Kitakado’s Lecture Series

Clustering methods: Part (2) Exercise

Toshihide Kitakado (TUMSAT)

Relased on 2021/01/17

Abel is still thinking. I should be doing so.
Photo taken at University of Oslo in 2003.

このスライドは部分的に次の授業素材として共通利用しています．一部，日本語と英語が混ざっていますが御容赦下さい．

- 「生物資源モデリング」（学部）
- 「生物資源解析学演習」（学部）
- 「資源動態・管理学(E)」（大学院）
- 「データサイエンス概論(E)」（卓越大学院）

Example analysis: morphometric data of horse mackerel

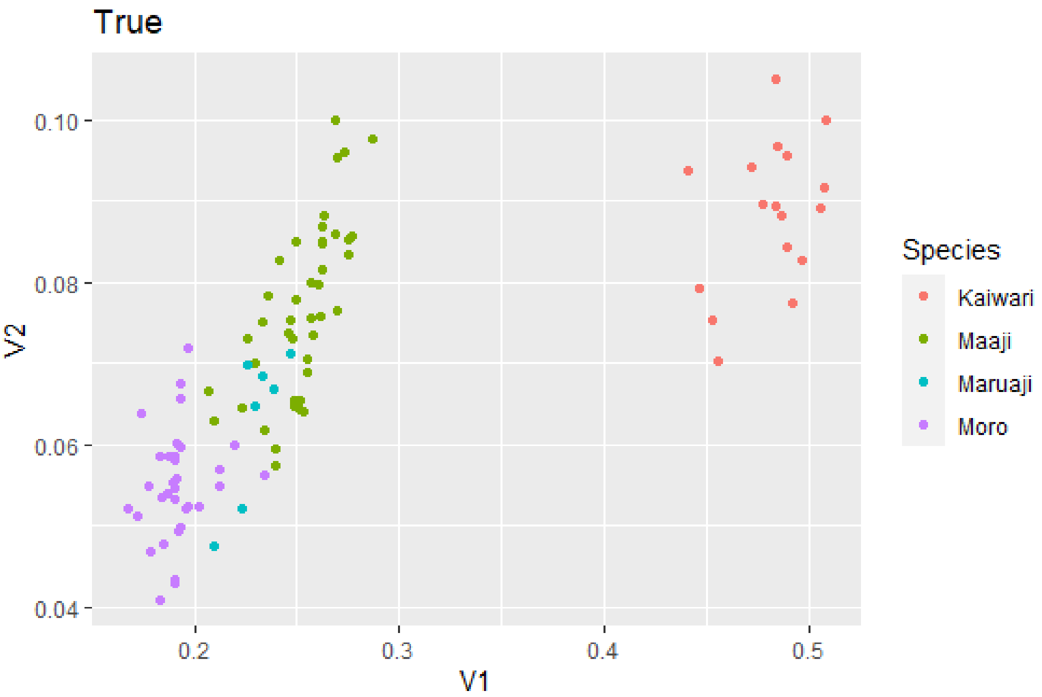
Data preparation

```
library(tidyverse)
library(gridExtra)
```

```
tmp <- read.csv("Data/Morph_horsemackerel.csv",
header=T)
Species <- tmp$Species
V1 <- tmp$BD/tmp$SL
V2 <- tmp$ED/tmp$SL
Data <- data.frame(V1,V2)
Data.gg <- cbind(Data, Species)
Data <- scale(Data)
dim(Data)
```

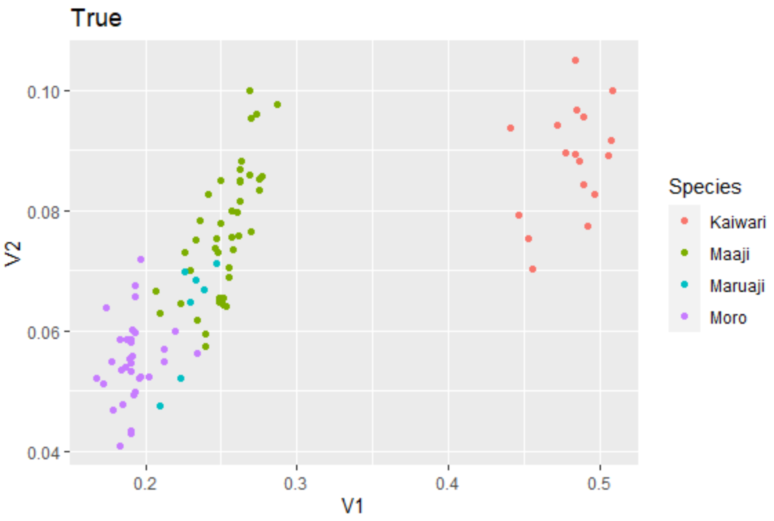
```
[1] 175   2
```

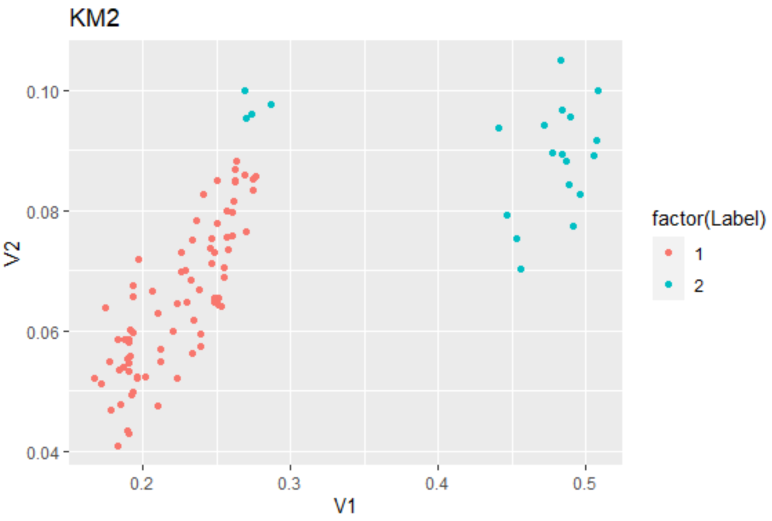
```
clust.true <- ggplot(Data.gg,
aes(x=V1,y=V2,col=Species)) + geom_point() +
ggtitle("True")
```


True
V2
0.10
0.08
0.06
0.04
V1
0.2
0.3
0.4
0.5
Species
Kaiwari
Maaji
Maruaji
Moro

Clustering: k-means method with K=2

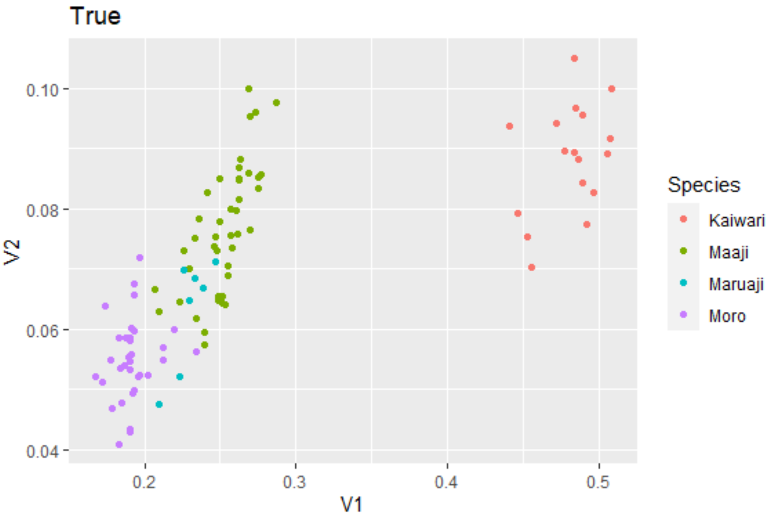
```
Res.km2 <- Res <- kmeans(Data, 2, nstart=10)
clust.km2 <- Data.gg %>% mutate(Label=Res$cluster) %>%
  ggplot(aes(V1,V2,col=factor(Label))) + geom_point()
+ ggtitle("KM2")

grid.arrange(clust.true, clust.km2, nrow=1)
```


True
0.10
0.08
0.06
0.04
V2
0.2
0.3
0.4
0.5
V1
Species
Kaiwari
Maaji
Maruaji
Moro


KM2
0.10
0.08
0.06
0.04
V2
0.2
0.3
0.4
0.5
V1
factor(Label)
1
2

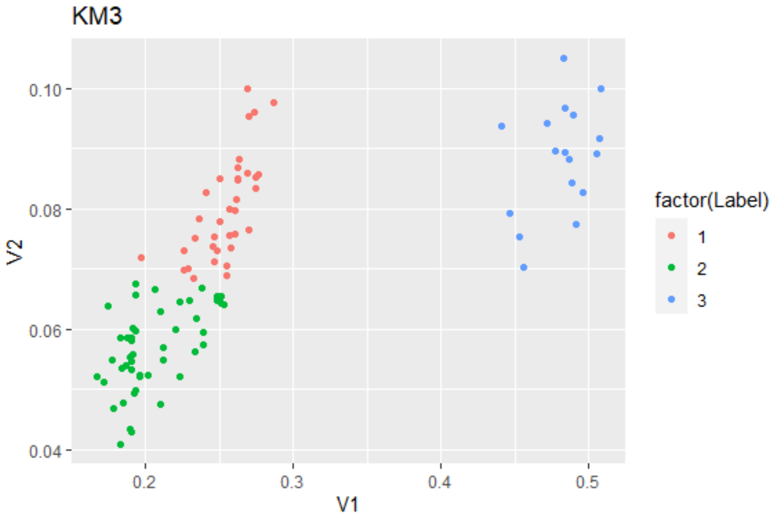
Clustering: k-means method with K=3

```
Res.km3 <- Res <- kmeans(Data, 3, nstart=10)
clust.km3 <- Data.gg %>% mutate(Label=Res$cluster) %>%
  ggplot(aes(V1,V2,col=factor(Label))) + geom_point()+
ggtitle("KM3")

grid.arrange(clust.true, clust.km3, nrow=1)
```



True

V2
0.10
0.08
0.06
0.04
0.2
0.3
0.4
0.5
V1
Species
● Kaiwari
● Maaji
● Maruaji
● Moro

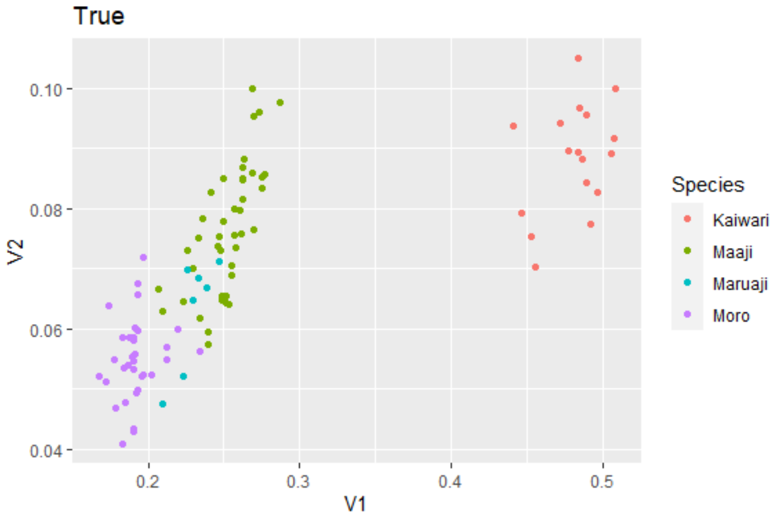


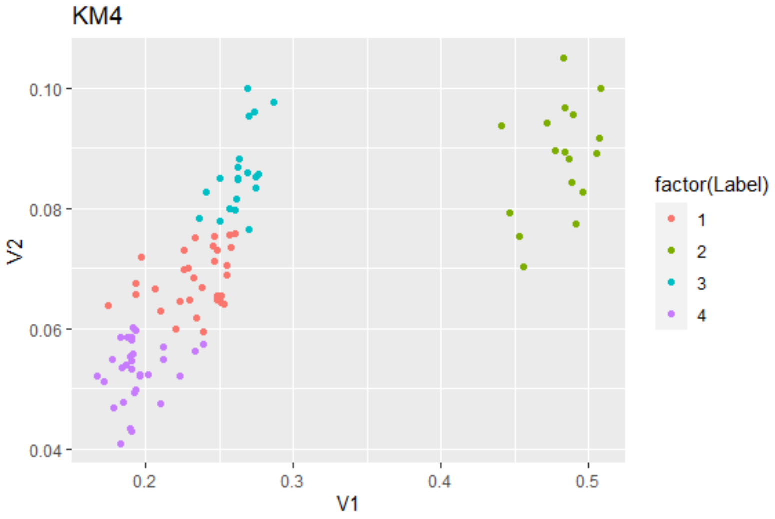
KM3

V2
0.10
0.08
0.06
0.04
0.2
0.3
0.4
0.5
V1
factor(Label)
● 1
● 2
● 3

Clustering: k-means method with K=4

```
Res.km4 <- Res <- kmeans(Data, 4, nstart=10)
clust.km4 <- Data.gg %>% mutate(Label=Res$cluster) %>%
  ggplot(aes(V1,V2,col=factor(Label))) + geom_point()
+ ggtitle("KM4")

grid.arrange(clust.true, clust.km4, nrow=1)
```

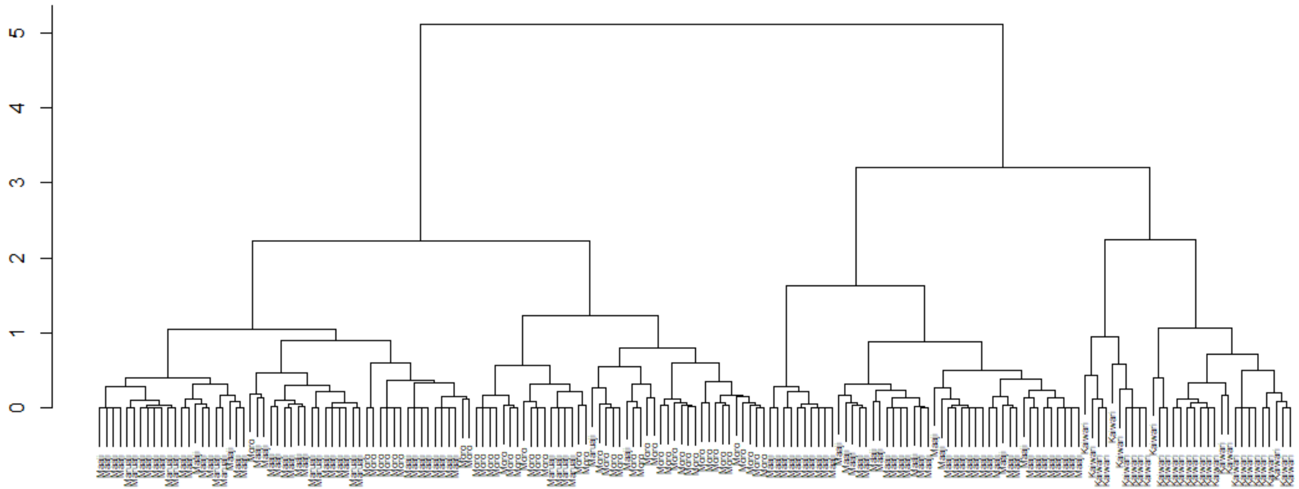

True
V2
0.10
0.08
0.06
0.04
0.2
0.3
0.4
0.5
V1
Species
Kaiwari
Maaji
Maruaji
Moro


KM4
V2
0.10
0.08
0.06
0.04
0.2
0.3
0.4
0.5
V1
factor(Label)
1
2
3
4

Hierarchical clustering with “complete” linkage

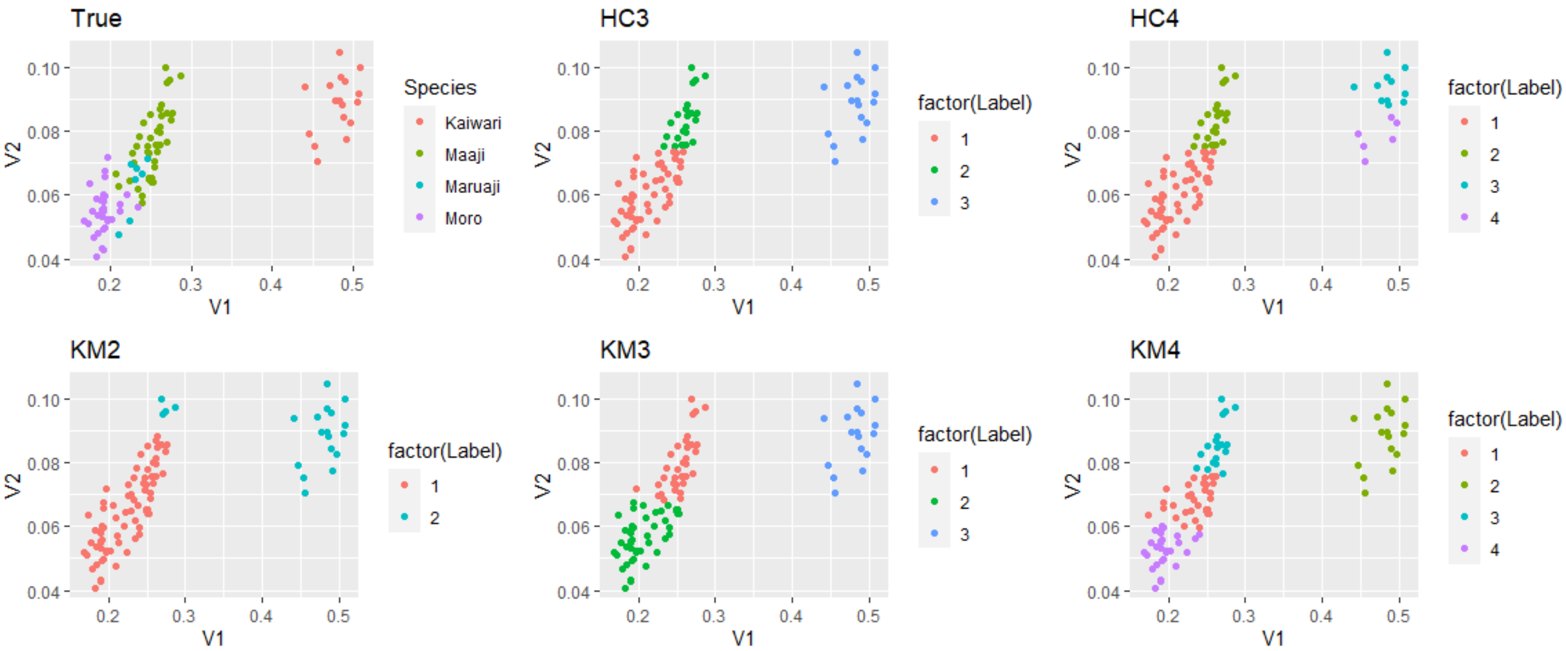
```
Res.hc.comp <- hclust(dist(Data), method="complete")
plot(Res.hc.comp, main="Complete Linkage",
labels=Species, xlab="", ylab="", sub="", cex=0.5)
```

Complete Linkage



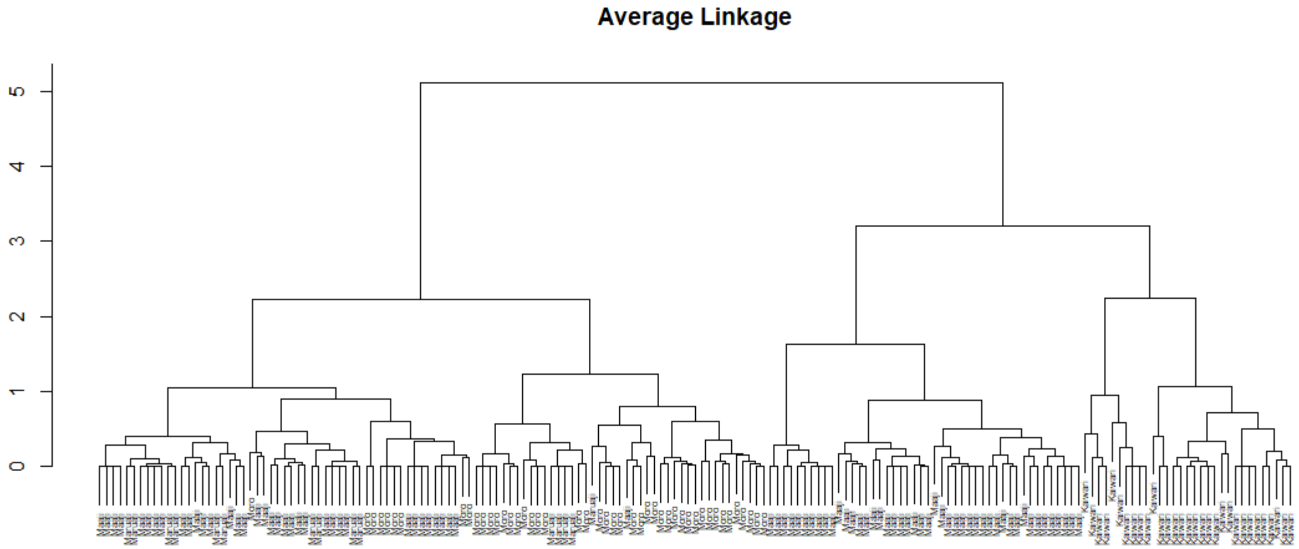
Hierarchical clustering with “complete” linkage with K=3, 4

```
clust.hc3 <- Data.gg %>%
mutate(Label=cutree(Res.hc.comp,k=3)) %>%
  ggplot(aes(V1,V2,col=factor(Label))) + geom_point()
+ ggtitle("HC3")

clust.hc4 <- Data.gg %>%
mutate(Label=cutree(Res.hc.comp,k=4)) %>%
  ggplot(aes(V1,V2,col=factor(Label))) + geom_point()
+ ggtitle("HC4")

grid.arrange(clust.true, clust.hc3, clust.hc4,
clust.km2, clust.km3, clust.km4, nrow=2)
```


True
Species
Kaiwari
Maaji
Maruaji
Moro
HC3
HC4
KM2
KM3
KM4
factor(Label)
1
2
3
4
V1
V2
0.2
0.3
0.4
0.5
0.04
0.06
0.08
0.10

Hierarchical clustering with “average” linkage

```
Res.hc.ave <- hclust(dist(Data), method="average")
plot(Res.hc.comp, main="Average Linkage",
labels=Species, xlab="", ylab="", sub="", cex=0.5)
```


Average Linkage
0
1
2
3
4
5
Maaji
Noro
Kaiwari

Exercise (1)

- Compare results of “complete” and “average” linkage in the hierarchical clustering.

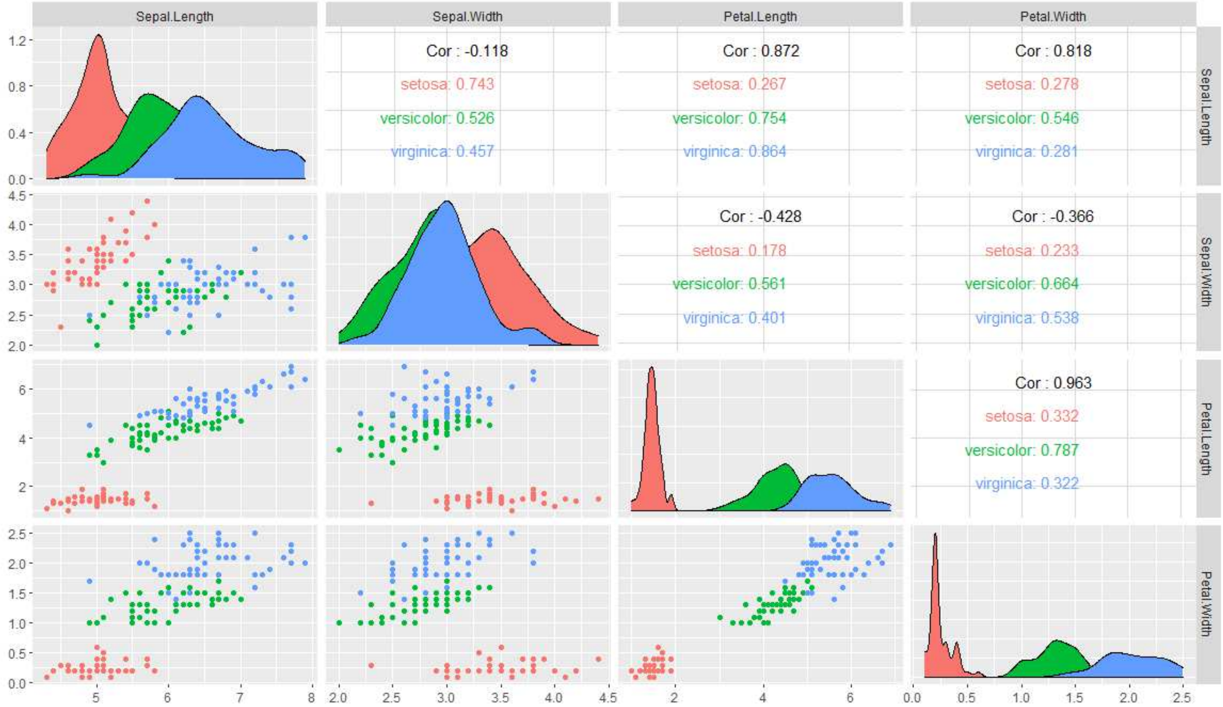
Example analyses: Iris data (アヤメデータ)

- Sepal (がく片)
- Petal (花弁)

```
head(iris, 3)

  Sepal.Length Sepal.Width Petal.Length Petal.Width
Species
1          5.1         3.5          1.4         0.2
setosa
2          4.9         3.0          1.4         0.2
setosa
3          4.7         3.2          1.3         0.2
setosa
```

Visual presentation of data

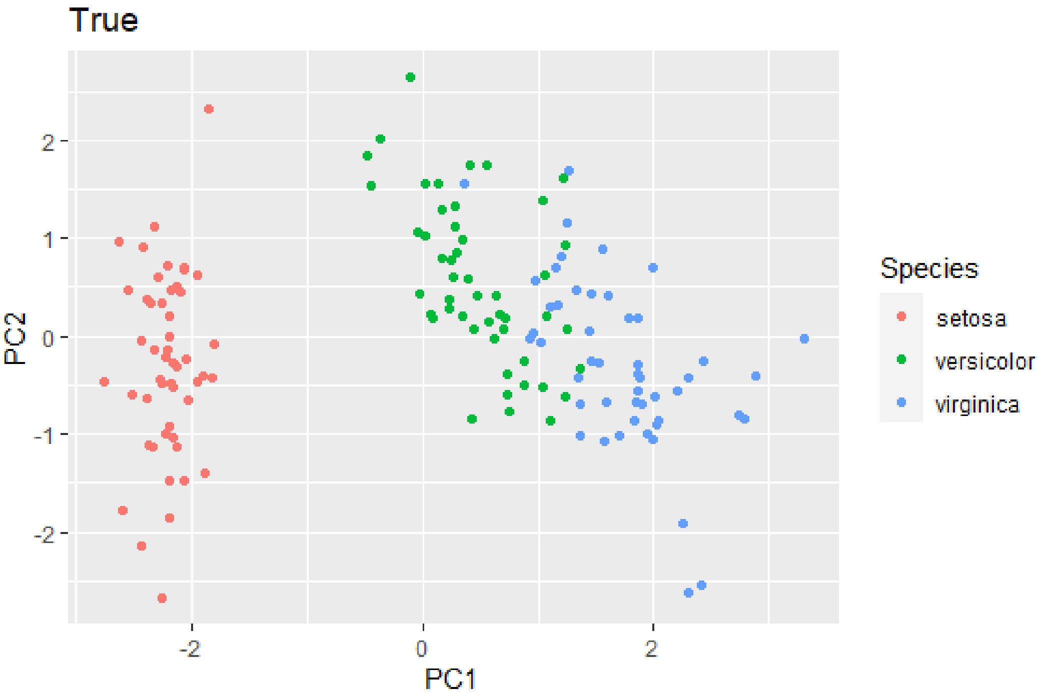
```r
library(GGally)
ggpairs(iris, columns=1:4, aes(col=Species))
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Cor : -0.118
setosa: 0.743
versicolor: 0.526
virginica: 0.457
Cor : 0.872
setosa: 0.267
versicolor: 0.754
virginica: 0.864
Cor : 0.818
setosa: 0.278
versicolor: 0.546
virginica: 0.281
Cor : -0.428
setosa: 0.178
versicolor: 0.561
virginica: 0.401
Cor : -0.366
setosa: 0.233
versicolor: 0.664
virginica: 0.538
Cor : 0.963
setosa: 0.332
versicolor: 0.787
virginica: 0.322
1.2
0.8
0.4
0.0
4.5
4.0
3.5
3.0
2.5
2.0
6
4
2
2.5
2.0
1.5
1.0
0.5
0.0
5
6
7
8
2.0
2.5
3.0
3.5
4.0
4.5
2
4
6
0.0
0.5
1.0
1.5
2.0
2.5

Dimension reduction by PCA

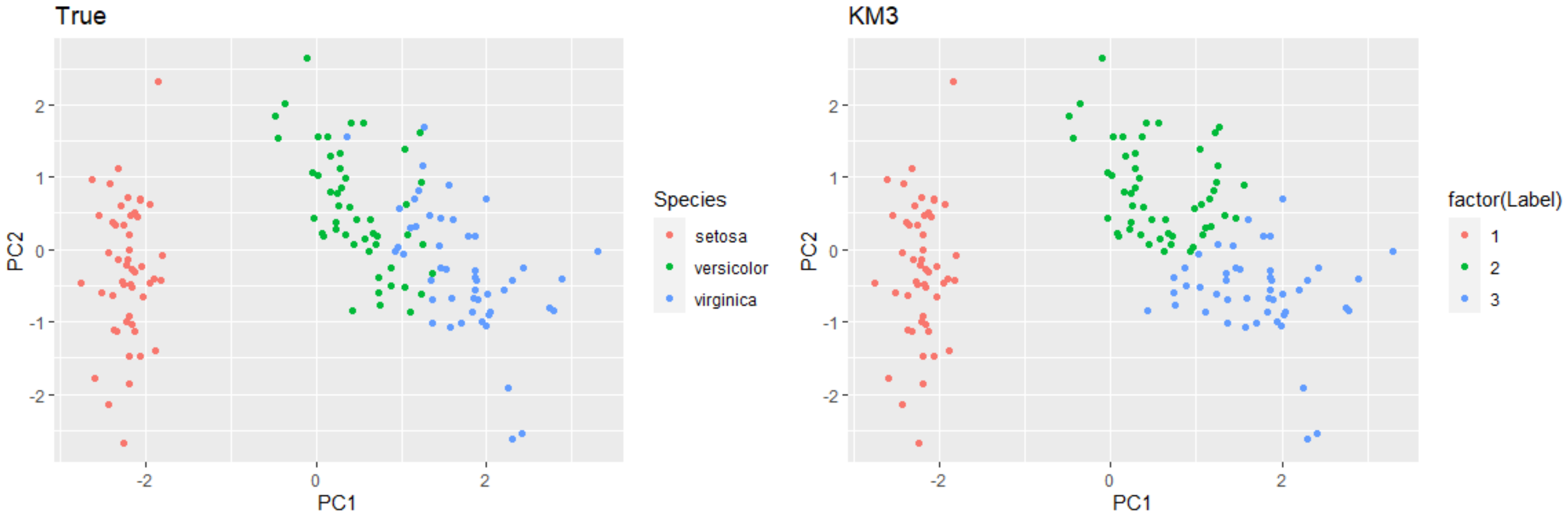
Let us reduce the dimension of data from 4 to 2 so that we can draw the data on the plain.

```
Data <- scale(iris[,-5])
Species <- iris$Species
Res.pca <- prcomp(Data)
DF <- data.frame(Res.pca$x[,1:2],
Species=iris$Species)
clust.true <- DF %>%
ggplot(aes(x=PC1,y=PC2,col=Species)) + geom_point() +
ggtitle("True")
clust.true
```


True
PC2
2
1
0
-1
-2
-2
0
2
PC1
Species
● setosa
● versicolor
● virginica

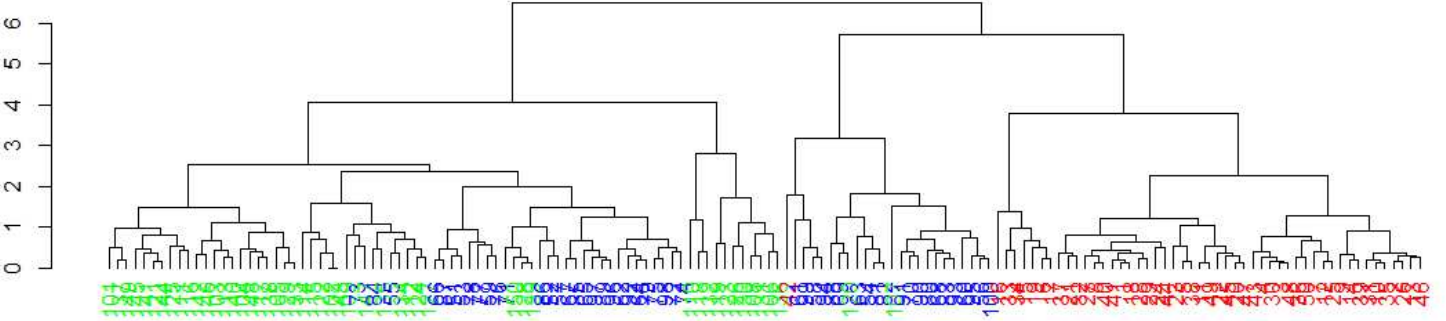
Clustering: K-means method with K=3

```
Res.km3 <- Res <- kmeans(Data, 3, nstart=10); Res.km3
K-means clustering with 3 clusters of sizes 50, 53, 47

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1  -1.01119138  0.85041372   -1.3006301  -1.2507035
2  -0.05005221 -0.88042696    0.3465767   0.2805873
3   1.13217737  0.08812645    0.9928284   1.0141287

Clustering vector:
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

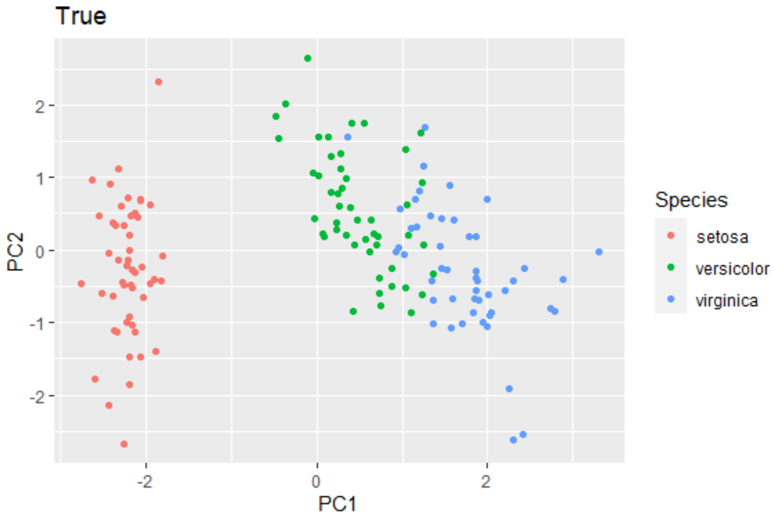

True
PC2
2
1
0
-1
-2
-2
0
2
PC1
Species
setosa
versicolor
virginica
KM3
PC2
2
1
0
-1
-2
-2
0
2
PC1
factor(Label)
1
2
3

Hierarchical clustering with “complete” linkage with K=3

```
Res.hc.comp <- Res <- hclust(dist(Data),
method="complete")
Den <- as.dendrogram(Res)
COL <- c("red","blue","green")
Label.col <- COL[Species][order.dendrogram(Den)]
Den %>% dendextend::set("labels_colors",
value=Label.col) %>% plot()
```

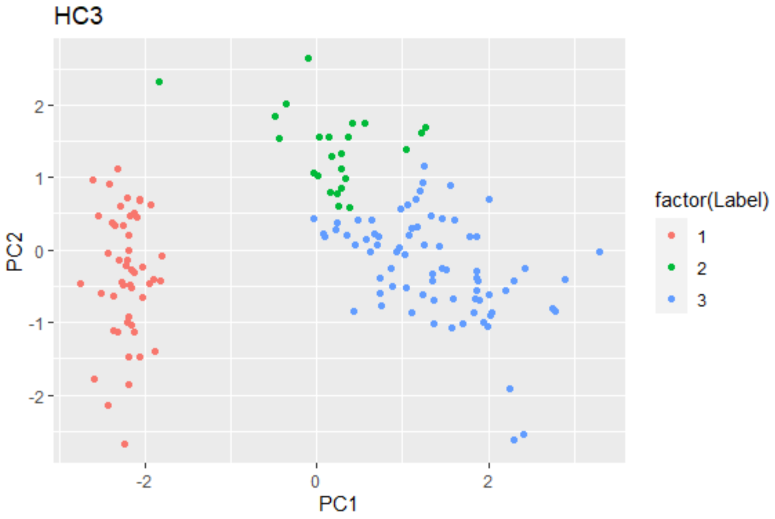

0
1
2
3
4
5
6

```
clust.hc3 <- DF %>%
mutate(Label=cutree(Res.hc.comp,k=3)) %>%
 ggplot(aes(PC1,PC2,col=factor(Label))) +
geom_point() + ggtitle("HC3")

grid.arrange(clust.true, clust.hc3, nrow=1)
```



True

PC2
2
1
0
-1
-2
-2
0
2
PC1
Species
● setosa
● versicolor
● virginica

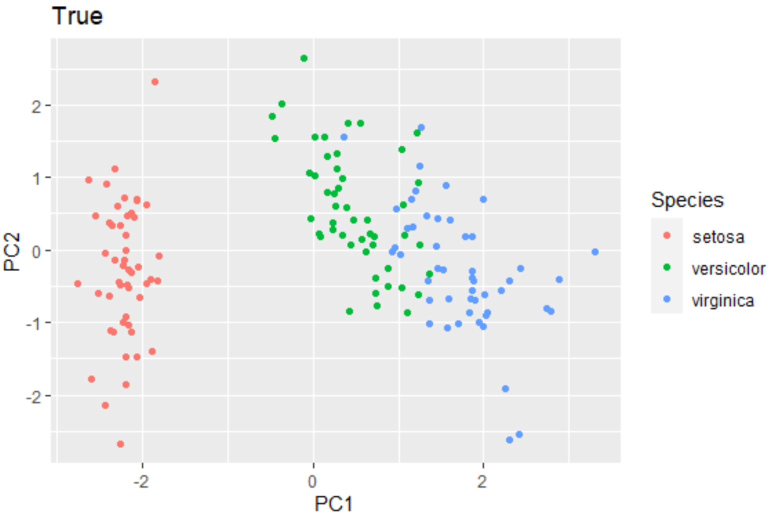


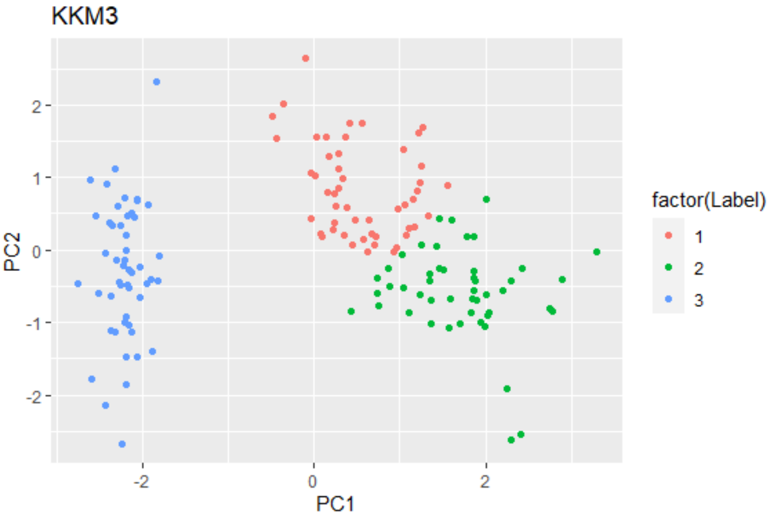
HC3

PC2
2
1
0
-1
-2
-2
0
2
PC1
factor(Label)
● 1
● 2
● 3

Kernel K-means method

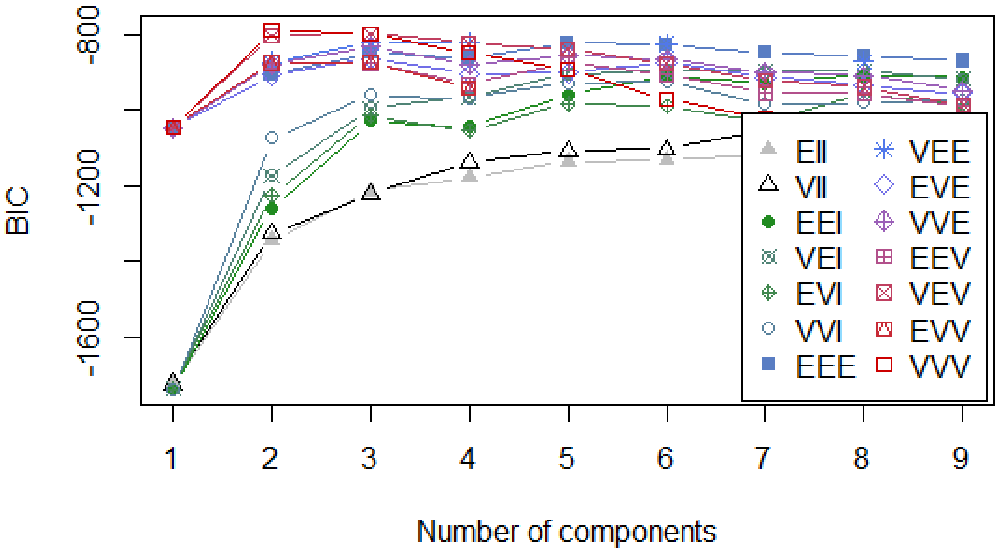
```
library(kernlab)
Res.kkm3 <- Res <- kkmeans(Data, 3, kernel="rbfdot");
Res.kkm3

Using automatic sigma estimation (sigest) for RBF or
laplace kernel

Spectral Clustering object of class "specc"


 Cluster memberships:


3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 1
1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 2 2 2 1 1 1
1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2
```


True
PC2
PC1
2
1
0
-1
-2
-2
0
2
Species
setosa
versicolor
virginica


KKM3
PC2
PC1
2
1
0
-1
-2
-2
0
2
factor(Label)
1
2
3

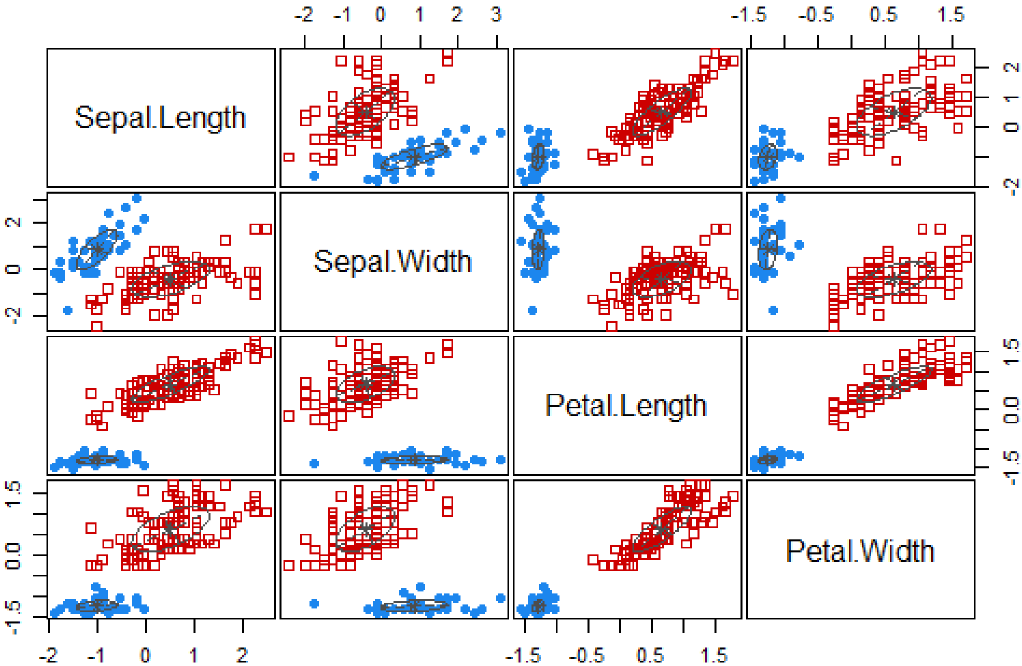
Mixture models (1) Best model (G=2)

```
library(mclust)
Res.mc <- Mclust(Data);
plot(Res.mc, what="BIC")
```


BIC
-800
-1200
-1600
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV

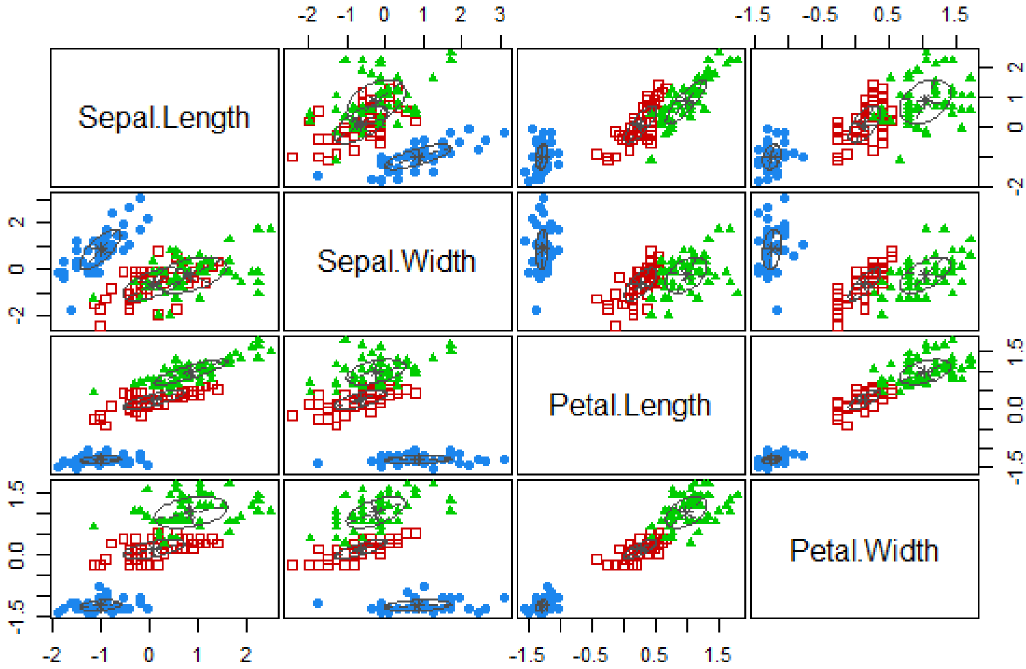
Mixture models (1 Continued) Best model (G=2)

```
plot(Res.mc, what="classification")
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
-2
-1
0
1
2
3
-1.5
-0.5
0.5
1.5
0.0

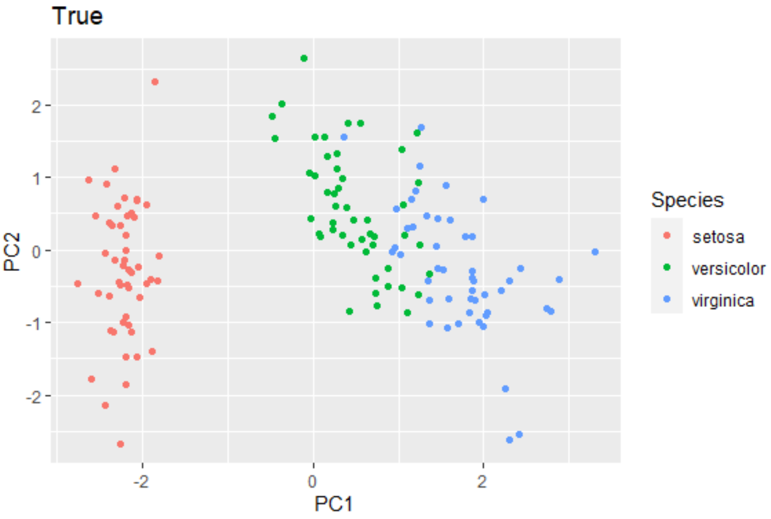
Mixture models (2) Model with G=3

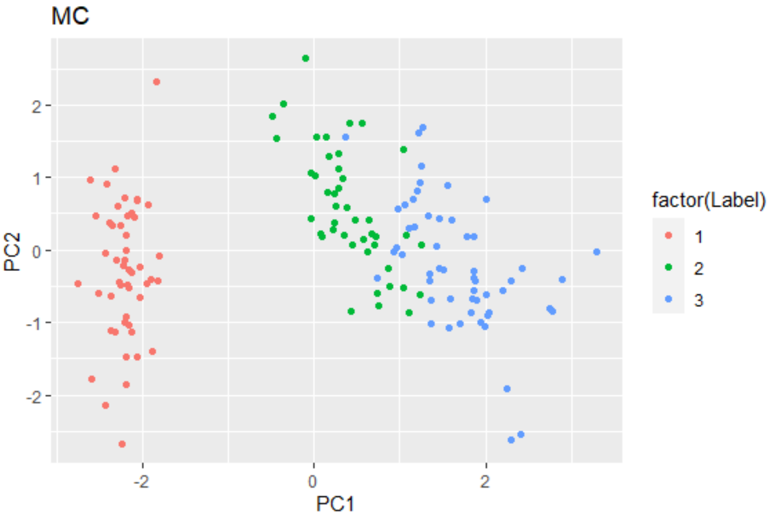
```
Res.mc3 <- Mclust(Data, G=3);
plot(Res.mc3, what="classification")
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
-2
-1
0
1
2
3
-1.5
-0.5
0.5
1.5
2
1
0
-2
0.0

Mixture models (3) plots

```
clust.mc3 <- DF %>% mutate(Label=Res.mc3$class) %>%
  ggplot(aes(PC1,PC2,col=factor(Label))) +
geom_point() + ggtitle("MC")
grid.arrange(clust.true, clust.mc3, nrow=1)
```


True
PC2
2
1
0
-1
-2
-2
0
2
PC1
Species
setosa
versicolor
virginica


MC
PC2
2
1
0
-1
-2
-2
0
2
PC1
factor(Label)
1
2
3

Example analyses: gene measurements between healthy and deseased individuals

Data preparation

Data consists of 40 tissue samples with measurements on a total of 1000 genes (from Chapter 10, ISL). Among 40 tissues, the first 20 samples are from a healthy group, while the second 20 samples are from a diseased group.

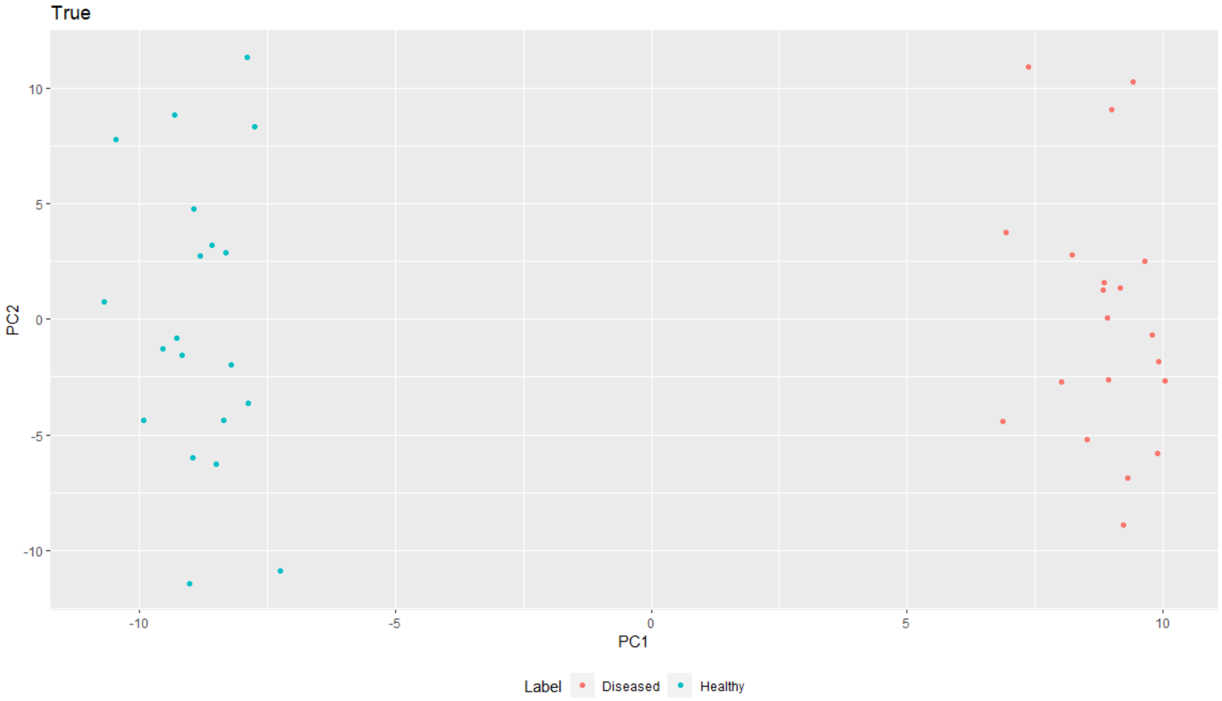
```
Data <- read.csv("Data/ISL_Ch10Ex11_cancer.csv",
header=T)
ID <- colnames(Data)
Label <- c(rep("Healthy",20), rep("Diseased",20))
Data <- t(Data)
Data <- scale(Data)
dim(Data)
```

```
[1]   40 1000
```

PCA analysis

To reduce 1000 dimension data to 2 dimension data by PCA to show the information of data.

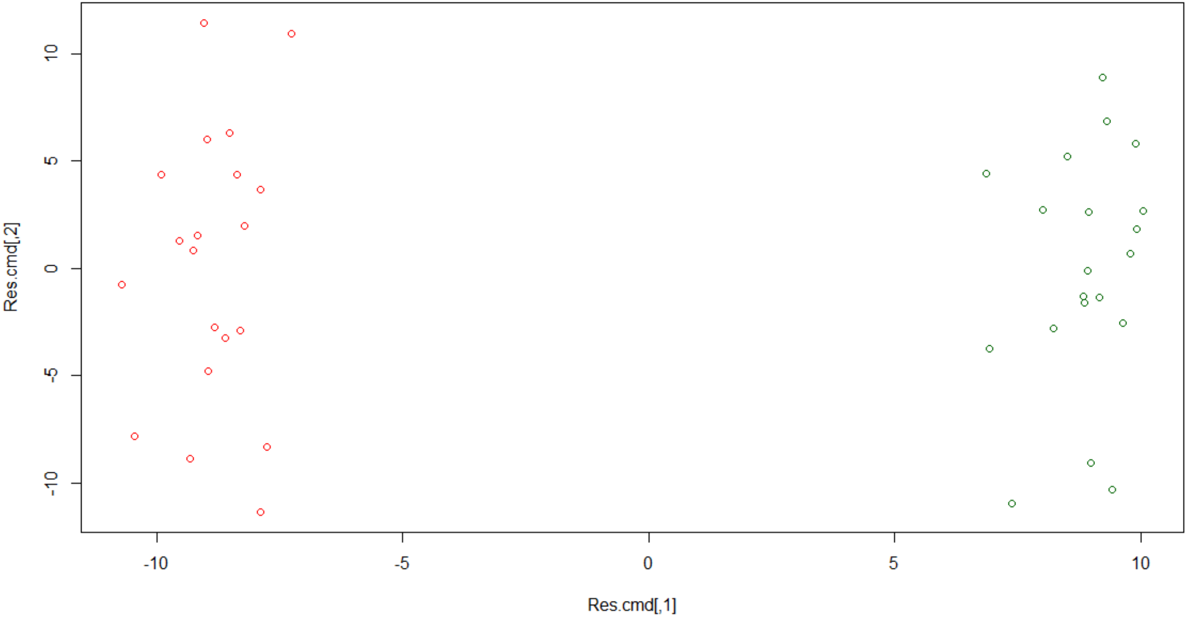
```
Res.pca <- prcomp(Data)
DF <- data.frame(Res.pca$x[,1:2], Label)
clust.true <- DF %>%
ggplot(aes(x=PC1,y=PC2,col=Label)) + geom_point() +
  ggtitle("True") + theme(legend.position="bottom")
clust.true
```


True
PC2
10
5
0
-5
-10
PC1
-10
-5
0
5
10
Label
● Diseased
● Healthy

MDS analysis

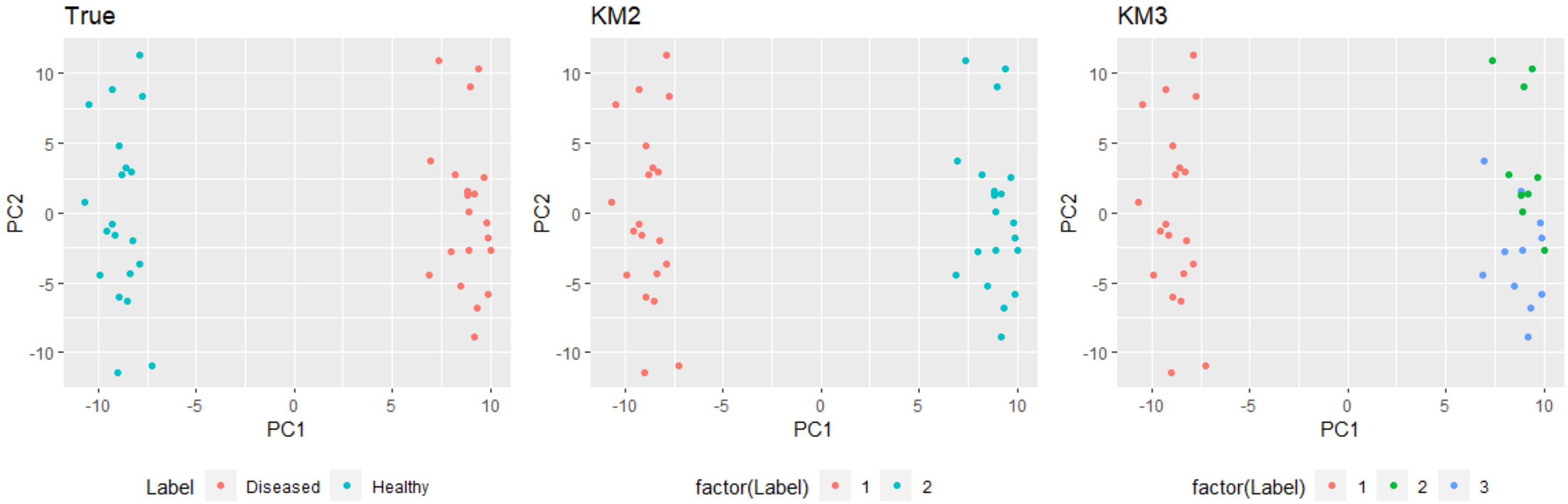
Another methods for dimension reduction.

```
Dist.Data <- dist(Data, method = "euclidean")
Res.cmd <- cmdscale(Dist.Data, 2)
COL = c(rep("red",20), rep("darkgreen",20))
plot(Res.cmd, col=COL)
```


Res.cmd[,2]
10
5
0
-5
-10
-10
-5
0
5
10
Res.cmd[,1]

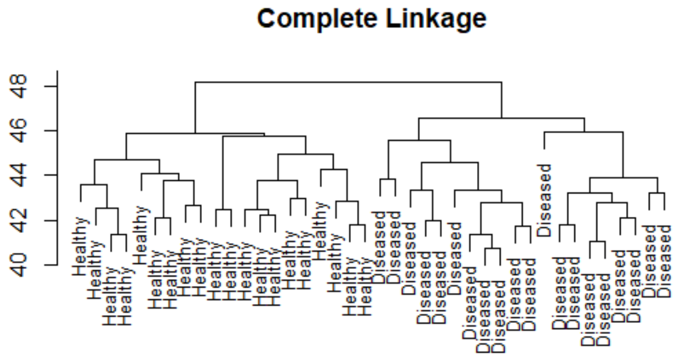
Clustering: k-means method with K=2 and 3

```
Res.km2 <- Res <- kmeans(Data, 2, nstart=10)
clust.km2 <- DF %>% mutate(Label=Res$cluster) %>%
  ggplot(aes(PC1,PC2,col=factor(Label))) +
geom_point() +
  ggtitle("KM2") + theme(legend.position="bottom")

Res.km3 <- Res <- kmeans(Data, 3, nstart=10)
clust.km3 <- DF %>% mutate(Label=Res$cluster) %>%
  ggplot(aes(PC1,PC2,col=factor(Label))) +
geom_point() +
  ggtitle("KM3") + theme(legend.position="bottom")

grid.arrange(clust.true, clust.km2, clust.km3, nrow=1)
```

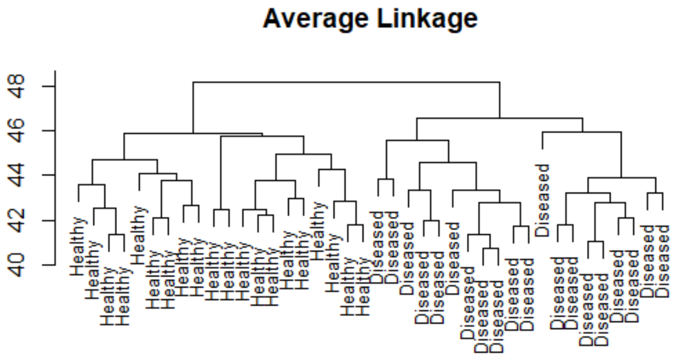

True
KM2
KM3
PC2
PC1
10
5
0
-5
-10
-10
-5
0
5
10
Label
● Diseased
● Healthy
factor(Label)
● 1
● 2
● 3

Hierarchical clustering with “complete” and “average” linkage

```
Res.hc.comp <- hclust(dist(Data), method="complete")
Res.hc.ave <- hclust(dist(Data), method="average")
par(mfrow=c(1,2))
plot(Res.hc.comp, main="Complete Linkage",
labels=Label, xlab="", ylab="", sub="", cex=0.8)
plot(Res.hc.comp, main="Average Linkage",
labels=Label, xlab="", ylab="", sub="", cex=0.8)
```



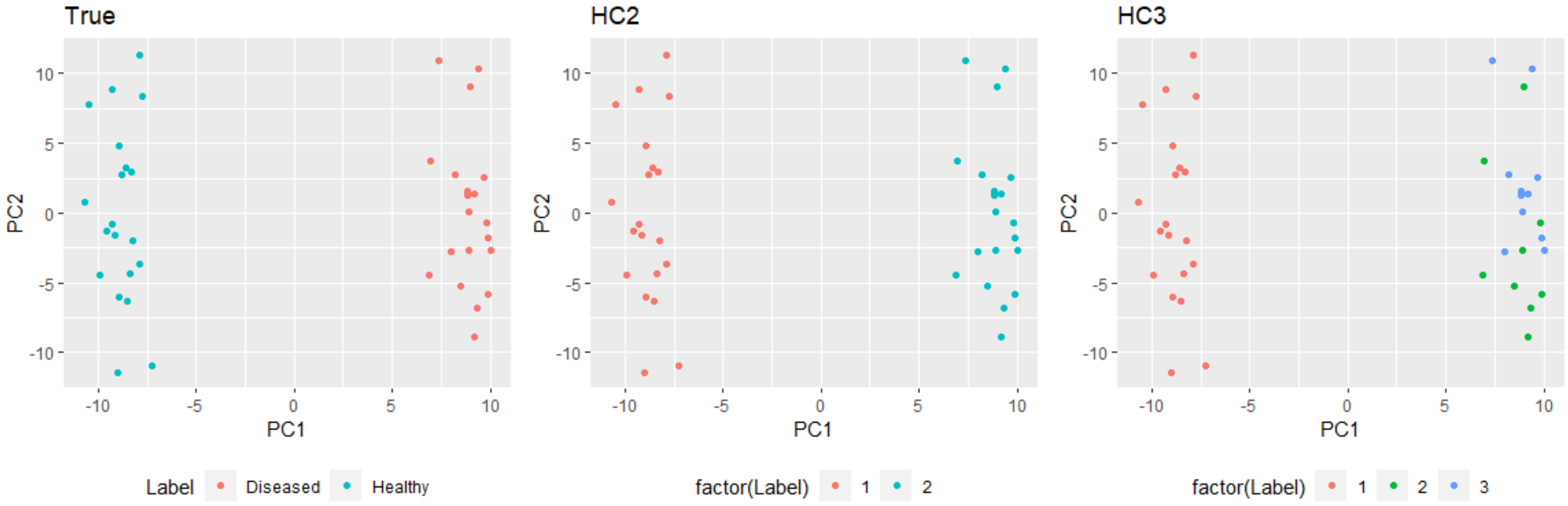
Complete Linkage



Average Linkage

Hierarchical clustering with “complete” linkage with K=2

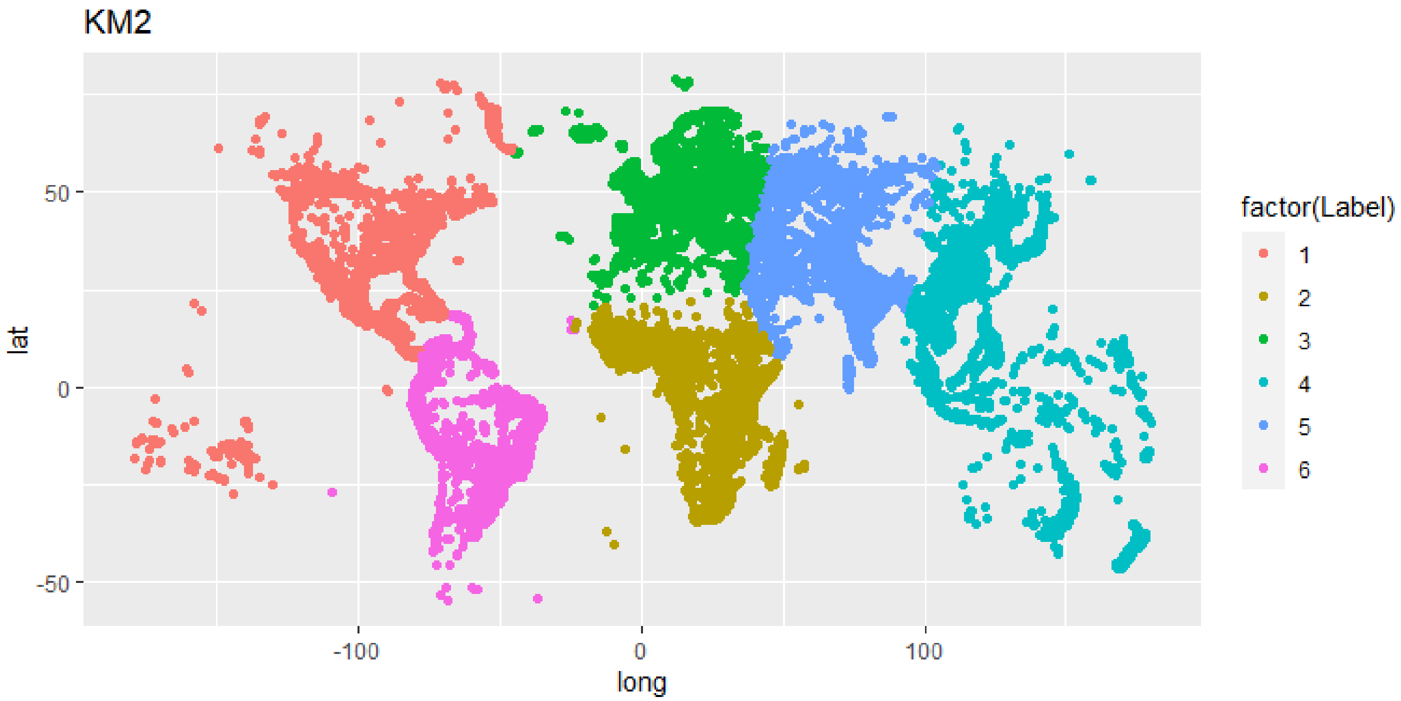
```
clust.hc2 <- DF %>%
mutate(Label=cutree(Res.hc.comp,k=2)) %>%
  ggplot(aes(PC1,PC2,col=factor(Label))) +
geom_point() + ggtitle("HC2") +
  theme(legend.position="bottom")
clust.hc3 <- DF %>%
mutate(Label=cutree(Res.hc.comp,k=3)) %>%
  ggplot(aes(PC1,PC2,col=factor(Label))) +
geom_point() + ggtitle("HC3") +
  theme(legend.position="bottom")
grid.arrange(clust.true, clust.hc2, clust.hc3, nrow=1)
```


True
HC2
HC3
PC2
PC1
10
5
0
-5
-10
-10
-5
0
5
10
Label
Diseased
Healthy
factor(Label)
1
2
3

Extra: world map

```
library(maps)
str(world.cities)
```

```
'data.frame':   43645 obs. of  6 variables:
 $ name       : chr  "'Abasan al-Jadidah" "'Abasan al-
Kabirah" "'Abdul Hakim" "'Abdullah-as-Salam" ...
 $ country.etc: chr  "Palestine" "Palestine"
"Pakistan" "Kuwait" ...
 $ pop        : int  5629 18999 47788 21817 2456 3434
9198 5492 22706 41731 ...
 $ lat        : num  31.3 31.3 30.6 29.4 32 ...
 $ long       : num  34.3 34.4 72.1 48 35.1 ...
 $ capital    : int  0 0 0 0 0 0 0 0 0 0 ...
```


KM2
lat
50
0
-50
-100
0
100
long
factor(Label)
● 1
● 2
● 3
● 4
● 5
● 6

Extra: from fishbase database

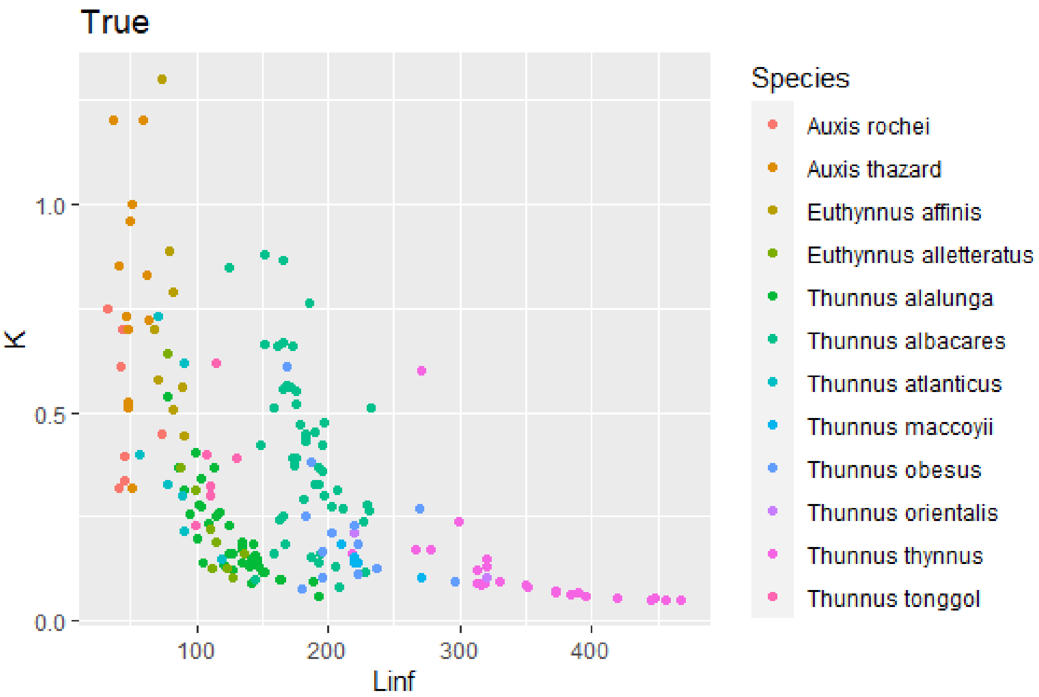
Data reading

```
#remotes::install_github("ropensci/rfishbase")
#library(rfishbase)

DF <- read.csv("Data/Growth_tuna.csv", header=T)
Species <- DF$Species
Linf <- DF$Linf
K <- DF$K
Data <- DF[,2:3]
Data <- scale(Data)
dim(Data)
```

```
[1] 193   2
```

True species

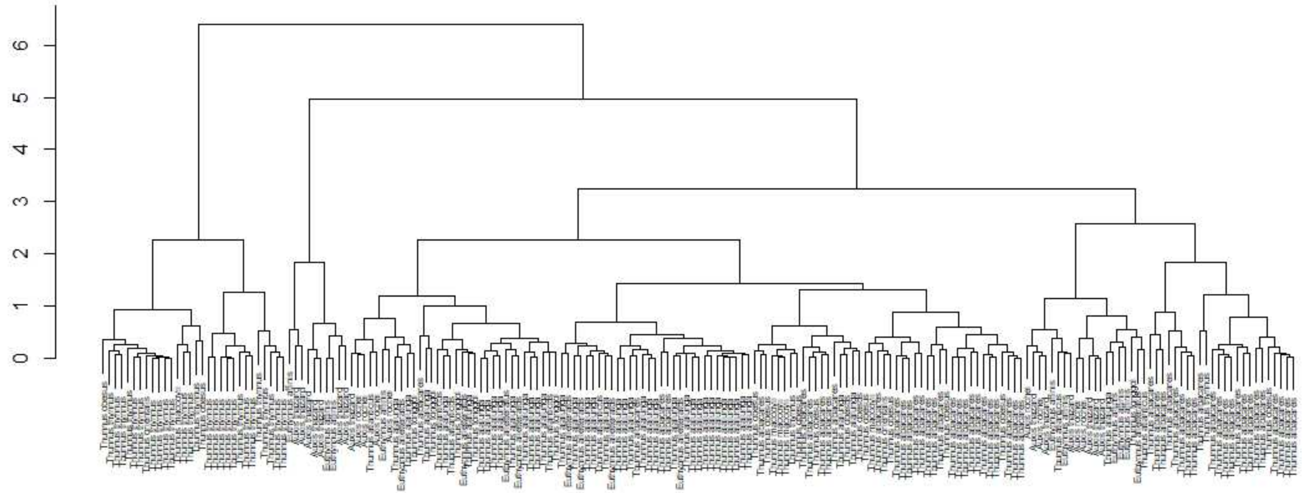
```
clust.true <- ggplot(DF,aes(x=Linf,y=K,col=Species)) +
geom_point() + ggtitle("True")
clust.true
```


True
K
1.0
0.5
0.0
100
200
300
400
Linf
Species
Auxis rochei
Auxis thazard
Euthynnus affinis
Euthynnus alletteratus
Thunnus alalunga
Thunnus albacares
Thunnus atlanticus
Thunnus maccoyii
Thunnus obesus
Thunnus orientalis
Thunnus thynnus
Thunnus tonggol

Hierarchical clustering

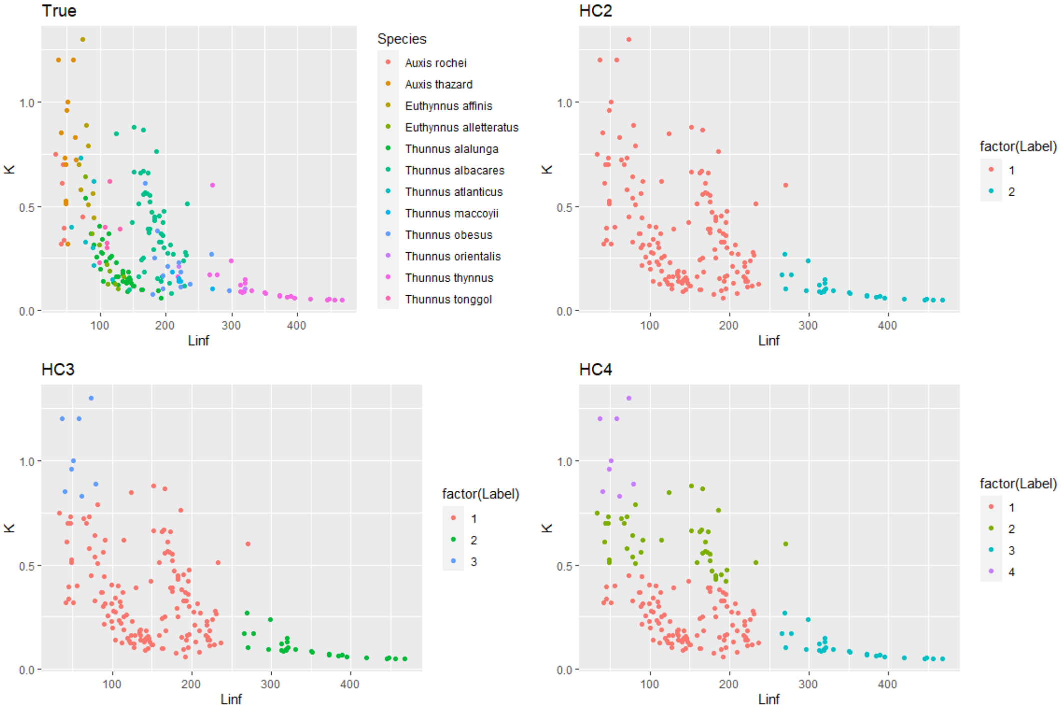
```
Res.hc.comp <- hclust(dist(Data), method="complete")
plot(Res.hc.comp, main="Complete Linkage",
labels=Species, xlab="", ylab="", sub="", cex=0.5)
```

Complete Linkage



Grouping

```
clust.hc2 <- DF %>%
mutate(Label=cutree(Res.hc.comp,k=2)) %>%
  ggplot(aes(Linf,K,col=factor(Label))) + geom_point()
+ ggtitle("HC2")
clust.hc3 <- DF %>%
mutate(Label=cutree(Res.hc.comp,k=3)) %>%
  ggplot(aes(Linf,K,col=factor(Label))) + geom_point()
+ ggtitle("HC3")
clust.hc4 <- DF %>%
mutate(Label=cutree(Res.hc.comp,k=4)) %>%
  ggplot(aes(Linf,K,col=factor(Label))) + geom_point()
+ ggtitle("HC4")
grid.arrange(clust.true, clust.hc2, clust.hc3,
clust.hc4, nrow=2)
```


True
Species
Auxis rochei
Auxis thazard
Euthynnus affinis
Euthynnus alletteratus
Thunnus alalunga
Thunnus albacares
Thunnus atlanticus
Thunnus maccoyii
Thunnus obesus
Thunnus orientalis
Thunnus thynnus
Thunnus tonggol
HC2
factor(Label)
1
2
HC3
factor(Label)
1
2
3
HC4
factor(Label)
1
2
3
4
K
0.0
0.5
1.0
Linf
100
200
300
400

References

- Baumer et al. (2017) “Modern Data Science with R”
- Hastie et al. (2016) “The Elements of Statistical Learning”
- James et al. (2013) "An Introduction to Statistical Learning with applications in R (2nd ed). [Freely downloadable from https://statlearning.com/]
- 金森敬文 (2017) “Rによる機械学習入門”
- 金明哲 (2017) “Rによるデータサイエンス(第2版)”

Announcement for students (大学院)

授業レポートは，このクラスター解析か

1. アジのデータに関してExercise (1)を行うこと.
2. クラスター解析の例題データ(なんでもよい)を用いて，いずれかの方法でクラスター解析を行いこと.

1と2の結果をA4紙に2枚にまとめ，メールで提出すること．締切は2月5日17時.

Announcement for students (演習)

- 「生物資源モデリング」のみを受講の方は，今回の内容を1ページのレポートにまとめて次週までに学務システムにて提出してください

- 「生物資源解析学演習」のみを履修の方は，この教材を参考にしつつ，演習資料の課題を行い提出してください

- 「生物資源モデリング」と「生物資源解析学演習」を両方履修している方は，上記のレポートに加え，演習課題も別途取り組んで下さい．提出はそれぞれ別々にお願いします

- いずれも提出期限は1月19日(火) 17時とします

- 授業や課題に関して質問のある方は，1月19日(火) 13.00-14.00にWebexを立ち上げますので，遠慮なく入室し質問をしてください
- 次回もオンデマンド教材は1月19日12時に学務システムを通して共有します