STAT 3019 ICA1 Part2

Student ID: 14006787

2018

## Part2:

**Q1.**

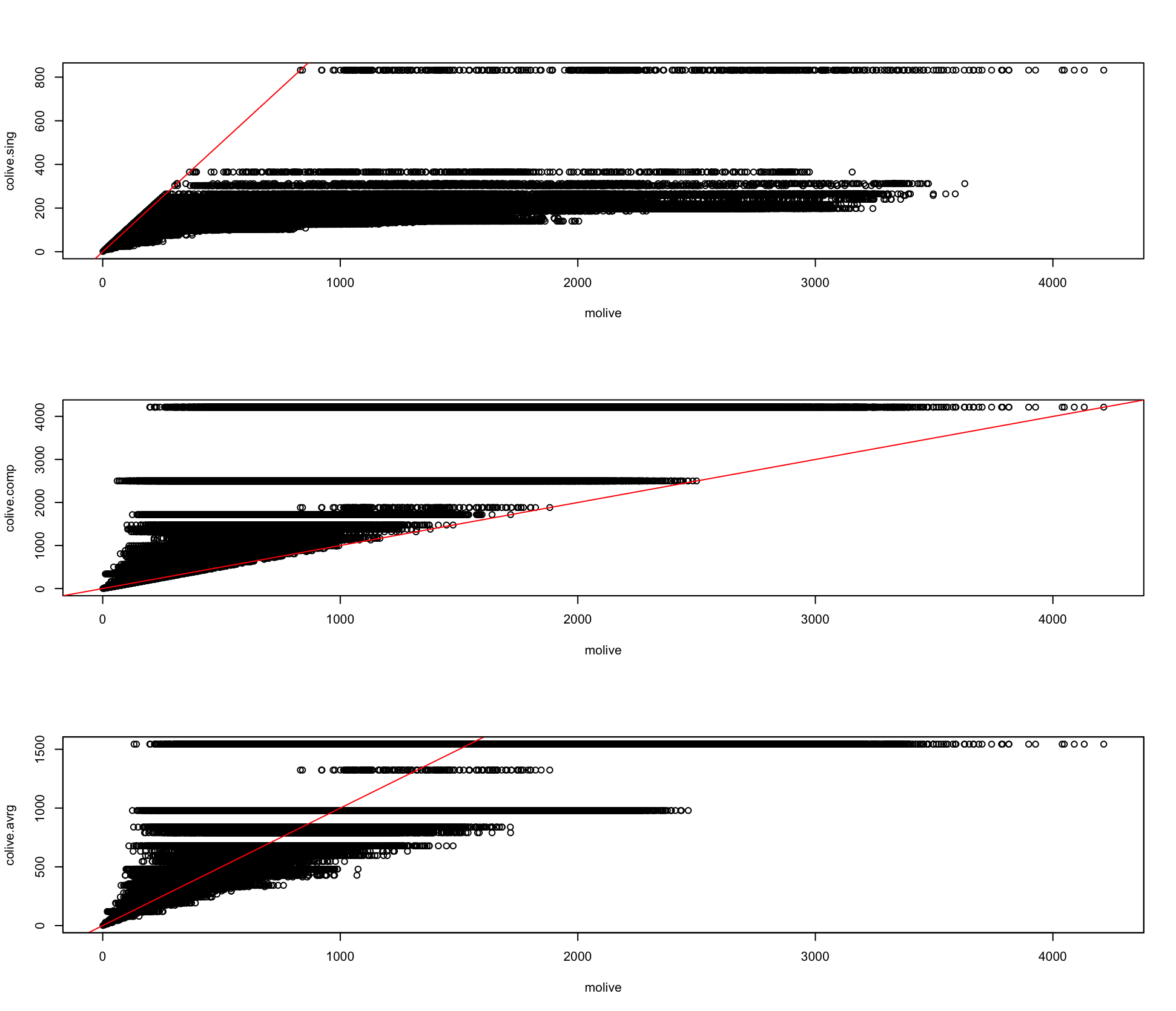
**(d)** The ranking for the Cophenetic correlations of the three hierarchical clusterings (Single Linkage, Complete Linkage, Average Linkage) is:

Average Linkage (0.739) > Complete Linkage (0.730) >> Single Linkage (0.458)

The Average Linkage clusterings gives the best result, following complete linkage which is only 0.009 lower than average.

Below (Figure 1) is the plot comparing the original (Manhattan) distances with the Cophenetic distances for single, complete and average linkages respectively, which the points will be closer to the red line (1,1) if the clustering measures have better captured the original distances.

Figure 1



The above results show that the cophenetic correlation values of average linkage and complete linkage is good, which indicates that the clusterings are quite fit, and single linkage is worse but not “too bad”.

However, the incorrect choice of distance can significantly affect the clustering result and may lead to a bad decision. Scaled data might show different result and so does using other distances measures.

More discussions are below combined with question (e).

**(e)** The ranking for the Adjusted Rand Index comparing the three linkages with the real region given in the dataset (macro area) is:

Average Linkage (0.365) > Complete Linkage (0.2856) >> Single Linkage (-0.0047)

The result of Single Linkage is so close to zero that does not indicate any similarity to the real regions. We could even hypothesize that these two clusterings are independently random.

The overall patterns of the ranking are consistent with the Cophenetic correlations one, which the Average Linkage clusterings gives the best result at all measures. And both Cophenetic correlations and ARI indicate that results for average linkage and complete linkage are more similar, and are much better than Single Linkage.

However, the three linkages shown for ARI all shows that these are not good fit of the data, which are all somewhat close to randomness (ARI=0), or say different to the real regions. And this obeys the result that we have gained for question (d). This may be due to the fact that we used Manhattan distance based on unstandardized variables as d, which could be a bad decision. So, we may conclude that, Cophenetic correlation itself is indeed a good index for comparing the goodness of fit of different clusterings, and for indicating the similarity of distances to observations on the same cluster. However, the choice of distance measures is also crucial for the results.

**Q2.**

**(a)**

Table 1: the best k with maximum asw for the 20 datasets for the 5 methods.

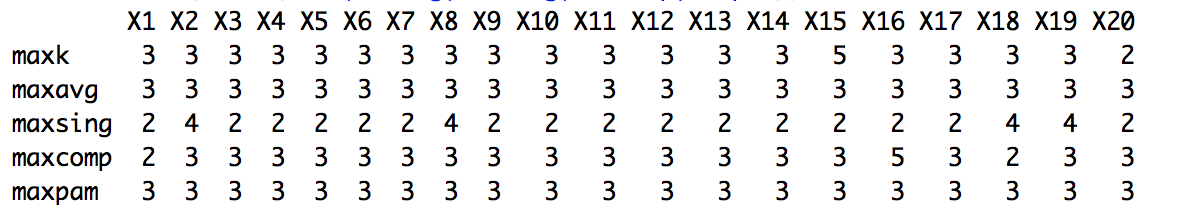


Table 2: the ARI value comparing the true clustering for the 20 datasets for the 5 methods.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methods** | **Kmeans** | **Single Linkage** | **Complete Linkage** | **Average Linkage** | **PAM** |
| **Frequency** | 18 | 0 | 17 | 20 | 20 |

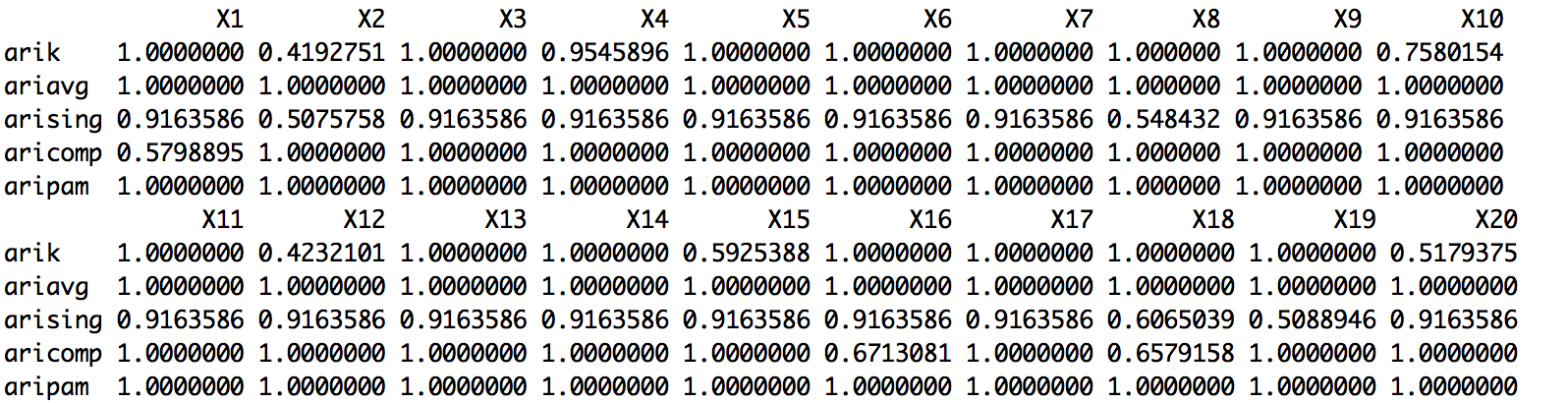


Table 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mean of the ARI for 20 datasets** | 0.8832783 | 0.8416572 | 0.9454557 | 1 | 1 |

Table 1 illustrates the best k with maximum ASW for the 20 simulated datasets. We can see that Partitioning Around Medoids and Average Linkage perform stably and perfectly with all the 20 datasets, whereas Single Linkage varies the most and has no result giving of K=3. Table 2 shows us the ARI value of these data for the 5 methods comparing the true clustering, and table 3 averages them. The patterns remain the same, PAM and Average Linkage are the most stable two methods, whereas there is no dataset clustered using Single Linkage that could give a ARI being 1.00.

Figure 1 and 2 are the sample plots for X20 of average silhouette width for K from 2 to 10 and the visualization of dataset coloured by the best clustering suggested for the different methods respectively. Comparing the true clustering of how the data was originally simulated in Figure 3, and astonishingly, there is no clustering method that shows exactly the same as the true data do, while there are lots of ARI=1.

It is also interesting to note that, although K-means gives lots of 3, the ARIs of some datasets that has predicted K=3 is fairly low. For example, dataset X2 has K=3 for Kmeans however the corresponding ARI is 0.423. Conversely, despite the fact that there is no correct prediction of K=3 using Single Linkage, the ARI for most of the datasets are quite high. And some even higher than K-means when K-means models K=3. And the frequency of Complete Linkage to give arise of K=3 is 17, but the ARI still works quite good. I assume the interesting results might be due to, first of all, the inappropriate choice of Manhattan distance when calculating the silhouette width, especially for assessing K-means, which Euclidean distance might be a better choice.

Figure 1:

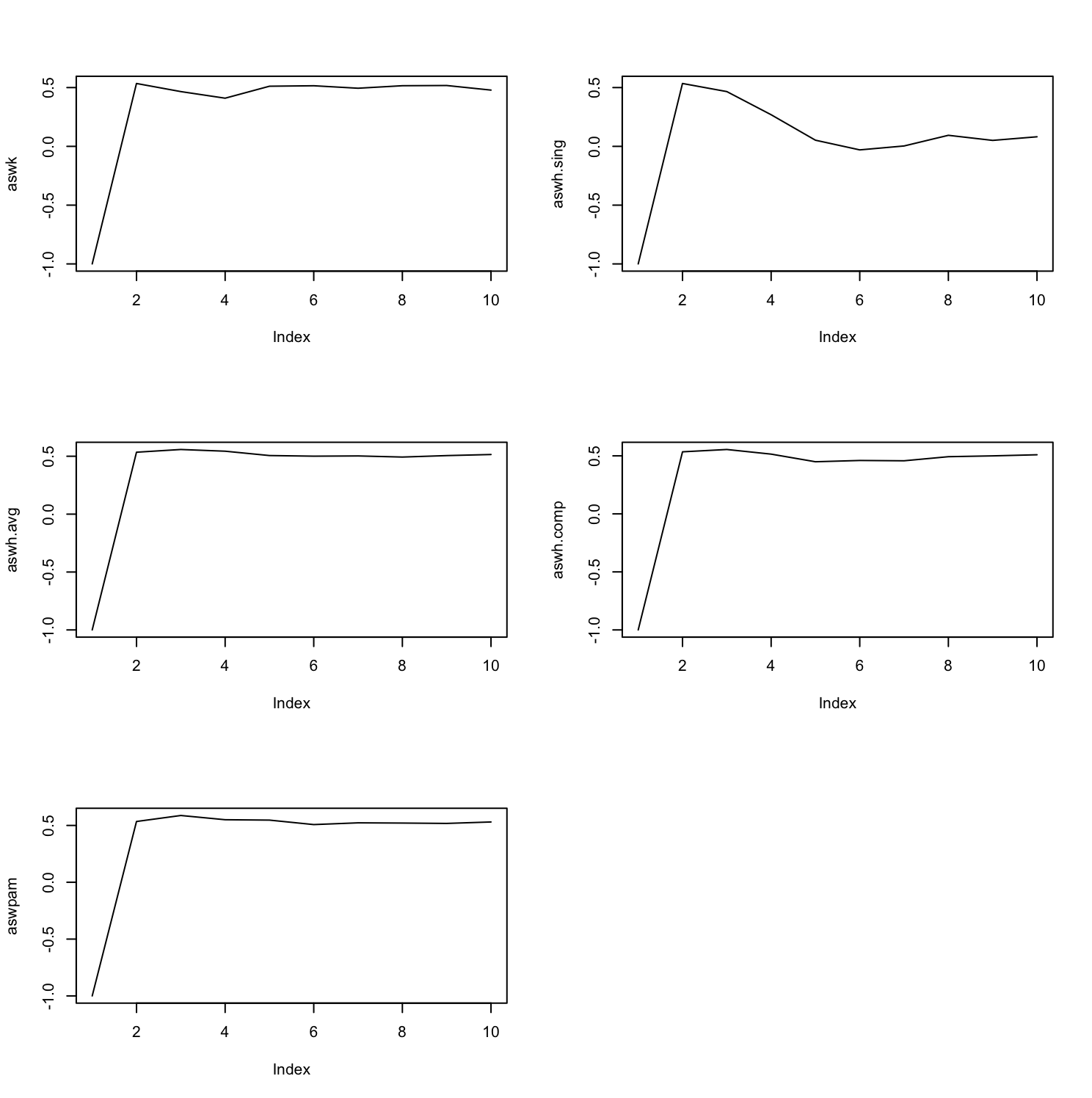


Figure 2:

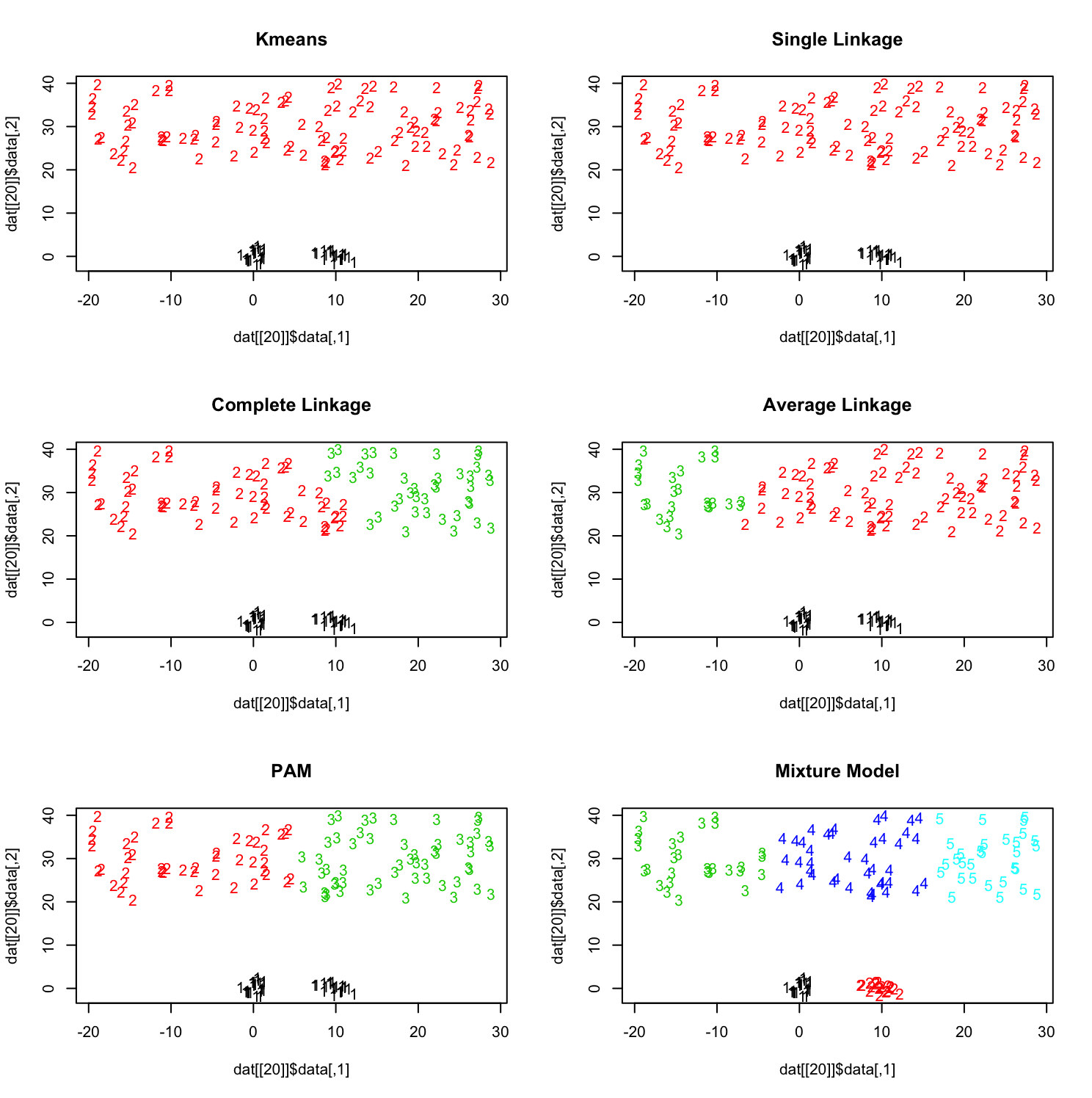
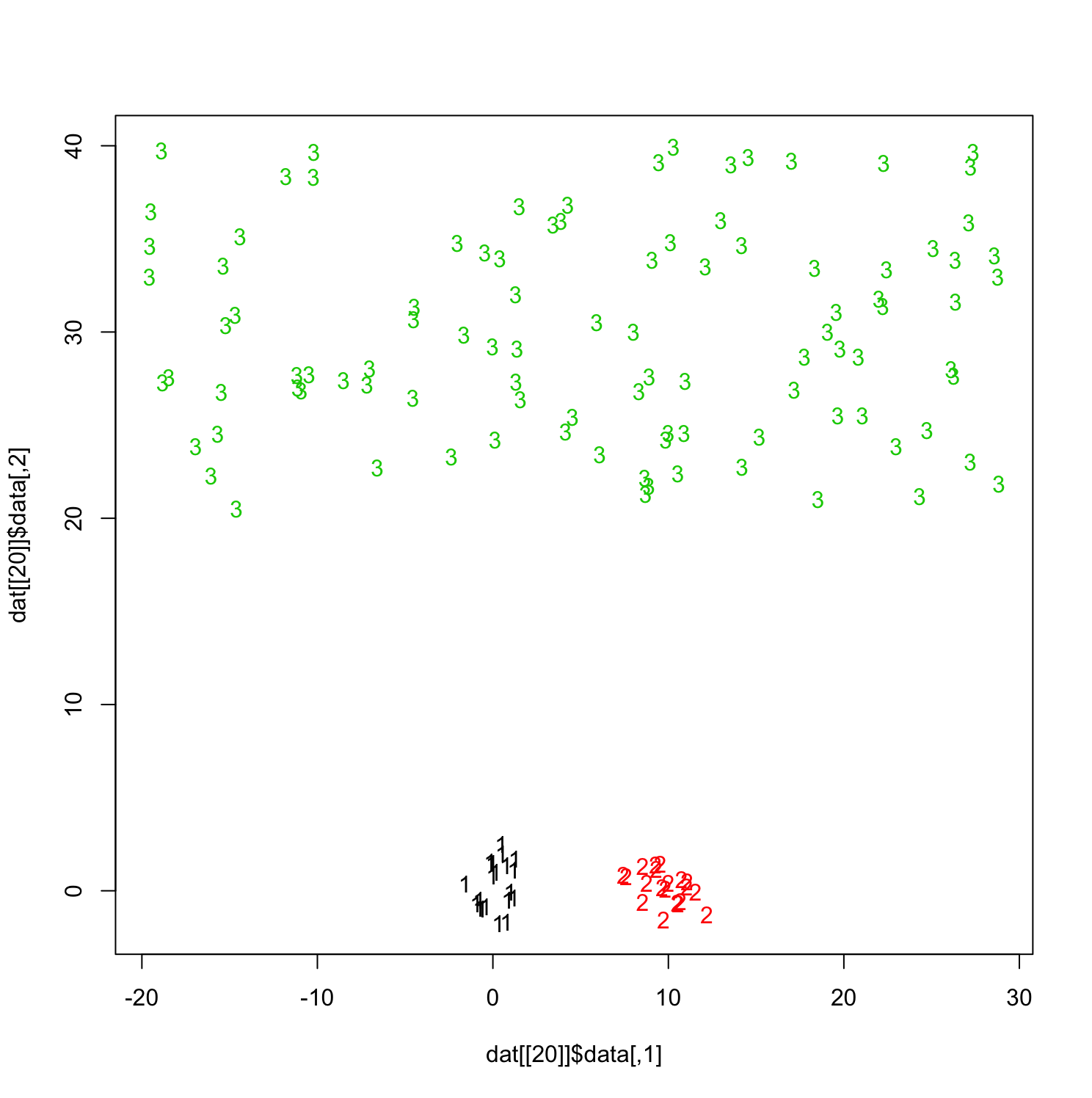


Figure 3: true clustering



**(b)**

The following pages are the outputs of the Gaussian mixtures clusterings. Figure 2 gives a visualization of the best clustering estimated by Mclust for dataset 20.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Data*** | ***X1*** | ***X2*** | ***X3*** | ***X4*** | ***X5*** | ***X6*** | ***X7*** | ***X8*** | ***X9*** | ***X10*** |
| ***K*** | 5 | 4 | 5 | 5 | 4 | 5 | 6 | 5 | 4 | 5 |
| ***Model*** | *VII* | *VEI* | *VII* | *VII* | *VVE* | *VII* | *VII* | *VII* | *VII* | *VII* |
| ***ARI*** | *0.281* | *0.310* | *0.263* | *0.262* | *0.546* | *0.330* | *0.712* | *0.520* | *0.310* | *0.279* |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Data*** | ***X11*** | ***X12*** | ***X13*** | ***X14*** | ***X15*** | ***X16*** | ***X17*** | ***X18*** | ***X19*** | ***X20*** |
| ***K*** | 4 | 5 | 4 | 4 | 4 | 4 | 5 | 5 | 4 | 5 |
| ***Model*** | *VEI* | *VII* | *VII* | *VII* | *VII* | *VEI* | *VII* | *VII* | *VII* | *VII* |
| ***ARI*** | *0.329* | *0.280* | *0.339* | *0.325* | *0.315* | *0.325* | *0.543* | *0.277* | *0.319* | *0.276* |

Where VII stands for spherical, unequal volume. VEI stands for diagonal, varying volume, equal shape. And VVE stands for ellipsoidal, equal orientation. We can see that none of the Mixture model clustering for the 20 datasets gives K=3, and model VII appears the most frequently. Most of the studies gives K=4 or 5.

The values of ARI are somewhat lower than almost all of the clustering methods conducted in (a). However, as shown in Figure 2 that the clusters themselves are fairly spherical.