Clustering Stellar Pairs to Detect Extended Stellar Structures

Sergey Sapozhnikov[1][0000−0002−1295−4464]

Institute of astronomy RAS, Russia, Moscow
thestriks@gmail.com

Abstract. Gaia data allows for search for extended stellar structures in phase (coordinates plus velocities) space. We describe a method of using DBSCAN clustering algorithm, which is used to group closely-packed-together data points, to a list of preliminary selected pairs of stars, with parameters expected to be found within stellar streams and comoving groups: loose structures in which stars are not gravitationally bound, but do share motion and evolutionary properties. To test our approach, we construct a model population of background stars, and use pair-constructing and clustering algorithms on it. Results show that transitioning to a list of pairs sharply reveals structures not presented in background model, which then become more apparent targets in coordinates-velocities phase space for DBSCAN algorithm thanks to now increased relative density of the extended stellar structure.

Keywords: Star clusters · Stellar associations · Data analysis

1 Introduction

Astrometric space mission Gaia of European Space Agency provides data on positions, proper motions, and parallaxes of stars with previously unseen precision and scale (> 1.5 billion sources) [2, 3]. Significant quantitative improvement over previous all-sky star surveys allow for qualitative improvement and opens new possibilities for research. In particular, this data is used for the task of determining the members of stellar clusters. High quality of data also allows to discern extended low-density structures within the stellar background. Such extended structures, comoving groups and stellar streams, form as a result of dissipation of open star clusters or associations under the influence of Galactic tidal disruption. While stars in such structures might be not gravitationally bound anymore, they do share common genesis, which makes them a useful tool in understanding both large-scale Galactic structure and stellar evolution. Methods for locating such structures are still in development and many different approaches by various scientific groups can be seen. No other stellar catalogue currently comes close to Gaia level of depth and completeness of astrometric data, which allows researchers to use methods previously deemed unfeasible. One approach often seen is to pick an already known stellar cluster or group, and search in phase space in Gaia data around it for an extended structures kinematically related to it, for example, [5, 7]. Convergent-point methods, which look for the stars

aiming at the same point at the sky where the proper motions (apparent motions on the sky) of the most stars in cluster do converge suit well for this task. Another approach is applying clustering methods like DBSCAN/HDBSCAN to stellar population in general [4]. Search for clusters in phase space (coordinates + velocities) allow to search for new structures, since this method does not rely on finding the starting point first, and works with overdensities in phase space. Another notable direction of research in Gaia data is an all-sky searches for an ultra-wide binary stars [1, 8].

We aim to combine the binary star search and clustering approaches: we use our algorithm designed for search of ultrawide binaries described in [8] to look not for the binary stars, but for a comoving stellar pairs (stars that are not gravitationally bound like binary stars are, but do share motion properties) and use the DBSCAN clustering algorithm to this preliminary list of stellar pairs (instead of using it on star catalogue directly). Selection of DBSCAN is based on this work [4] highlighting DBSCAN and HDBSCAN as the best algorithms for clustering such data, and we choose DBSCAN over HDBSAN due to possibility to explicitly set expected distance as a clustering parameter (so-called epsilon parameter in DBSCAN). Constructing list of comoving pairs would allow us to constrain the parameters of relative movement of stars before applying the clustering algorithms. By decreasing the relative amount of pairs with unfavorable parameters, we aim to increase the contrast with which stars (and pairs) constituting extended stellar structures appear in the phase space, and hopefully, improve the sensitivity of clustering algorithms thanks to that.

This paper describes this work still in progress by presenting the results of using such approach to a region of 30x30 deg in the sky, for stars between 100 and 1000 pc from the Sun. Paper is divided in two sections: first describes the principles behind creating the preliminary list of pairs, and the second one describes creating a model distribution of stars and applying the clustering algorithm to real and model data.

2 Assembling Pairs Catalogue

Assembling a preliminary pairs catalogue involves picking several parameters of stellar pairs and putting a constraints of them. Search for pairs then is done by filtering lists of all possible pairs in the catalogue to fit the constraints. These constraints are based on what values of these parameters are expected to be met in pairs of stars constituting extended stellar structures.

Most obvious first criteria is stars being spatially close to one another. We limit projected separation between stars to 1 parsec. This limit also allow us to significantly optimise pairs catalogue assembly process by subdividing search region into smaller regions on the celestial sphere, sizes of which is determined by maximum possible coordinate separation of a pair and then searching for pairs just within these sub-regions and in neighbouring ones. This allows to decrease computation complexity from $C * N^2$ to $C * N^2/M$, where N is the number of stars and M is the number of regions (M = 1537 in reviewed case, which

meant a transition from impractically long to quite fast computation time). This optimisation is reviewed in more detail in article [8].

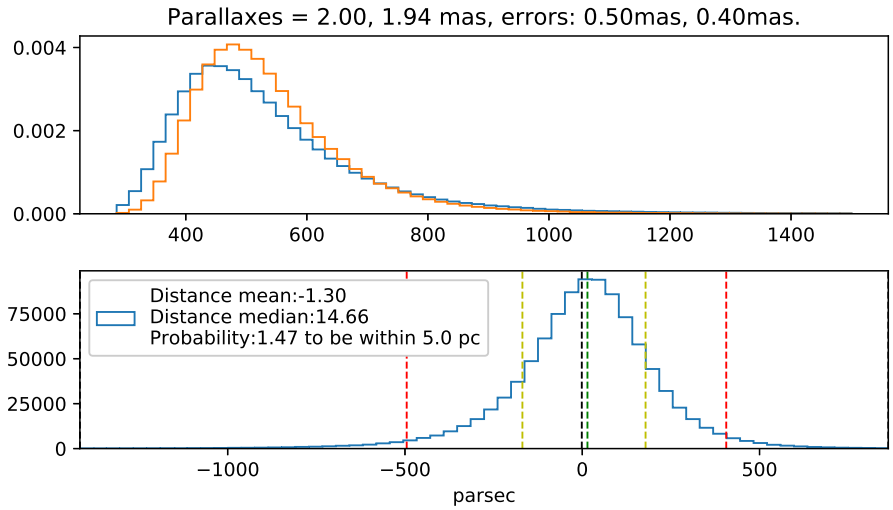

Fig. 1: Illustrating the modelling of distance probability distribution. Above is PDF for distance to Sun for two stars at 500 and 515 pc. Below is probability distribution for possible values of their distance from one another along the line-of-sight. Note that median and mean values of this distribution might differ from difference of "naively" $(1/\pi)$ defined distances.

In regards to data on stellar positions there is a significant difference between treating errors of position on stellar sphere (right ascension and declination) and radial distance (parallax). Errors of parallax are larger, also, the way distance to the star is determined as inverse of parallax, coupled with the fact that absolute error of parallax is not directly related to value of parallax, means that:

1. relative errors of parallax become larger for distant stars
2. symmetric errors of parallax values lead to asymmetric errors in distance determination.

For Gaia sample we use in this article (distances of 100 - 1000 parsec from the Sun) mean error of parallax is 0.58 mas (milliarcseconds), which corresponds to uncertainty of -370/+1380 parsec for star with actual distance of 1000 pc. Uncertainty of position on celestial sphere is negligible compared to distance uncertainties.

Accounting for radial distance between stars thus cannot be done by simply placing a constraint on apparent measured distances difference. We use two approaches simultaneously. First approach is imposing limits on parameter labelled

"parallax consistency":

$$\pi_{cons} = 3(\sigma_{\pi_1} + \sigma_{\pi_2}) - |(\pi_1 - \pi_2)| \quad (1)$$

Here π_1, π_2 are parallaxes of components, and $\sigma_{\pi_1}, \sigma_{\pi_2}$ are their errors. Negative values indicate that parallaxes are too different for stars to be close even if we allow overlap within 3 standard deviations for both components. Highly positive values indicate that errors are too large to tell with certainty that stars are close together. Values that are positive, but close to zero indicate either that we can say that stars are close together with relative certainty, or that the sum of errors happened to be close to parallax difference (even if it is large).

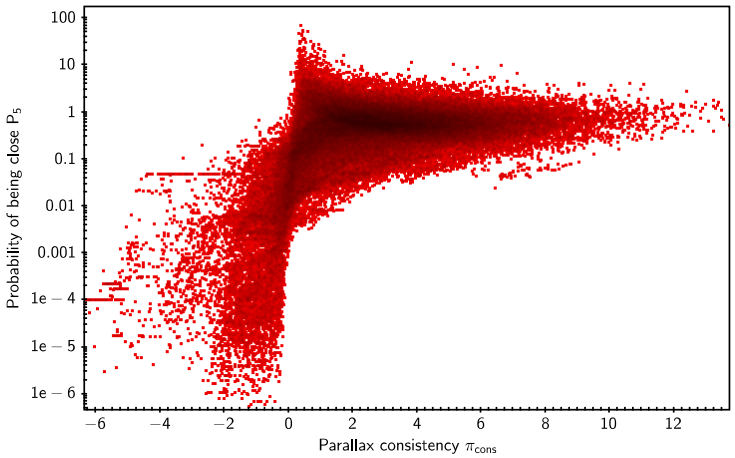

Fig. 2: Scatterplot illustrating relation between consistency of parallaxes within errors and modelled probability for pair components to be within 5pc of one another along line of sight (in percent). Scale for probability is logarithmic. This scatterplot shows data for small subsample (180000 pairs) of an original sample of pairs.

To differentiate between these two scenarios, we use a second approach: a model to simulate probability density distribution of radial distance between stars by modelling probability density of distance to Sun for both stars, based on their given parallaxes and their errors. We then integrate over limit (-5pc, +5pc) for each pair to determine probability of them being within 5 pc of one another along line-of-sight (P_5) (See Fig. 1). We do not model this directly for each actual pair, instead values for P_5 in pairs are interpolated from grid of 26000 pre-computed points in 4-parameter $(\pi_1, \triangle_\pi, \sigma_{\pi_1}, \sigma_{\pi_2})$ space covering all the possible values in our sample densely enough.

These two metrics do somewhat correlate (See Fig. 2): most of the pairs with negative parallax consistencies have $P_5 < 0.2\%$, and high P_5 is mostly

associated with positive but not very high π_{cons}. We also noted an increased density of pairs with $0 < \pi_{cons} < 1$ in region with proper motion difference of stars $< 2mas/yr$ and projected relative motion of $< 3km/s$. Resulting constrains on pairs parameters used are:

- projected separation $<$ 1 pc
- projected relative motion < 3 km/s
- proper motion difference < 6 mas/yr
- parallax consistency within $-0.1 < \pi_{cons} < 1$ and $P_5 > 0.5\%$.

When deciding on constraints on pairs properties, we guide ourselves with theoretical and model considerations from [6]. Selected constraints are what we roughly expect from stellar pair of neighbours in the same extended stellar structure. Radial velocities are not considered since they are present only for the small fraction of all Gaia EDR3 sources. Applying this constrains leaves us with list of 104 000 pairs, distributed very non-uniformly at the celestial sphere.

3 Stellar Background Model and Clustering

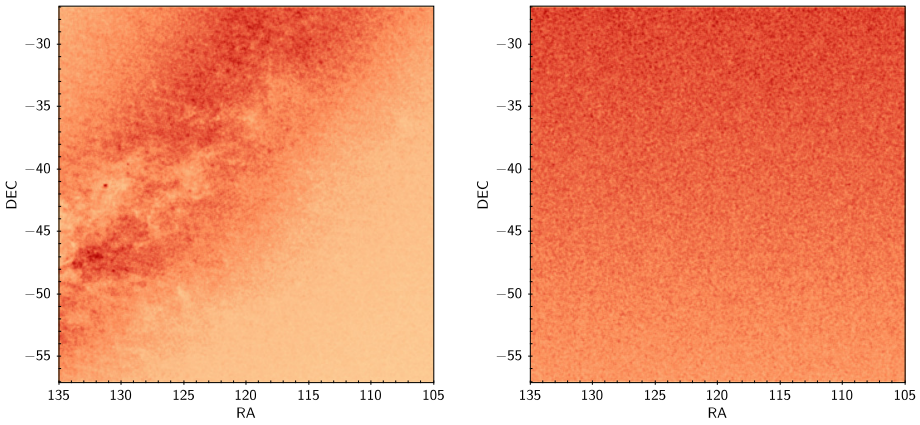

Fig. 3: Scatterplots showing the sky positions of stars from real (left) and background model (right) data. More densely populated areas are closer to orange color.

Aim of the stellar background model is to produce "mock" catalogue of stars which would be known to not contain any clusters, streams, or other stellar structures. This catalogue then undergoes the same process of constructing preliminary pairs and their clustering as a real catalogue, to check our algorithm for not treating random overdensities in field stars distribution as an actual extended stellar system.

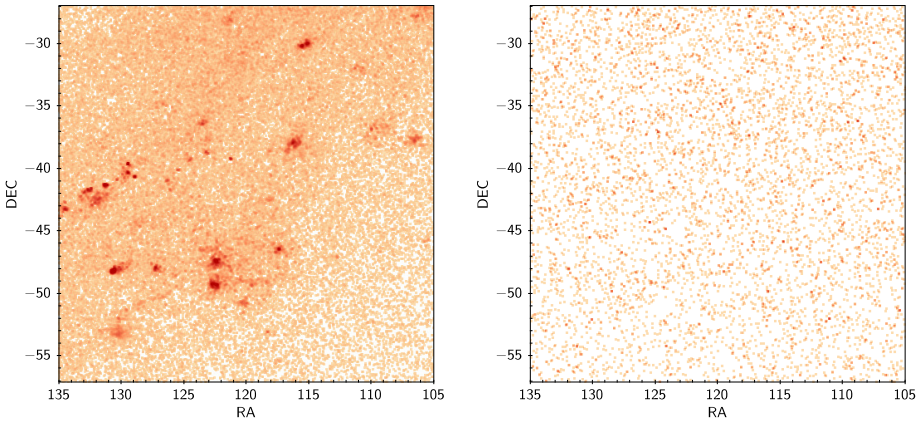

Fig. 4: Scatterplots showing the sky positions of pairs from real (left) and background model (right) data. More densely populated areas are closer to orange color. Generated pairs are notably more scarce and uniformly distributed.

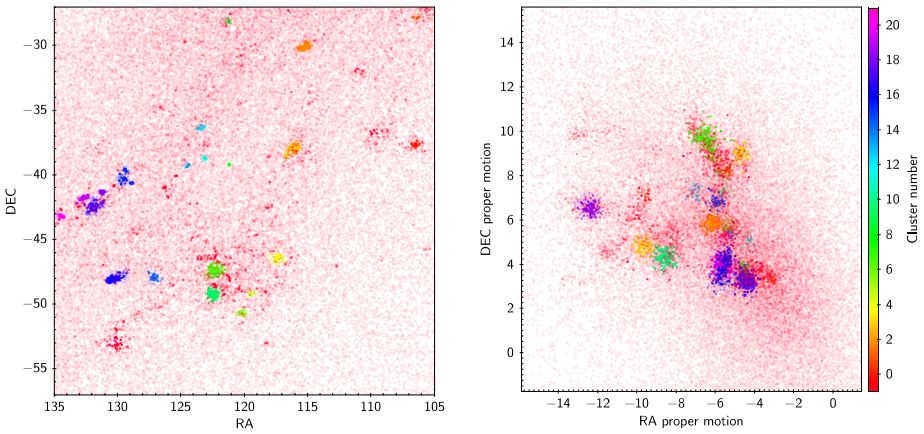

Fig. 5: Scatterplots showing the sky positions (left) and proper motions (right) of pairs from clustered real data, colored in according to which cluster they were allocated to. Light red dots are pairs which are considered to be part of background (cluster number = -1).

Model we use is the following: stars are first generated uniformly within same region of sky as the real sample, but in a broader range of distances: 50 - 2000 pc instead of 100 - 1000. To simulate the effect of parallax errors on measuring actual parallaxes, each star is assigned a "real" parallax in mas (1000./distance) and a parallax measurement error. Distribution of parallax errors was made to mimic that of the real sample (a combination of two Gaussian distributions and an exponential distribution was found to be representing the real distribution the best). "Measured" star parallaxes are obtained by modifying their "real" parallaxes by a random number drawn from normal distribution with the standard deviation equal to parallax error. After that the dataset is cropped to 100 - 1000 pc of "measured" distance. Stars get a random galactocentric velocity components (U, V, W) with dispersions of 35, 30, and 20 km/s respectively to simulate actual movement of stars in solar neighbourhood [9]. Density of stars in the model sample is adjusted to match the average density of the real EDR3 data (4,5 million stars in volume in question). Applying the algorithm to find stellar pairs to this model with the same parameters and constrains results in a sample of 10000 model pairs. Comparison of stellar density distribution on the sky for real and model star samples can be seen at Fig. 3, and such a distribution for pairs at Fig. 4.

We apply a DBSCAN clustering algorithm to both real and model data. Clustering is done in 4-coordinate space (2 for positions on the sky and 2 proper motions). Radial distance is not considered due to parallax errors stretching any stellar structure in line-of-sight direction too much, restrictions on π_{cons} and P_5 on pair composition phase are considered to be sufficient to take distances into account. Varying the DBSCAN clustering parameters shows that for majority of parameter values where real data is subdivided into reasonable amount of clusters, model data is all marked as background. Fig. 5 shows clustering result for possible pair of parameters, subdividing the data into 21 clusters and "background" pairs.

4 Summary and Conclusions

We used Gaia EDR3 data for constructing a stellar pairs catalogue with properties for pairs expected to be found within clusters and extended stellar structures. With selected parameters, pairs distribution display structures of high density which are easily seen against the 'background" pairs. Comparison with using the same approach to stellar background model of similar density shows that this dense structures cannot be attributed to random overdensities of uniform, featureless distribution of stars. High contrast of such areas compared to background eases the clustering task for algorithms of unsupervised learning, like DBSCAN which we use in this article.

Further work will include comparison of these structures with known extended stellar structures and clusters, tuning parameters of pairing and clustering to better recognize stellar structures, using other means for determining the existence of evolutinary relationship between found structures and comparison

with models of stellar systems evolution to explain the observed stellar system qualities.

Acknowledgements. This work is supervised by Dr. Dana Kovaleva, Department of Physics of stellar systems, Institute of Astronomy of Russian Academy of Sciences. The work has made use of data from the European Space Agency (ESA) mission Gaia processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. The study was partially supported by Russian Foundation for Basic Research (Project no. 20-52-12009).

References

1. El-Badry, K., Rix, H.W., Tian, H., Duchêne, G., Moe, M.: Discovery of an equal-mass 'twin' binary population reaching 1000 + au separations. Mon. Not. R. Astron. Soc**489**(4), 5822–5857 (Nov 2019). https://doi.org/10.1093/mnras/stz2480
2. Gaia Collaboration: The Gaia mission. Astron. Astrophys**595**, A1 (Nov 2016). https://doi.org/10.1051/0004-6361/201629272
3. Gaia Collaboration: Gaia Early Data Release 3. Summary of the contents and survey properties. Astron. Astrophys**649**, A1 (May 2021). https://doi.org/10.1051/0004-6361/202039657
4. Hunt, E.L., Reffert, S.: Improving the open cluster census. I. Comparison of clustering algorithms applied to Gaia DR2 data. Astron. Astrophys**646**, A104 (Feb 2021). https://doi.org/10.1051/0004-6361/202039341
5. Jerabkova, T., Boffin, H.M.J., Beccari, G., de Marchi, G., de Bruijne, J.H.J., Prusti, T.: The 800 pc long tidal tails of the Hyades star cluster. Possible discovery of candidate epicyclic overdensities from an open star cluster. Astron. Astrophys**647**, A137 (Mar 2021). https://doi.org/10.1051/0004-6361/202039949
6. Kamdar, H., Conroy, C., Ting, Y.S., Bonaca, A., Smith, M.C., Brown, A.G.A.: Stars that Move Together Were Born Together. Astrophys. J. Lett**884**(2), L42 (Oct 2019). https://doi.org/10.3847/2041-8213/ab4997
7. Röser, S., Schilbach, E.: Praesepe (NGC 2632) and its tidal tails. Astron. Astrophys**627**, A4 (Jul 2019). https://doi.org/10.1051/0004-6361/201935502
8. Sapozhnikov, S.A., Kovaleva, D.A., Malkov, O.Y., Sytov, A.Y.: Binary Star Population with Common Proper Motion in Gaia DR2. Astronomy Reports **64**(9), 756–768 (Sep 2020). https://doi.org/10.1134/S1063772920100078
9. Schönrich, R., Binney, J., Dehnen, W.: Local kinematics and the local standard of rest. Mon. Not. R. Astron. Soc**403**(4), 1829–1833 (Apr 2010). https://doi.org/10.1111/j.1365-2966.2010.16253.x