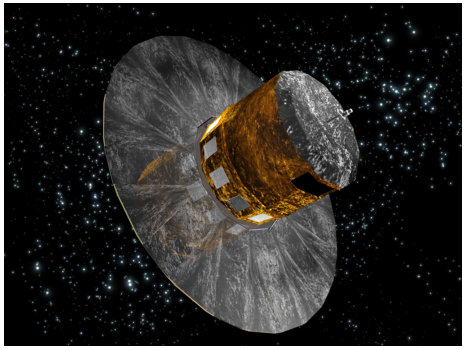
Membership of Stars in Open Clusters using Random Forest with Gaia Data

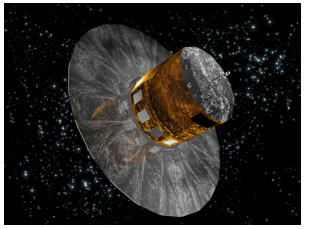


Priya Hasan
Maulana Azad National Urdu University Hyderabad,500032
priya.hasan@gmail.com

With Md Mahmudonobe, Mudasir Raja, S N Hasan
Maulana Azad National Urdu University Hyderabad,500032
Minerva Schools at KGI, San Francisco, California 94103, USA.

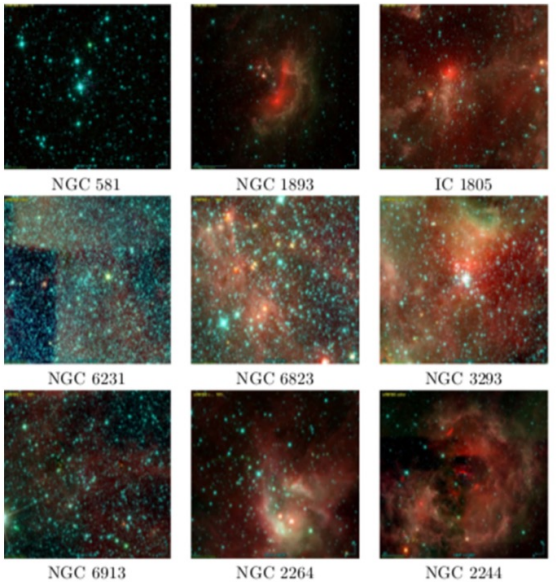
Star Clusters: the Gaia Revolution

Membership in clusters (Supervised method)



- *GAIA DR2 has a very strong influence on the membership of star clusters. This is one of the most crucial parameters in studies of star clusters. In the present study, we use membership data from Cantat-Gaudin et al(2018) based on GAIA DR2 as a training set.*
- *Random Forest (RF), which is a supervised classification method, is applied to the Gaia DR2 data in this paper. We use the results from Cantat et al as our training data to find new members in a sample of nine open clusters (NGC 581, NGC 1893, IC 1805, NGC 6231, NGC 6823, NGC 3293, NGC 6913, NGC 2264, NGC 2244).*
- *The sample has clusters with ages ranging from 1.3–20 Myr, at galactocentric distanceRGCranging from 7.3–14.5 kpc and at varying galactic latitudes land longitudes b*

Data, Training



A&A 618, A93 (2018)
https://doi.org/10.1051/0004-6361/201833476

Astronomy & Astrophysics

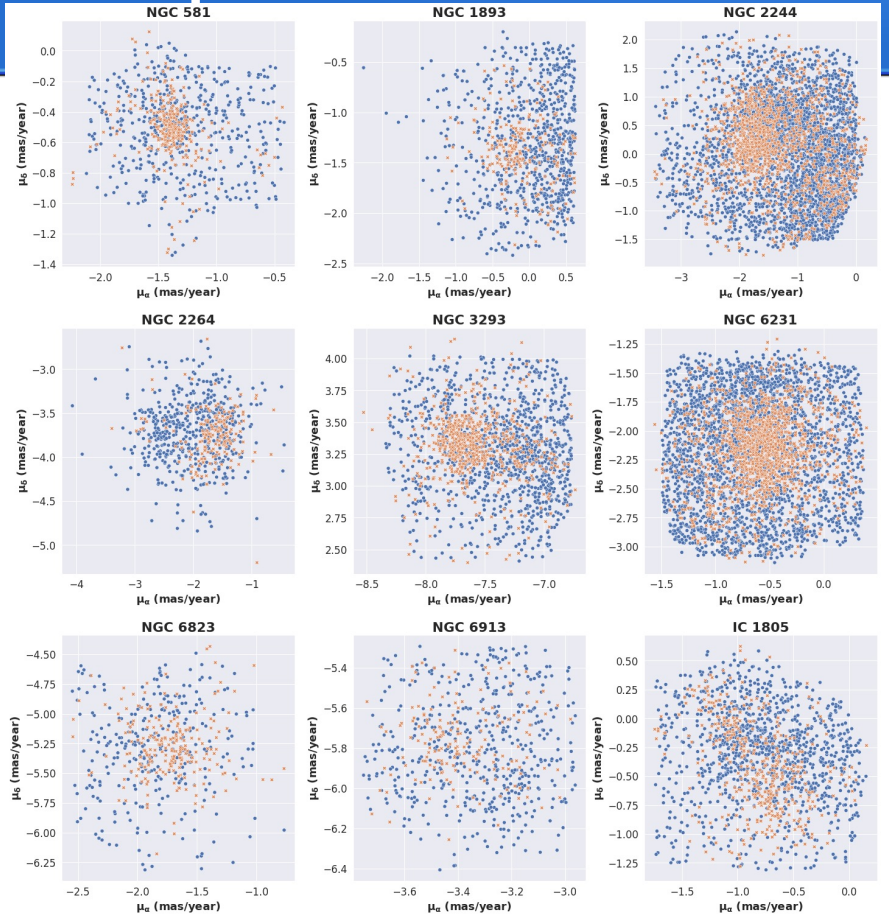
A *Gaia* DR2 view of the open cluster population in the Milky Way★

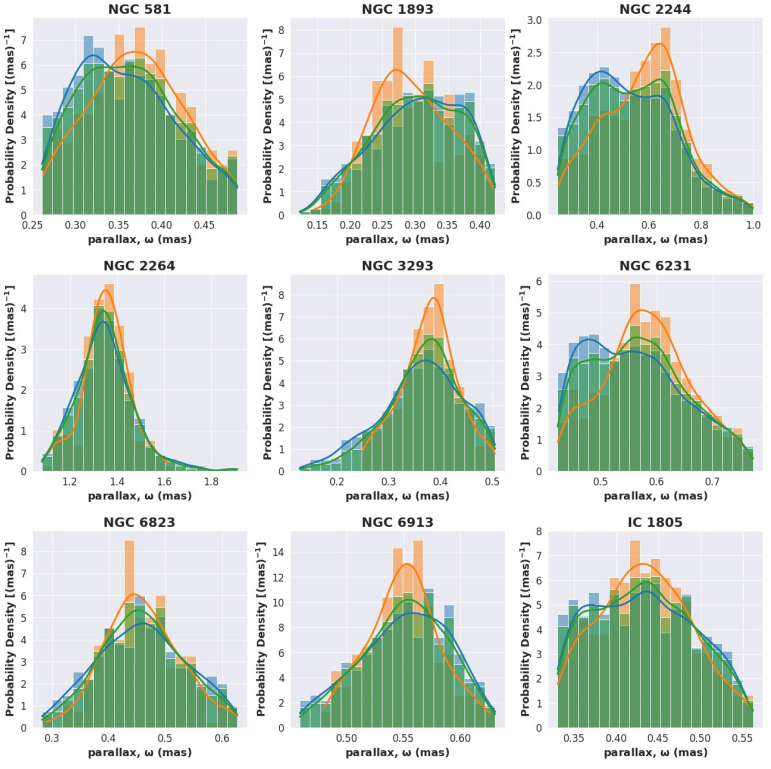
T. Cantat-Gaudin[1], C. Jordi[1], A. Vallenari[2], A. Bragaglia[3], L. Balaguer-Núñez[1], C. Soubiran[4], D. Bossini[2], A. Moitinho[5], A. Castro-Ginard[1], A. Krone-Martins[5], L. Casamiquela[4], R. Sordo[2], and R. Carrera[2]

Table 1 Basic cluster parameters

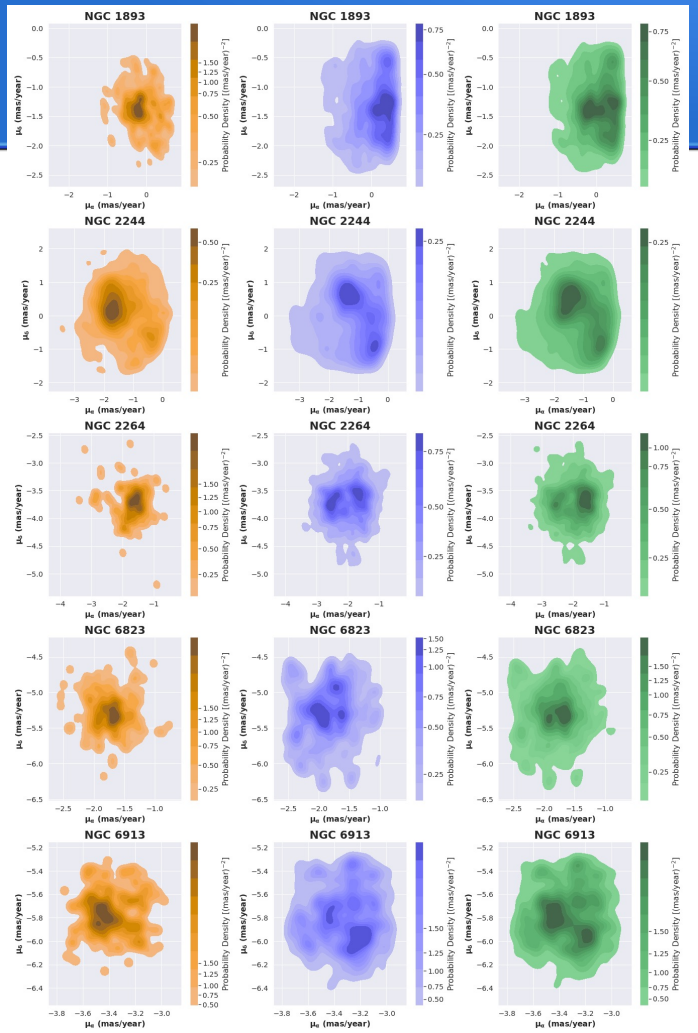
Cluster	l deg	b deg	Ang.Dia arc min	Distance pc	$E(B-V)$ mag	log(age) log(yr)	R_{GC} kpc
NGC 581	128.05	– 01.80	5.0	2194	0.38	7.3	10.0
NGC 1893	173.59	– 1.68	25	6000	0.45	6.5	14.5
IC 1805	134.73	0.92	20	2344	0.87	6.1	10.3
NGC 6231	343.46	+01.18	14.0	1243	0.85	6.5	7.4
NGC 3293	285.86	+00.07	6.0	2327	0.26	7.0	8.2
NGC 6913	76.91	+00.59	10.0	1148	0.74	7.1	8.3
NGC 2264	202.94	+02.2	39.0	667	0.05	6.9	9.1
NGC 2244	206.31	– 02.07	29.0	1445	0.46	6.9	9.8

Proper Motion & Parallax Plots





PM



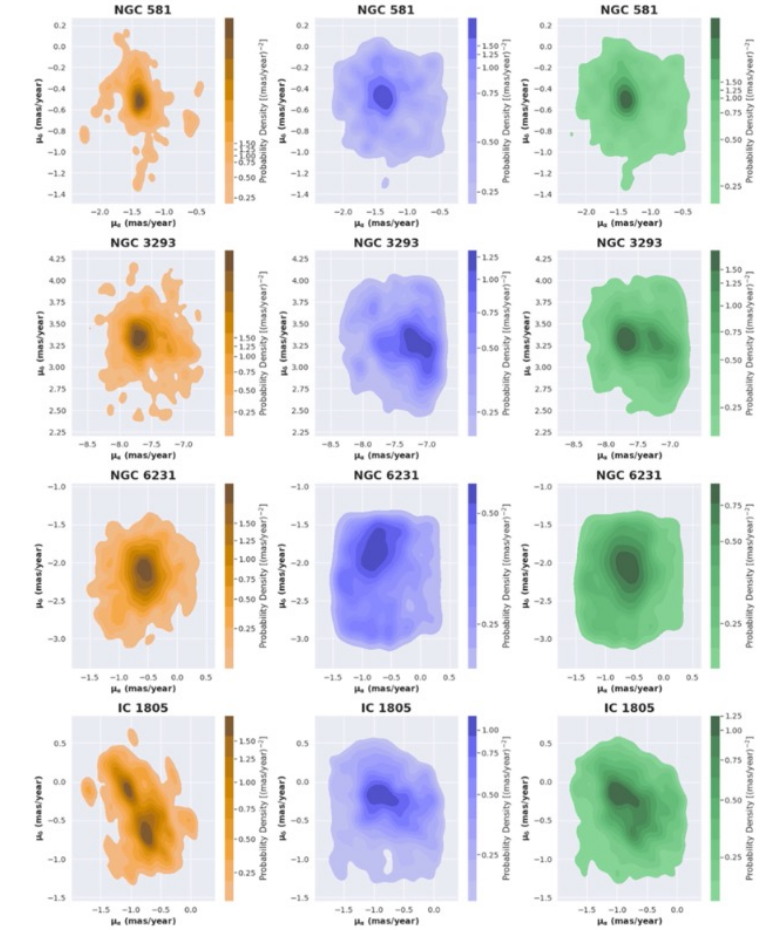
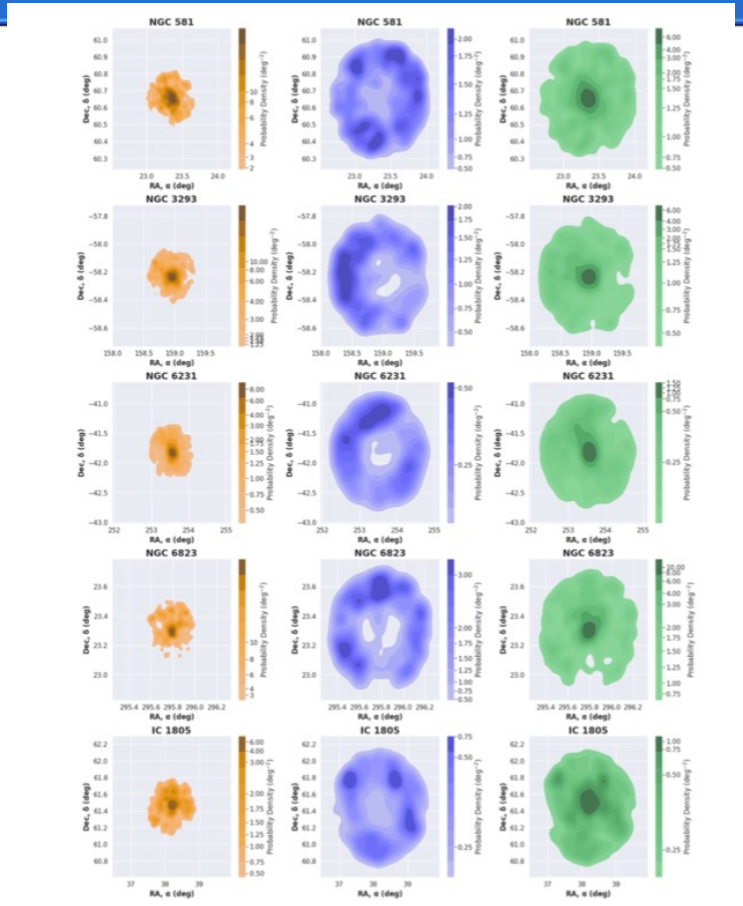
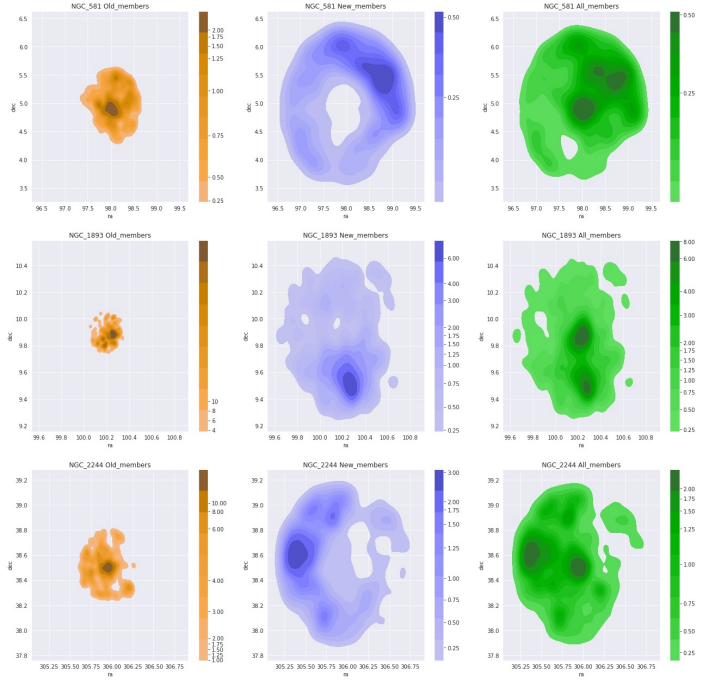


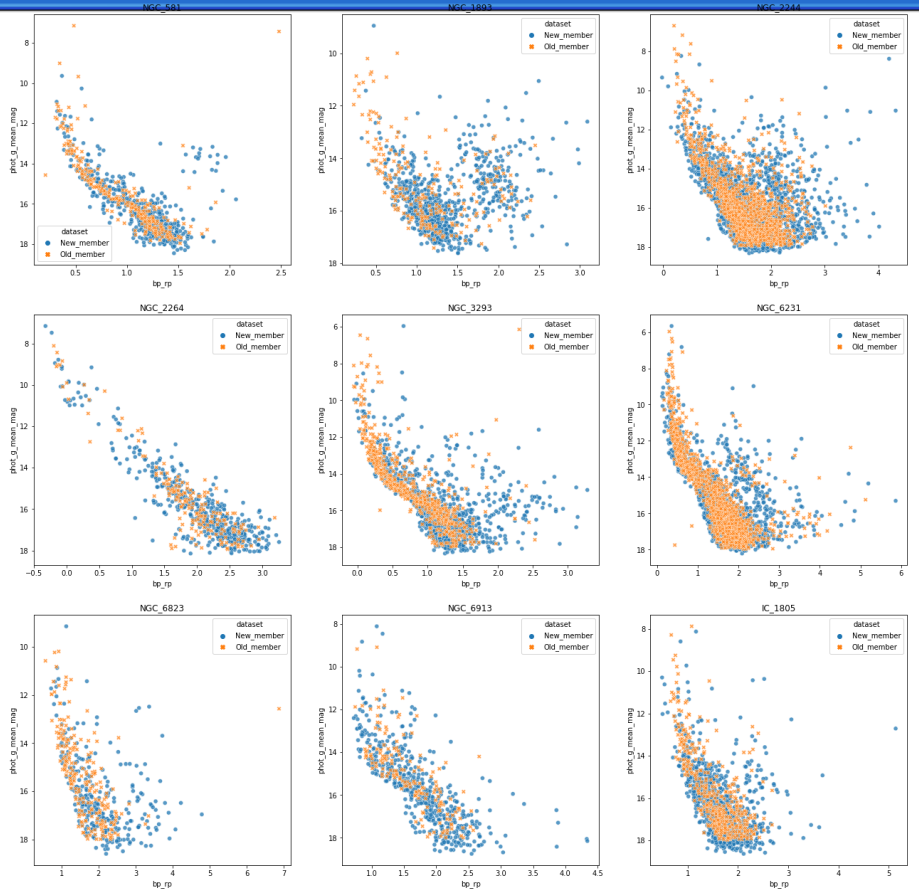
Fig. 9 Proper Motion Plot of CG (orange), new members (blue) and combined (green) members of the sample clusters (contd)

Sky plots

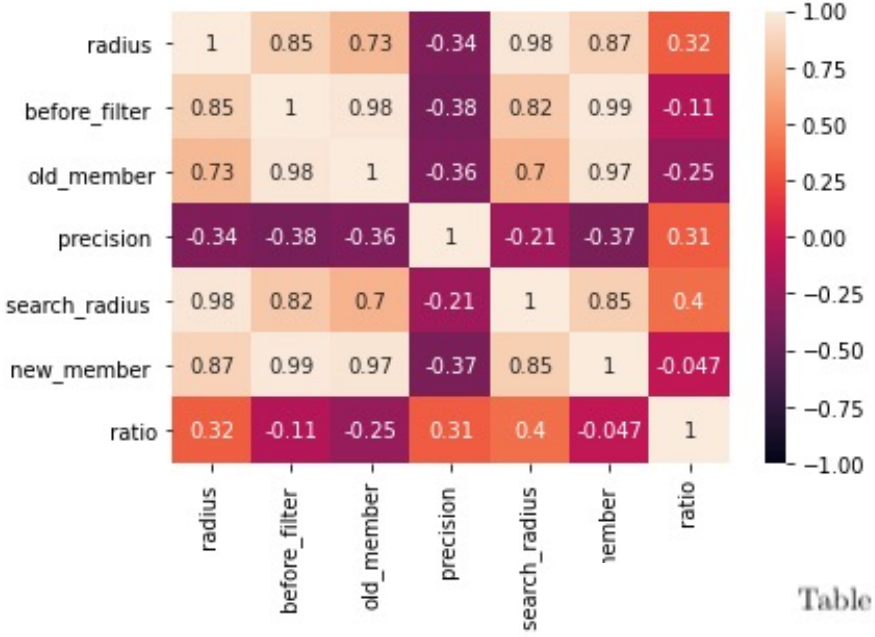

NGC 581
NGC 581
NGC 581
NGC 3293
NGC 3293
NGC 3293
NGC 6231
NGC 6231
NGC 6231
NGC 6823
NGC 6823
NGC 6823
IC 1805
IC 1805
IC 1805
RA, α (deg)
Dec, δ (deg)
Probability Density (deg⁻²)


NGC_581 Old_members
NGC_581 New_members
NGC_581 All_members
NGC_1893 Old_members
NGC_1893 New_members
NGC_1893 All_members
NGC_2244 Old_members
NGC_2244 New_members
NGC_2244 All_members
ra
dec

Color-Magnitude Diagrams



Validation



Divide the training data in a ratio of 30 : 70.
We made a grid with the possible range of values for important model parameters (i.e. number of trees in RF, maximum depth of a tree, minimum samples needed for a split, minimum sample for a leaf node etc)
Then we applied a randomized search 5-fold cross validation in the train subset with 100 iteration which in total builds 500 models with randomly chosen parameters from the grid and select the model which resulted in maximum precision.

Table 3: Prediction from the Random Forest Model

Cluster	Radius deg	Members before filter	Members after filter	Non-Member radius deg	Search radius deg	New Members	Precision %	Ratio of new to CG
NGC 581	0.17	306	290	0.7-0.8	0.34	525	86	1.81
NGC 1893	0.41	494	218	1.0-1.1	0.82	774	93	3.55
NGC 2244	0.67	1701	1192	1.4-1.5	1.33	3043	88	2.55
NGC 2264	0.19	186	179	1.0-1.1	0.60	514	99	2.87
NGC 3293	0.20	657	617	0.7-0.8	0.40	1089	94	1.76
NGC 6231	0.47	1580	1354	0.95-1.0	0.94	2710	92	2.00
NGC 6823	0.2	236	220	0.7-0.8	0.40	304	93	1.38
NGC 6913	0.3	170	170	0.7-0.8	0.60	536	95	3.15
IC 1805	0.33	456	430	0.7-0.8	0.66	1104	90	2.57

Results

Eur. Phys. J. Spec. Top.
https://doi.org/10.1140/epjs/s11734-021-00205-x

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Membership of stars in open clusters using random forest with gaia data

Md Mahmudunnobe[1], Priya Hasan[2,a], Mudasir Raja[2], and S. N. Hasan[2]

[1] Minerva Schools at KGI, San Francisco, CA 94103, USA
[2] Maulana Azad National Urdu University, Gachibowli, Hyderabad 500 032, India

- Members increased by 2--3 times. Improves accuracy in determining various parameters of a star cluster ranging from distance, extinction and mass function.
- The sizes revised
- Likely cluster members, escaped members
- find sub-structure in velocity space as well as spatial distribution of the cluster unresolved binary sequences (NGC~6231) as well as all other possible non main-sequence members of the cluster.

Supervised Learning ?

Supervised methods (where we NEED good training data)
•Pro: It can perform better or give good accuracy or prediction even with a high number of data
•Cons: Its accuracy depends on how good the training set is

Unsupervised Learning

Unsupervised method (to get the training set from the raw data)

Which UM is better? Is there any single UM which works well for all or does it depends on the cluster?

•Which SM is better? Is there any single UM which works well for all or does it depends on the cluster?

Gaussian Mixture Modelling/DBSCAN

Does it work for all clusters?

Field/Cluster ratio?

Gao 2018:

- Stars within 20' of cluster radius
- Normalizing the features
- the stars must lie within the proper motion range of $\left|\mu_{\alpha} \cos\delta\right| \leq 20\ mas/yr$ and $\left|\mu_{\delta}\right| \leq 20\ mas/yr$
- within a distance range of 300 ~ 700 pc from the Sun

Agarwal 2020:

- Stars with its G-mag error less than 0.005,
- The chosen range of proper motions has a width between 3 and 5 mas/yr around their median value
- The selected range of parallax has a width between 0.4 and 2.5 mas depending on their distance (i.e. wide for nearby clusters and narrow for distant clusters)
- Normalizing the features