SCLUDAM: SOFTWARE FOR ASTROMETRIC MEMBERSHIP PROBABILITIES CALCULATION IN STAR CLUSTERS

Simón P. González†, Myiram Herrera† and Susana Ruiz‡

†Instituto de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Av. José Ignacio de la Roza Oeste 590, 5400 San Juan, Argentina
‡Departamento de Geofísica, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Av. José Ignacio de la Roza Oeste 590, 5400 San Juan, Argentina

Abstract: Understanding stellar and galactic evolution relies heavily on the study of star clusters. However, detecting them and identifying their members from astrometric data remains a challenging task. We introduce SCLUDAM, a Python library designed to detect star clusters and estimate membership probabilities from astrometric data. It combines density peak detection in multidimensional histograms with a probabilistic analysis pipeline based on HDBSCAN and KDE. Simulated datasets were used to evaluate performance, showing high precision.

Keywords: Scientific Software, Membership Probabilities, Star Clusters

1 INTRODUCTION

Identifying member stars is a key preliminary step for analyzing the mass, composition, age, and kinematics of star clusters [1]. This requires distinguishing between two populations: the cluster and the surrounding field stars, and estimating the probability of each star belonging to either group. A prior detection phase is usually needed to isolate the cluster from the broader celestial region.

Several open-source tools address parts of this problem, such as SHiP [2], UPMASK [3], ASTECA [4], CLUSTERIX 2.0 [1], and pyUPMASK [5], but often handle only one task. SCLUDAM (Star Cluster Detection and Membership Estimation) is a modular, open-source Python library designed to offer a more complete and customizable analysis flow. It can be used independently or integrated into larger Python scripts, making it suitable for diverse applications beyond stellar astronomy.

SCLUDAM includes features for querying and downloading GAIA catalog data [6], cluster detection via star counts, clusterability tests, membership estimation using HDBSCAN and KDE, data simulation, and plotting tools. It is designed for use with astrometric data, such as sky coordinates, proper motions, and parallaxes, which are known to be highly discriminative for separating stellar populations [3].

The query builder in SCLUDAM targets GAIA catalogs due to their precision and completeness in astrometry and widespread adoption in the field. It includes common data quality filters, such as photometric excess [7] and astrometric noise excess [8].

2 ANALYSIS

The proposed analysis method (Fig. 1) starts with a data matrix covering a wide celestial region, potentially containing an unknown number of clusters. First, density peaks are identified as indicators of clusters. A sample is then selected from the variable ranges around each overdensity. Statistical tests assess whether the sample contains cluster structures. If so, membership probabilities are computed, and the final output is a probability matrix summarizing the results.

2.1 CLUSTER DETECTION

Detecting density peaks is a common task in astronomy, and several methods have been proposed to address it [9]. In SCLUDAM, a variant of the Star Counts (SC) algorithm was implemented. This method builds a multidimensional frequency histogram by counting stars in bins and comparing counts to identify overdensities. SC was chosen for its simplicity, scalability to large datasets, and detection performance comparable to more complex alternatives [9].

The SCLUDAM implementation introduces several improvements. Low-count bins are first removed to reduce histogram size without discarding relevant regions. A mean smoothing filter is then applied, following Alejo, González, and González [10], to estimate the field star density. Subtracting this smoothed histogram from the original highlights local overdensities. Additionally, a local dispersion measure is used to differentiate potential clusters from fluctuations in field density. The final detection histogram is calculated as $H_{score}=(H-H_{blurred})/H_{\sigma}$, where H is the original histogram, $H_{blurred}$ is the smoothed version and H_{σ} is the result of applying a standard deviation filter. As shown in Fig. 2, this score map enhances the visibility of cluster peaks. Finally, the detection process is repeated for all possible bin shifts, following the Nyquist spatial sampling criterion [9]. For each peak, the shift that yields the most distinct detection is retained.

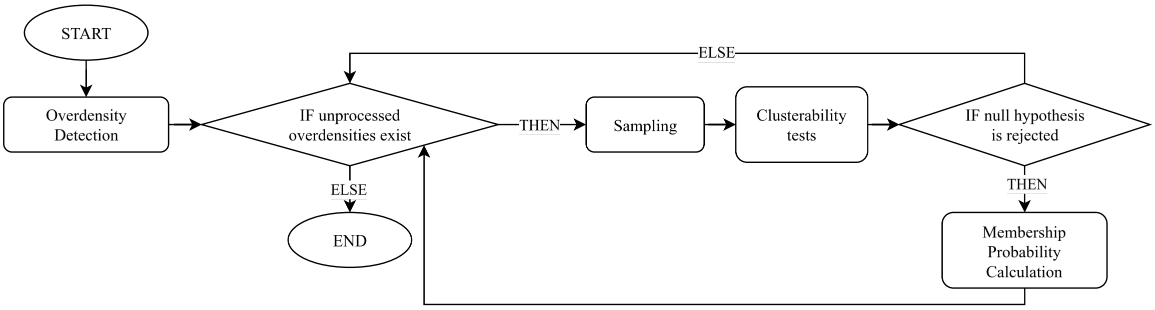
START
Overdensity Detection
IF unprocessed overdensities exist
THEN
Sampling
Clusterability tests
IF null hypothesis is rejected
ELSE
ELSE
END
THEN
Membership Probability Calculation

Figure 1: Analysis pipeline.

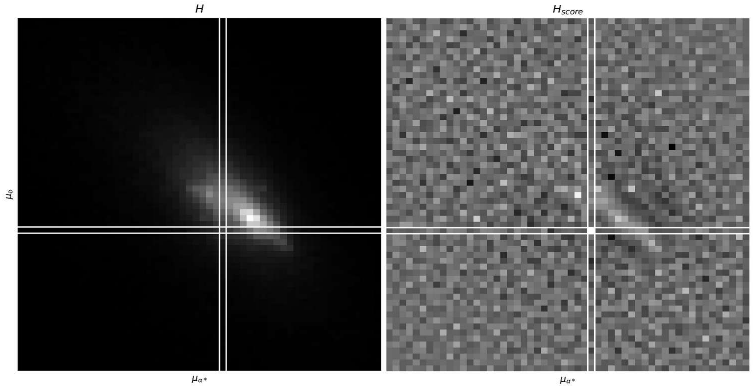

Figure 2: From left to right: the initial histogram computed from proper motions, and the corresponding score histogram. In both, the location of the region's most prominent cluster is indicated

2.2 Clusterability Tests

The process continues by sampling around the obtained density peaks. Various tests are conducted on these samples to determine if there is actual evidence of clustering structure [11]. After analyzing different options, three tests considered suitable for this application were implemented: the Hopkins test [12], the Dip-dist test [13], and Ripley's k-Test [14, 15]. Generally, these tests assess the existence of evidence against the null hypothesis that the dataset adheres to the property of Complete Spatial Randomness (CSR). This property implies that the data correspond to an underlying model of a homogeneous Poisson Point Process (PPP).

2.3 Membership Probabilities

If the previous tests fail to reject the null hypothesis, the program proceeds to compute membership probabilities within the sample. A two-step algorithm, inspired by the methods of Sampedro [16] and Krone-Martins & Moitinho [3], was implemented with several modifications.

First, initial labels are assigned to each star using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [17, 18], an extension of DBSCAN that produces a hierarchy and selects the final clustering based on group stability. HDBSCAN is well-suited for this task as it clusters based on density, a key criterion for identifying star clusters, and requires only one parameter: the minimum group size. This can be estimated from the excess number of stars in the bin where the cluster was initially detected.

These initial labels are used to estimate the probability density functions (PDFs) of the two populations, cluster and field, via Kernel Density Estimation (KDE). A distinct bandwidth matrix is used for each observation, based on a plug-in selector [19] and a variance-covariance matrix constructed from the catalog's uncertainties and correlations [20]. This ensures that the estimation accurately reflects the data's uncertainty structure.

Finally, Bayes' Theorem is applied to compute posterior membership probabilities. The end result is a probability matrix, which can be visualized with SCLUDAM (Fig. 3).

3 Results

To test the program, 155 simulated datasets were generated using five astrometric variables, with distribution parameters, uncertainties, and correlations based on real clusters from the catalog by Días et al. [21] and GAIA Data Release 3 [6]. The program was run on these datasets to compute membership probabilities, and stars were classified using the Bayes classification rule.

Binary classification metrics (precision, recall, F1 score, and Matthews correlation coefficient) were computed, along with the Brier and logarithmic scoring rules for probabilistic forecasts. All metrics range from 0 to 1, except the logarithmic score, which ranges from −∞ to 1. The program successfully detected the clusters, and as shown in Table 1, its classification performance is satisfactory.

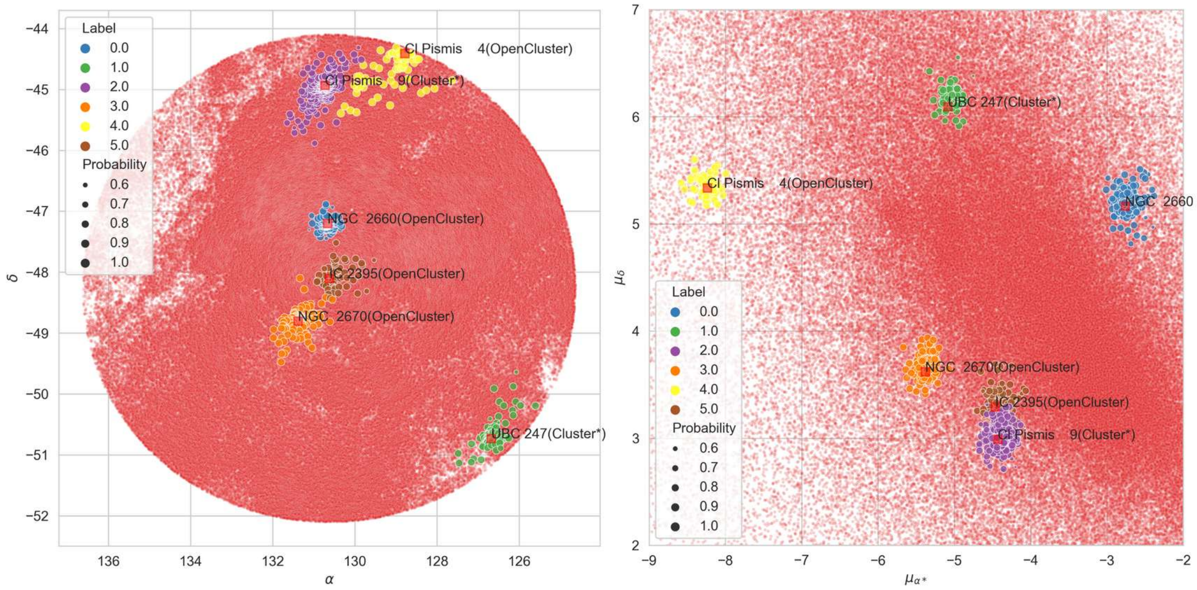

Figure 3: Results of the membership probability calculation for the same dataset shown in Fig. 2. The plot on the left shows celestial coordinates, and the plot on the right shows proper motions.

	PR	*EX*	*F1*	*MCC*	BSL^+	LSR^+
μ	0.988	0.992	0.989	0.990	0.993	0.822
σ	0.016	0.011	0.010	0.009	0.011	0.273

Table 1: Means and standard deviations of each metrics calculated on the results obtained for 155 simulated data sets.

4 Conclusions

Preliminary results indicate that the implemented program performs as expected, achieving an average precision of 0.988 and recall of 0.992 in the classification based on the calculated probabilities. The library is already available for download [22], all its functionalities are documented [23], and the code is accessible through a public repository [24].

As future work, we propose releasing some of SCLUDAM's functionalities as standalone packages to facilitate their reuse in other applications. Additionally, we aim to develop a web-based tool that allows users to access the library's features without requiring Python knowledge.

Acknowledgments

I would like to extend my sincerest thanks to Dr. Jorge F. González and Lic. Antonio D. Alejo from the Institute of Astronomical, Earth, and Space Sciences (ICATE) for their interest and collaboration in the development of this work.

References

1 L. Balaguer-Núñez, M. López del Fresno, E. Solano, D. Galadí-Enríquez, C. Jordi, F. Jimenez-Esteban, E. Masana, J. Carbajo-Hijarrubia, y E. Paunzen. *Clusterix 2.0: a virtual observatory tool to estimate cluster membership probability,* Monthly Notices of the Royal Astronomical Society, 492, (2020), pp. 5811–5843.

2 L. Liu y P. Xiaoying, *A Catalog of Newly Identified Star Clusters in Gaia DR2,* The Astrophisics Journal Supplement Series, (245), 32, (2019).

3 A. Krone-Martins y A. Moitinho, *UPMASK: Unsupervised photometric membership assignment in stellar clusters,* Astronomy and Astrophysics, (561), 10 (2013).

4 G. I. Perren, R. A. Vázquez y A. E. Piatti, *ASteCA: Automated Stellar Cluster Analysis,* Astronomy and Astrophysics, (576), A6, (2015).

5 M. S. Pera, G. I. Perren, A. Moitinho, H. D. Navone y R. A. Vázquez, *pyUPMASK: an improved unsupervised clustering algorithm,* Astronomy and Astrophysics, (650), A109, (2021).

6 Gaia Collaboration, *Gaia Data Release 3,* (2022). Retrieved from https://www.cosmos.esa.int/web/gaia/dr3.

7 F. Arenou, X. Luri, C. Babusiaux, C. Fabricius, A. Helmi, T. Muraveva, A. Robin, F. Spoto, A. Vallenari, T Antoja, T. Cantat-Gaudin, C. Jordi, N. Leclerc, C. Reyle, M. Romero-Gomez, I.-C. Shih, S. Soria, C. Barache y D. Bossini, *Gaia Data Release 2 catalogue validation,* Astronomy and Astrophysics, (616) A17, (2018).

8 GAIA Team, *GAIA EDR3 data model,* (2021). Retrieved from https://gea.esac.esa.int/archive/documentation/GEDR3/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html.

9 S. Schmeja, *Identifying star clusters in a field: A comparison of different algorithms,* Astronomische Nachrichten, 332, (2011) pp. 172-184.

10 A. D. Alejo, J. F. González y S. P. González, *Estudio de membresía de cúmulos estelares utilizando Gaia DR2,* Cuaderno de Resúmenes 62a Reunión Anual Asociación Argentina de Astronomía, Rosario, Provincia de Santa Fe, (2020), pp. 64.

11 A. Adolfssona, M. Ackermana y N. C. Brownsteinb, *To Cluster, or Not to Cluster: An Analysis of Clusterability Methods,* (2018). DOI: 10.1016/j.patcog.2018.10.026

12 B. Hopkins y J.G. Skellam, *A new method of determining the type of distribution of plant individuals,* Annals of Botany, 18, (2), (1954), pp.213-227. Retrieved from https://doi.org/10.1093/oxfordjournals.aob.a083391.

13 A. Kalogeratos y A. Likas, *Dip-means: an incremental clustering method for estimating the number of clusters,* Advances in Neural Information Processing Systems (25), (2012), pp. 2393–2401.

14 B. D. Ripley, *Tests of Randomness for Spatial Point Patterns*. J. R. Statist. Soc. B (41), (1979), pp. 368-374. Retrieved from https://doi.org/10.1111/j.2517-6161.1979.tb01091.x.

15 P. M. Dixon, *Ripley's K Function*. Encyclopedia of Environmetrics (3), John Wiley & Sons, (2002), pp. 1796–1803.

16 L. M. Sampedro y E. J. Alfaro, *Caracterización de Sistemas Estelares en Espacios de N-Dimensiones: Simulaciones y Aplicación al Catálogo Astrométrico UCAC4,* [Tesis doctoral, Universidad de Granada, Granada, España], (2016).

17 L. McInnes, J. Healy y S. Astels, *HDBSCAN: Hierarchical density based clustering,* Journal of Open Source Software, (2), 11, (2017).

18 R.J.G.B. Campello, D. Moulavi y J. Sander, *Density-Based Clustering Based on Hierarchical Density Estimates,* Advances in Knowledge Discovery and Data Mining, PAKDD 2013, Lecture Notes in Computer Science, 7819, (2013). DOI: 10.1007/978-3-642-37456-2_14

19 J. E. Chacón y T. Duong, *Multivariate plug-in bandwidth selection with unconstrained pilot matrices,* Test Journal, (19), (2010), pp. 375-398.

20 X. Luri, A. G. A. Brown, L. M. Sarro, F. Arenou, C. A. L. Bailer-Jones, A. Castro-Ginard, J. De Bruijne, T. Prusti, C. Babusiaux y H. E. Delgado, *Gaia Data Release 2: using Gaia parallaxes,* Astronomy and Astrophysics, 616, A9, (2018). DOI: 10.1051/0004-6361/201832964

21 W. S. Dias, H. Monteiro, A. Moitinho, J. R. D. Lépine, G. Carraro, E. Paunzen, B. Alessi y , L. Villela, *Updated parameters of 1743 open clusters based on Gaia DR2,* Monthly Notices of the Royal Astronomical Society (504), (2021), pp. 356–371. Retrieved from https://doi.org/10.1093/mnras/stab770.

22 S.P. González, *SCLUDAM (Star CLUster Detection And Membership estimation) Python Package,* (2022a). Retrieved from https://pypi.org/project/scludam/.

23 S.P. González, *SCLUDAM (Star CLUster Detection And Membership estimation) documentation,* (2022b). Retrieved from https://simonpedrogonzalez.github.io/scludam-docs/index.html.

24 S.P. González, *SCLUDAM (Star CLUster Detection And Membership estimation),* (2022c). Retrieved from https://github.com/simonpedrogonzalez/scludam.