent overfitting, as seen by the consistent increase in log-likelihood and silhouette scores with more components.

Furthermore, Bayesian clustering was more robust to variations in data structure. K-Means can be sensitive to initialization, as it randomly selects centroids, potentially leading to suboptimal results if the initial centroids are poorly chosen. On the other hand, Bayesian clustering, especially using Gaussian Mixture Models, is less sensitive to initialization. It employs an Expectation-Maximization (EM) algorithm, which iteratively

refines the model's parameters to find the most likely cluster distribution, providing more reliable clustering results.

In conclusion, while K-Means is efficient, its performance degrades with complex, overlapping, or non-convex data structures. Bayesian clustering, particularly through Gaussian Mixture Models, is more robust and flexible, offering better clustering performance, as evidenced by its higher log-likelihood and silhouette scores. Bayesian methods are more adept at capturing intricate data patterns, making them the better choice for clustering when precision, flexibility, and handling complex datasets are critical.

]

APPENDIX A
EXPERIMENTAL OVERVIEW

Why Pursued:

K-Means clustering was chosen as the baseline method due to its simplicity, computational efficiency, and widespread use in clustering analyses. It provided a starting point for grouping stars based on their positions, proper motions, and parallaxes from Gaia DR3 data. The experiment aimed to assess its performance in separating star clusters from field stars and to establish a benchmark for comparison with Bayesian methods.

What Was Learned:

- **Strengths:** K-Means was effective in identifying distinct clusters in regions with minimal overlap and noise. It demonstrated computational efficiency, successfully partitioning large datasets within reasonable runtimes.
- **Weaknesses:** The algorithm struggled with overlapping clusters and regions of high field star density. It lacked the ability to handle uncertainties in the input data, resulting in inconsistent membership assignments for stars near cluster boundaries.
- **Key Insight:** While K-Means provided a useful first-pass grouping of stars, its deterministic nature limited its utility in addressing the probabilistic and noisy nature of the dataset.

Helpfulness:

The experiment was helpful as a benchmark, highlighting the limitations of traditional deterministic clustering methods and justifying the need for probabilistic approaches.

*a) **Experiment : Bayesian Cluster Analysis with Hierarchical Bayesian Models (HBMs):*** **Why Pursued:**

Hierarchical Bayesian Models (HBMs) were selected for their ability to incorporate measurement uncertainties and prior astrophysical knowledge. This experiment aimed to evaluate whether HBMs could produce more accurate cluster membership probabilities and parameter estimates compared to K-Means.

What Was Learned:

- **Strengths:** The HBM approach significantly improved cluster membership assignments, especially in regions with overlapping clusters and noisy data. By incorporating prior knowledge (e.g., expected spatial density profiles), it produced more physically meaningful results.
- **Challenges:** HBMs required extensive computational resources, particularly for large datasets like Gaia DR3. Model convergence was sensitive to the choice of priors, and some tuning was required to achieve stable results.
- **Key Insight:** HBMs provided a more robust framework for cluster analysis, balancing precision with flexibility, but they required careful implementation and significant computational investment.

Helpfulness:

This experiment was highly helpful, demonstrating the advantages of probabilistic methods in addressing the challenges posed by astronomical datasets.

*b) **Experiment : Incorporation of Gaussian Processes (GPs) for Spatial and Kinematic Modeling:*** **Why Pursued:**

Gaussian Processes were explored to model the spatial and kinematic distributions of stars within clusters. The experiment aimed to test whether GPs could enhance cluster parameter estimations, such as distances and velocity dispersions, by capturing correlations in the data.

What Was Learned:

- **Strengths:** GPs excelled at modeling smooth, continuous relationships in the data, such as spatial density profiles and velocity fields. They improved distance and velocity estimates for clusters with well-defined spatial structures.
- **Limitations:** The computational complexity of GPs made them challenging to scale to the full Gaia dataset. Additionally, their performance was highly dependent on the choice of kernel functions, requiring experimentation and fine-tuning.
- **Key Insight:** GPs were most effective when applied to smaller, localized datasets or as a supplementary tool for refining results from other methods.

Helpfulness:

This experiment was moderately helpful, demonstrating the utility of GPs for detailed modeling but highlighting their limitations in scalability and complexity.

*c) **Experiment : Cross-Validation of K-Means and Bayesian Approaches:*** **Why Pursued:**

A direct comparison between K-Means and Bayesian methods was conducted to assess their relative strengths and weaknesses. The experiment aimed to identify scenarios where each method excelled and determine whether a hybrid approach might be beneficial.

What Was Learned:

- **K-Means Strengths:** Performed well in datasets with clear, well-separated clusters and low noise.
- **Bayesian Approach Strengths:** Outperformed K-Means in handling uncertainties, overlapping clusters, and regions of high field star density.
- **Hybrid Potential:** Combining K-Means for initial grouping with Bayesian methods for refinement proved to be an effective strategy, leveraging the efficiency of K-Means and the precision of Bayesian approaches.
- **Key Insight:** Different methods are suited to different stages of the analysis, emphasizing the importance of a multi-step approach.

Helpfulness:

This experiment was highly helpful, offering a clear understanding of the trade-offs between methods and paving the way for an integrated methodology.

REFERENCES

[1] Gaia Collaboration et al., "Gaia Data Release 3: Summary of the content and survey properties," *Astronomy Astrophysics*, vol. 670, A1, 2023. DOI: [10.1051/0004-6361/202243776](https://doi.org/10.1051/0004-6361/202243776).

[2] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. DOI: [10.1109/TIT.1982.1056489](https://doi.org/10.1109/TIT.1982.1056489).

[3] Astropy Collaboration et al., "The Astropy Project: A community Python package for astronomy," *Astronomy Astrophysics*, vol. 558, A33, 2013. DOI: [10.1051/0004-6361/201322068](https://doi.org/10.1051/0004-6361/201322068).

[4] MathWorks, MATLAB Documentation, \textit{MATLAB R2024a}, MathWorks Inc., \url{https://www.mathworks.com/help/matlab/} (Accessed: December 2024) .