Applied Statistics in R

Elena Parilina

Master's Program
Game Theory and Operations Research

Saint Petersburg State University

2019

Agenda

1. Cluster analysis
   - Distance between clusters
2. Hierarchical clustering
3. Non-hierarchical clustering: k-means

Cluster analysis

Dissimilarity Measures

Dissimilarity Measures d:

- $d(x_i, x_j) \geqslant 0$;
- $d(x_i, x_j) = 0$ iff $x_i = x_j$;
- $d(x_i, x_j) = d(x_j, x_i)$;
- (not necessarily) $d(x_i, x_j) \leqslant d(x_i, x_k) + d(x_k, x_j)$.

Let $x_i, x_j \in \mathbb{R}^p$, $x_i = (x_{i1}, \ldots, x_{ip})^\intercal$, $x_j = (x_{j1}, \ldots, x_{jp})^\intercal$.

Minkowski metric:

$$d(x_i, x_j) = \left(\sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|^m \right)^{1/m}.$$

Special cases:

- Manhattan or city block metric $m = 1$: $d(x_i, x_j) = \sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|$.
- Euclidean distance $m = 2$: $d(x_i, x_j) = \sqrt{\sum_{\ell=1}^{p} (x_{i\ell} - x_{j\ell})^2}$.
- Chebyshev distance $m = \infty$: $d(x_i, x_j) = \max_{\ell=1,\ldots,p} |x_{i\ell} - x_{j\ell}|$.

Similarity measure

Similarity measure s:

- $0 \leqslant s(x_i, x_j) \leqslant 1$;
- $s(x_i, x_j) = 1$ iff $x_i = x_j$;
- $s(x_i, x_j) = s(x_j, x_i)$.

The relationship between s and d:

$$d(x_i, x_j) = 1 - s(x_i, x_j), \qquad s(x_i, x_j) = \frac{1}{1 + d(x_i, x_j)}.$$

Similarity measure: examples

Let $x_i, x_j \in \mathbb{R}^p$, $x_i = (x_{i1}, \ldots, x_{ip})^\intercal$, $x_j = (x_{j1}, \ldots, x_{jp})^\intercal$:

1. Using Pearson correlation:

$$r(x_i, x_j) = \frac{\sum_{\ell=1}^{p} (x_{i\ell} - \bar{x}_{i\cdot})(x_{j\ell} - \bar{x}_{j\cdot})}{\sqrt{\sum_{\ell=1}^{p} (x_{i\ell} - \bar{x}_{i\cdot})^2 \cdot \sum_{\ell=1}^{p} (x_{j\ell} - \bar{x}_{j\cdot})^2}} \in [-1, 1],$$

where $\bar{x}_{i\cdot} = \frac{1}{p} \sum_{\ell=1}^{p} x_{i\ell}$, $\bar{x}_{j\cdot} = \frac{1}{p} \sum_{\ell=1}^{p} x_{j\ell}$.

Thus $s(x_i, x_j) = |r(x_i, x_j)|$ or $s(x_i, x_j) = r^2(x_i, x_j)$.

2. Using cosine of the angle between vectors x_i and x_j:

$$\cos(x_i, x_j) = \frac{x_i^\intercal x_j}{\sqrt{x_i^\intercal x_i} \cdot \sqrt{x_j^\intercal x_j}} \in [-1, 1].$$

Distance matrix

$$\begin{pmatrix} 0 & \cdots & d_{1j} & \cdots & d_{1n} \\ \vdots & & \vdots & & \vdots \\ d_{i1} & \cdots & d_{ij} & \cdots & d_{in} \\ \vdots & & \vdots & & \vdots \\ d_{n1} & \cdots & d_{nj} & \cdots & 0 \end{pmatrix}$$

The distance matrix can be calculated using either

- dissimilarity measure d,
- or similatiry measure s and the relationship between s and d.

Using R...

dist function in `stats` package

dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p=2)

Arguments

- x — a data frame, or a list. A vector will be converted into a column matrix.
- method — measure of distance. The default for dist is "Euclidean" and for simil "correlation". E.g., "maximum", "manhattan", "minkowski".
- diag — logical value indicating whether the diagonal of the distance/similarity matrix should be printed by print.dist/print.simil. Note that the diagonal values are never stored in dist objects.
- upper — logical value indicating whether the upper triangle of the distance/similarity matrix should be printed by print.dist/print.simil.

Using R...

p the power of Minkowski distance (when method="minkowski").

simil

simil(x, method = "euclidean", diag = FALSE, upper = FALSE, p=2)

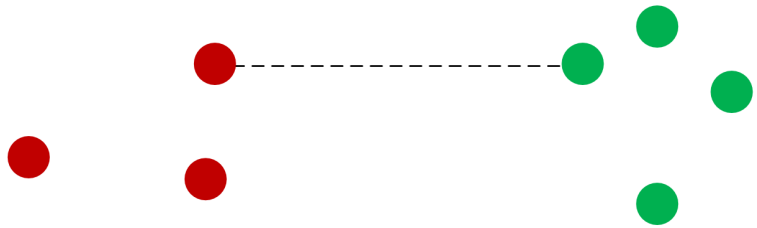
Distance measures can be used with `simil`, and similarity measures with dist. In these cases, the result is transformed accordingly using the specified coercion functions (default: $pr\_simil2dist(x) = 1 - abs(x)$ and $pr\_dist2simil(x) = 1/(1 + x)$). Objects of class simil and dist can be converted one in another using as.dist and as.simil, respectively.
The default for dist is "Euclidean", and for simil "correlation".

Distance between clusters

Let $C_1, C_2 \subset \{x_1, \ldots, x_n\}$, $C_1 \cap C_2 = \varnothing$.

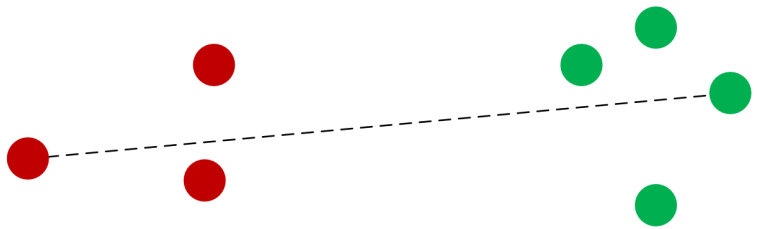
1. Single link (nearest-neighbor) method:

$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y).$$



2. Complete link (farthest-neighbor) method

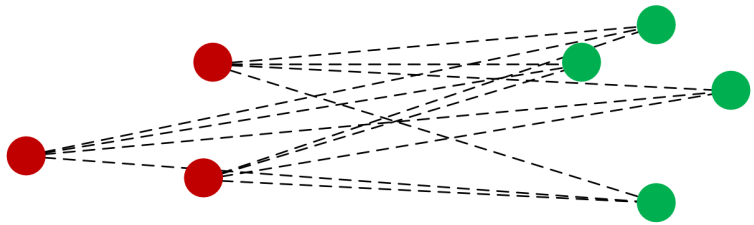
$$d(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y).$$



Distance between clusters

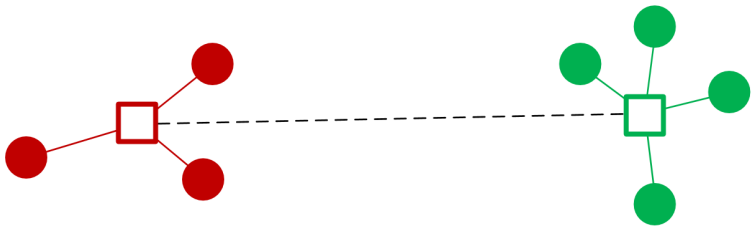
3. Average link method

$$d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y).$$



4. Centroid method

$$d(C_1, C_2) = d^2(\bar{x}_{C_1}, \bar{y}_{C_2}), \quad \text{ãäå} \quad \bar{x}_{C_1} = \frac{1}{|C_1|} \sum_{x \in C_1} x, \quad \bar{y}_{C_2} = \frac{1}{|C_2|} \sum_{y \in C_2} y.$$



Distance between clusters

5. Ward's (incremental sum of squares) method

$$d(C_1, C_2) = \frac{|C_1| \cdot |C_2|}{|C_1| + |C_2|} \cdot d^2(\bar{x}_{C_1}, \bar{y}_{C_2}),$$

where $\bar{x}_{C_1} = \frac{1}{|C_1|} \sum_{x \in C_1} x, \quad \bar{y}_{C_2} = \frac{1}{|C_2|} \sum_{y \in C_2} y.$

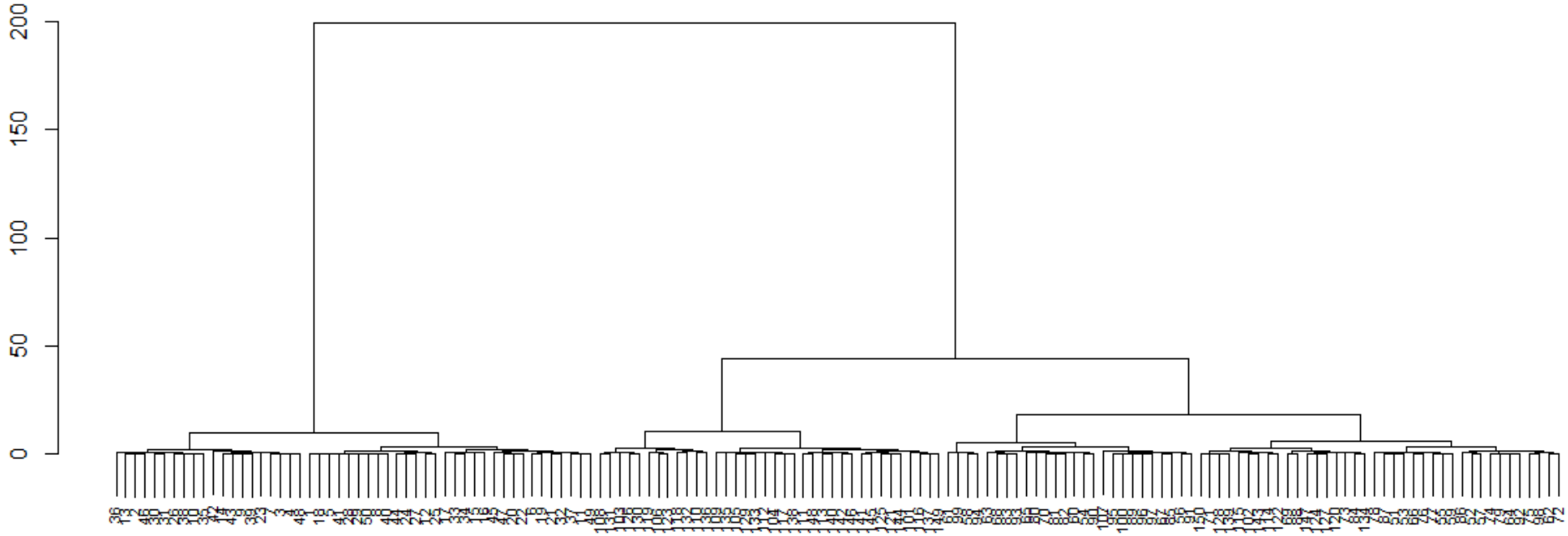
Hierarchical clustering

Hierarchical clustering

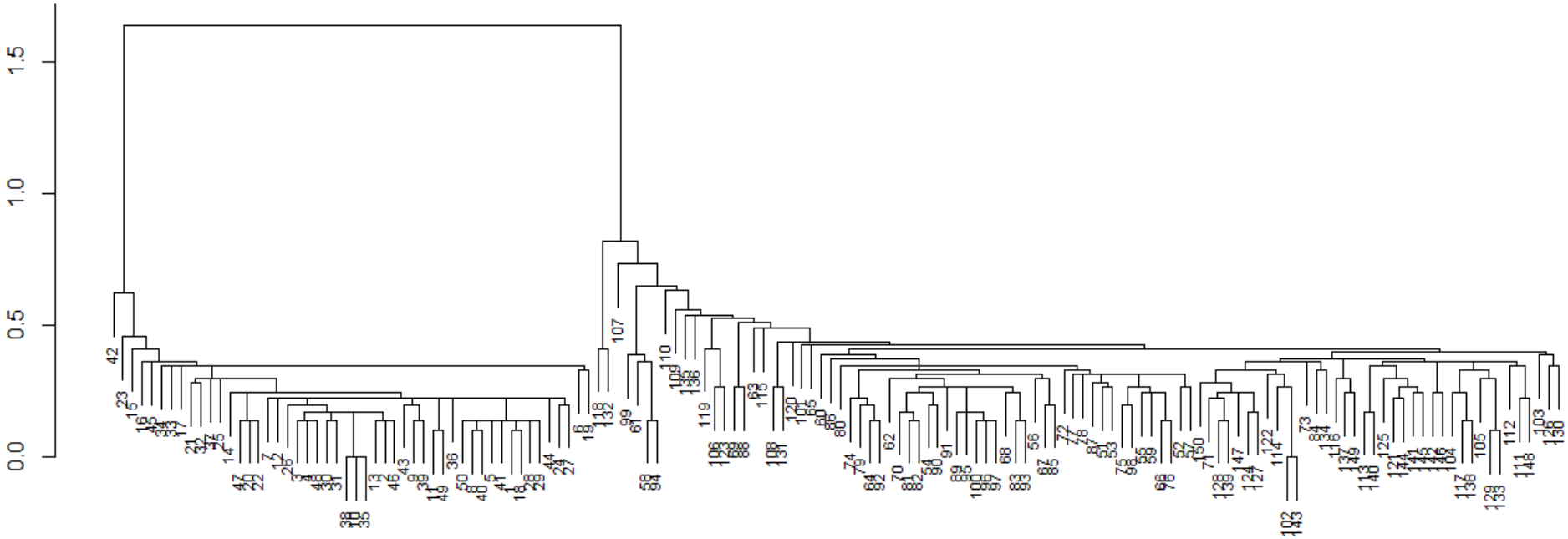
Agglomerative hierarchical clustering methods use the elements of a distance matrix to generate a tree diagram.

1. Begin with n clusters, each containing only a single object.
2. Search the dissimilarity matrix D for the most similar pair. Let the pair chosen be associated with element $d(x_i, x_j)$ so that object i and j are selected.
3. Combine objects i and j into a new cluster (ij) employing some criterion and reduce the number of clusters by 1 by deleting the row and column for objects i and j. Calculate the dissimilarities between the cluster (ij) and all remaining clusters, using the criterion, and add the row and column to the new dissimilarity matrix.
4. Repeat steps 2 and 3, $(n-1)$ times until all objects form a single cluster. At each step, identify the merged clusters and the value of the dissimilarity at which the clusters are merged.

Example (iris dataset, Ward's method)



Example (iris dataset, single link method)



Using R...

hclust

hclust(d, method = "complete", members = NULL)

Arguments

d	a dissimilarity structure as produced by dist.
method	the agglomeration method to be used. This must be (an unambiguous abbreviation of) one of "single", "complete", "average", "mcquitty", "ward.D", "ward.D2", "centroid" or "median".
members	NULL or a vector with length the number of observations.

```
data("USArrests")
my_data <- USArrests
head(my_data)
```

result

```
Murder Assault UrbanPop Rape
Alabama 13.2 236 58 21.2
Alaska 10.0 263 48 44.5
Arizona 8.1 294 80 31.0
Arkansas 8.8 190 50 19.5
California 9.0 276 91 40.6
Colorado 7.9 204 78 38.7
```

```
my_data <- na.omit(my_data)
my_data <- scale(my_data)
install.packages("factoextra")
head(my_data)
```

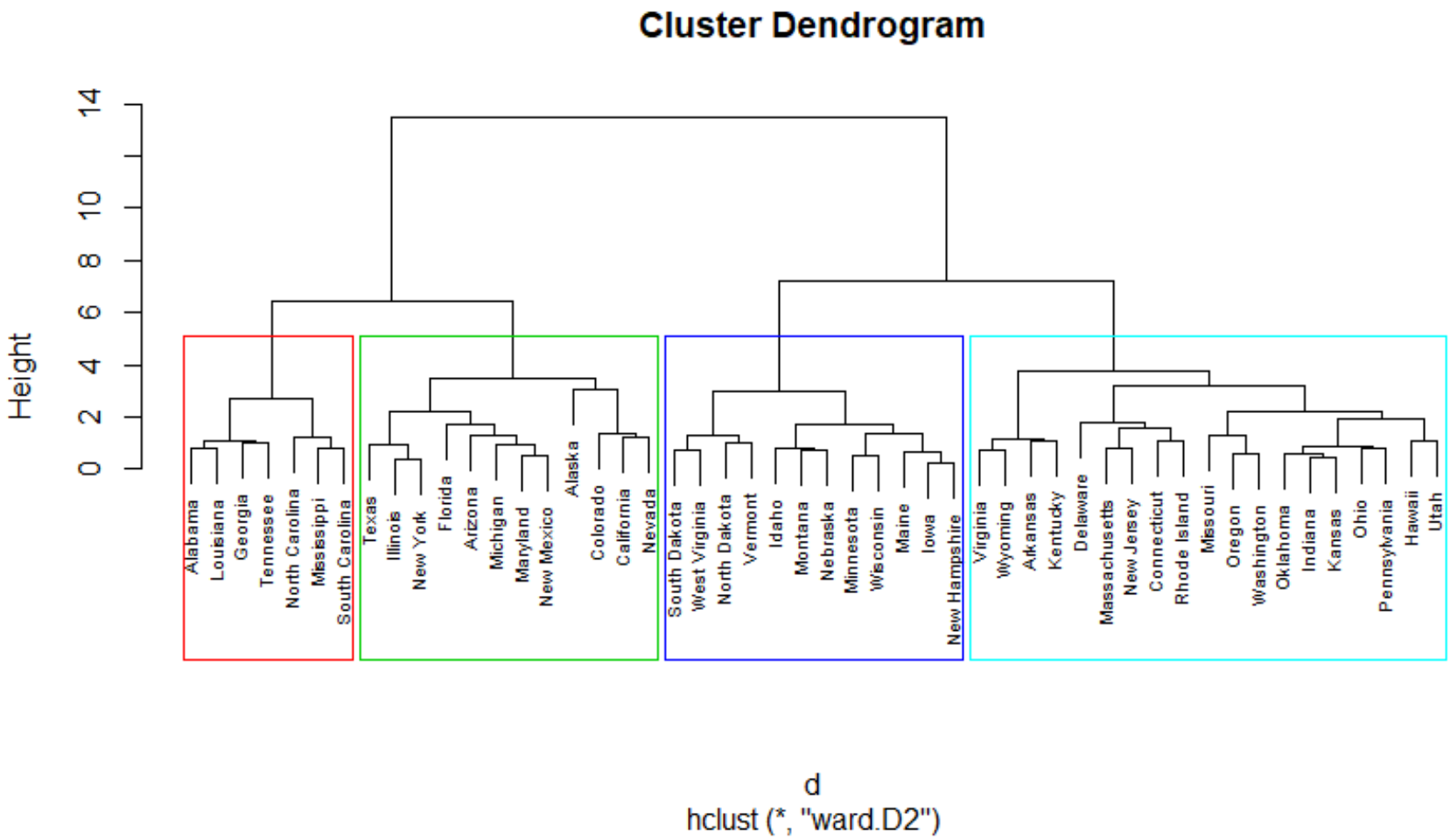
result

Murder Assault UrbanPop Rape
Alabama 1.24256408 0.7828393 -0.5209066 -0.003416473
Alaska 0.50786248 1.1068225 -1.2117642 2.484202941
Arizona 0.07163341 1.4788032 0.9989801 1.042878388
Arkansas 0.23234938 0.2308680 -1.0735927 -0.184916602
California 0.27826823 1.2628144 1.7589234 2.067820292
Colorado 0.02571456 0.3988593 0.8608085 1.864967207

hierarchical clustering

```
library("factoextra")
d <- dist(my_data, method = "euclidean")
res.hc <- hclust(d, method = "ward.D2" )
grp <- cutree(res.hc, k = 4)
plot(res.hc, cex = 0.6) # plot tree
rect.hclust(res.hc, k = 4, border = 2:5) # add rectangle
```

Plot


Cluster Dendrogram
Height
0
2
4
6
8
10
14
Alabama
Louisiana
Georgia
Tennessee
North Carolina
Mississippi
South Carolina
Texas
Illinois
New York
Florida
Arizona
Michigan
Maryland
New Mexico
Alaska
Colorado
California
Nevada
South Dakota
West Virginia
North Dakota
Vermont
Idaho
Montana
Nebraska
Minnesota
Wisconsin
Maine
Iowa
New Hampshire
Virginia
Wyoming
Arkansas
Kentucky
Delaware
Massachusetts
New Jersey
Connecticut
Rhode Island
Missouri
Oregon
Washington
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Hawaii
Utah
d
hclust (*, "ward.D2")

Non-hierarchical clustering: k-means

Notation

Within-cluster scatter matrix: $\mathbf{S}_W = \sum_{j=1}^{k} \mathbf{S}_j$, where

$$\mathbf{S}_j = \sum_{x \in C_j} \left(x - \bar{x}_{C_j}\right) \left(x - \bar{x}_{C_i}\right)^\intercal, \quad \bar{x}_{C_j} = \frac{1}{|C_j|} \sum_{x \in C_j} x, \quad i = 1, \ldots, k.$$

(Scatter matrix for the ith cluster)

Between-cluster scatter matrix:

$$\mathbf{S}_B = \sum_{j=1}^{k} |C_j| \left(\bar{x}_{C_j} - \bar{x}\right) \left(\bar{x}_{C_j} - \bar{x}\right)^\intercal, \quad \bar{x} = \frac{1}{n} \sum_{j=1}^{k} \sum_{x \in C_j} x.$$

Total scatter matrix:

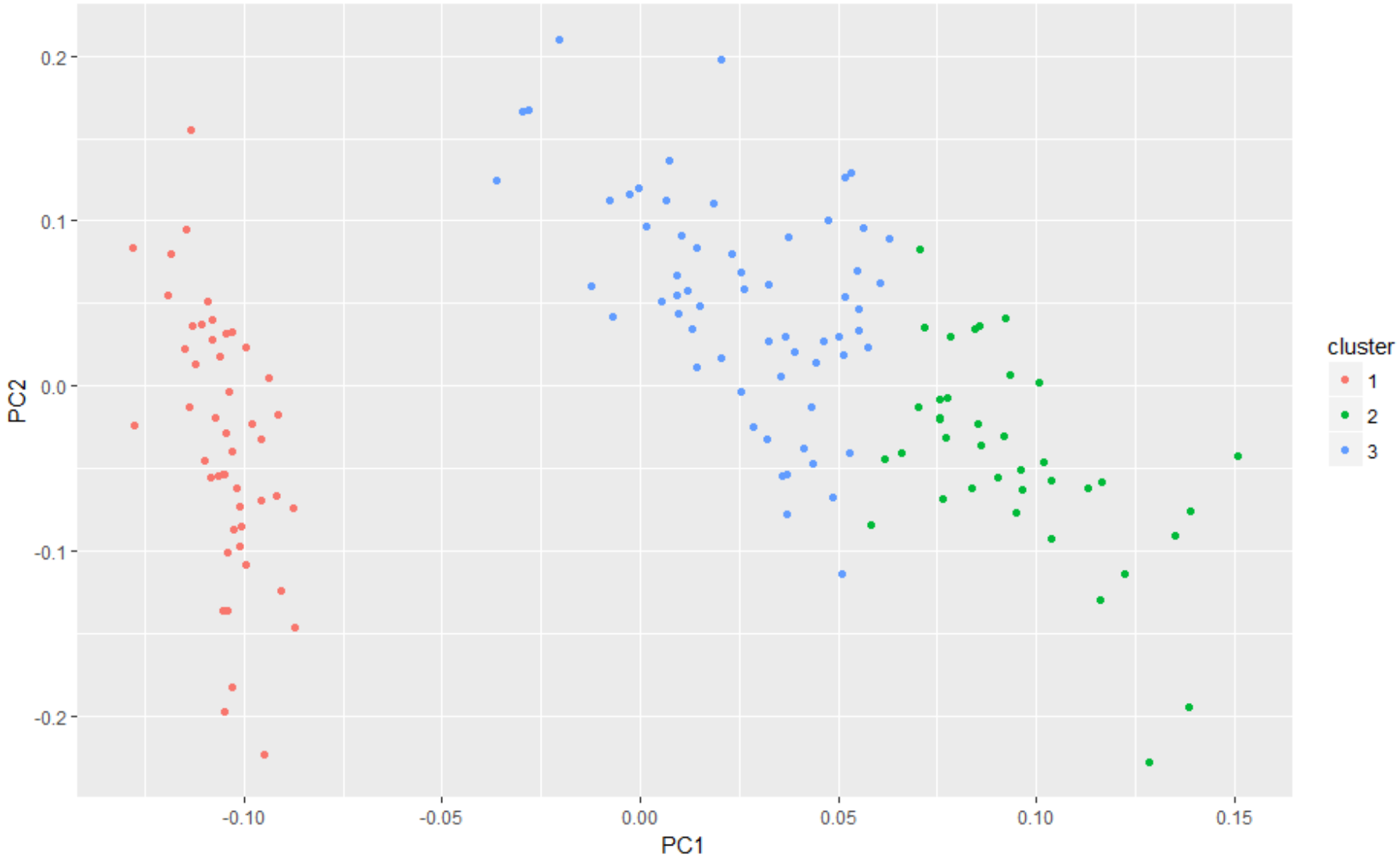
$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B = \sum_{x} \left(x - \bar{x}\right) \left(x - \bar{x}\right)^\intercal.$$

k-means

$$\min_{C_1,\ldots,C_k} \operatorname{tr} \mathbf{S}_W = \max_{C_1,\ldots,C_k} \operatorname{tr} \mathbf{S}_B = \min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{x \in C_j} d^2(x, \bar{x}_{C_j}).$$

1. Select k p-dimensional centroids or seeds (clusters).
2. Assign each observation to the nearest centroid using some L_p-norm, usually the Euclidean distance.
3. Reassign each observation to one of the k clusters based upon some criterion.
4. Stop if there is no reallocation of observations or if reassignment meets some convergence criterion; otherwise, return to Step 2.

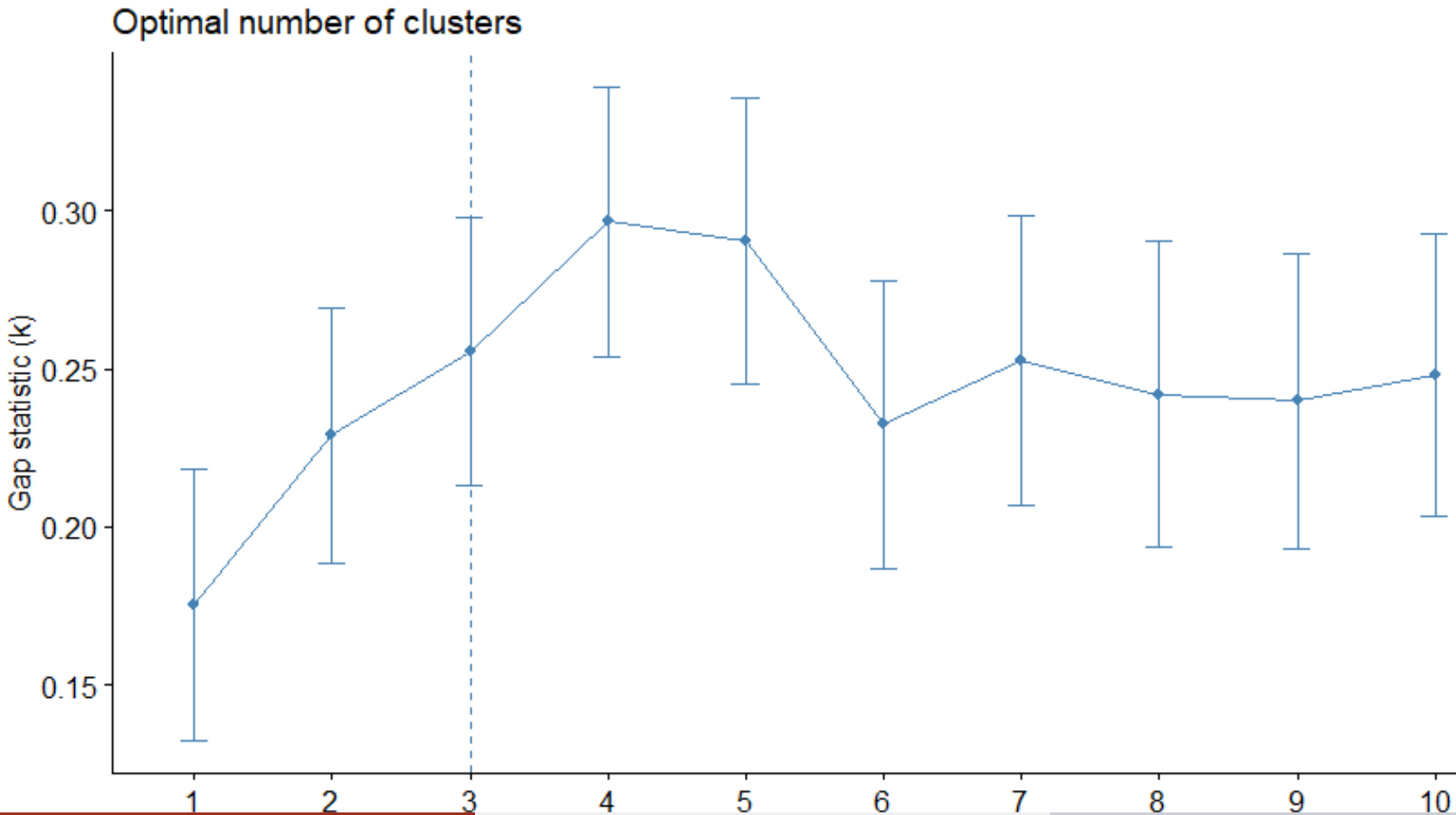
Example (iris dataset)



Using R...

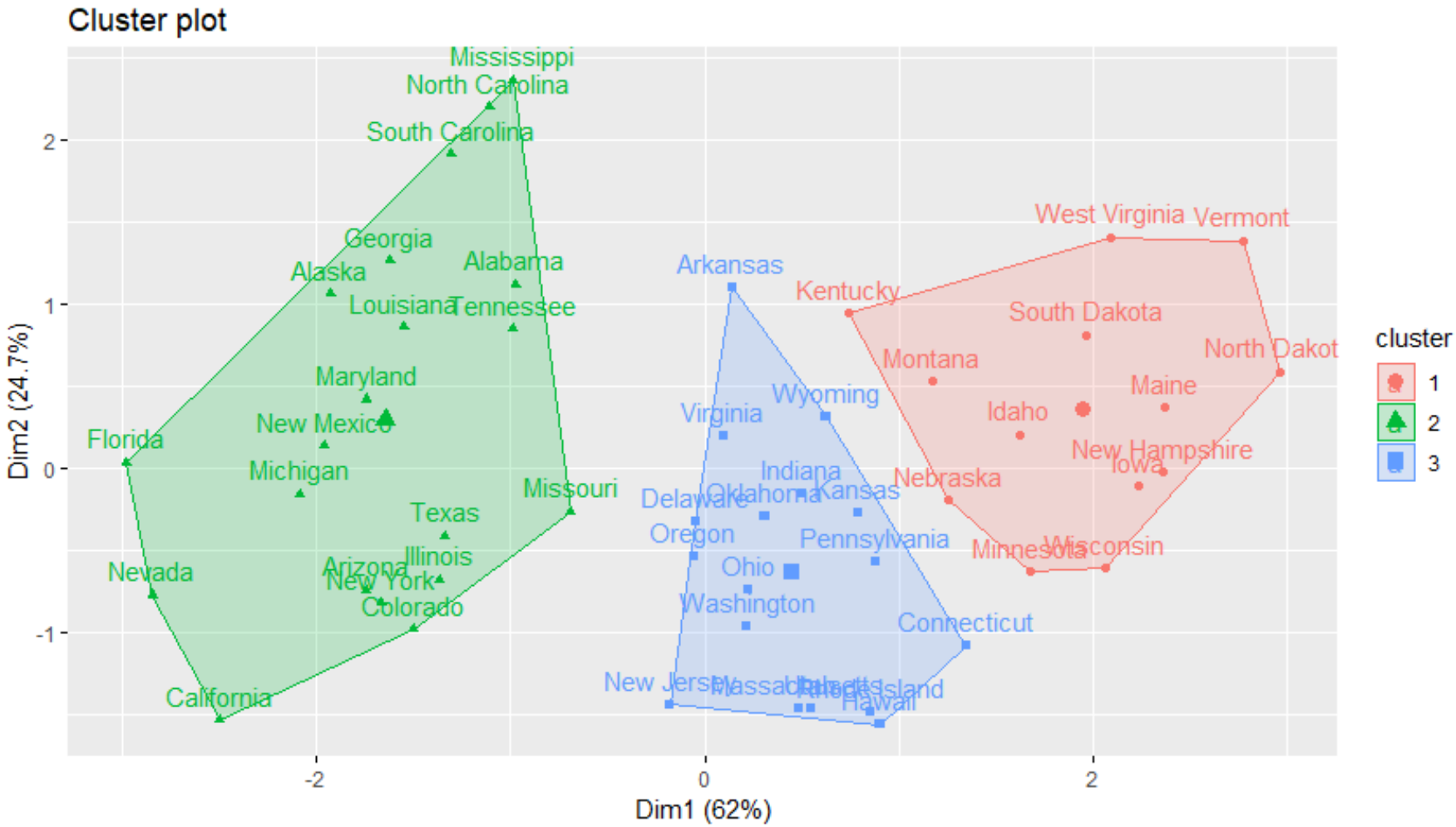
Optimal number of clusters

fviz_nbclust(my_data, kmeans, method = "gap_stat")



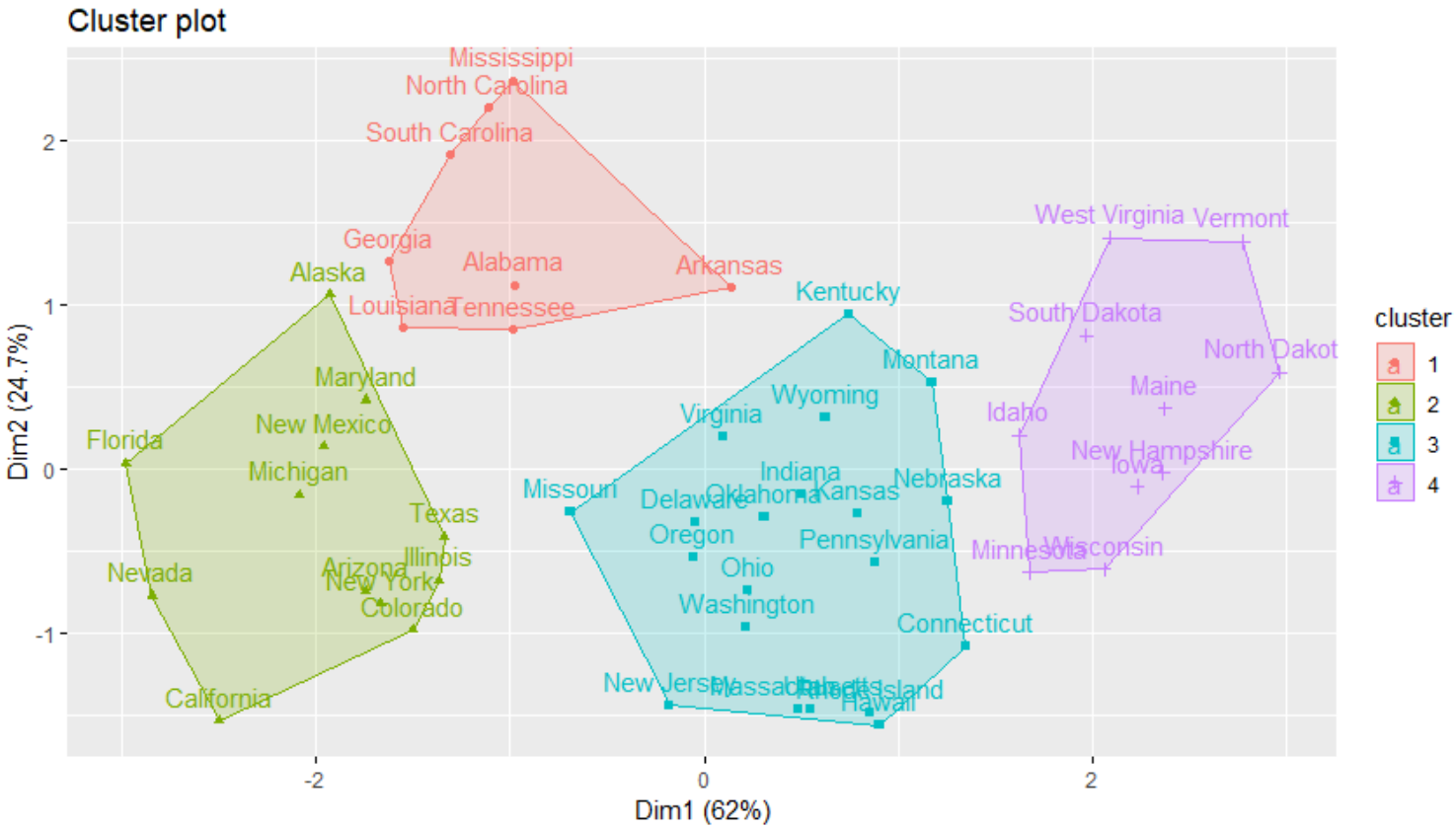
kmeans

```
km.res <- kmeans(my_data, 3, nstart = 25)
fviz_cluster(km.res, data = my_data, frame.type = "convex")
```



PAM

```
# Compute PAM library("cluster")
pam.res <- pam(my_data, 4)
# Visualize
fviz_cluster(pam.res)
```



Example

K-Means Cluster Analysis

```
fit <- kmeans(mydata, 5) # 5 cluster solution
# get cluster means
aggregate(mydata,by=list(fit$cluster),FUN=mean)
# append cluster assignment
mydata <- data.frame(mydata, fit$cluster)
```

Ward Hierarchical Clustering

```
d <- dist(mydata, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
plot(fit) # display dendogram
groups <- cutree(fit, k=5) # cut tree into 5 clusters
# draw dendogram with red borders around the 5 clusters
rect.hclust(fit, k=5, border="red")
```

Model Based

Model based approaches assume a variety of data models and apply maximum likelihood estimation and Bayes criteria to identify the most likely model and number of clusters. Specifically, the `Mclust( )` function in the `mclust` package selects the optimal model according to BIC for EM initialized by hierarchical clustering for parameterized Gaussian mixture models. One chooses the model and number of clusters with the largest BIC. See `help(mclustModelNames)` to details on the model chosen as best.

Model Based Clustering

```
library(mclust)
fit <- Mclust(mydata)
plot(fit) # plot results
summary(fit) # display the best model
```

Plotting Cluster Solutions

It is always a good idea to look at the cluster results.

plotting

```
# K-Means Clustering with 5 clusters
fit <- kmeans(mydata, 5)
# Cluster Plot against 1st 2 principal components
# vary parameters for most readable graph
library(cluster)
clusplot(mydata, fit$cluster, color=TRUE, shade=TRUE, labels=2,
lines=0)
# Centroid Plot against 1st 2 discriminant functions
library(fpc)
plotcluster(mydata, fit$cluster)
```

Validating cluster solutions

The function `cluster.stats( )` in the `fpc` package provides a mechanism for comparing the similarity of two cluster solutions using a variety of validation criteria (Hubert's gamma coefficient, the Dunn index and the corrected rand index).

Comparing 2 cluster solutions

```
library(fpc)
cluster.stats(d, fit1$cluster, fit2$cluster)
```

where `d` is a distance matrix among objects, and `fit1$cluster` and `fit$cluster` are integer vectors containing classification results from two different clusterings of the same data.

Useful links:

https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html

http://www.sthda.com/english/wiki/print.php?id=234

Datasets

file01.txt Sightings of Minor Planets, 5 columns, 19 rows. Some minor planets may have been sighted more than once. In the table, sightings thought to be of the same planet are listed together.

file02.txt Animal Milk Constituent Percentages. 5 columns, 16 rows. A list of animals, and the constituents of their milk.

file07.txt Expectations of Life by Country, Age and Sex. 8 columns, 31 rows.

iris Some characteristics of three types of iris.