STAT 636, Fall 2015 - Assignment 7

SOLUTIONS

1. Consider the matrix of distances for four items

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 11 & 2 & 0 & \\ 5 & 3 & 4 & 0 \end{bmatrix}$$

For each of single, complete, and average linkage:

(a) List all intermediate distance matrices involved in the hierarchical clustering routine.

ALL THREE LINKAGE METHODS SHARE THE SAME MERGE EVENT, OF ITEMS 1 AND 2 (DISTANCE = 1). THE REVISED DISTANCE MATRICES ARE THEN AS FOLLOWS (THE CLUSTER CONTAINING ITEMS 1 AND 2 IS REPRESENTED BY COLUMN 1, ITEM 2 BY COLUMN 2, AND ITEM 3 BY COLUMN 3):

$$\text{SINGLE}: \begin{bmatrix} 0 & & & \\ 2 & 0 & & \\ 3 & 4 & 0 & \end{bmatrix}$$

$$\text{COMPLETE}: \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 5 & 4 & 0 & \end{bmatrix}$$

$$\text{AVERAGE}: \begin{bmatrix} 0 & & & \\ \frac{13}{2} & 0 & & \\ 4 & 4 & 0 & \end{bmatrix}$$

AT THIS STAGE, SINGLE LINKAGE MERGES ITEM 3 WITH CLUSTER $(1, 2)$, WITH A DISTANCE OF 2. THE DISTANCE BETWEEN THE NEW CLUSTER $(1, 2, 3)$ AND ITEM 4 IS 3. COMPLETE LINKAGE MERGES ITEMS 3 AND 4, WITH A DISTANCE OF 4. THE DISTANCE BETWEEN THE CLUSTERS $(1, 2)$ AND $(3, 4)$ IS 11. AVERAGE LINKAGE HAS TO MAKE A DECISION, SINCE THERE ARE TWO POSSIBLE MERGES (ITEM 4 WITH CLUSTER $(1, 2)$ OR ITEM 3 WITH ITEM 4). WITH THE SAME DISTANCE (4). WE COULD CONSIDER PERFORMING BOTH MERGES. THIS WOULD RESULT IN ONE CLUSTER CONTAINING ALL FOUR ITEMS. HOWEVER, WHILE THE DISTANCE BETWEEN ITEM 4 AND CLUSTER $(1, 2)$ IS 4, AND THE DISTANCE BETWEEN ITEMS 3 AND 4 IS 4, THE DISTANCE BETWEEN CLUSTER $(1, 2)$ AND ITEM 3 IS $13/2 > 4$. BECAUSE OF THIS, WE PICK ONE OF THE TWO MERGES. `R` PICKED THE FIRST POSSIBILITY (MERGING ITEM 4 WITH CLUSTER $(1, 2)$). THE DISTANCE BETWEEN ITEM 3 AND THE NEW CLUSTER $(1, 2, 4)$ IS $17/3$.

(b) Draw the dendrograms and compare the results of the different linkage methods.

SEE FIGURE 1. THE RESULTS DO DIFFER SOMEWHAT DEPENDING ON WHICH LINKAGE METHOD IS USED. ALL THREE METHODS JOIN ITEMS 1 AND 2 FIRST. SINGLE AND AVERAGE LINKAGE ARE SIMILAR, WITH THE ONLY DIFFERENCE BEING THE ORDER WITH WHICH ITEMS 3 AND 4 WERE ADDED TO THE REST. ACCORDING TO COMPLETE LINKAGE, ITEMS 1 AND 2 BELONG TOGETHER, AND ITEMS 3 AND 4 BELONG TOGETHER.

2. Suppose we measure two variables X_1 and X_2 for four items A, B, C, and D. The data are as follows:

	Observations	
Item	x_1	x_2
A	5	4
B	1	-2
C	-1	1
D	3	1

(a) Starting with the initial groups (AB) and (CD), write out all steps to the K-means clustering routine with $K = 2$. Use Euclidean distance on the unstandardized variables.

STEP 1 OF THE K-MEANS ALGORITHM REQUIRES THE CENTROIDS OF THE INITIAL CLUSTER LOCATIONS, CALL THEM $\mathbf{c}_{AB}$ AND $\mathbf{c}_{CD}$. WE HAVE $\mathbf{c}'_{AB} = [3, 1]$ AND $\mathbf{c}'_{CD} = [1, 1]$. IN STEP 2, WE COMPUTE EUCLIDEAN DISTANCES BETWEEN EACH ITEM AND THESE CENTROIDS, RE-ASSIGNING TO THE NEAREST CENTROID AS NEEDED. ITEM B IS RE-ASSIGNED TO THE SECOND CLUSTER, AND ITEM D IS RE-ASSIGNED TO THE FIRST, SO WE NOW HAVE THE TWO CLUSTERS (AD) AND (BC). REPEATING, CLUSTER MEMBERSHIP DOES NOT CHANGE, SO WE STOP.

(b) Repeat, now starting with the initial groups (AC) and (BD). Compare the results with those in the part (a).

WITH THIS INITIAL CHOICE OF CLUSTERING, THERE ARE TIES IN THE DISTANCE CALCULATIONS. SPECIFICALLY, ITEMS C AND D ARE EQUIDISTANT FROM THE TWO CENTROIDS. R APPEARS TO PICK THE FIRST POSSIBILITY BY DEFAULT. FOLLOWING THAT CONVENTION, WE ITERATE ONCE, RESULTING IN THE CLUSTERS (ACD) AND (B).

3. Consider the breakfast cereal data in textbook Table 11.9 ($n = 43$ cereals and $p = 8$ variables). In what follows, use Euclidean distance on the unstandardized variables.

(a) Carry out complete linkage hierarchical clustering of the cereal brands. Report a dendrogram. How many clusters would you say there are? Comment on the composition of those clusters.

A PLOT OF MERGE HEIGHTS IS SHOWN IN FIGURE 2. VIEWING THIS PLOT LIKE A "SCREE" PLOT, I WOULD CONCLUDE THAT THERE ARE 6 CLUSTERS, SINCE THERE IS A "BEND" IN THE CURVE AT THE CORRESPONDING MERGE POINT. FIGURE 3 SHOWS THE DENDROGRAM. CLUSTER 1 (THE FIRST FROM THE LEFT IN THE DENDROGRAM) CONSISTS OF MOSTLY SUGARY KIDS CEREALS, CLUSTER 2 CONSISTS MOSTLY OF CEREALS LOW IN CALORIES, FAT, AND SODIUM, ETC. ONE INTERESTING CASE IS ALL BRAN, WHICH WAS KEPT SEPARATE FROM ALL OTHER CEREALS. THIS PARTICULAR CEREAL IS VERY LOW ON CARBOHYDRATES AND VERY HIGH ON POTASSIUM.

(b) Carry out K-means clustering of the cereal brands, with $K = 3$. Plot the first two principal components and color-code by cluster membership. How do the K-means results compare with the hierarchical clustering results?

SEE FIGURE 4. THE FIRST CLUSTER CONTAINS A WIDE VARIETY OF CEREALS. THE SECOND CLUSTER CONTAINS HIGH-CALORIE CEREALS, ALTHOUGH ALL BRAN HAS BEEN

lumped in here, and it has low calorie content. The third cluster contains mostly low-calorie cereals.

4. Consider the pottery data in textbook Table 12.8 ($n = 7$ sites and $p = 4$ variables).

   (a) Using Euclidean distance, carry out multidimensional scaling to find the "best" representation of the data in $q = 2$ dimensions; you can use the `cmdscale` function in `R` for this. Make a scatterplot of the resulting variables, using the site names as plotting characters. Comment.

   See Figure 5. Sites $\mathtt{P}_0$ and $\mathtt{P}_3$ are similar to one another, and site $\mathtt{P}_6$ is distinct from all others.

   (b) Construct a biplot and interpret. How does the biplot compare to the MDS plot in part (a)?

   See Figure 6. The biplot tells a story that is similar to that told by MDS. We again have that $\mathtt{P}_0$ and $\mathtt{P}_3$ are similar to one another, and site $\mathtt{P}_6$ is distinct from all others. With the biplot, we are given further insight into what is driving the apparent clustering. In particular, $\mathtt{P}_0$ and $\mathtt{P}_3$ are mostly similar in terms of their measurements on variable D, and $\mathtt{P}_6$ is mostly distinct in terms of variable C.

```
####
#### (1)
####

dd <- as.dist(matrix(c(0, 1, 11, 5, 1, 0, 2, 3, 11, 2, 0, 4, 5, 3, 4, 0), nrow = 4))

pdf("figures/1.pdf", width = 6, height = 3)
par(mfrow = c(1, 3))
plot(hc_sng <- hclust(dd, method = "single"))
plot(hc_com <- hclust(dd, method = "complete"))
plot(hc_avg <- hclust(dd, method = "average"))
dev.off()

####
#### (2)
####

X <- matrix(c(5, 1, -1, 3, 4, -2, 1, 1), nrow = 4)

##
## K-means with K = 2. Euclidean distance, no standardization. We end up with clusters
## (1, 4) and (2, 3).
##

class_0 <- factor(c(1, 2, 1, 2))
cntrd_0 <- by(X, class_0, colMeans)

## Let R do it first.
out_kmeans <- kmeans(X, centers = cbind(cntrd_0[[1]], cntrd_0[[2]]),
  algorithm = "MacQueen")

## Then we match manually.
max_iter <- 100
for(i in 1:max_iter) {
  cat(".")

  ## Distances to centroids.
  d_1 <- colSums((t(X) - cntrd_0[[1]]) ^ 2)
  d_2 <- colSums((t(X) - cntrd_0[[2]]) ^ 2)

  class_1 <- rep(NA, 4)
  for(j in 1:4)
    class_1[j] <- ifelse(d_1[j] <= d_2[j], 1, 2)
  class_1 <- factor(class_1)
```

```
  ## Check convergence.
  if(all(class_1 == class_0))
    break;

  ## Prepare for next iteration.
  cntrd_0 <- by(X, class_1, colMeans)
  class_0 <- class_1
}

####
#### (3)
####

## Load data.
dta <- read.table("T11-9.DAT", header = FALSE)
colnames(dta) <- c("Brand", "Manufacturer", "Calories", "Protein", "Fat", "Sodium",
  "Fiber", "Carbohydrates", "Sugar", "Potassium", "Group")

## Pull out the numerical variables.
X <- as.matrix(dta[, 3:10])

##
## Hierarchical clustering with complete linkage.
##

hc <- hclust(dist(X), method = "complete")

pdf("figures/3a.pdf")
plot(hc$height)
dev.off()

pdf("figures/3b.pdf")
plot(hc)
rect.hclust(hc, k = 6)
dev.off()

## Cluster membership.
cluster_membership <- list(c(15, 19, 24, 21, 36), c(26, 43, 41, 42), c(2, 6, 33, 7, 5,
  20, 35), c(17, 1, 4, 28, 3, 8, 30, 12, 25, 37, 23, 38, 39, 10, 32, 40, 31, 9, 14, 16),
  c(18), c(29, 11, 22, 27, 13, 34))

dta[cluster_membership[[1]], ]
dta[cluster_membership[[2]], ]
dta[cluster_membership[[3]], ]
dta[cluster_membership[[4]], ]
```

```
dta[cluster_membership[[5]], ]
dta[cluster_membership[[6]], ]

##
## K-Means clustering, with K = 3.
##

out_kmeans <- kmeans(X, centers = 3, algorithm = "MacQueen")

## Principal component plot.
pca <- prcomp(X)
Y <- X %*% pca$rotation[, 1:2]

pdf("figures/3c.pdf")
plot(Y, pch = 20, col = out_kmeans$cluster)
dev.off()

## Cluster membership.
dta[out_kmeans$cluster == 1, ]
dta[out_kmeans$cluster == 2, ]
dta[out_kmeans$cluster == 3, ]

####
#### (4)
####

## Load data.
X <- as.matrix(read.table("T12-8.DAT", header = FALSE))
pot_type <- c("A", "B", "C", "D")
pot_site <- paste("P", 0:6, sep = "_")
rownames(X) <- pot_site
colnames(X) <- pot_type

##
## Multidimensional scaling.
##

dd <- dist(X)

mds_fit <- cmdscale(dd, k = 2)

pdf("figures/4a.pdf")
plot(mds_fit[, 1], mds_fit[, 2], xlab = "", ylab = "", type = "n")
text(mds_fit[, 1], mds_fit[, 2], labels = pot_site, cex = 0.7)
dev.off()
```

```
##
## Biplot.
##

pdf("figures/4b.pdf")
biplot(prcomp(X))
dev.off()
```

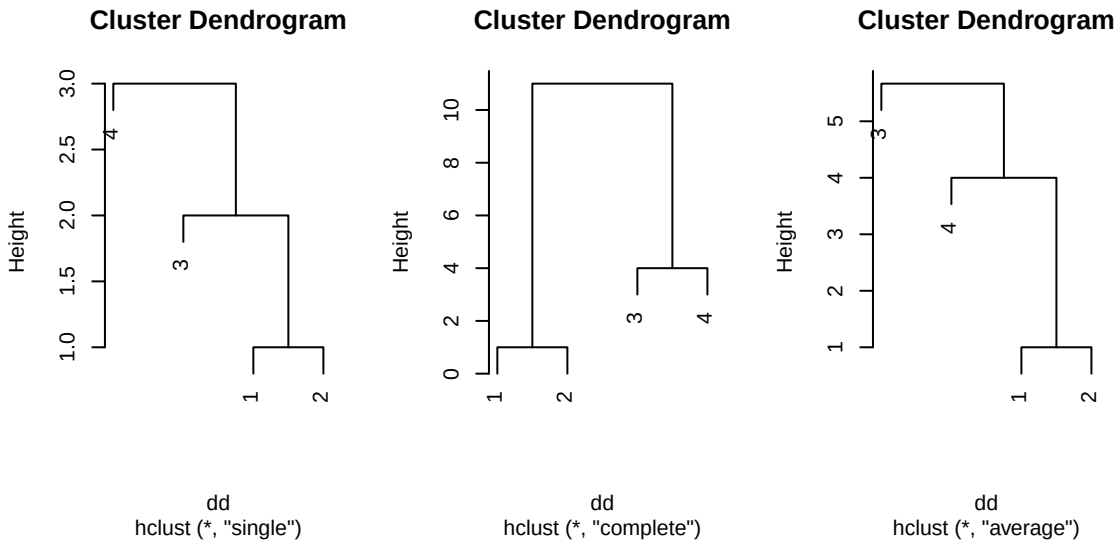

Figure 1: Clustering dendrograms for number 1, using each of single, complete, and average linkage.

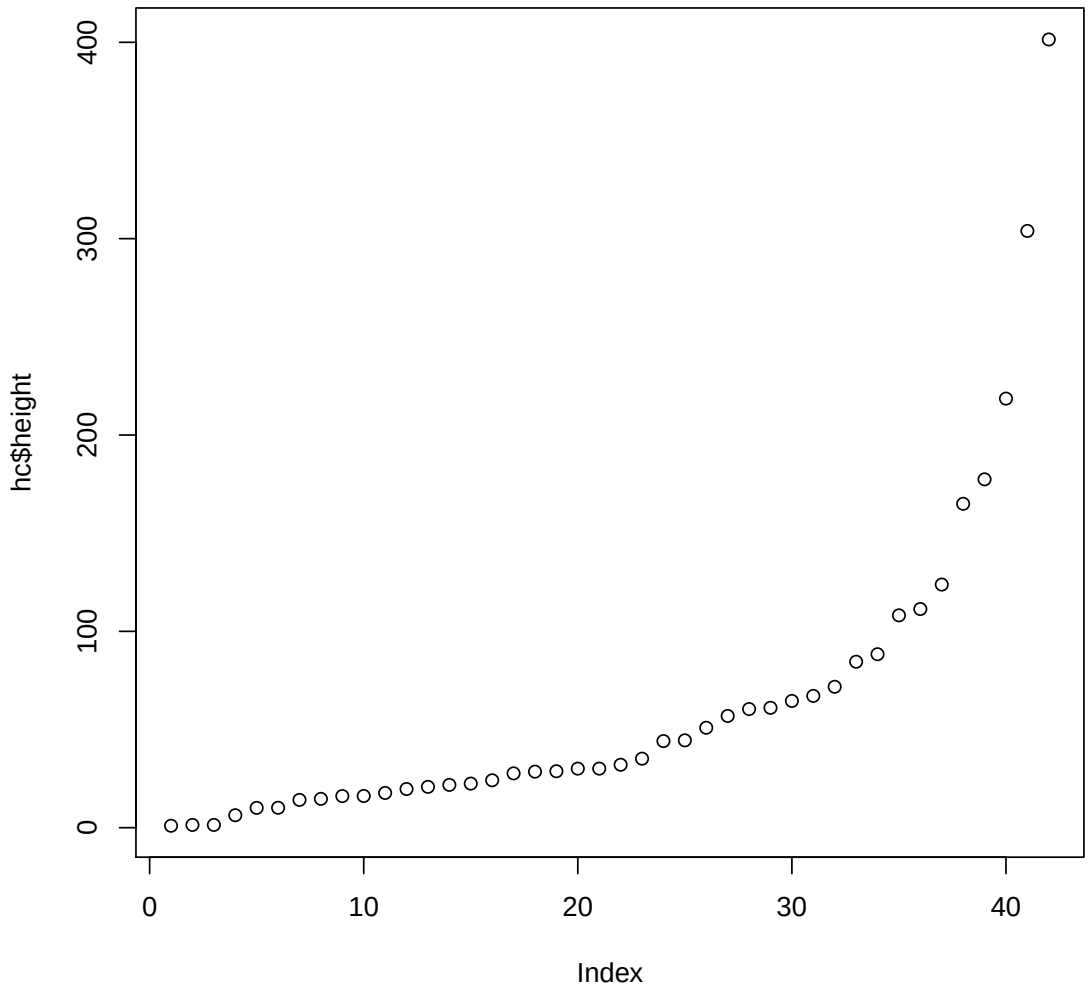

Figure 2: Plot of merge heights for number 3.

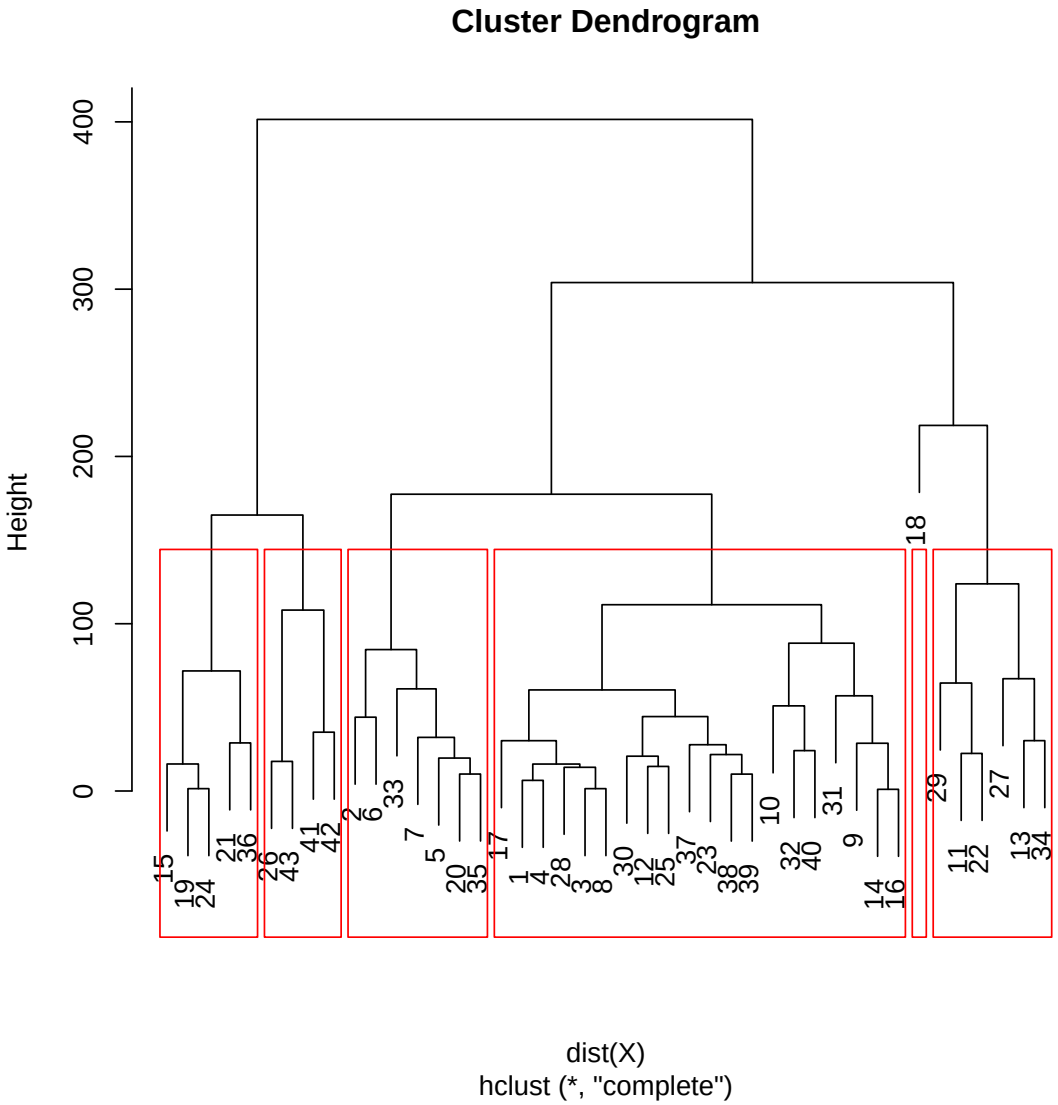

Figure 3: Plot of dendrogram for number 3.

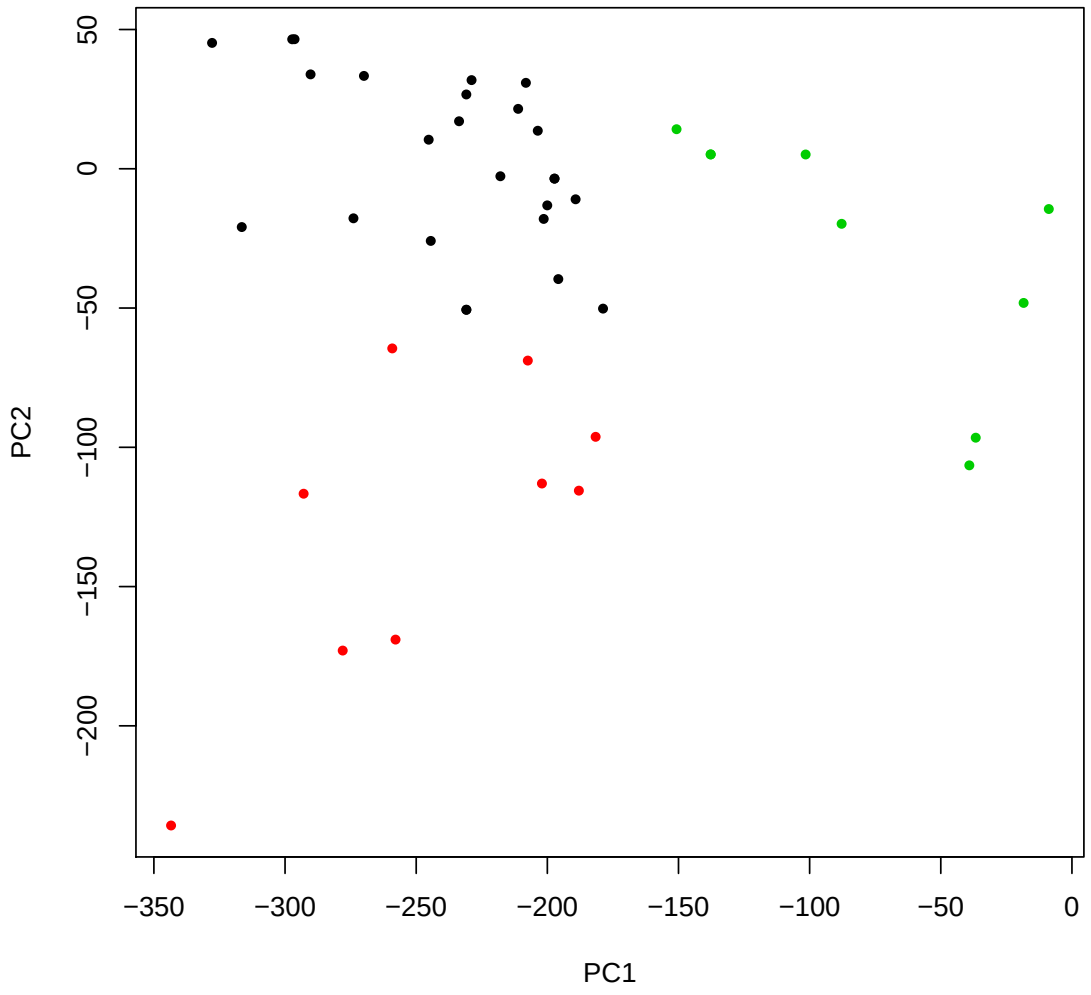

Figure 4: Plot of first two principal components, color-coded by K-means cluster membership, for number 3.

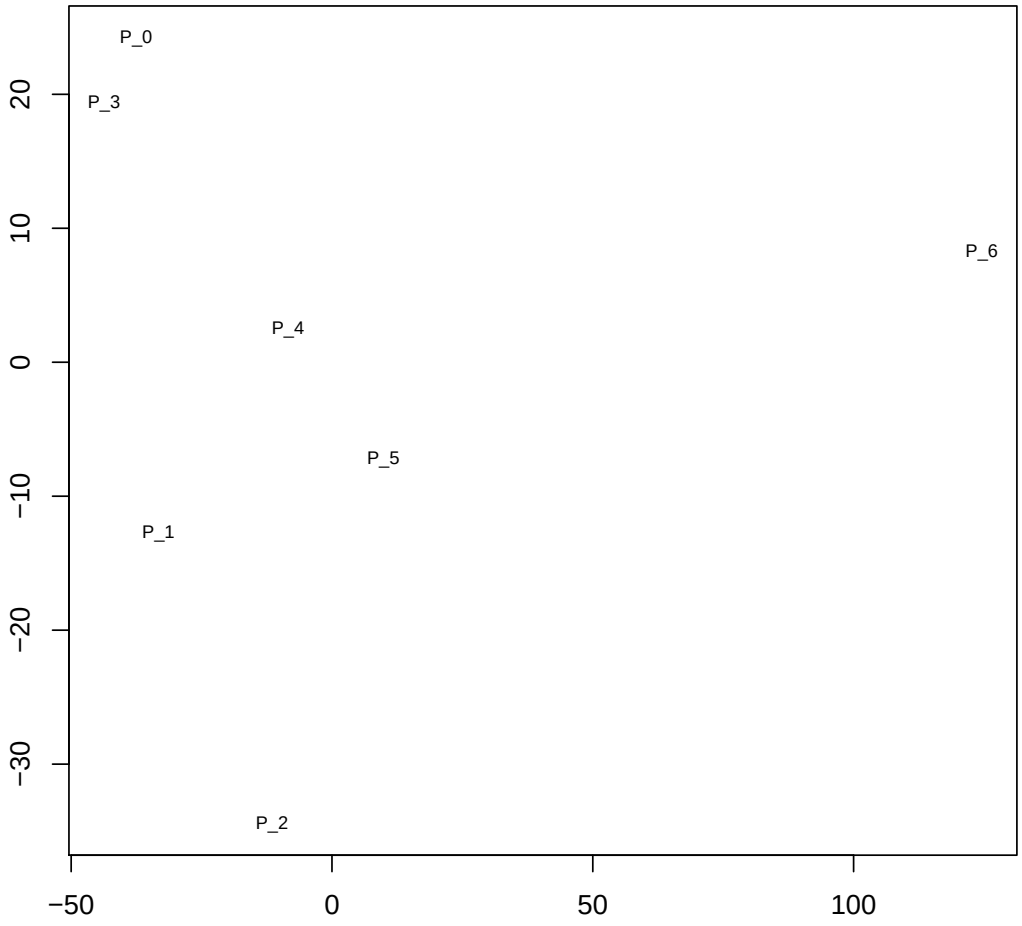

Figure 5: MDS plot in 2 dimensions for number 4.

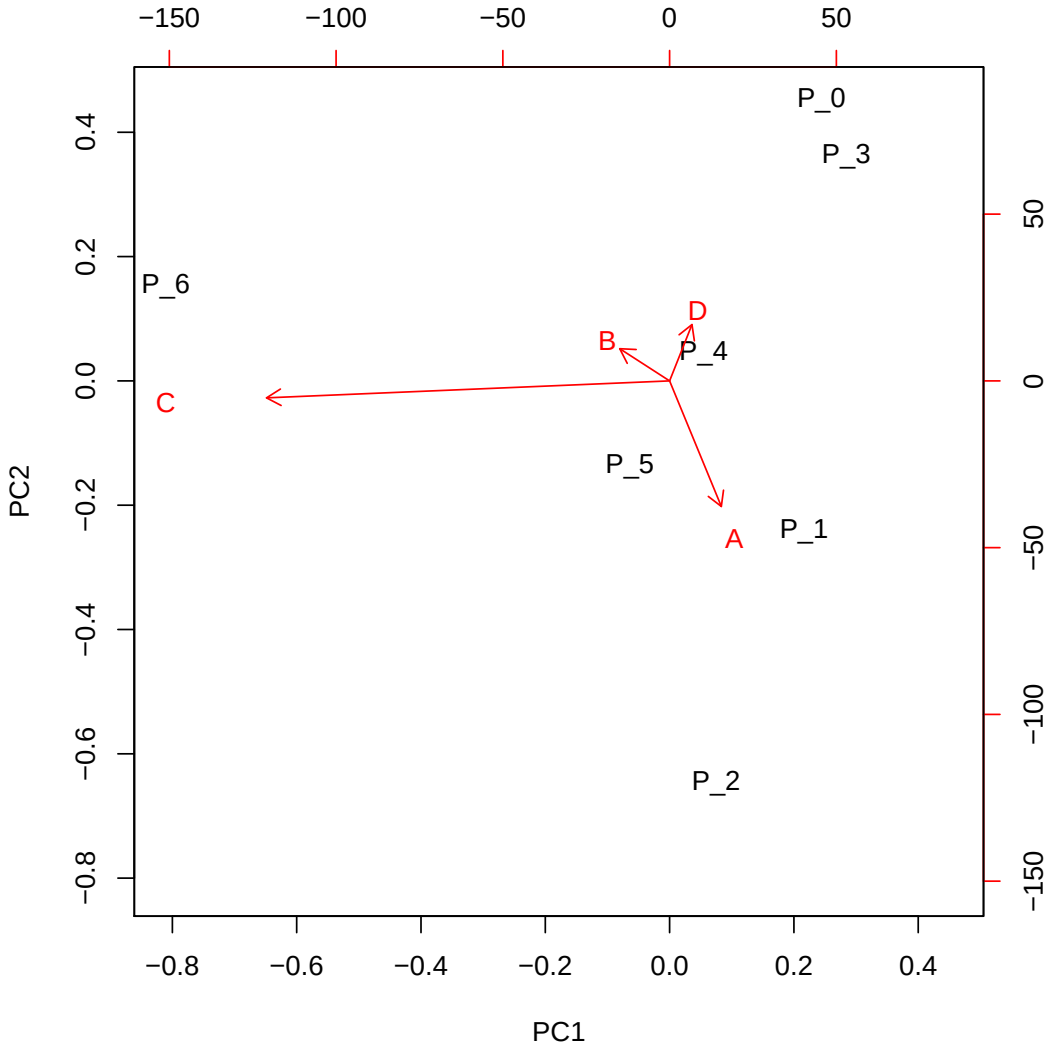

Figure 6: Biplot for number 4.