Task 2

The objective of this task was to investigate to what extent we can recover the true class labels using unsupervised clustering techniques with 4 clusters. This will be done using the Adjusted Rand Index (ARI). First, the data was loaded then centered and the true labels were extracted.

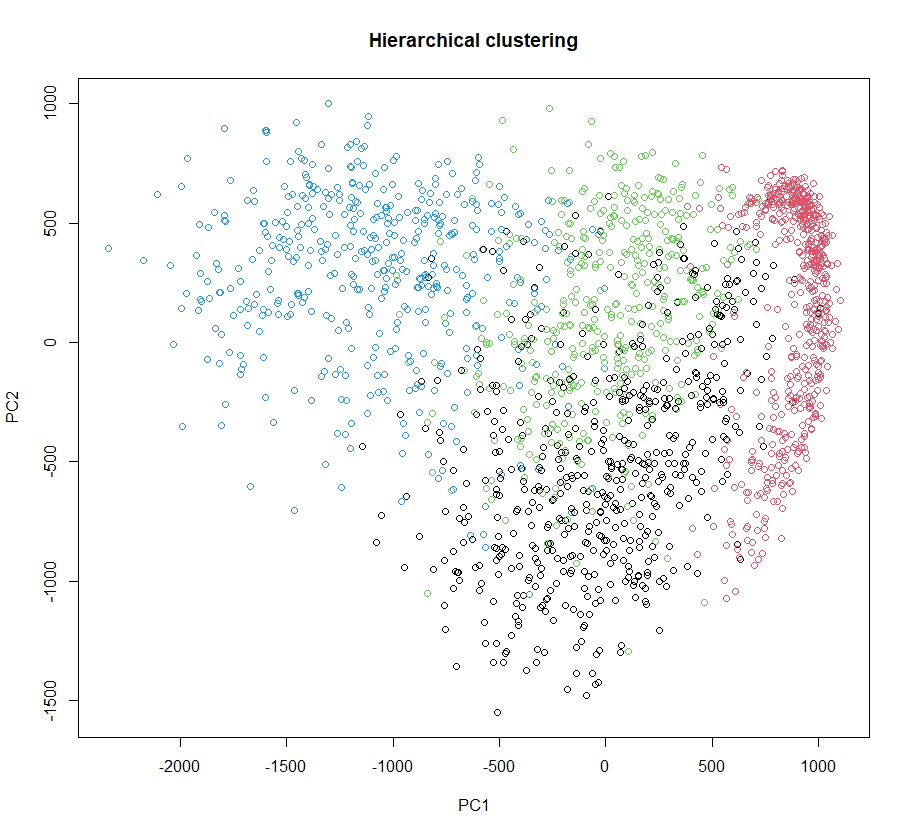
The asked analyses were executed:

* Hierarchical clustering on squared Euclidean distances using the method of Ward
* K-means clustering
* HDDC with model AkjBkQkD and initialized with the results of hierarchical and K-means clustering.
* HDDC with model AkjBQkD and initialized with the results of hierarchical and K-means clustering.

The ARI was calculated for each method to evaluate performance and the results are shown in this table:

|  |  |
| --- | --- |
| *Method* | *ARI* |
| Hierarchical | 0.9076422 |
| K-means | 0.6843655 |
| HDDC AkjBkQkD Hierarchical | 0.7778358 |
| HDDC AkjBQkD Hierarchical | 0.7632849 |
| HDDC AkjBkQkD K-means | 0.7778358 |
| HDDC AkjBQkD K-means | 0.7785905 |

With a value of 0.91, the hierarchical clustering yielded an excellent recovery. All the other models have a moderate recovery. Using the first two principal components we can visualize the observed and predicted class labels with the hierarchical model.



The results indicate that hierarchical clustering outperforms the other methods in recovering the true class labels. This could be attributed to the dataset's nature and the images' inherent structure. K-means clustering showed lower performance, possibly due to its sensitivity to the initial choice of centroids and its tendency to find spherical clusters. The HDDC models, while not outperforming hierarchical clustering, still showed reasonable performance.