ML-MOC: Machine Learning (kNN and GMM) based Membership determination for Open Clusters

Manan Agarwal [1]★ Khushboo K. Rao,[1] Kaushar Vaidya [1] and Souradeep Bhattacharya [2]
[1]*Department of Physics, Birla Institute of Technology and Science – Pilani, Rajasthan 333031, India*
[2]*Inter University Centre for Astronomy and Astrophysics, Ganeshkhind, Post Bag 4, Pune 411007, India*

Accepted 2021 January 12. Received 2021 January 3; in original form 2020 November 16

ABSTRACT

The existing open-cluster membership determination algorithms are either prior dependent on some known parameters of clusters or are not automatable to large samples of clusters. In this paper, we present ML-MOC, a new machine-learning-based approach to identify likely members of open clusters using the *Gaia* DR2 data and no a priori information about cluster parameters. We use the *k*-nearest neighbour (kNN) algorithm and the Gaussian mixture model (GMM) on high-precision proper motions and parallax measurements from the *Gaia* DR2 data to determine the membership probabilities of individual sources down to $G \sim$ 20 mag. To validate the developed method, we apply it to 15 open clusters: M67, NGC 2099, NGC 2141, NGC 2243, NGC 2539, NGC 6253, NGC 6405, NGC 6791, NGC 7044, NGC 7142, NGC 752, Blanco 1, Berkeley 18, IC 4651, and Hyades. These clusters differ in terms of their ages, distances, metallicities, and extinctions and cover a wide parameter space in proper motions and parallaxes with respect to the field population. The extracted members produce clean colour–magnitude diagrams and our astrometric parameters of the clusters are in good agreement with the values derived in previous work. The estimated degree of contamination in the extracted members ranges between 2 per cent and 12 per cent. The results show that ML-MOC is a reliable approach to segregate open-cluster members from field stars.

Key words: methods: data analysis – open clusters and associations: general – methods: statistical – astrometry.

2021.3.5

优势 : no prior information of parameters of cluster(kNN and GMM) & G~20 mag & probability

Various methods have been used for membership determination based on the analysis of the positions, proper motions, parallaxes, radial velocities, photometry, and combinations thereof (Vasilevskis, Klemola & Preston 1958; Sanders 1971; Cabrera-Cano & Alfaro 1990; Zhao & He 1990; Galadi-Enriquez, Jordi & Trullols 1998; Balaguer-Núñez et al. 2020). In the last few years, machine-learning algorithms such as DBSCAN (Gao 2014; Bhattacharya et al. 2017a), HDBSCAN (Hunt & Reffert 2020), KMEANS (El Aziz, Selim & Essam 2016), the Gaussian mixture model (Gao 2020), RANDOMFOREST (Gao 2018a), UPMASK (CG18), and artificial neural networks (Castro-Ginard et al. 2018) have been put to the task of separating the true members of open clusters from the field stars. Most of these previous studies have only been applied to a few old open clusters (e.g. Gao 2014 studied NGC 188; El Aziz et al. 2016 studied NGC 188 and NGC 2266) or are highly sensitive to the initial sample selection, making them unsuitable for being scalable (e.g. Gao 2018a, 2020). Methods developed by CG18 and Castro-Ginard et al. (2018) have been applied to a large number of open clusters, but they limited their membership analysis to sources brighter than $G \sim 18$ mag and $G \sim 17$ mag, respectively. Furthermore, CG18 needs a priori information (distance and radius) about the cluster. Castro-Ginard et al. (2018), on the other hand, did not obtain the membership probability for individual stars.

Our approach to membership determination only uses astrometric measurements from *Gaia* DR2 and does not require any a priori information about the cluster parameters. The method is independent

In this paper, we propose ML-MOC, a new probabilistic membership determination algorithm for open-cluster members down to $G \simeq 20$ mag using only *Gaia* DR2. Our algorithm is based on the *k*-nearest neighbour algorithm (kNN, Cover & Hart 1967) and the Gaussian mixture model (GMM, McLachlan & Peel 2000). We apply it to the high-precision *Gaia* DR2 proper motions and parallaxes to determine the membership probability of individual sources. We aim to facilitate homogeneous analysis of open-cluster populations by developing a robust algorithm that works reliably on a large number of open clusters. The method is applied to 15 open clusters: M67, NGC 2099, NGC 2141, NGC 2243, NGC 2539, NGC 6253, NGC 6405, NGC 6791, NGC 7044, NGC 7142, NGC 752, Blanco 1, Berkeley 18, IC 4651, and Hyades, which cover a range of ages, distances, metallicities, and extinctions.

以 M67 为例：

First stage: extract the sample sources

quite robust to the choice of this initial radius, as a rule of thumb, we generally use the value of radius that is 1.5 times the tidal radius. For M67, we download sources within a radius of 150 arcmin from the cluster centre. Next, we select the sources that satisfy the following criteria:

(i) each source must have the five astrometric parameters, positions, proper motions, and parallax as well as valid measurements in the three photometric passbands G, G_{BP}, and G_{RP} in the *Gaia* DR2 catalogue;
(ii) their parallax values must be non-negative;
(iii) the errors in their G-mag must be less than 0.005. This last criterion allows us to eliminate sources with high uncertainty while still retaining a fraction of sources down to $G \sim 21$ mag.

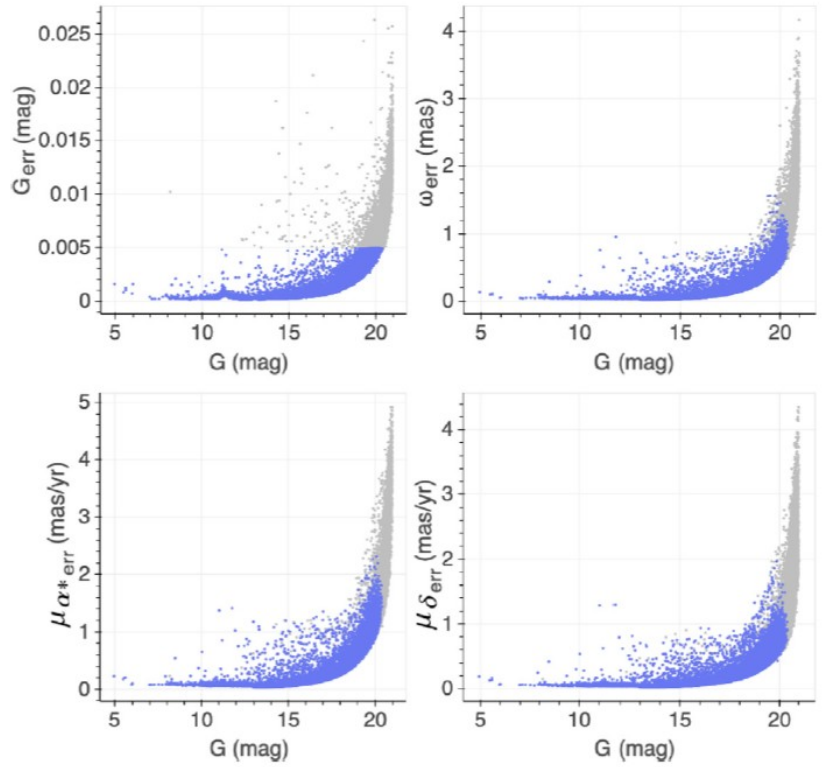


Figure 1. The correlation of errors (photometry, parallax, and proper motions) with the G-mag of sources. The grey points are all *Gaia* DR2 sources for M67 within a radius of 150 arcmin from the cluster centre. The sources in blue are those with $G_{err} < 0.005$ mag.

'All sources': 56135 sources

First stage: extract the sample sources

kNN

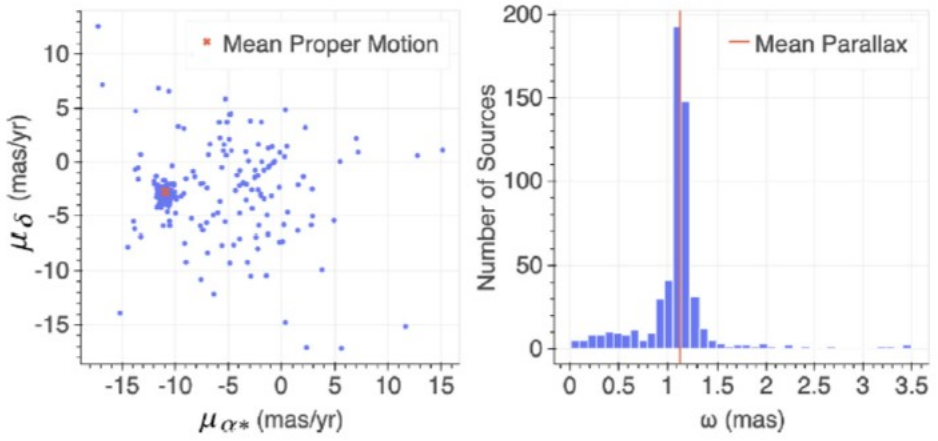


Figure 2. The distributions of proper motions and parallaxes of sources brighter than $G = 18$ mag lying within 10 arcmin from the cluster centre. The means of the proper motions and parallaxes, determined by the kNN algorithm, are indicated with a red dot and a red line, respectively.

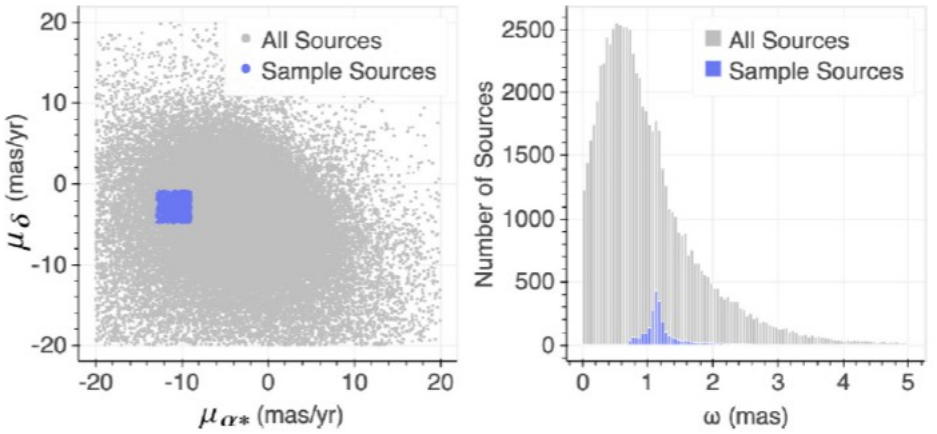


Figure 3. The result of the first-stage analysis. The extracted *Sample sources* for the cluster M67 are shown in blue.

'Sample sources': 2427 sources

Second stage: identify the member sources

The model is fitted by maximizing the likelihood estimates of the distribution parameters using the expectation maximization (EM) algorithm (Dempster, Laird & Rubin 1977). Given N data points $x = \{x_1, x_2, x_3, \ldots, x_N\}$ in an M-dimensional parameter space, the K-component GMM is defined as

$$P(x) = \sum_{i=1}^{K} w_i G(x \mid \mu_i, \Sigma_i), \text{ such that } \sum_{i=1}^{K} w_i = 1, \quad (2)$$

where $P(x)$ denotes the probability distribution of data points x and w_i is the mixture weight of the ith Gaussian component. $G(x|\mu_i, \Sigma_i)$, defined as

$$G(x \mid \mu_i, \Sigma_i) = \frac{\exp[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)]}{(2\pi)^{M/2}\sqrt{|\Sigma_i|}}. \quad (3)$$

In equation (3), μ_i and Σ_i are the mean vector and the full covariance matrix of the ith Gaussian component, respectively. The GMM assigns each data point a soft membership probability for each cluster, i.e. how likely the data point is to be described by each cluster. The GMM has been previously used to determine the membership probabilities of the sources of open clusters (M67 by Uribe, Barrera & Brieva 2006 using proper motion; NGC 6791 by Gao 2020 using 5D astrometry data, i.e. position, proper motions, and parallax).

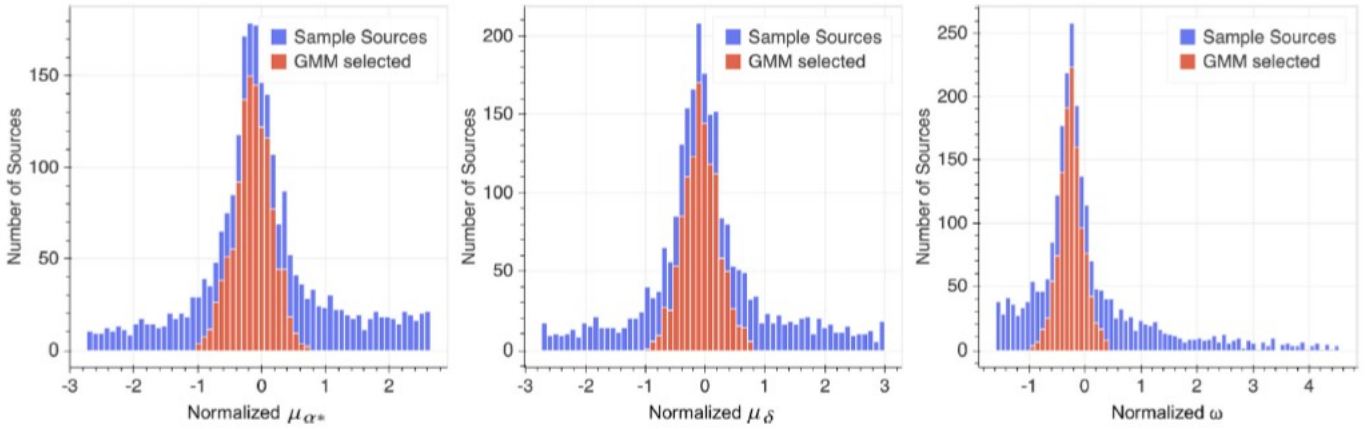
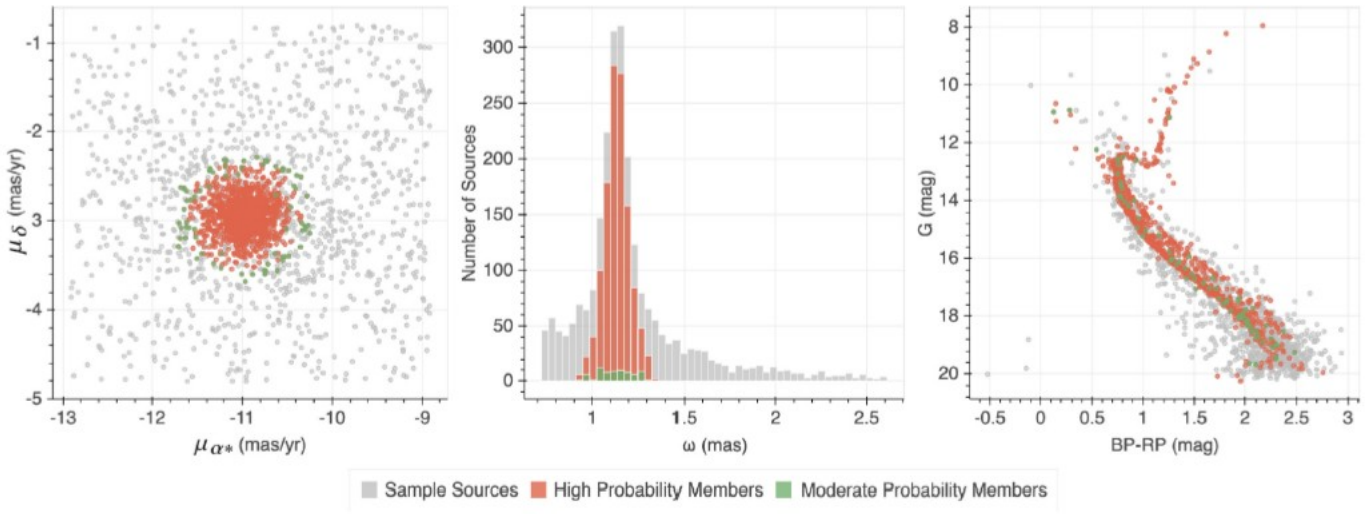


Figure 4. The frequency distributions of normalized proper motions and parallaxes of the *Sample sources* (blue). The sources with membership probability greater than 0.6, as determined by the GMM, are shown in red bars.

'Sample sources': 1150 sources (Proba>=0.6)

Third stage: including moderate-probability sources



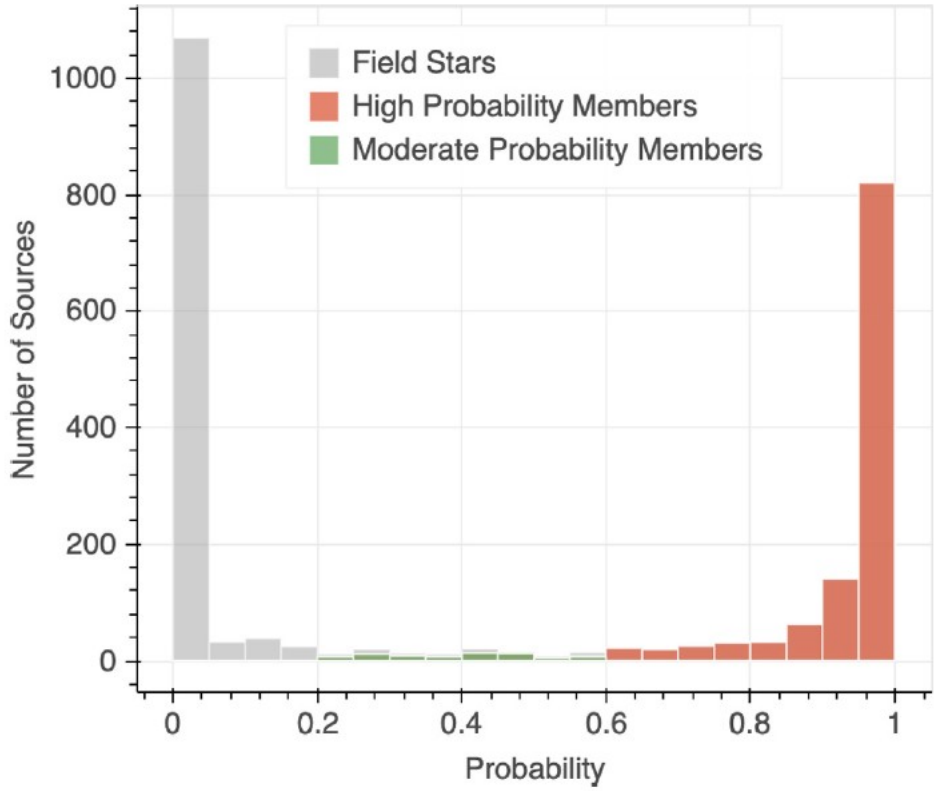


Figure 6. Distribution of the membership probabilities assigned by the GMM to the *Sample sources* of M67.

'Sample sources'(1221): 1150 sources (Proba>=0.6)+71 (parallax range & Proba_rv>=0.8)

Degree of contamination: 2-10 percent

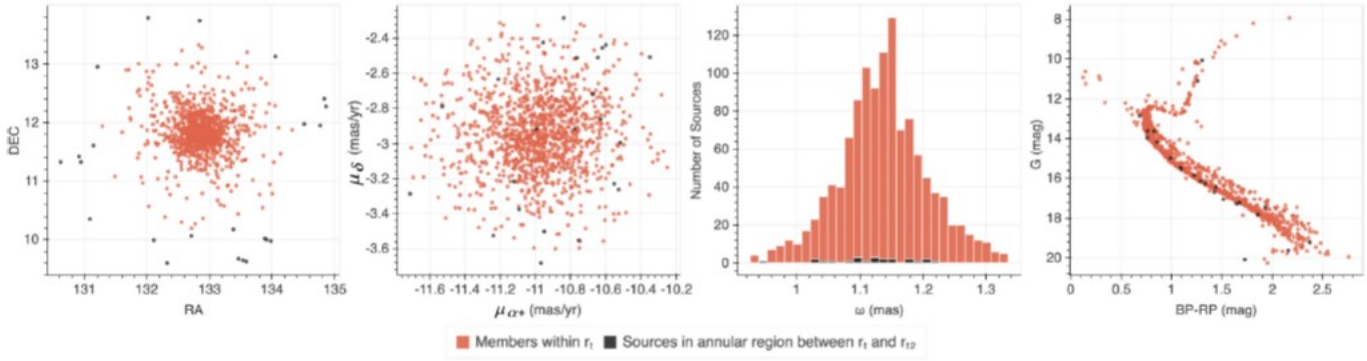


Figure 9. Comparing the member sources (in red) of M67 within r_t with the sources (in black) in an annular region of the same area as the area up to r_t from the cluster centre. Out of the 1217 sources within r_{t2} we find 23 sources beyond r_t.

Table 2. An estimate of the degree of contamination in the member sources extracted by our algorithm. Column 2 gives the cluster radius, columns 3 and 4 give core and tidal radii, respectively, as estimated from the King's profile fitting, column 5 gives the radius that encloses twice the area enclosed by the tidal radius, column 6 gives the number of member stars for each cluster, column 7 gives the number of sources identified by the algorithm between the tidal radius of the cluster and the radius that encloses twice the area of the tidal radius, and column 8 gives the degree of contamination for each cluster. See the text for an explanation of the clusters marked with '∗'.

Clusters	r (arcmin)	r_c (arcmin)	r_t (arcmin)	r_{t2} (arcmin)	N to r_t	N between r_t and r_{t2}	Degree of contamination (%)
M67	53	5.603	98.686	139.563	1194	23	1.93
NGC 2099	41	5.249	57.471	81.276	1640	103	6.28
NGC 2141∗	11	3.811	16.883	23.876	828	102	12.32
NGC 2243∗	14	1.207	32.911	46.543	583	12	2.06
NGC 2539∗	23	5.808	39.508	55.873	466	38	8.15
NGC 6253∗	12	3.364	23.651	33.448	743	47	6.33
NGC 6405	53	16.511	58.869	83.253	688	39	5.67
NGC 6791	14	3.041	20.507	29.001	2422	134	5.53
NGC 7044∗	9	1.597	16.982	24.016	693	46	6.64
NGC 7142∗	12	2.842	22.719	32.130	316	21	6.65

Comparison with other clustering algorithms

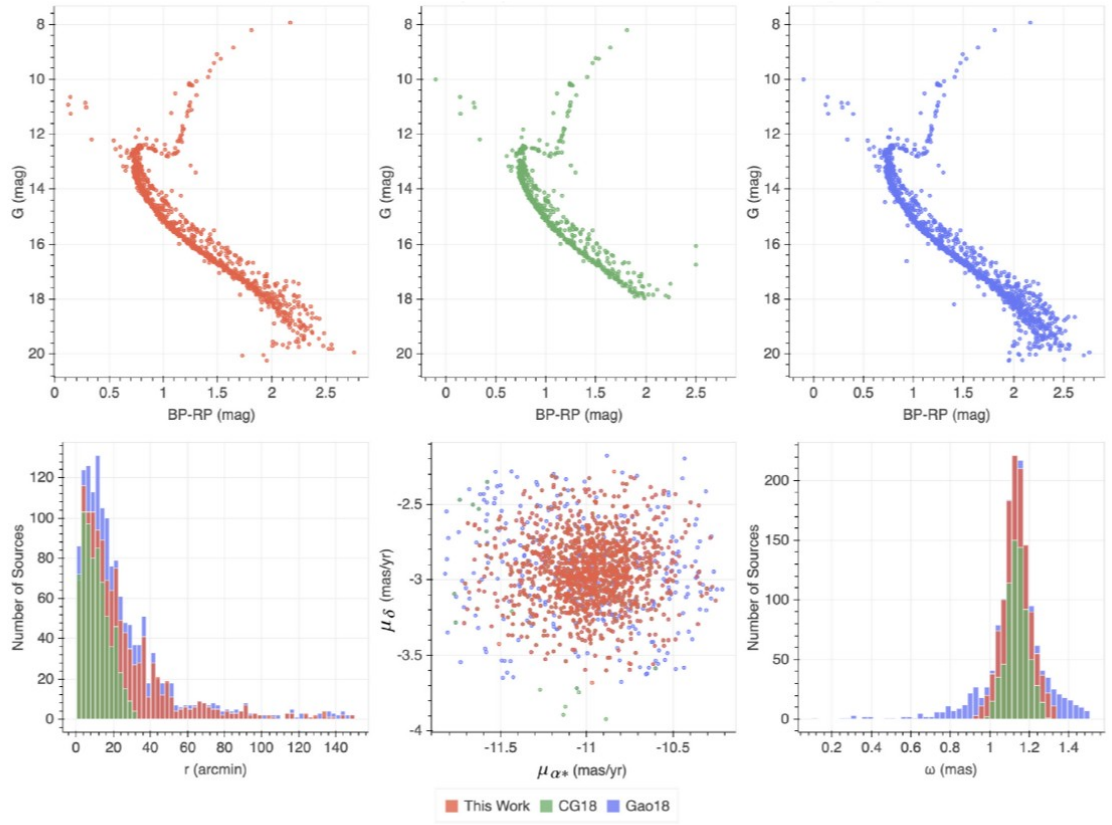


Figure 10. Top: The CMDs of members identified for the cluster M67 by the three algorithms. Bottom: The radial, proper-motion, and parallax distributions of members by the three algorithms.

This work: kNN and GMM
CG18: UPMASK
Gao 18: GMM and RANDOMFOREST method