Assignment 3: Task 1 - Cluster Analysis

Group 11

DATE

1. Problem Statement

The presented data contains students' enrolment activites for 72 universities (observations), which is further subdivided into 13 faculties (variables). The task is to perform several cluster methods in order to identify for which universites students have a similar enrolment behavior. The analysis starts with some descriptive statistics, followed by a short comment regarding the assumtions. Afterwards hierarchical methods as well as non-hierarchical methods and others are applied. Finally, a conclusion regarding the best performing clustering method is made.

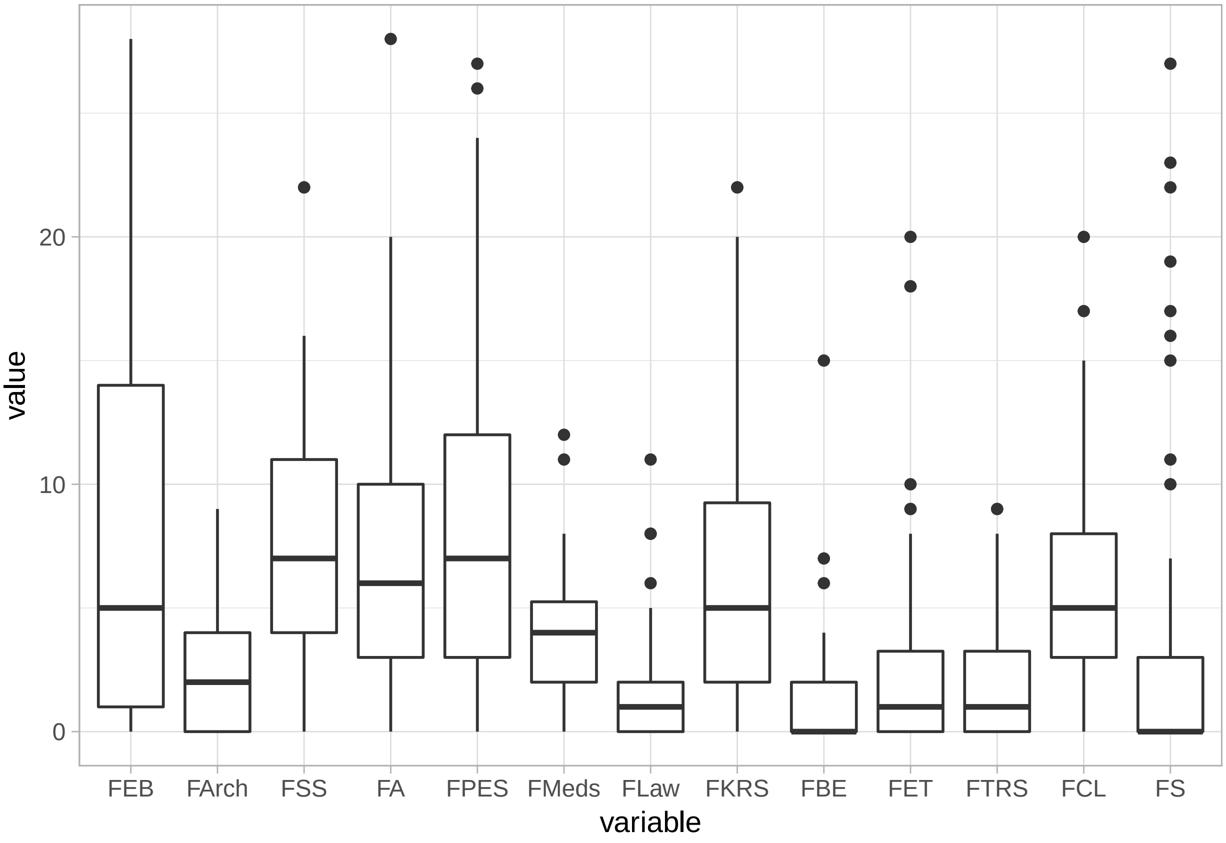
2. Descriptive Statistics

```
# Load data
unistudis <- as.tibble(read.table("unistudis.txt", header=T))

# Descriptive statistics
descr(unistudis[,-length(unistudis)], style = "rmarkdown",
      stats = c("mean", "sd", "min", "q1", "med", "q3", "max", "pct.valid"))
```

```
## ### Descriptive Statistics
## **Data Frame:** unistudis
## **N:** 72
##
## |        &nbsp; |    FEB |  FArch |    FSS |     FA |   FPES |  FMeds |   FLaw |   FKRS |
## |--------------:|-------:|-------:|-------:|-------:|-------:|-------:|-------:|-------:|
## |      **Mean** |   8.31 |   2.15 |   7.99 |   7.10 |   8.04 |   3.89 |   1.62 |   6.47 |
## |   **Std.Dev** |   8.60 |   2.09 |   4.65 |   5.73 |   6.18 |   2.60 |   2.15 |   6.13 |
## |       **Min** |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
## |        **Q1** |   1.00 |   0.00 |   4.00 |   3.00 |   3.00 |   2.00 |   0.00 |   2.00 |
## |    **Median** |   5.00 |   2.00 |   7.00 |   6.00 |   7.00 |   4.00 |   1.00 |   5.00 |
## |        **Q3** |  14.00 |   4.00 |  11.00 |  10.00 |  12.00 |   5.50 |   2.00 |   9.50 |
## |       **Max** |  28.00 |   9.00 |  22.00 |  28.00 |  27.00 |  12.00 |  11.00 |  22.00 |
## | **Pct.Valid** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
##
## Table: Table continues below
##
##
##
## |        &nbsp; |    FBE |    FET |   FTRS |    FCL |     FS |
## |--------------:|-------:|-------:|-------:|-------:|-------:|
## |      **Mean** |   1.14 |   2.75 |   1.92 |   5.90 |   3.04 |
## |   **Std.Dev** |   2.23 |   4.20 |   2.52 |   4.24 |   6.10 |
## |       **Min** |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
## |        **Q1** |   0.00 |   0.00 |   0.00 |   3.00 |   0.00 |
## |    **Median** |   0.00 |   1.00 |   1.00 |   5.00 |   0.00 |
## |         **Q3** |   2.00 |   3.50 |   3.50 |   8.00 |   3.00 |
## |        **Max** |  15.00 |  20.00 |   9.00 |  20.00 |  27.00 |
## | **Pct.Valid** | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
```

First of all there are no missing values in the data, which is seen in the Pct.Valid row. Other than that the variables seem to differ in terms of mean and spread. FEB, FSS, FA, FPES, FKRS and FCL have relatively high mean values compared to the remaining ones. This means that students of the explicitly mentioned facultes on average enrl more frequently into language coureses. In terms of standard deviations FEB, FSS, FA, FPES, FKRS, FET, FCL and FS seem to have a high spread compared to the other faculties. So for these faculties students enrolment behavior differs strongly from university to university, whereas for other faculties it doesn't.

```
# Detecting Outliers
melt(unistudis[,-length(unistudis)]) %>% ggplot(aes(x = variable, y = value)) +
  geom_boxplot() +
  theme_light()
```



```
subset(unistudis, uniID == "16" | uniID == "46")
```

```
## # A tibble: 2 x 14
##     FEB FArch   FSS    FA  FPES FMeds  FLaw  FKRS   FBE   FET  FTRS   FCL
##   <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1     1     9    16     5    11     4     2     5    15     1     8     3
## 2    14     5     5    11    27     1     0     6     0     1     0     3
## # ... with 2 more variables: FS <int>, uniID <int>
```

In terms of outliers one can clearly see that some are present. Especially observation 46 and 16 seem to be problematic. Whereas observation 46 has outlier values at variable FS and FPES, observation 16 has outlier values at variable FArch, FSS, FBE and FTRS.

3. Assumptions

Even though there are no explicit assumptions when using cluster algorithms one has to consider the fact that variables with a higher spread will have a higher importance in hierarchical cluster algorithms. This argumentation goes along with outlier values, as they might be clustered in a cluster containing only the outlier value. Therefore, observation 16 and 46 are removed before centering and standardizing the data.

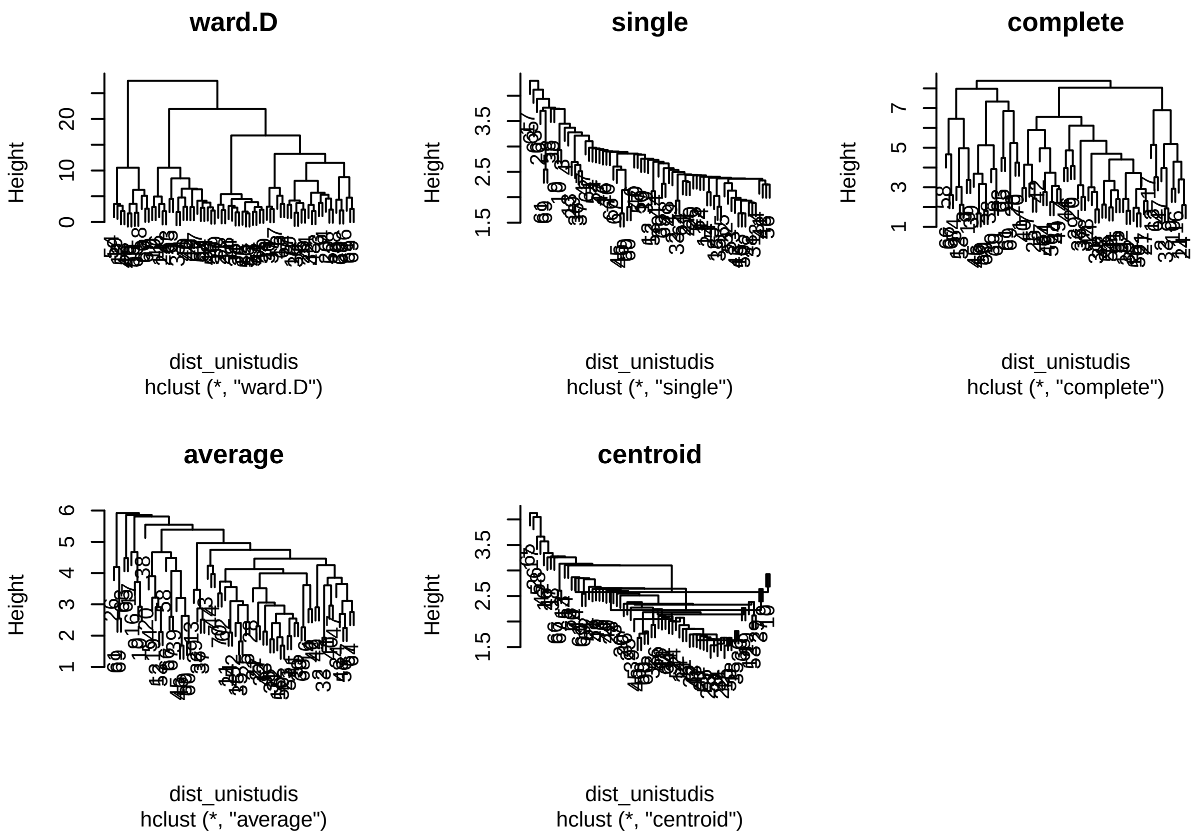
```
# Remove observation 16 and 46
unistudis <- unistudis %>% filter(uniID != "16", uniID != "46")

# Standardize values
std_unistudis <- as.tibble(unistudis[,1:ncol(unistudis)-1] %>% scale(center = T, scale = T))
std_unistudis <- std_unistudis %>% mutate(uniID = unistudis$uniID) %>% dplyr::select(uniID, 1:13)
```

4. Method and Interpretation

4.1 Hierarchical Methods

The hierarchical methods applied in this section are single linkage, complete linkage, average linkage, centroid method and Ward's error sum of squares. The first step is to identify the correct amount of clusters based on either the visual analysis of the dendrograms or by investigating the drop in R^2. Analysing the dendrograms means examining the sizes of the changes in height in the dendrograms. A large change indicates the appropriate number of clusters. The authors decide to evaluate the dendrograms.

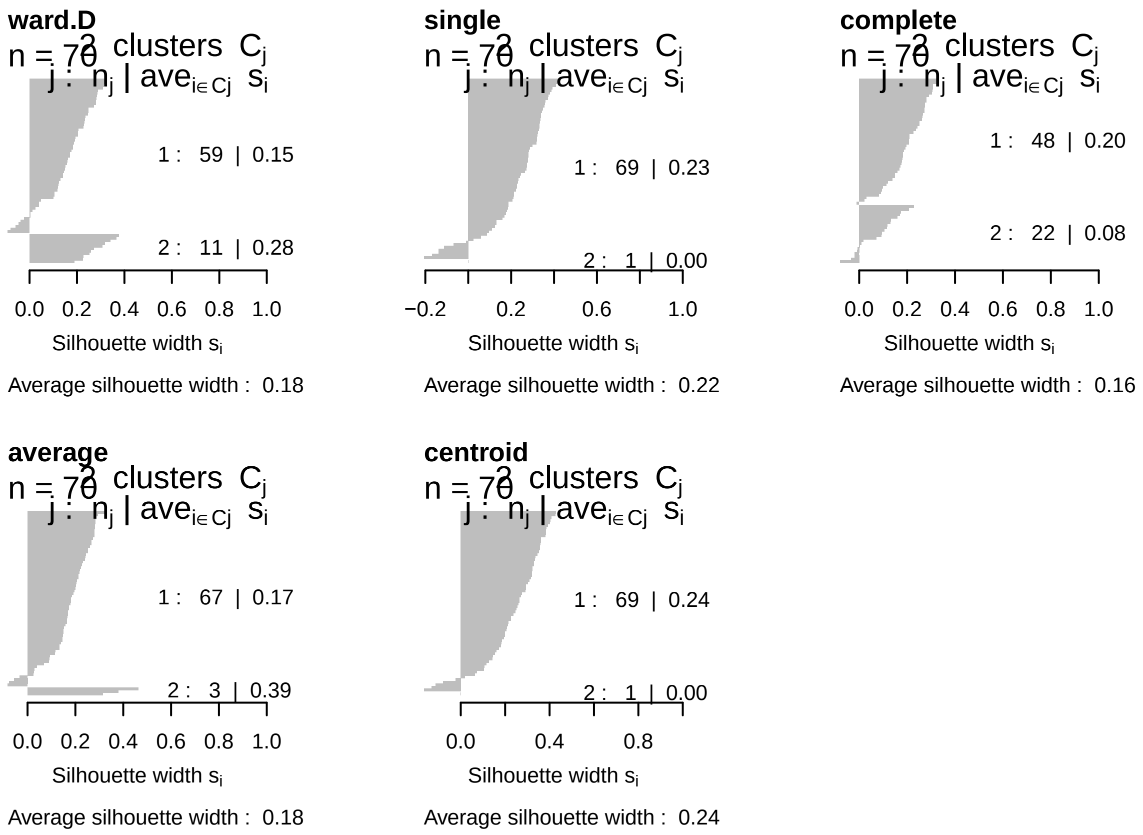
```
# Calculate euclidean distance based on standardized values
dist_unistudis <- dist(std_unistudis[,-1])

# Apply hierarchical cluster methods
hclust_methods <-
  c("ward.D", "single", "complete", "average", "centroid")
hclust_results <- lapply(hclust_methods, function(m) hclust(dist_unistudis, m))
names(hclust_results) <- hclust_methods

# Plot dendrograms
par(mfrow = c(2,3))
hclust_dendro <- lapply(names(hclust_results), function(m) plot(hclust_results[[m]],
                                                               main = m))
names(hclust_dendro) <- hclust_methods
par(mfrow = c(1,1))
```



The dendrograms of Ward's method and complete indicate a two cluster solution, whereas based on the remaining dendrograms on could also decide for more clusters. The authors decide to continue with a two cluster solution.

Hence, the trees of each method are each cut into two clusters. To further evaluate whether the right amount of clusters is chosen the silhouette plots need to be analysed. On the other hand, silhouette plots not only provide information about the right amount of clusters chosen, but also give an indication about the degree of homogeneousity in each cluster. Silhouette values range from -1 to 1, whereas values close to -1 indicate observations poorly classified and values close to 1 vice versa. Observations with values close to 0 are intermediate cases, which can be assigned to one or another cluster equally likely.

```r
# Assign observations to two clusters
noclust <- 2
hclust_cutree <- mapply(cutree, hclust_results, noclust)

# Create silhouette plots
par(mfrow = c(2,3))
sapply(colnames(hclust_cutree), function(m) plot(silhouette(hclust_cutree[,m],
                                                             dist_unistudis), main = m))
par(mfrow = c(1,1))
```



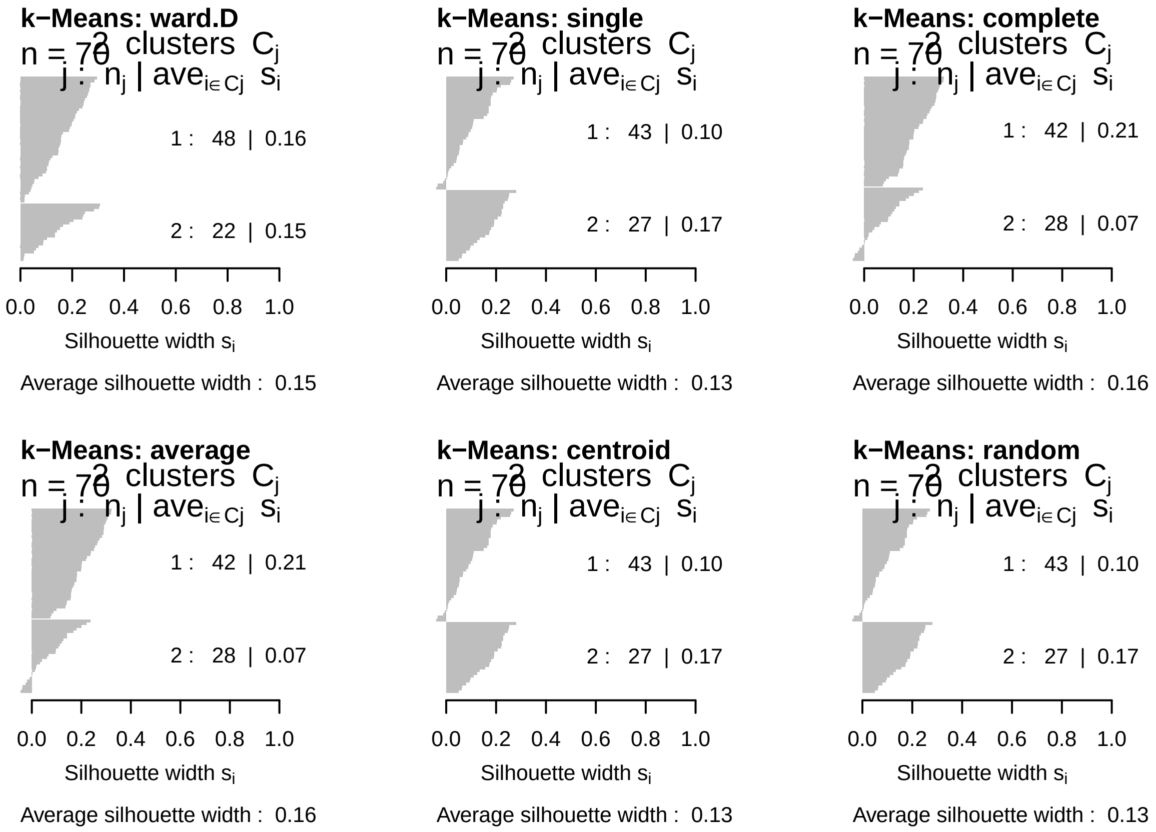
The silhouette plots show that the single linkage and centroid method cluster 69 in one cluster, which is a result of the chaining effect and gives basically no gain in information. The remaining models are better interpretalbe, as the relative frequency in the second cluster is higher. Nonetheless, silhouette scores of bigger than 0.4 are rare in each method, meaning that most of the observations are close to be intermediate cases. Therefore one might try to use non-hierarchical methods using input values from hierarchical methods to come to a better solution.

4.2 Non-Hierarchical and Model Based Methods

The non-hierarchical method used in this analysis is k-Means. As the result of k-Means is strongly dependent on the initial seeds, we are using random seeds as well as the results from each hierarchical cluster method as initial seeds. To evaluate the goodness of the methods the silhouette plots are again analysed.

```r
# Attach clusters to observations
std_unistudis_EXT <- as.tibble(cbind(std_unistudis[,-1], hclust_cutree))

# Calculate initials
initials <- lapply(hclust_methods,
                   function(m) aggregate(std_unistudis[,-1],
                                         list(as.vector(t(std_unistudis_EXT[,m]))), mean))

# k-Means with initials
clus_kmeans <- lapply(seq(1:length(hclust_methods)),
                      function(m) kmeans(std_unistudis[,-1], centers = initials[[m]][,-1]))

# k-means with random seeds
```

```r
set.seed(1)
clus_kmeans[[length(clus_kmeans)+1]] <- kmeans(std_unistudis[,-1], centers = noclust)

# Create silhouette plots
par(mfrow = c(2,3))
pnames <- paste("k-Means:", c(hclust_methods, "random"))
lapply(seq(1:length(clus_kmeans)),
       function(m) plot(silhouette(clus_kmeans[[m]]$cluster, dist_unistudis),
                        main = pnames[m]))
```



k−Means: ward.D

k−Means: single

k−Means: complete

k−Means: average

k−Means: centroid

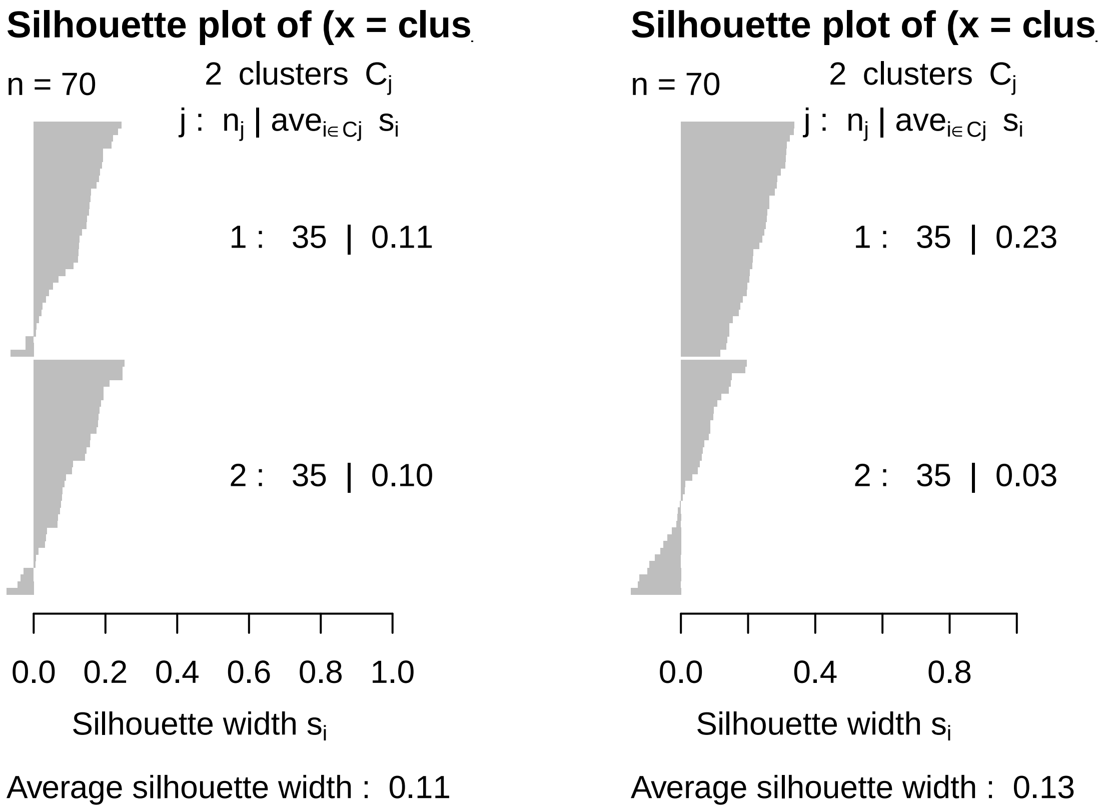
k−Means: random

First of all each method clusters approximataly the same observations in each cluster. Other than that the silhouette plots of k-Means with Ward and average linkage seeding seem to be the best k-Means cluster solutions. This can also be validated when comparing the average silhouette width. K-Means with complete seeding performs bad in specifying the second cluster, whereas the remaining methods have difficulties in specifying the first cluster.

5. Alternative solutions

```r
# k-Mediods (Partitioning Around Mediods)
clus_pam <- pam(std_unistudis[,-1], noclust)

# Model based clustering (selection based on BIC)
clus_mod <- Mclust(std_unistudis[,-1], G = 2:9)
```

```r
par(mfrow = c(1,2))
plot(silhouette(clus_pam$cluster, dist_unistudis))
plot(silhouette(clus_mod$classification, dist_unistudis))
```



```r
# Cluster ensemble
```

6. Conclusion

Appendix