# DSM3004
# Spatial Statistics for Remotely Sensed Images

Mohammad Wasiq

## Table of Contents

1 Practical: 5 .......... 1
1.1 Objective : .......... 1

# 1 Practical: 5

## 1.1 Objective :

Using **ClusterR** package in R, estimate the optimal number of clusters and plot the results using silhouette or 2-dimensional plot for the given data set.

*Cluster Analysis* or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is the main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

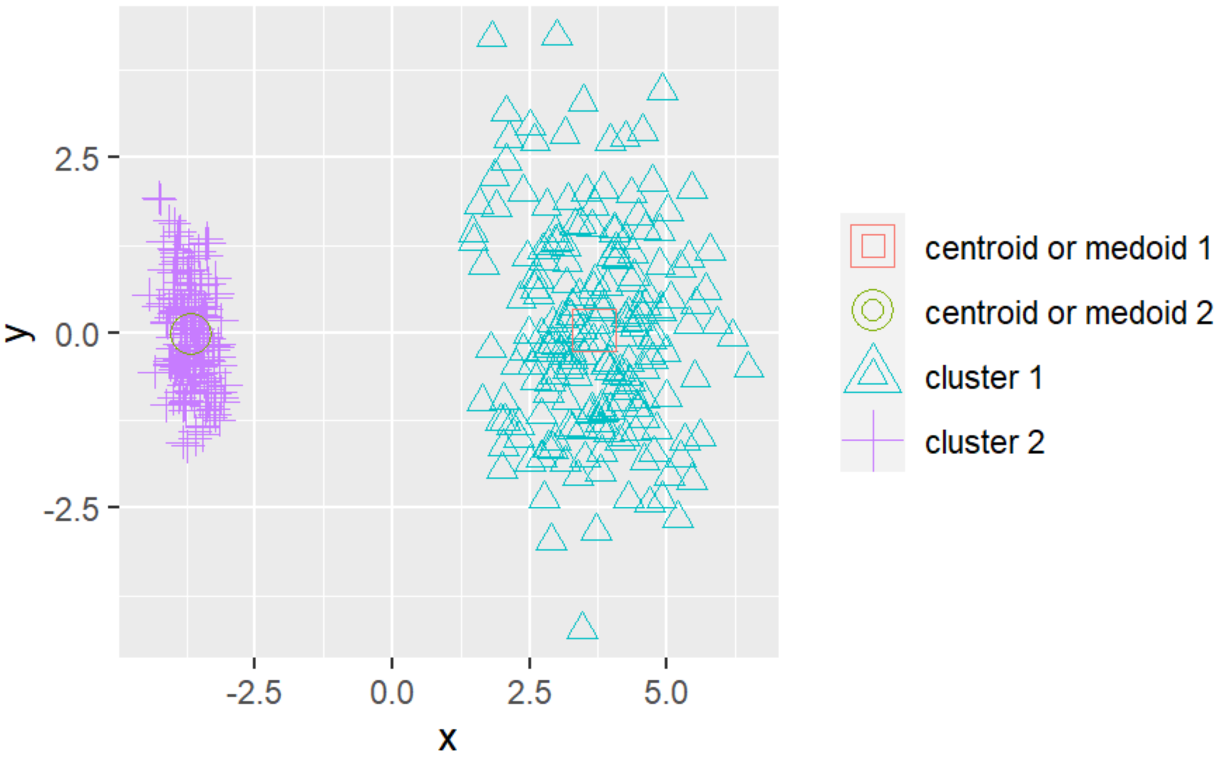
The ClusterR package consists of centroid-based (k-means, mini-batch-kmeans, k-medoids) and distribution-based (GMM) clustering algorithms. Furthermore, the package offers functions to

- Validate the output using the true labels,
- Plot the results using either a silhouette or a 2-dimensional plot,
- Predict new observations,
- Estimate the optimal number of clusters for each algorithm separately.

The following example explains the functionality of the clustering algorithms, which are part of the **ClusterR** package.

```
library(ClusterR)

data(dietary_survey_IBS)
dim(dietary_survey_IBS)

## [1] 400  43
```

```
X<- dietary_survey_IBS[, -ncol(dietary_survey_IBS)]
Y<- dietary_survey_IBS[, ncol(dietary_survey_IBS)]

dat<- center_scale(X, mean_center = T, sd_scale = T)

gmm<- GMM(dat, 2, dist_mode = "maha_dist",
          seed_mode = "random_subset",
          km_iter = 10, em_iter=10, verbose=F)

pr<- predict(gmm, newdata = dat)

opt_gmm<- Optimal_Clusters_GMM(dat, max_clusters= 10,
                               criterion= "BIC", dist_mode= "maha_dist",
                               seed_mode= "random_subset", km_iter= 10,
                               em_iter= 10, var_floor= 1e-10, plot_data= T)

pca_dat<- stats::princomp(dat)$scores[, 1:2]

km<- KMeans_arma(pca_dat, clusters= 2,
                 n_iter= 10, seed_mode= "random_subset",
                 verbose= T, CENTROIDS= NULL)

## kmeans(): generating initial means
## kmeans(): n_threads: 1
## kmeans(): iteration:    1   delta: 13.0164
## kmeans(): iteration:    2   delta: 3.49254
## kmeans(): iteration:    3   delta: 0.0784074
## kmeans(): iteration:    4   delta: 0

pr<- predict_KMeans(pca_dat, km)

table(dietary_survey_IBS$class, pr)

##    pr
##       1   2
##   0   0 200
##   1 200   0

class(km)<- 'matrix'
class(km)

## [1] "matrix" "array"

plot_2d(data= pca_dat, clusters= as.vector(pr),
        centroids_medoids= as.matrix(km))
```



The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters.