AHA 7: Clustering | k-mediods

2024-02-21

Load packages (and set seed)

```r
# Load packages
library(cluster); library(factoextra)
library(NbClust); library(igraph)
set.seed(42)
```

Load data

```r
# Load data
bp_data <- read.csv("../data/bipolar_depression/bipolar_depression_clean.csv")
```

Data wrangling

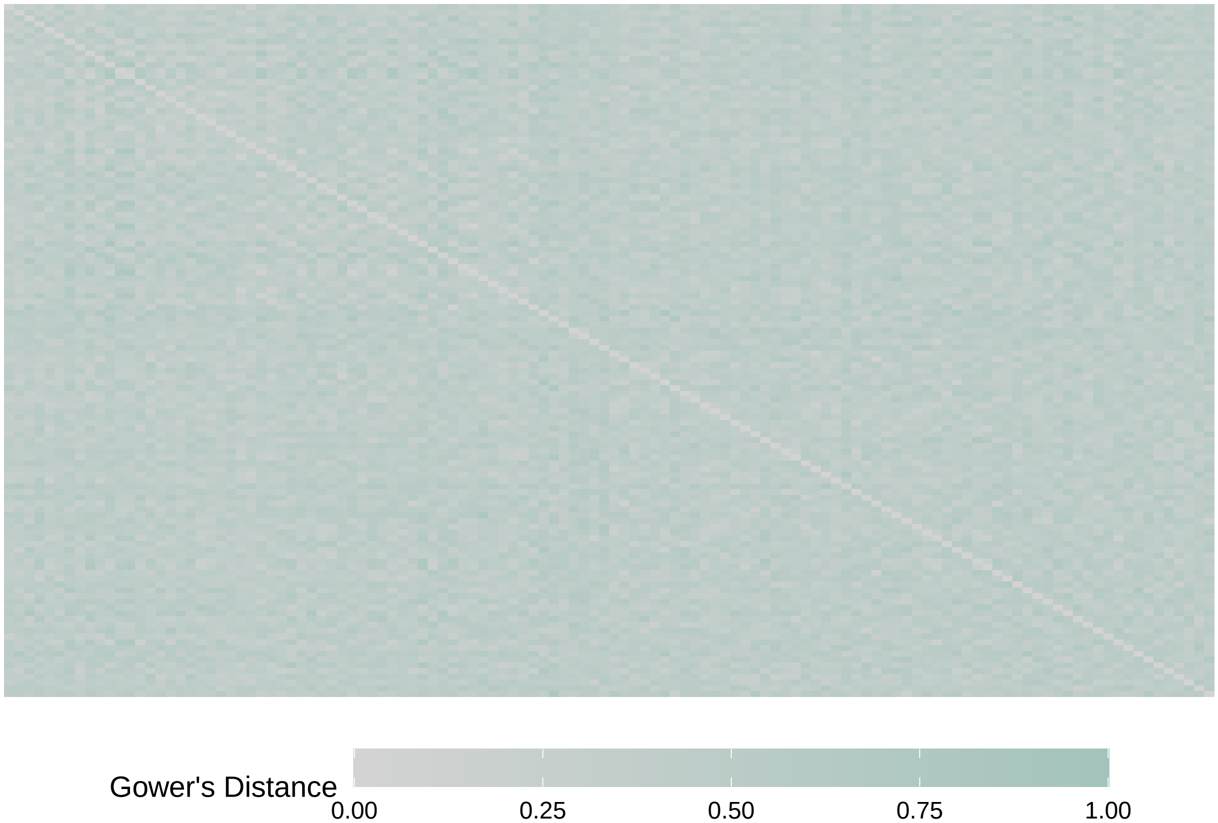
```r
# Get expert diagnosis
expert <- bp_data$Expert.Diagnose

# Extract variables of interest
bp_voi <- apply(bp_data[,-c(1,19)], 2, as.numeric)
```

Compute Gower's distance

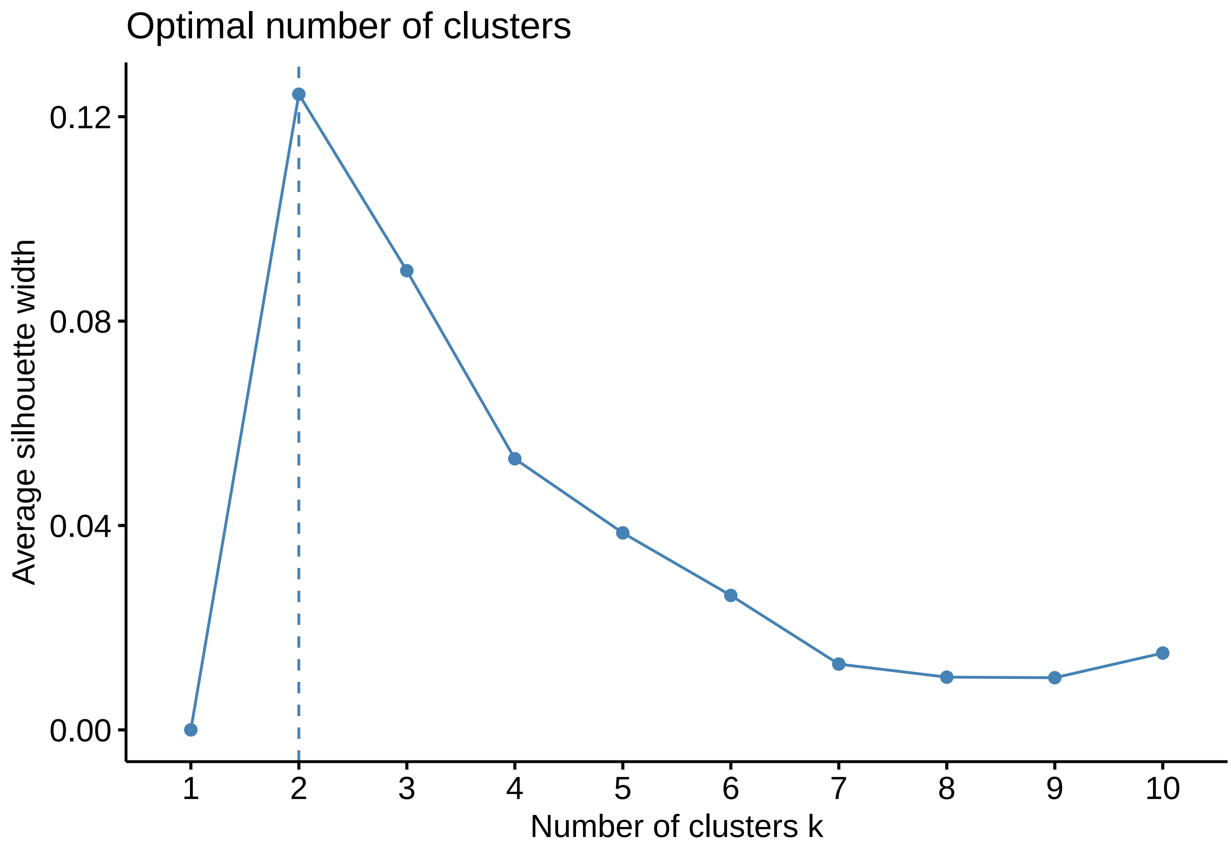
```r
# Compute Gower's distance
bp_distance <- daisy(x = bp_voi, metric = "gower")

# Produce heatmap
EGAnet:::ggheatmap(bp_distance) +
  scale_fill_gradient(
    name = "Gower's Distance",  limits = c(0, 1),
    low = "lightgrey", high = "#A3C4BC"
  ) + theme(
    axis.text = element_blank(), axis.title = element_blank(),
    axis.ticks = element_blank(), legend.position = "bottom",
    legend.key.width = unit(2, "cm"),
    legend.key.height = unit(0.5, "cm")
  )
```

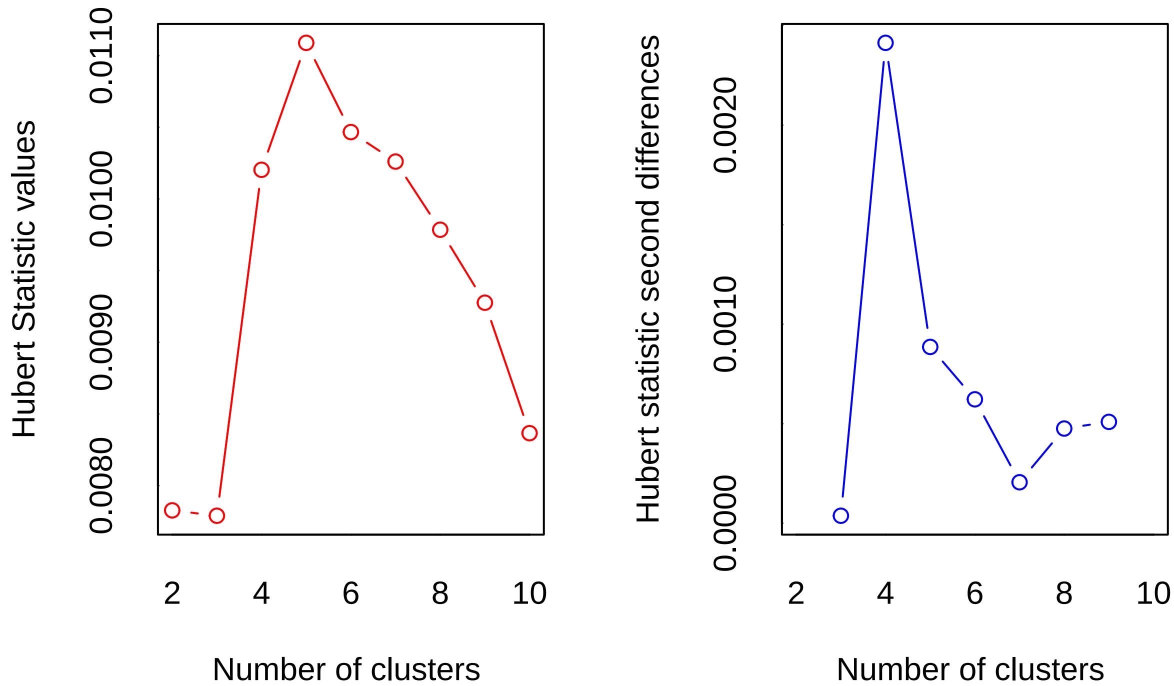


Identify number of clusters with k-mediods

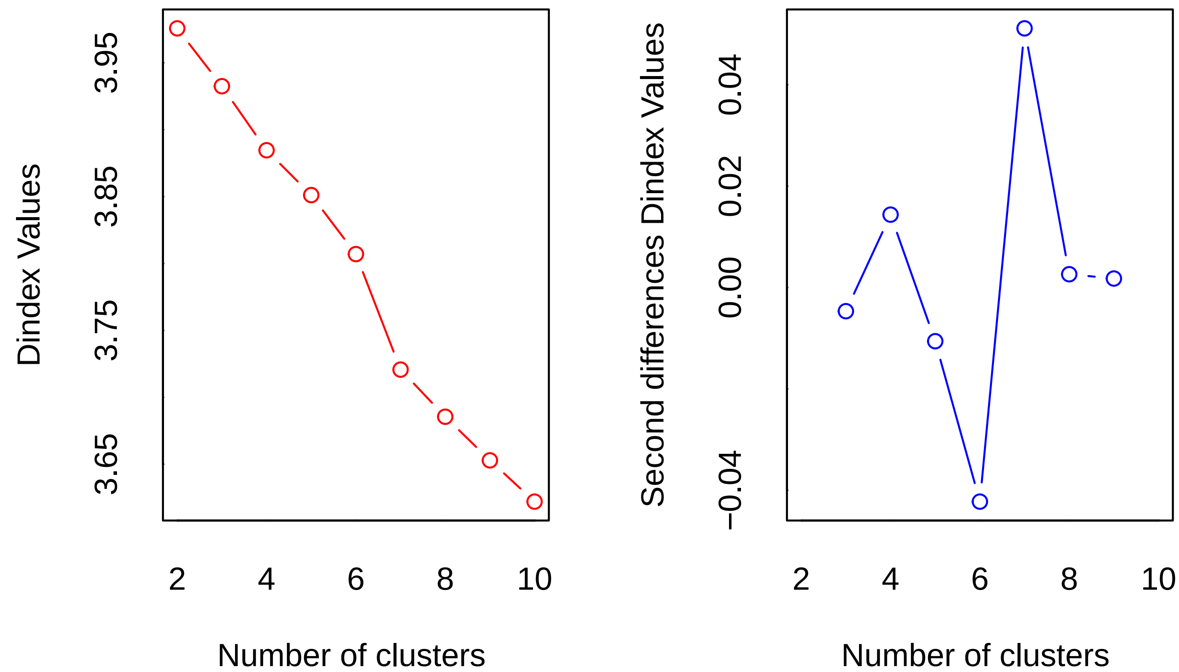
```
# Plot silhouette method
fviz_nbclust(
  x = bp_voi, # supply data
  FUNcluster = pam, # cluster function
  diss = bp_distance, # supply distance
  method = "silhouette", # silhouette
  k = 10,  # maximum number of clusters
  nstart = 25 # same as our k-mediods setup
)
```




```r
# {NbClust} has over 30 different metrics to evaluate
# the number of clusters -- majority approach:
majority <- NbClust(
  data = bp_voi, # supply data
  diss = bp_distance, # supply distance
  distance = NULL, # using our own distance
  max.nc = 10, # maximum number of clusters
  method = "median", # perhaps more consistent with mediods
  index = "all" # all metrics
)
```




```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.
```




```
*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 1 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 10 proposed 7 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  7


*******************************************************************
```

7 is the most but 2 is provided by Silhouette and suggested by nearly as many as methods as 7. I'll proceed with 2 clusters.

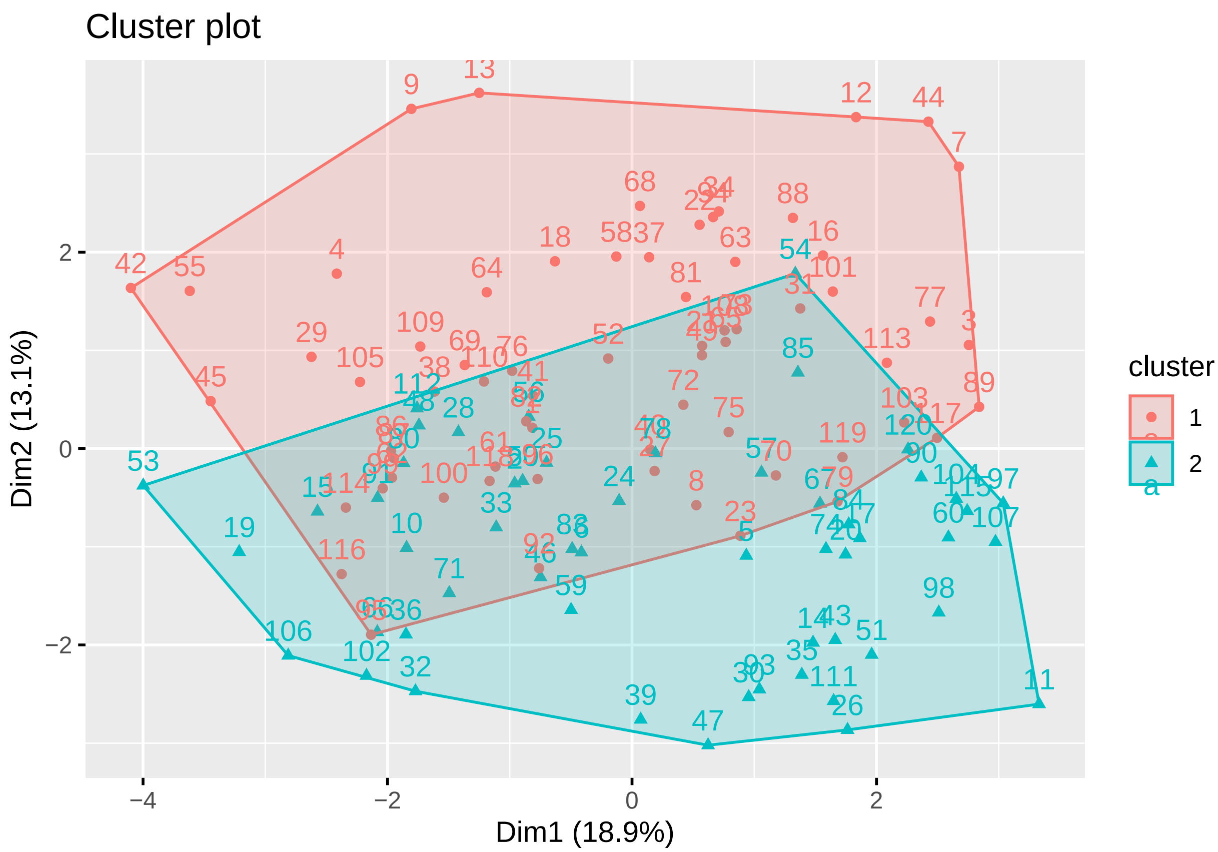
Perform k-mediods

```
# Perform k-mediods with silhouette
silhouette_run <- pam(
  x = bp_distance, # supply distance
  k = 2, # number of clusters
  nstart = 25 # number of random starting values
)
```

Get results

```
# Need to supply data back to object
silhouette_run$data <- bp_voi

# Plot
fviz_cluster(silhouette_run, bp_voi)
```




```
# Median observations
bp_voi[silhouette_run$medoids,]
```

```
     Sadness Euphoric Exhausted Sleep.dissorder Mood.Swing Suicidal.thoughts
[1,]       3        2         2               2          1                 1
[2,]       2        1         3               2          0                 0
     Anorxia Authority.Respect Try.Explanation Aggressive.Response
[1,]       1                 0               1                   1
[2,]       0                 1               0                   0
```

```
     Ignore...Move.On Nervous.Break.down Admit.Mistakes Overthinking
[1,]                0                  0              0            1
[2,]                1                  0              1            0
     Sexual.Activity Concentration Optimisim
[1,]               6             4         6
[2,]               5             5         3
```

It seems like there is a cluster with higher suicidality, mood swings, and overthinking.

Compare clusters with expert's opinion

```r
# Adjusted Rand Index
compare(silhouette_run$clustering, expert, method = "adjusted.rand")
```

```
[1] 0.15171
```

```r
# Normalized Mutual Information
compare(silhouette_run$clustering, expert, method = "nmi")
```

```
[1] 0.1740543
```

There is some similarity but not much between these clusters and the expert's diagnoses. k-mediods is less similar to experts than hierarchical clustering.