**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 hierarchical and K-means clustering results.
* HDDC with model AkjBQkD and initialized with hierarchical and K-means clustering results.

For the Hierarchical clustering, we first calculated the Euclidean distances and use the Ward.D2 to create the clusters. After this, we selected 4 clusters

> #Hierarchical clustering on squared Euclidean distances using the method of Ward

> dist\_data <- dist(data, method = "euclidean",diag = TRUE, upper = TRUE)

> hier\_clust <- hclust(dist\_data,"ward.D2")

>

> #Cutting the tree to 4 clusters

> clust\_hier\_4 <- cutree(hier\_clust, k = 4)

For K means clustering we used 500 different starting points with a max iteration of 2000 steps. Again 4 clusters centroids were used:

> #K-means clustering

> kmeans\_result <-kmeans(data,4,nstart=500,iter.max=2000)

For the HDDC, 4 different models were created using the following code:

> #HDDC clustering AkjBkQkD with hierarchical clustering

> hddc\_AkjBkQkD\_hier <- hddc(data, K = 4, model = "AkjBkQkD",d\_select = "Cattell" ,init.vector = clust\_hier\_4, threshold = 0.05)

>

> #HDDC clustering AkjBkQkD with kmeans

> hddc\_AkjBkQkD\_means <- hddc(data, K = 4, model = "AkjBkQkD",d\_select = "Cattell" ,init.vector = kmeans\_result, threshold = 0.05)

>

> #HDDC clustering AkjBQkD with hierarchical clustering

> hddc\_AkjBQkD\_hier<-hddc(data, K = 4, model = "AkjBQkD",d\_select = "Cattell" ,init.vector = clust\_hier\_4, threshold = 0.05)

>

> #HDDC clustering AkjBkQkD with kmeans

> hddc\_AkjBQkD\_means<-hddc(data, K = 4, model = "AkjBQkD",d\_select = "Cattell" ,init.vector = kmeans\_result, threshold = 0.05)

After creating all the models, 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.

> # Visualizing the higest ARI with 2 PCA on centerd data

> prcomp<- prcomp(data, center = TRUE, scale. = FALSE)

> pc2\_data <- prcomp$x[, 1:2]

> plot(pc2\_data, col = clust\_hier\_4, main = "Hierarchical clustering", xlab = "PC1", ylab = "PC2")

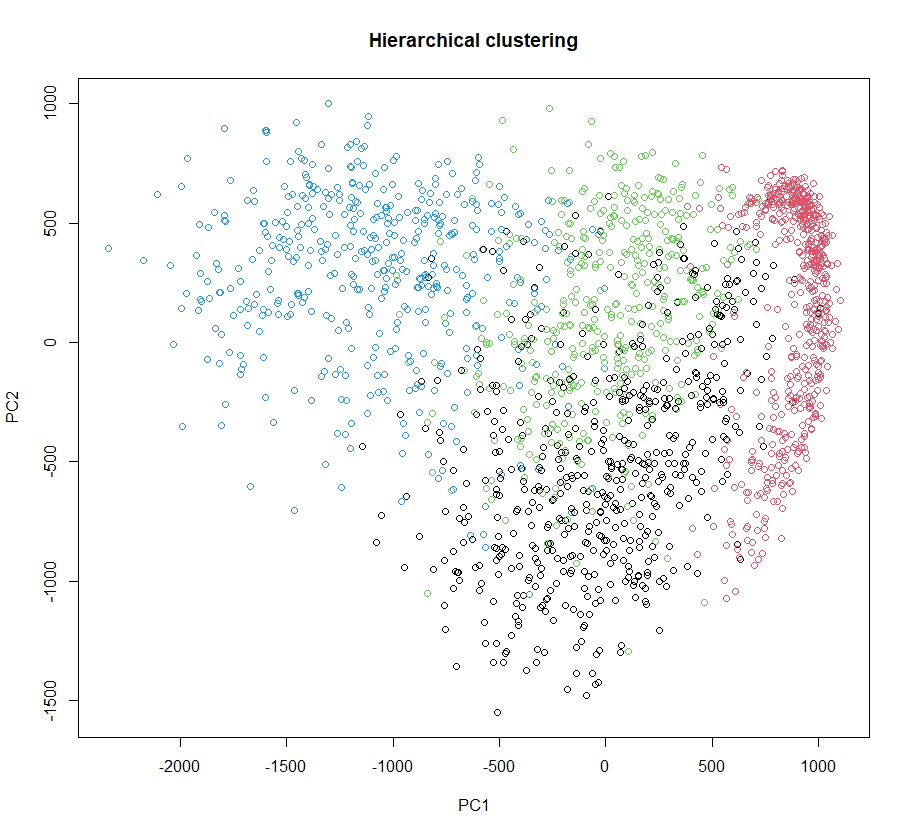
The results indicate that hierarchical clustering outperforms the other methods in recovering the actual class labels. This could be attributed to the dataset's nature and the images' inherent structure. K-means clustering showed lower performance, possibly due to its sensitivity to the initial choice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

Figure 2. Hierarchical clustering