IOM-clustering, the short version

Sofie Lövdal

May 24, 2021

1) Quality cuts: `ruwe` > 1.4, `parallax_over_error` > 5, `vlos_error` < 20 km/s. We consider halo as $v_{toomre} > 210$ km/s.

2) E, $L_\perp$, L_z and η (circularity) are used as clustering features. These are scaled to an equal range, implying that they will be considered equally important in a distance-based clustering algorithm. Circularity is defined as

$$\eta = L_z / L_z^{max}(E) \tag{1}$$

where $L_z^{max}(E)$ is the largest possible L_z for a given energy (corresponding to a circular orbit). As it is derived based on L_z and correlated with this quantity, there will be an implicit double emphasis on L_z in the clustering. This is intentional as we expect a cluster to have a smaller spread in L_z than in E and $L_\perp$.

3) The single linkage algorithm is applied on our data set. For each step of the algorithm, it connects the two groups with the smallest distance between each other. The distance between two groups in this case is the smallest (Euclidean) distance between a data point in group one and a data point in group two (see Figure 1). Each data point is considered a singleton group initially. Each merge corresponds to a new connected component in the data set. The algorithm is also closely related to graph theory, as the result after the last merge is equivalent to the minimum spanning tree. Correspondingly, the series of connected components obtained also corresponds to the set of every potential cluster in the data, under the assumption that the most likely clusters are the groups of data points with the smallest distance between each other, without assuming any specific cluster shape.

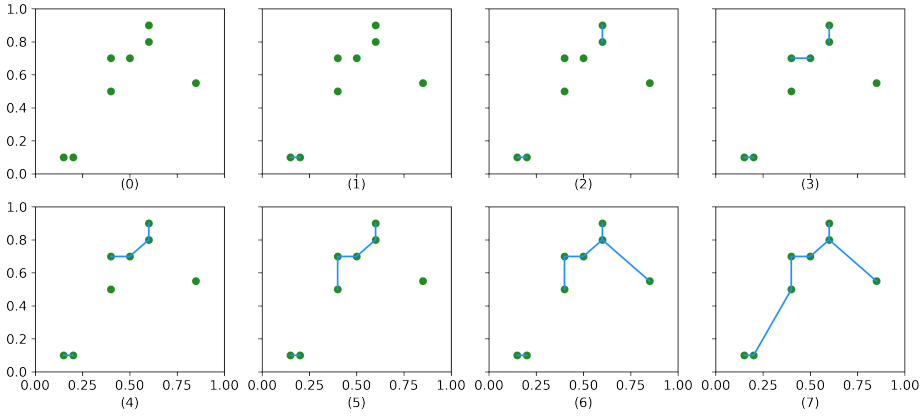

Figure 1: The single linkage algorithm applied on a two-dimensional data set. Each step of the algorithm, labelled above by the number under the corresponding panel, connects two components into a group.

4) Next, we can check the validity of every potential cluster obtained by the merging process of the single linkage algorithm, corresponding to every new connected component in the panels of Figure 1. This is done by comparing the number of stars in a potential cluster towards the number of stars in the same region in an artificially generated smooth halo. The artificial halo is obtained by scrambling the velocity components of the existing data and recomputing the clustering features. We generate 100 artificial halos to get an average representation.

In order to compare the observed versus expected number of stars in a region, we define a 4D-ellipsoid around a potential cluster that we want to investigate by applying PCA on the members. The lengths of the ellipsoid axes are chosen to be two standard deviations of spread along each axis. We then assess the statistical significance of each potential cluster as:

$$N_{C_i} - N_{C_i}^{art} > 3\sigma_i \tag{2}$$

corresponding to

$$\frac{N_{C_i} - N_{C_i}^{art}}{\sqrt{N_{C_i} + (\sigma_{C_i}^{art})^2}} > 3 \tag{3}$$

where N_{C_i} is the number of real stars that fall within the ellipsoidal cluster boundary, $N_{C_i}^{art}$ is the number of stars from the artificial halo that fall within the boundary, and $\sigma_{C_i}^{art}$ is the standard deviation of the latter quantity across our 100 realizations of the artificial halo.

We now have a set of statistically significant connected components in the data. In many cases these significant structures are iteratively larger subsets of the parent cluster, representing the core of a cluster growing out to its full extent in the merging hierarchy. In this case, when we produce the final catalogue, we always choose the largest significant structure and assign a single label to the stars of this cluster.

5) The above procedure gives 76 significant clusters for the extended sample within 2.5 kpc.

6) In order to identify potential new members of a structure we also give an indication of membership probability, $S(D^2) = s$, where S is the complement (or survival function) of the cumulative chi-square distribution with four degrees of freedom and D is a star's Mahalanobis distance.

Mahalanobis distance can loosely be described as distance between a data point and a distribution in terms of standard deviations along an axis intersecting the data point and the mean of the distribution. The theoretical distribution of squared Mahalanobis distances of a n-dimensional Gaussian is known: it is the chi-square distribution with n degrees of freedom. We can then use the Mahalanobis distance of each star to estimate whether it is a good fit to the distribution or not. Specifically, a constant Mahalanobis distance defines a specific confidence ellipse around the distribution and we measure how much of the distribution has a better fit than a specific star by consulting the complementary cumulative chi-square distribution.

For example, if theoretically 40% of the cluster distribution has a smaller Mahalanobis distance than some particular star, this star will get $s = 0.6$. In Figure 2 we visualize an example in 2D: Say that the cluster from merge number five of Figure 1 has been extracted as a significant cluster. In the figure we have plotted a confidence ellipse with axis lengths corresponding to two standard deviations of spread along each axis. In 2D this area covers 86% of a Gaussian distribution, so a star that falls exactly on the edge of the confidence ellipse will in this case get $s = 0.14$. When looking for potential new members we will be residing in the outskirts of the Gaussian tails, so will be using looking at small values for s.

Figure 2: A two sigma confidence ellipse drawn around the cluster extracted at merge (5) of the single linkage algorithm in Figure 1.