When is GMM efficient in member determination of open clusters?

Md Mahmudunnobe[1], Priya Hasan[2]

1. Minerva University, California, USA, 2. Maulana Azad National Urdu University, Hyderabad, India

mahmud.nobe@uni.minerva.edu

1. Abstract

The unprecedented precision of Gaia data opens up a new paradigm shift in membership determination of open clusters where several machine learning (ML) models are used for this purpose[1]. However, a lack of a quantitative evaluation metric for unsupervised clustering methods hampers the comparison among various models. In this paper, we develop a quantifiable metric Modified Silhouette Score (MSS) to evaluate the performance of the unsupervised models in membership determination. Specifically, we analyze the efficiency of the Gaussian Mixture Model (GMM) to find members from Gaia EDR3 data. We study the dependence of MSS on age, distance, extinction, galactic latitude and longitude, and other parameters to find the particular cases when GMM seems to be more efficient than other methods. We find that GMM is most effective for closer and relaxed clusters.

2. Gaussian Mixture Model

The primary assumption of GMM is that the data is generated from two or more Gaussian distributions. Given the data and the number of components, k, GMM first estimate the mean and standard deviation of the k Gaussian components. Then using those parameters, GMM assigns a probability of being into each of the k cluster for each data point.

The previous work on membership determination using GMM includes [2] and [3].

As the field stars do not follow a normal distribution, we need to do some additional preprocessing to ensure the member to field star ratio is not very low which includes using a smaller search radius and optimal range of features [2].

3. Modified Silhouette Score

MSS can be used to measure the performance of an unsupervised model, which outputs two sets of stars: members and field stars. For star cluster membership problem, we assumed a normal distribution (low SD) for members and an uniform distribution (high SD) for field stars in the feature space with K features.

$$MSS = \frac{1}{K}\sum_{i=1}^{K}\frac{SD_{member} - SD_{field}}{max(SD_{member}, SD_{field})}$$

MSS can range from -1 to 1. The higher the MSS value, the better the performance of the model to distinguish member (normal) and field star (uniform) group. To validate the metric, we created multiple simulated datasets of member and field-star groups and calculate their MSS value. [4] shows how MSS value increases for better separation of member and field star groups.

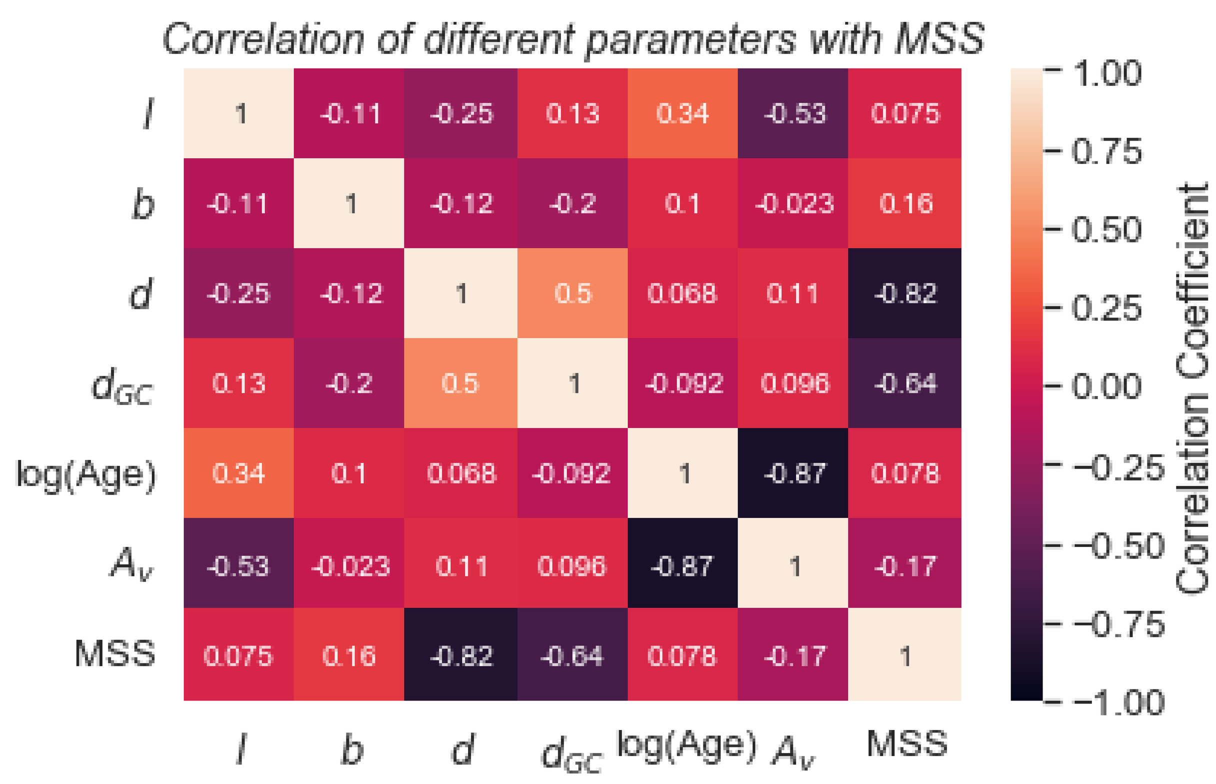
4. Methods

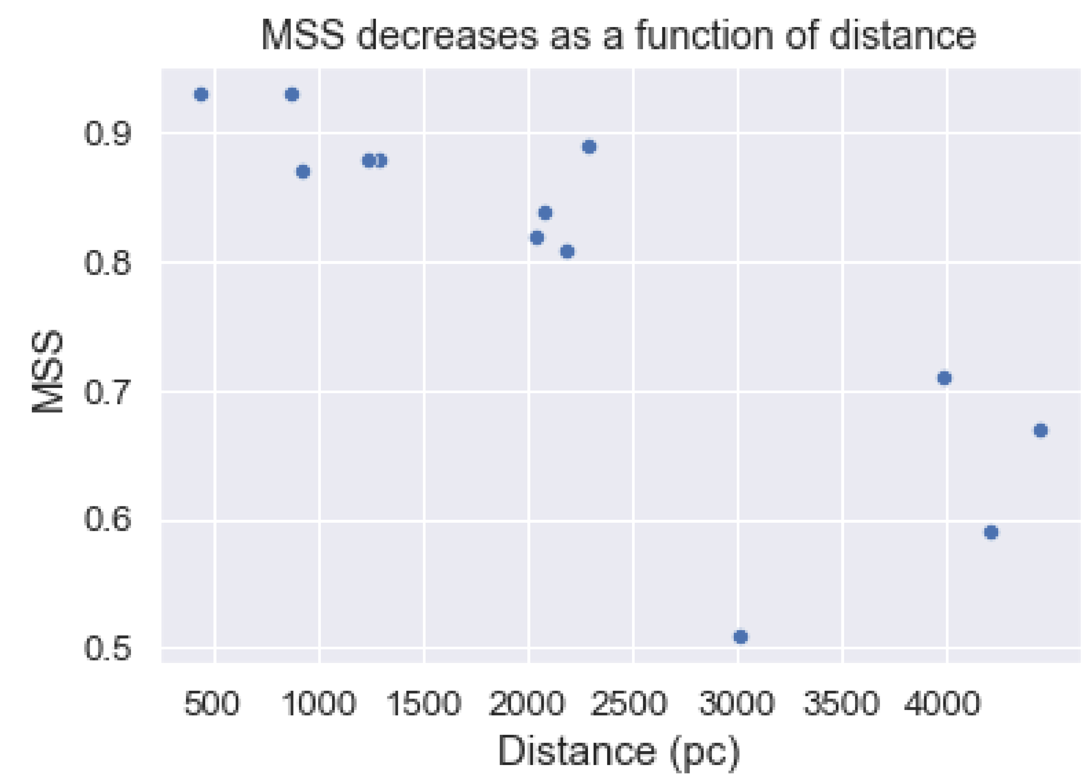
The general workflow to find members using GMM model are the following:

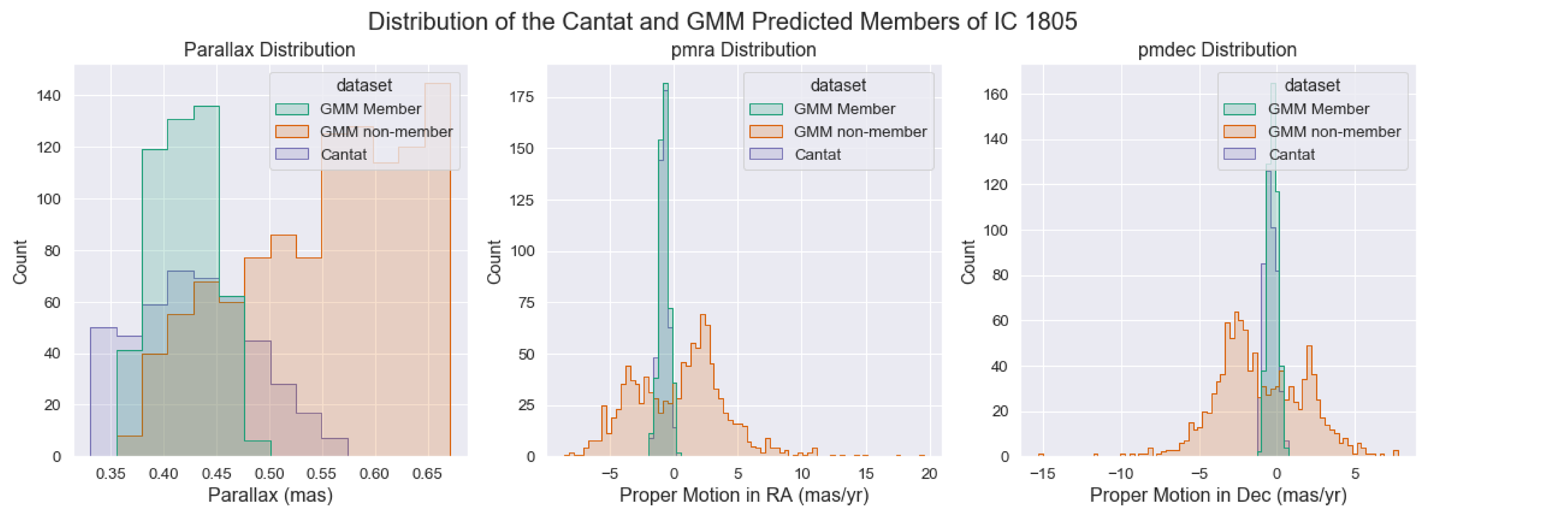
- Data Selection: From Gaia EDR3
- Pre-processing: Remove noisy data ($pmra\_error,\ pmdec\_error < 1\ mas/yr,\ parallax\_over\_error >= 3$),
- Normalization
- Feature Selection: pmra, pmdec, parallax
- Running a 2-component GMM and assign the group with lower SD as member group
- Find optimal range of distance and member threshold using MSS

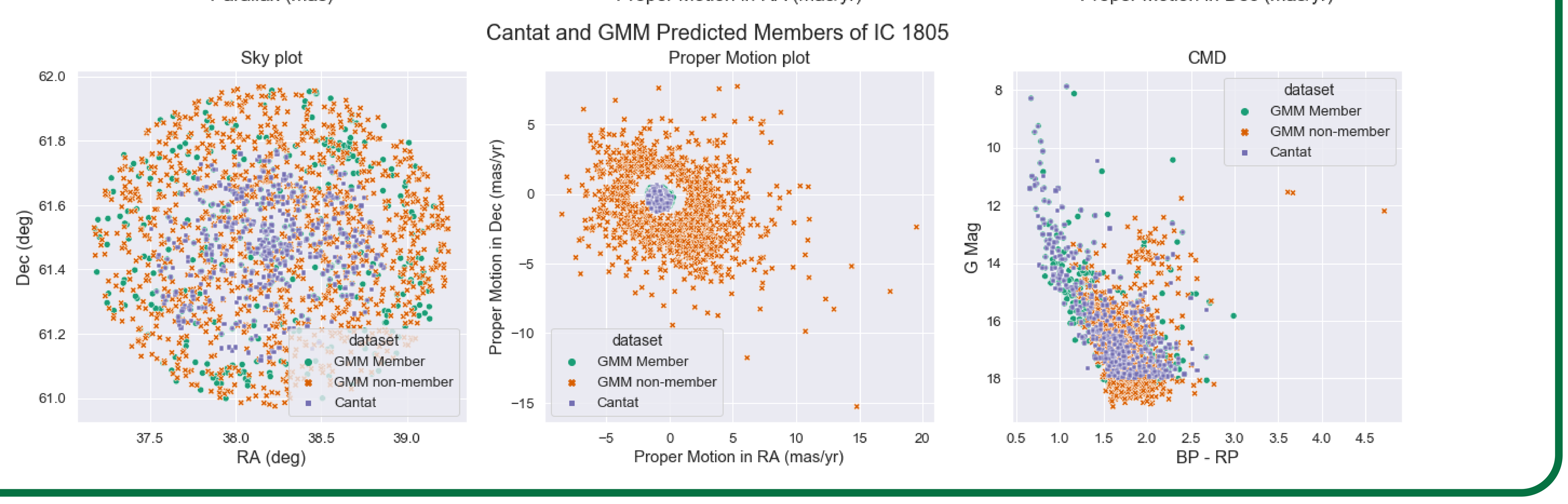
5. Results

Cluster	GLON l	GLAT b	Dist d (pc)	Dist$_{GC}$ d$_{GC}$(pc)	log(Age)	A$_v$	MSS	Member GMM	Member Cantat
NGC 752	136.96	-23.29	441	8640.50	9.18	0.16	0.93	149	240
NGC 2682	215.69	31.92	865	8942.40	9.57	0.13	0.93	1033	691
IC 4651	340.10	-7.90	920	7488.40	9.31	0.35	0.87	754	854
NGC 2539	233.72	11.11	1243	9137.80	8.88	0.22	0.88	481	518
NGC 2099	177.64	3.09	1299	9775.00	8.78	0.92	0.88	1405	1710
NGC 7142	105.35	9.49	2040	9241.20	9.55	1.29	0.82	269	401
NGC 6823	59.42	-0.14	2081	7492.10	6.84	2.52	0.84	486	158
IC 1805	134.73	0.94	2187	9922.90	6.84	2.33	0.81	495	136
NGC 581	128.05	-1.80	2292	10064.70	7.47	1.45	0.89	198	152
NGC 1893	173.58	-1.63	3019	11697.00	7.04	1.69	0.51	592	169
NGC 2243	239.48	-18.01	3996	10901.40	9.54	0.26	0.71	300	515
NGC 2141	198.04	-5.80	4213	12615.70	9.46	0.83	0.59	284	831
NGC 6791	69.96	10.91	4447	7995.30	9.86	0.31	0.67	235	1654









6. Conclusions

1. MSS can be used to measure the performance of any unsupervised model in membership determination problem. As it is a quantitative metric, we can also optimize any model hyperparameter(s) and compare between different models using MSS.
2. In this small sample, we found a moderate correlation between the distance of the cluster and the model performance measured by, where GMM works better for closer clusters. All other parameters have very weak to no correlation with GMM performance. A more comprehensive analysis using a larger number of cluster sample is discussed in our upcoming paper.

7. References

[1] Cantat-Gaudin & Anders. *AA*, 633:A99, 2020.

[2] Agarwal et al. *MNRAS*, 502(2):2582–2599, 2021.

[3] Gao. *ASS*, 365(2):1–8, 2020.

[4] Mahmudunnobe. youtu.be/i5sXHLx5apo.