Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **489,** 2189–2194 (2019) doi:10.1093/mnras/stz2290

Evaluation of nearby young moving groups based on unsupervised machine learning

Jinhee Lee★ and Inseok Song★
Department of Physics and Astronomy, The University of Georgia, Athens, GA 30602, USA

Accepted 2019 August 15. Received 2019 August 15; in original form 2019 May 9

ABSTRACT

Nearby young stellar moving groups have been identified by many research groups with different methods and criteria, giving rise to caution regarding the reality of some groups. We aim to utilize moving groups in an unbiased way to create a list of unambiguously recognizable moving groups and their members. For the analysis, two unsupervised machine-learning algorithms (K-means and Agglomerative Clustering) are applied to previously known bona fide members of nine moving groups from our previous study. As a result of this study, we recovered six previously known groups (AB Doradus, Argus, β Pic, Carina, TWA and Volans–Carina). Three other known groups are recognized as well; however, they are combined into two new separate groups (ThOr + Columba and TucHor + Columba).

Key words: methods: data analysis – open clusters and associations: general – solar neighbourhood.

1 INTRODUCTION

Nearby young stellar moving groups (hereafter NYMGs) are gravitationally unbound loose stellar associations; in this article, we focus primarily on NYMGs that are younger than 100 Myr and have mean distances of less than 100 pc from the Sun. The first NYMG, the TW Hya association (TWA), was discovered in late 1990 (Kastner et al. 1997). New NYMGs were then identified by discoveries of comoving groups of stars in a small region of space (within a few tens of pc) showing signs of similar ages. Even though NYMGs have a relatively short history (about 20 years), many NYMGs ($N \sim 10$) and their members ($N > 2000$) have been discovered (e.g. Zuckerman et al. 2001; Zuckerman & Song 2004; Mamajek 2007; Torres et al. 2008; Malo et al. 2013; Elliott et al. 2016; Riedel et al. 2017; Gagné et al. 2018).

Recent studies of NYMGs (e.g. Faherty et al. 2018; Gagné et al. 2018; Lee & Song 2019; Zuckerman 2019) recognize nine NYMGs: the TW Hya association (TWA), β Pic moving group (BPMG), 32 Ori association (ThOr), Tucana–Horologium association (TucHor), Carina association (Carina), Columba association (Columba), Argus association (Argus), AB Doradus moving group (ABDor) and Volans–Carina association (VCA). Other NYMGs have been claimed (e.g. GAYA 1 and 2 and AnA by Torres et al. 2003; Carina–Vela by Makarov & Urban 2000), but were later redefined or disappeared. The Tucana association (Zuckerman & Webb 2000) and Horologium association (Torres et al. 2000) were merged as a single group by Zuckerman (2001): the Tucana Horologium association (TucHor). Liu (2015) points out that several moving groups may be merged or rejected, or new groups could be discovered.

In this study, we re-evaluate the nine well-known NYMGs without relying on any previous knowledge. For this task, cluster analysis is an ideal technique. Cluster analysis is a branch of unsupervised machine learning that finds groups within a dataset without prior reference to the outcome. To identify a star belonging to a certain NYMG, the required information is Galactic position (*XYZ*), Galactic velocity (*UVW*) and age. Stars occupying a similar location in age–spatio-kinematic space (i.e. sharing characteristics in 7D space) can be classified as members of the same group. On the other hand, a star located far from this group of stars in 7D space can be assessed as a non-member.

2 METHOD

2.1 Data

As input data, we use bona fide members of nine NYMGs from Lee & Song (2019; Paper I), which enable us to evaluate NYMGs by comparing clustering results with previously known NYMGs. The nine groups considered are TWA, BPMG, ThOr, TucHor, Carina, Columba, Argus, ABDor and VCA While the spatio-kinematic information for the stars, *XYZ* and *UVW*, is straightforward to calculate, their age is difficult to obtain and generally has a large uncertainty. In the age range of NYMGs (∼8 to ∼100 Myr), age is evaluated using Li equivalent width, position on colour–magnitude diagrams, near-UV excess or X-ray brightness (e.g. Zuckerman & Song 2004; Soderblom 2010; Rodriguez et al. 2011; Malo et al.

★ E-mail: jinhee@uga.edu (JL); song@uga.edu (IS)

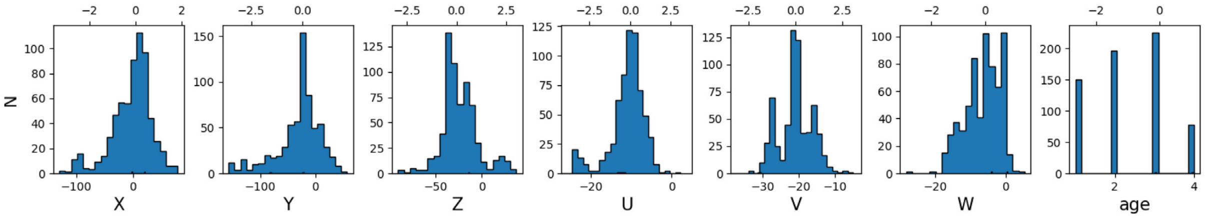


Figure 1. Distribution of input data. Raw and transformed scales are displayed along the bottom and top *x*-axis, respectively. Units for raw data are pc (*X*, *Y* and *Z*) and km s^{-1} (*U*, *V* and *W*). Age (age class) has no unit.

2013). The bona fide members of these nine NYMGs are youth-confirmed members. Instead of using numerical, continuous ages in Myr, we use categorical ages (i.e. age classes) to ensure more efficient groupings in our cluster analysis. Using common age-dating methods mentioned previously, we age-dated all input stars ($N = 652$) into five age classes: 1 for 8–10 Myr (TWA), 2 for 12–20 Myr (BPMG and ThOr), 3 for 30–40 Myr (TucHor, Carina, Columba and Argus), 4 for 90–150 Myr (ABDor and VCA) and 5 for older (field stars).

We do not consider chemical abundance or metallicity in our analysis, because metallicity has a large uncertainty and it is known that young nearby stars should all have similar abundance (Viana Almeida et al. 2009).

2.2 Clustering algorithms

Clustering algorithms find groups in data in such a way that members in the same group are more similar to each other than to those in other groups. There are various clustering algorithms, with different strategies and criteria for finding groups. In this study, we use two algorithms in the SKLEARN package of PYTHON (Pedregosa et al. 2011). Each algorithm has its own specific parameters.

2.2.1 K-means

The K-means algorithm finds groups in data by trying to separate samples into a desired number of groups while minimizing the within-cluster sum-of-squares. The within-cluster sum-of-squares, also known as the *inertia*, is calculated by the following equation:

$$\sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2, \qquad (1)$$

where C_k is the kth cluster and μ_k is the centroid (the mean) of C_k. The K-means algorithm requires the number of clusters (n_{clusters}) as an input parameter.

2.2.2 Agglomerative Clustering

The Agglomerative Clustering algorithm tries to build a hierarchy of clusters, starting with clusters consisting of an individual sample. Clusters are then successively merged together. To determine which neighbouring clusters will merge, there are several options for linking criteria (e.g. Ward, maximum, average and single linkage). We use Ward in this study. Ward has a merge strategy minimizing the sum of squared differences within all clusters. The number of clusters (n_{clusters}) is required as an input parameter, similarly to the K-means method.

2.3 Procedure

2.3.1 Preprocessing of data

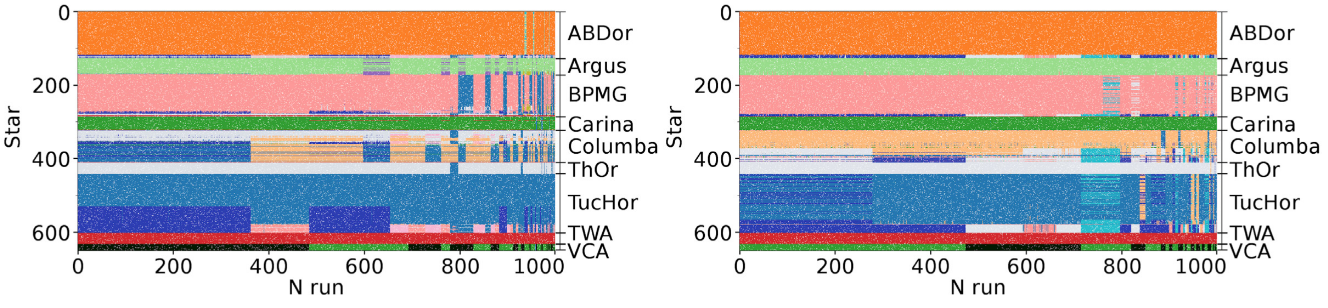
For reliable clustering, raw data should be transformed to make all variables have a similar range. Using raw data without proper transformation can cause biased results. For example, *XYZ* values are distributed over a range of 100 (in pc) and *UVW* values are distributed over a range of tens (in km s^{-1}; see Fig. 1). Parameters with more extensive ranges can have a more significant effect on clustering.

Preprocessing in cluster analysis is a process for homogenizing the range of data variables. There are various transformation functions in the SKLEARN package. StandardScaler in SKLEARN transforms data to make the distribution have a mean value of zero and standard deviation of one, by subtracting the mean and then dividing by the standard deviation of the whole dataset. Since outliers have a significant influence on the mean and standard deviation, StandardScaler would not be a proper choice as a transformation function. RobustScaler in SKLEARN scales variables using values within percentiles and therefore is not influenced by a small number of outliers. In this study, RobustScaler is applied for transforming *XYZUVW* values, resulting in the transformed variables having a range of approximately −2.5 to + 2.5. After transformation, two outliers (∼15σ in *U* and ∼7σ in *V*) are removed.

The age class has values in the range from +1 to + 5. To make an age class spanning a similar range to the transformed *XYZUVW* variables, we subtract 3.125 from the age class and multiply by 1.25. This procedure, as shown in Fig. 1, is to ensure that all variables are transformed to match similar scales.

2.3.2 Strategy for evaluating NYMGs

When the entire input data set is used, both Agglomerative Clustering and K-means algorithms generate nearly the same results with a given n_{clusters}. To find a general trend of groupings from multiple clustering results, for a given algorithm, we performed 1000 trial runs by varying the input data slightly for each run. In each run, 95 per cent of the input data are randomly selected. Because of the slight difference in the selected data for each run, a different outcome appears each time. For each combination of the chosen algorithm (K-means and Agglomerative Clustering) and n_{clusters} (from 3 to 10), we run the calculations 1000 times. From the clustering results from a total of 16,000 runs, we identify groups. Members of these newly identified groups are then tracked and members consistently assessed into a group (>70 per cent of trial runs between the two algorithms) can be recognized as 'bona fide' members of the group.



(a) K-means

(b) Agglomerative Clustering

Figure 2. Results from 1000 runs with $n_{\rm clusters} = 9$. Left and right panels present the results of K-means and Agglomerative Clustering algorithms, respectively. The *x* and *y*-axis represent the run number and stellar index, respectively. The results are sorted along the *x* and *y*-axis for better display and interpretation. Stars (the *y*-axis) are sorted based on previously known grouping. The 'N run' (the *x*-axis) is sorted to locate the most frequent grouping first (at the left side of the result figure). Therefore, a star placed in the 30th row in K-means results might be located in a different row in Agglomerative Clustering results. A vertical line represents a result from a single run and stars displayed with the same colour in the line belong to the same group. Colour figures are available online.

3 RESULTS

3.1 Finding a grouping trend

The most frequent groupings for each of 16 combinations of algorithms and $n_{\rm clusters}$ values are found and discussed. To identify the most dominant grouping with a given $n_{\rm clusters}$ and an algorithm, we need to identify the most frequent pattern of grouping through the 1000 runs. The tested input parameter, $n_{\rm clusters}$, is in the range 3–10. Fig. 2 presents the results with $n_{\rm clusters} = 9$. Results with other $n_{\rm clusters}$ values are presented in Appendix A.

In these figures, the *x*- and *y*-axes represent the indices of runs and stars, respectively. In each run, each star's clustering results are displayed as a colour-coded point. For an easier interpretation, the 2D results are sorted along both the *x*- and *y*-axes. The results along the *y*-axis (stellar indices) are sorted first. Stars are sorted by previously known group membership first, to compare the new clustering results with the previously known grouping. ABDor members are located at the top (index 0 to ∼100) and VCA members are located at the bottom. Then the results are sorted along the *x*-axis. This process sorts the results, placing similar grouping results together, which enables us to identify frequent groupings easily through the entire set of trial runs. When a star is not included in the 95 per cent sampling, such cases are represented as white dots.

If one extracts a slice along the *y*-axis (i.e. takes results from a single run), stars with the same colour belong to the same group. The number of colours in one trial run should be the same as $n_{\rm clusters}$. However, when we look at the results of the entire 1000 runs, generally more than $n_{\rm clusters}$ groupings appear. For example, results from K-means with $n_{\rm clusters} = 5$ have six colours. However, if one takes out any single vertical slice, the number of colours is always five for $n_{\rm clusters} = 5$.

3.2 The best results with $n_{\rm clusters} = 9$ and comparison with previously known NYMGs

The interpretation of clustering results can depend critically on the choice of $n_{\rm clusters}$. However, choosing a proper $n_{\rm clusters}$ value is difficult, because it requires acceptance of the number of 'true' NYMGs. Nonetheless, because we use nine NYMGs as the input, $n_{\rm clusters}$ close to 9 would be an appropriate choice to compare our results with the previously known NYMGs. To evaluate if $n_{\rm clusters}$ = 'number of input groups' is a suitable choice, we performed a test with a smaller input data set using only four NYMGs (ABDor, BPMG, TucHor and TWA). The two clustering algorithms were utilized in the same way as we used the entire data, but the tested $n_{\rm clusters}$ range is from 3–8. When $n_{\rm clusters} = 4$, both algorithms correctly recovered four NYMGs. Although this exercise does not guarantee that $n_{\rm clusters} = 9$ should be the best choice in considering the entire dataset, the choice of $n_{\rm clusters}$ may be suitably set to the number of 'expected' true groupings in the input data.

Table 1 and Fig. 2 present the dominant grouping with $n_{\rm clusters} = 9$. Both algorithms find ABDor, Argus, BPMG and TWA (orange, light green, pink and red in Fig. 2, respectively; see online version to view figures in colour) in more than 80 per cent of cases. Carina (green) and VCA (dark green) are found with both algorithms when considering a lower percentage (70 per cent with K-means and 35 per cent with Agglomerative Clustering).

Interesting results come from the remaining groups (ThOr, Columba and TucHor). Both algorithms find a new group, consisting of the entire ThOr and a subset of Columba (95 per cent of cases with K-means and 40 per cent of cases with Agglomerative Clustering). With the remaining members, while Agglomerative Clustering splits TucHor (blue) and the subset of Columba (light orange), the K-means algorithm makes two groups consisting of the mixture of TucHor and Columba (blue and dark blue).

Based on the results with $n_{\rm clusters} = 9$, six groups matching previously known NYMGs are recognized (ABDor, Argus, BPMG, TWA, Carina and VCA). A new group containing the entire ThOr and a subset of Columba members is frequently recognized. We suggest calling this group ThOr–Col. The remaining stars consist mainly of the entire TucHor and the majority of Columba. While each K-means and Agglomerative Clustering algorithm recognizes two groups in the remaining members, the two groups from these algorithms do not match well (i.e. the specific content of the two groups from both algorithms is different). Therefore, we define a single group enclosing the members of these two groups. This group is called THC (TucHor–Col). Even though the groups are taken from the results with $n_{\rm clusters} = 9$, the number of recognized groups is eight.

As explained in Section 2.3.2, stars consistently (i.e. >70 per cent) assigned to the same group from 1000 runs can be regarded as 'reliable' members of the group. From two such lists of members using the two algorithms, a set of common members is selected as bona fide members of the newly identified eight NYMGs. The number of bona fide members listed in Table 2 appears to be

Table 1. The most frequent grouping with $n_{\mathrm{clusters}} = 9$. Groupings with other n_{clusters} values are given in Appendix A. The per cent value indicates that a group is recognized in that percentage of runs. Groups with a single prime and double prime (′ and ″) indicate that not the entire group, but a fraction of the group (about 50–80 and 20–50 per cent, respectively), is enclosed in the new group.

n_{clusters}	K-means		Agglomerative Clustering	
	Recognized groups (per cent)	N_{members} ^{a}	Recognized groups (per cent)	N_{members} ^{a}
9	ABDor (99)	120	ABDor (100)	120
	Argus (85)	50	Argus (100)	50
	BPMG (85)	120	BPMG (80)	120
	TWA (100)	20	TWA (100)	20
	Carina (70)	40	Carina (35)	40
	VCA (70)	20	VCA (35)	20
	ThOr + Columba″ (95)	50	ThOr + Columba″ (40)	50
	Columba′ + TucHor′ (60)	100	Columba′ (80)	40
	Columba′ + TucHor″ (60)	70	TucHor′ (50)	150

Note. ^{a}Approximate number of members.

Table 2. Sample bona fide members of the newly defined groups. The full data are available online.

Group	Name	SpT	RA (hh:mm:ss)	Dec. (dd:mm:ss)
ABDor	PW And	K0Ve	00:18:20.89	+ 30:57:22.1
ABDor	2MASS J00192626 + 4614078	M8	00:19:26.27	+ 46:14:07.8
ABDor	2MASS J00489 + 4435AB	M4	00:48:58.18	+ 44:35:08.8
	...			
	...			
	...			
Argus	CD-29 2360	K3Ve	05:34:59.23	−29:54:04.0
Argus	CD-42 2906	K1V	07:01:53.41	−42:27:56.2
Argus	CD-48 2972	G8V	07:28:22.03	−49:08:37.7
	...			
	...			
	...			

considerably larger than those in the Bayesian Analysis for Nearby Young AssociatioNs (BANYAN) series (Malo et al. 2013; Gagné et al. 2014, 2018). The bona fide members referred to in those articles are 'classical' members used to build kinematic models for identifying new members. Our input data include not only these classical members but also newly identified members from recent studies (see Paper I).

Fig. 3 compares the distributions of bona fide members of the eight new groups (this study) and those of classical groups (the input data) in spatio-kinematic space. As explained earlier, THC and ThOr–Col enclose TucHor, Columba and ThOr, while the other six groups recovered are almost the same as in the previous NYMGs.

4 DISCUSSION AND CONCLUSION

In this study, we carried out pattern recognition in age–spatio-kinematic space for an unbiased evaluation of NYMGs using an unsupervised machine learning method. The grouping was evaluated by numerous trials of the clustering algorithms with 95 per cent random sampling of previously known NYMG members. The most frequent grouping pattern was found with a given n_{clusters}. While there are some differences between results with K-means and Agglomerative Clustering algorithms, there is a significant similarity between these results.

ABDor is the first recognized group using both algorithms at $n_{\mathrm{clusters}} = 3$ and Argus is the second recognized group ($n_{\mathrm{clusters}} = 4$). Carina and VCA tend to form a single group at a small n_{clusters}, which indicates that these two groups share similar characteristics (*XYZUVW* and age). However, with a large n_{clusters} value (n_{clusters} = 9–10), these two groups are clearly separated as in previously known groupings (i.e. recognized as Carina and VCA). This implies that these two groups share similar characteristics, but are probably not a single entity. Similar to the case of Carina and VCA, BPMG and TWA have similar age–spatio-kinematic characteristics, making them into a single entity with a smaller n_{clusters} value (<7). However, these two groups are clearly split with larger n_{clusters} values.

One interesting result comes from Columba and ThOr. Both algorithms create a merged group, ThOr–Col. The entire ThOr and the majority of Columba members form this new group. The THC group consists of almost the entire TucHor and half the Columba members. Zuckerman et al. (2011) proposed combining members of TucHor and Columba, because there is difficulty assigning some of them into either TucHor or Columba. Our result supports the claim of Zuckerman et al. (2011).

The input data include 18 planet-host stars. Among these 18 stars, 11 stars are retained as bona fide members via cluster analysis in this study, while the other seven stars are not consistently recognized as members of the newly defined groups. Table 3 shows the membership of these stars. Membership from this study and the Bayesian membership probability from Paper I are compared. In the case of the 11 retained bona fide members, their memberships are consistent with those in Paper I.

In this study, previously known NYMGs are evaluated using previously known bona fide members of the groups. If old field stars are included in the analysis, the results will be different. For

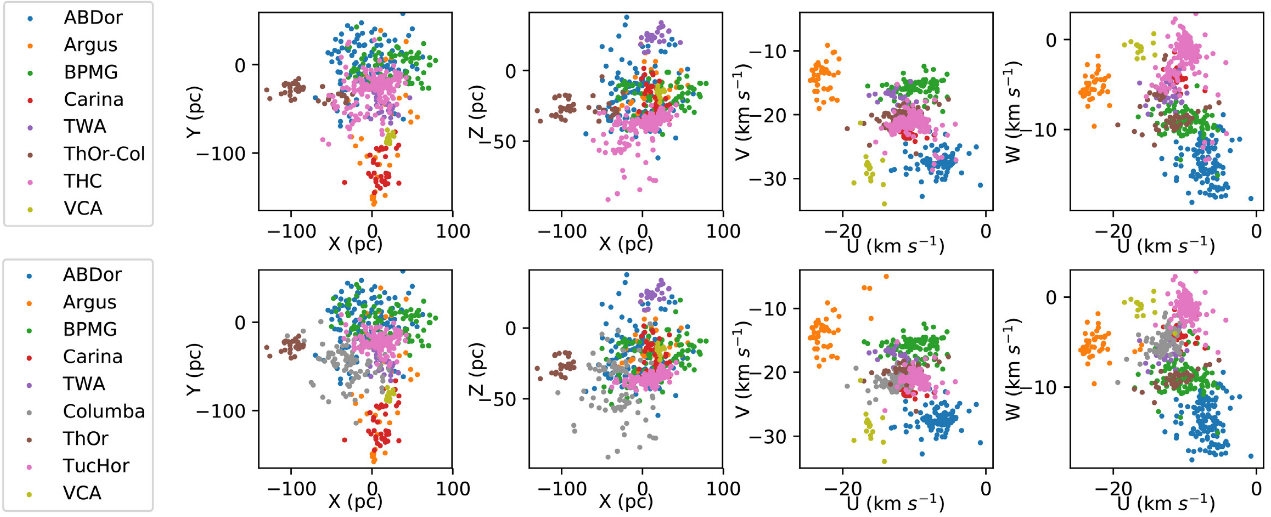


Figure 3. Distribution of bona fide members of the eight new groups (this study; top) and those of classical groups (input data; bottom). Colour figures are available online.

Table 3. Membership of planet-host stars. While bona fide members based on Bayesian membership probability in Paper I are used as input data, there are members not retained as bona fide members in the cluster analysis.

Name	New group^a	Old group (prob)^b
1RXS J235133.3 + 312720 (AB)	ABDor	ABDor (100)
2MASS J02192210-3925225B	THC	TucHor (100)
TWA 27	TWA	TWA (100)
2MASS J00413538-5621127	THC	TucHor (100)
2MASS J04373613-0228248	BPMG	BPMG (100)
CD-35 2722	ABDor	ABDor (100)
CD-52 381	THC	Columba (99)
2MASS J01123504 + 1703557	ABDor	ABDor (100)
TWA 5A	TWA	TWA (100)
β-Pic	BPMG	BPMG (100)
η Tel	BPMG	BPMG (97)
2MASS J01033563-5515561	–^{c}	TucHor (100)
TYC 9486-927-1	–^{c}	BPMG (100)
2MASS J22501512 + 2325342A	–^{c}	ABDor (100)
AB Pic	–^{c}	Columba (100)
2MASS J21140802-2251358	–^{c}	BPMG (100)
PZ Tel	–^{c}	BPMG (100)
SDSS J111010.01 + 011613.1	–^{c}	ABDor (100)

Notes. ^{a}Assigned group in this study.
^{b}Previously known groups in Paper I with a Bayesian membership probability (per cent).
^{c}These stars are in the input data, but not retained as bona fide members in Section 3.2.

example, ABDor and Argus are easily separated from other groups, because of their unique *XYZ* and *UVW*. However, because the two groups are relatively old themselves, any inclusion of field stars in the analysis will serve to undermine the appearance of separation. The number density of NYMGs is lower than that of field stars and there are many field stars sharing similar *XYZ* and *UVW* with ABDor and Argus members.

Expanding this study, young nearby stars and field stars should be analysed without knowledge about known NYMGs for the most unbiased and objective analysis of nearby stars. The new configuration of groups will provide more reliable and objective investigations of young nearby stars and the environment of the solar neighbourhood.

ACKNOWLEDGEMENTS

We acknowledge the important and critical review by the referee, which improved the manuscript significantly. We also thank Tara Cotten for valuable comments and suggestions.

REFERENCES

Elliott P., Bayo A., Melo C. H. F., Torres C. A. O., Sterzik M., Quast G. R., Montes D., Brahm R., 2016, A&A, 590, 13
Faherty J., Bochanski J., Gagné J., Nelson O., Coker K., Smithka I., Desir D., Vasquez C., 2018, ApJ, 863, 91
Gagné J. et al., 2018, ApJ, 856, 23
Gagné J., Lafrenière D., Doyon R., Malo L., Artigau É., 2014, ApJ, 783, 121
Kastner J. H., Zuckerman B., Weintraub D. A., Forveille T., 1997, Science, 277, 67
Lee J., Song I., 2019, MNRAS, 486, 3434
Liu M., 2015, in Kastner J., Stelzer B., Metchev S., eds, Proc. IAU Symp. 314, Young Stars and Planets Near the Sun. Cambridge Univ. Press, Cambridge, p. 290
Makarov V. V., Urban S., 2000, MNRAS, 317, 298
Malo L., Doyon R., Lafrenière D., Artigau É, Gagné J., Baron F., Riedel A., 2013, ApJ, 762, 88
Mamajek E. E., 2007, in Elmegreen B. G., Palous J., eds, Proc. IAU Symp. 237, Triggered Star Formation in a Turbulent ISM. Cambridge Univ. Press, Cambridge, p. 442
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Riedel A., Blunt S., Lambrides E., Rice E., Cruz K., Faherty J., 2017, AJ, 153, 95
Rodriguez D. R., Bessell M. S., Zuckerman B., Kastner J. H., 2011, ApJ, 727, 62
Soderblom D., 2010, ARA&A, 48, 581
Torres C. A. O., da Silva L., Quast G. R., de la Reza R., Jilinski E., 2000, AJ, 120, 1410
Torres C. A. O., Quast G. R., de La Reza R., da Silva L., Melo C. H. F., 2003, in De Buizer J. M., van der Bliek N. S., eds, ASP Conf. Ser. Vol. 287, Galactic Star Formation Across the Stellar Mass Spectrum. Astron. Soc. Pac., San Francisco, p. 439

Torres C. A. O., Quast G. R., Melo C. H. F., Sterzik M., 2008, in Reipurth B., ed., Handbook of Star Forming Regions, Volume II: The Southern Sky (ASP Monograph Publ., Vol. 5). Astron. Soc. Pac., San Francisco, CA, p. 757

Viana Almeida P., Santos N. C., Melo C., Ammler-von Eiff M., Torres C. A. O., Quast G. R., Gameiro J. F., Sterzik M., 2009, A&A, 501, 965

Zuckerman B., 2001, in Jayawardhana R., Greene T. P., eds, ASP Conf. Ser. Vol. 244, Young Stars Near Earth: Progress and Prospects, Astron. Soc. Pac., San Francisco, p. 122

Zuckerman B., 2019, ApJ, 870, 27

Zuckerman B., Song I., 2004, ARA&A, 42, 685

Zuckerman B., Webb R. A., 2000, ApJ, 535, 959

Zuckerman B., Song I., Bessell M. S., Webb R. A., 2001, ApJ, 562, 87

Zuckerman B., Rhee J., Song I., Bessell M. S., 2011, ApJ, 732, 61

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Table2.csv

Appendix A. Results of K-means and Agglomerative Clustering with several n_{clusters} values.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.