# Chapter 3 - Cluster

*张笑竹 / 201618070114*

*2018年10月28日*

第三章实验采用课本例3.5和例3.7中给出的数据进行聚类。例3.5是我国31个省、市及自治区的城镇居民消费水平，而例3.7是我国35家上市公司的2008年年报数据。
将本次的实验任务拆分如下：
1）编写系统聚类法的封装函数；
2）利用1）中编写的函数对例3.5数据分别进行R型聚类和Q型聚类；
3）利用K-means聚类法对例3.5数据进行Q型聚类；
4）编写模糊聚类函数，并利用其对例3.7数据进行Q型聚类。

## 1 系统聚类法

### 1.1 函数的封装

首先，将数据复制进入剪贴板，在R中读入本次实验的数据集，并对数据集结构进行微调，将各个省份名称标记为行名，从而为后续分析得到更好的结果做铺垫。

```
X=read.table("clipboard",header=T)
rownames(X)=X[,1]
X=X[,-1]
```

然后，对数据集进行标准化，分别利用dist()函数和hclust()函数对数据集进行系统聚类。需要注意的是，采用“系统聚类法”必须自行确定最终聚类数目，此处利用NbClust包中的NbClust()函数确定最优聚类数目。其原理是，利用26个指数（对应26种判定准则）进行计算；而这26个指数的结果可能不一致，因此通过“投票”的方式，选择出赞同判定准则数量最多的聚类数目。当多个聚类数目“票数”相同时，根据方便的原则，选择最小数目。

```
H.clust <- function(X,d,m,gmin,gmax,ind){ #输入向量c(数据集X,"距离种类","聚类方法",允许最小类数,允许最大类数,"分类指标")
  X.scaled <- scale(X)
  dis <- dist(X.scaled,method=d)
  hc <- hclust(dis,method=m)
  #寻找最佳聚类数
  library(NbClust)
  devAskNewPage(ask=FALSE)
  nc <- NbClust(X.scaled,distance=d,min.nc=gmin,max.nc=gmax,method=m,index=ind)
  t <- table(nc$Best.n[1,])
  tframe <- as.data.frame(t)
  tframe1 <- tframe[which(tframe$Freq==max(t)),]
  k0=as.numeric(as.character(tframe1$Var1))[1]
  #结果输出
  clusters <- cutree(hc,k=k0)
  par(mfrow=c(1,1))
  barplot(t,xlab="Number of Clusters",ylab="Number of Criteria",main="Clusters Chosen")
  plot(hc,hang=-1)
  rect.hclust(hc,k=k0)
  return(list(clusters=table(clusters),median=aggregate(X,by=list(clusters=clusters),median)))
}
```

H.clust()函数的输入参数包括数据集X,“距离种类”,“聚类方法”,允许最小类数,允许最大类数,“分类指标”。最终将输出“聚类数目投票图”，聚类树状图，聚类频数表，以及各个类别的中位数。

## 1.2 R型聚类（变量聚类）

首先，将数据集进行转置。

然后，调用H.clust()函数，选择欧几里得距离、类平均法，允许最小类别数目为2、最大数目为7，选用silhouette和gap指数（准则），对变量进行聚类。

```
X1=t(X)
H.clust(X1,"euclidean","average",2,7,c("silhouette","gap"))
```

```
## Warning in if (is.na(indice)) stop("invalid clustering index"): 条件的长度
## 大于一，因此只能用其第一元素
```

```
## Warning in if (indice == -1) stop("ambiguous index"): 条件的长度大于一，因
## 此只能用其第一元素
```

```
## Warning in if (indice < 31) {: 条件的长度大于一，因此只能用其第一元素
```

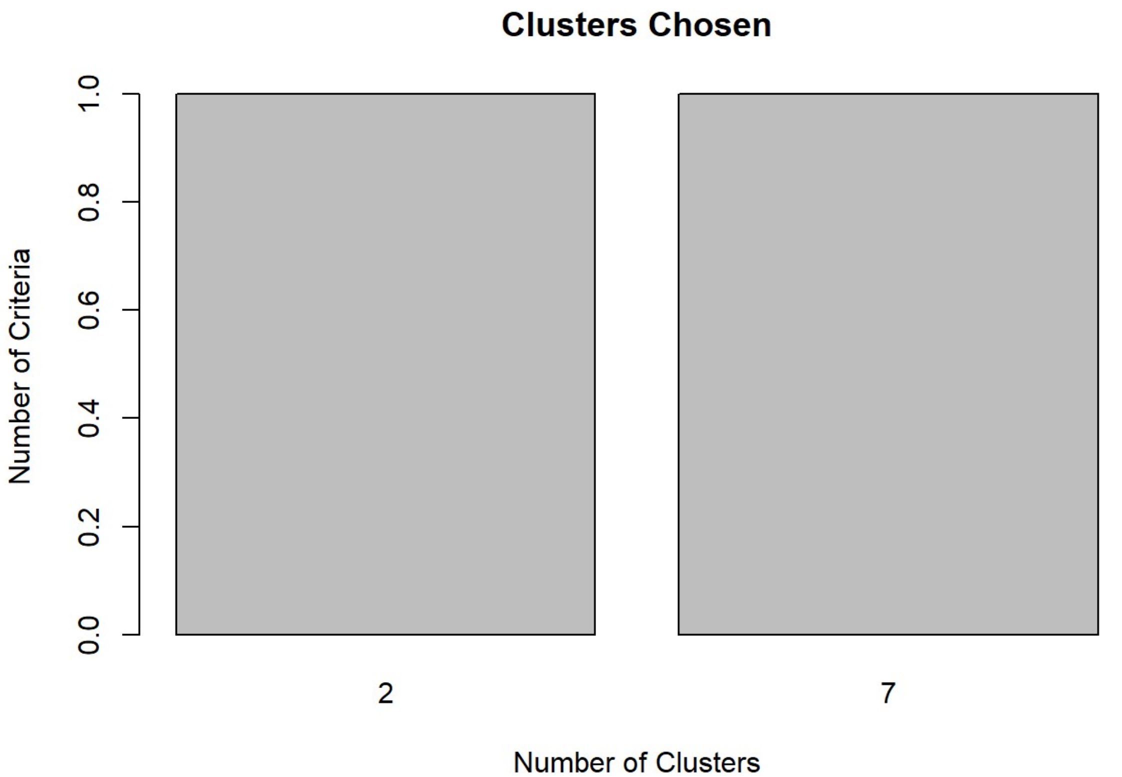
```
## Warning in if (indice == 14) {: 条件的长度大于一，因此只能用其第一元素
```

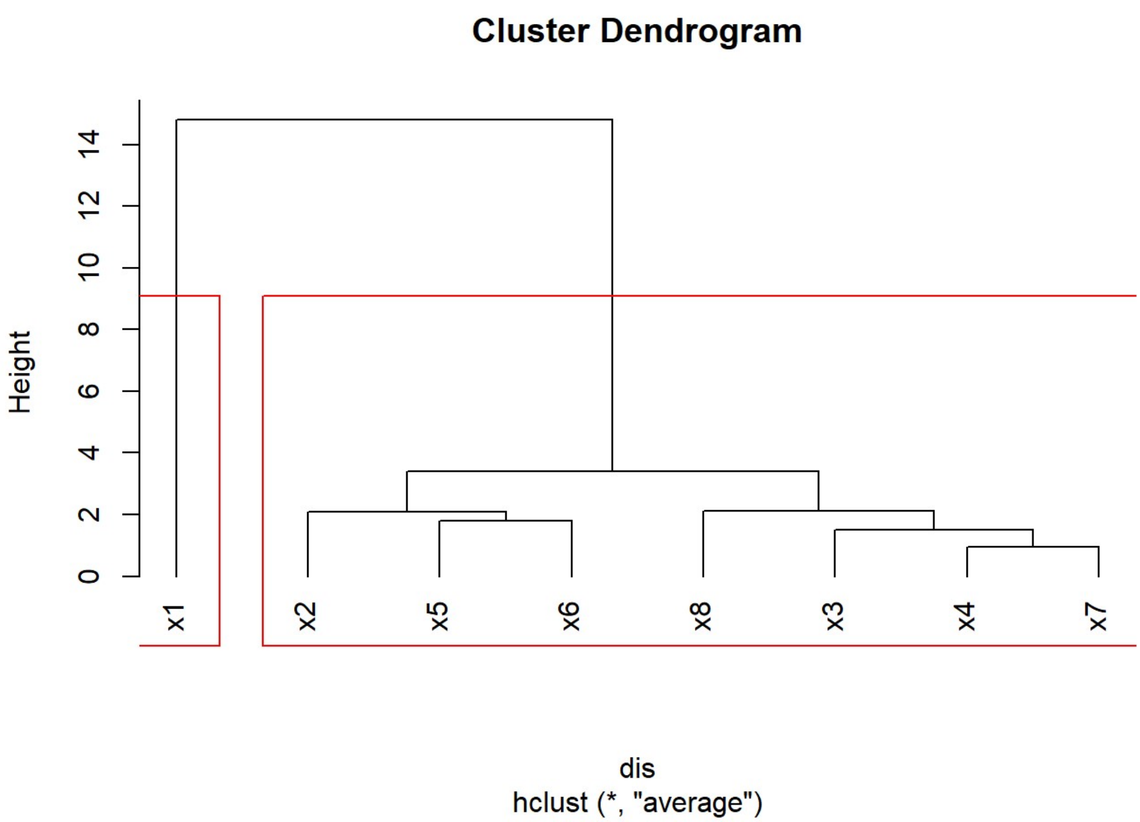
```
## Warning in if (indice == 15) {: 条件的长度大于一，因此只能用其第一元素
```

```
## Warning in if (indice == 16) {: 条件的长度大于一，因此只能用其第一元素
```

```
## Warning in if (indice == 20) {: 条件的长度大于一，因此只能用其第一元素
```

```
## Warning in if (indice == 31) {: 条件的长度大于一，因此只能用其第一元素
```


Clusters Chosen
Number of Criteria
0.0
0.2
0.4
0.6
0.8
1.0
2
7
Number of Clusters


Cluster Dendrogram
Height
0
2
4
6
8
10
12
14
x1
x2
x5
x6
x8
x3
x4
x7
dis
hclust (*, "average")

```
## $clusters
## clusters
## 1 2
## 1 7
##
## $median
##   clusters    北京    天津    河北    山西  内蒙古    辽宁    吉林  黑龙江
## 1        1 7535.29 7343.64 4211.16 3855.56 5463.18 5809.39 4635.27 4687.23
## 2        2 1970.94 1854.22 1203.80 1438.88 1583.56 1433.28 1594.14 1216.56
##       上海    江苏    浙江    安徽    福建    江西    山东    河南    湖北
## 1 9655.60 6658.37 7552.02 5814.92 7317.42 5071.61 5201.32 4607.47 5837.93
## 2 1906.49 1437.08 1551.69 1396.97 1634.21 1173.91 1572.35 1190.81 1371.15
##       湖南    广东    广西    海南    重庆    四川    贵州    云南    西藏
## 1 5441.63 8258.44 5552.56 6556.10 6870.23 6073.86 4992.85 5468.17 5517.69
## 2 1301.60 1520.59 1146.46  993.24 1196.03 1284.09 1013.53  973.76  580.05
##       陕西    甘肃    青海    宁夏    新疆
## 1 5550.71 4602.33 4667.34 4768.91 5238.89
## 2 1322.22 1287.93 1097.21 1193.37 1166.59
```

根据输出结果，可以看出，“silhouette”和“gap”准则各“投”了2类和7类一票。根据方便的原则，最终确定了聚类数目为2类。当然，由输出的树状图可以看出，下面的分类方法也是可以的：变量$X_1$为一类，变量$X_2$、$X_5$、$X_6$为一类，变量$X_8$、$X_3$、$X_4$、$X_7$为一类。

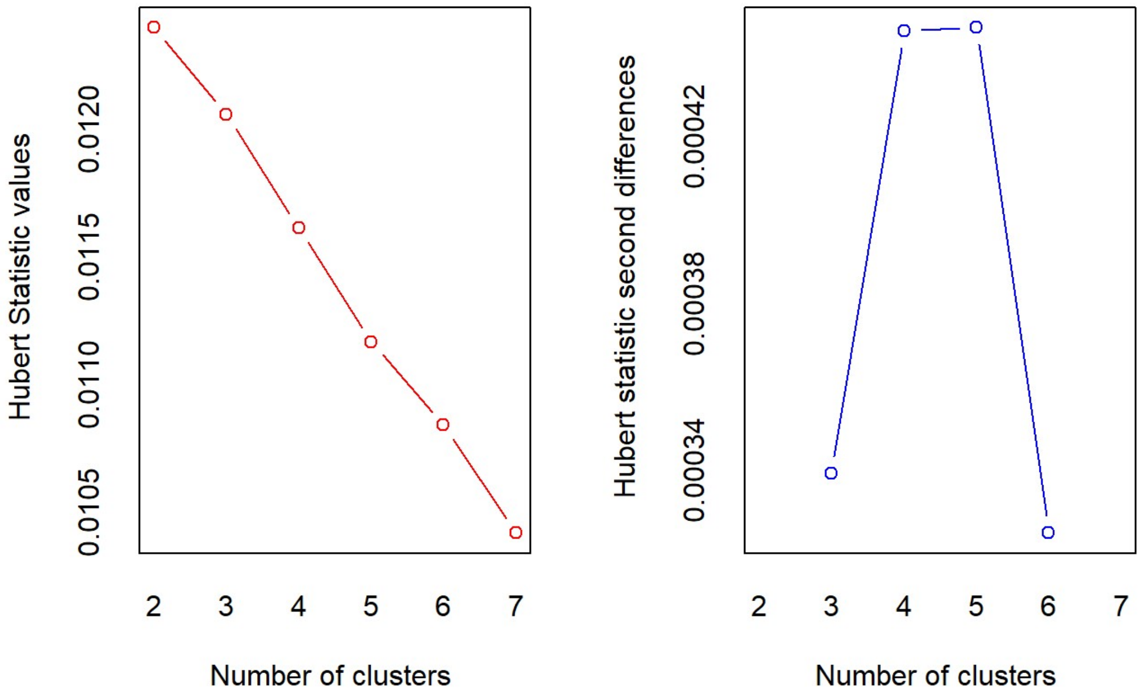
## 1.3 Q型聚类（样品聚类）

本节分别采用最短距离法、最长距离法、重心法、类平均法和Ward法对样品进行聚类。

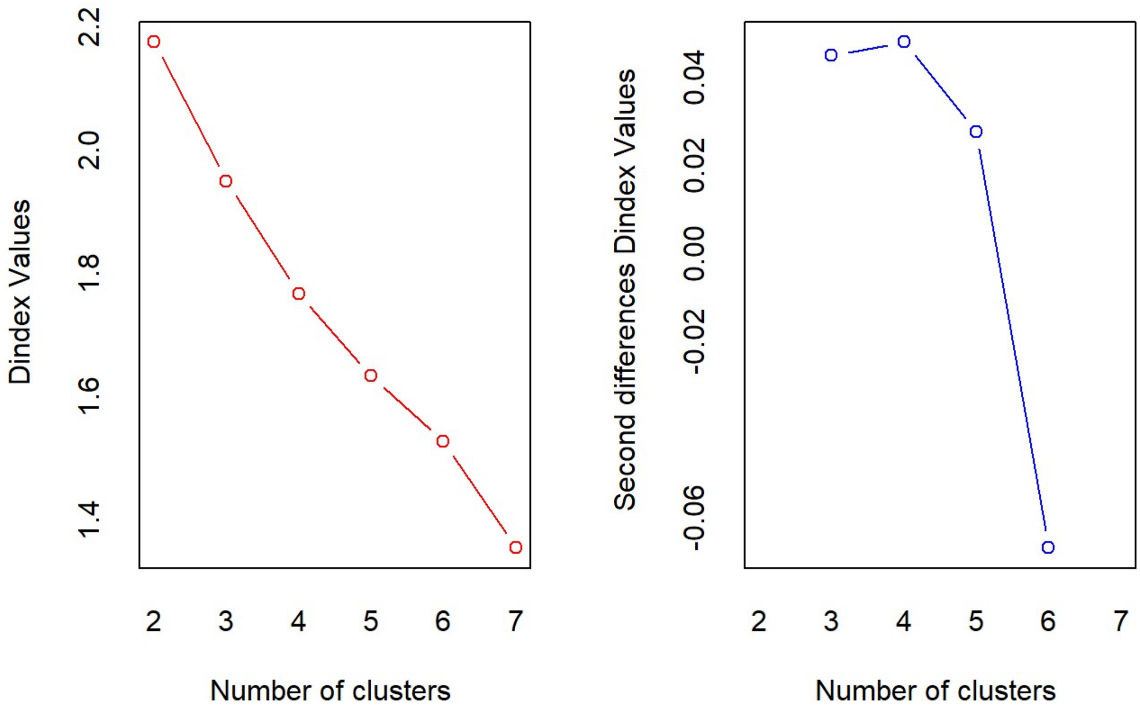
### 1.3.1 最短距离法

调用H.clust()函数，选择最短距离法，对样品进行聚类。其中距离采用欧几里得距离，聚类数目准则选择全部(“all”)。

```
H.clust(X,"euclidean","single",2,7,"all")
```

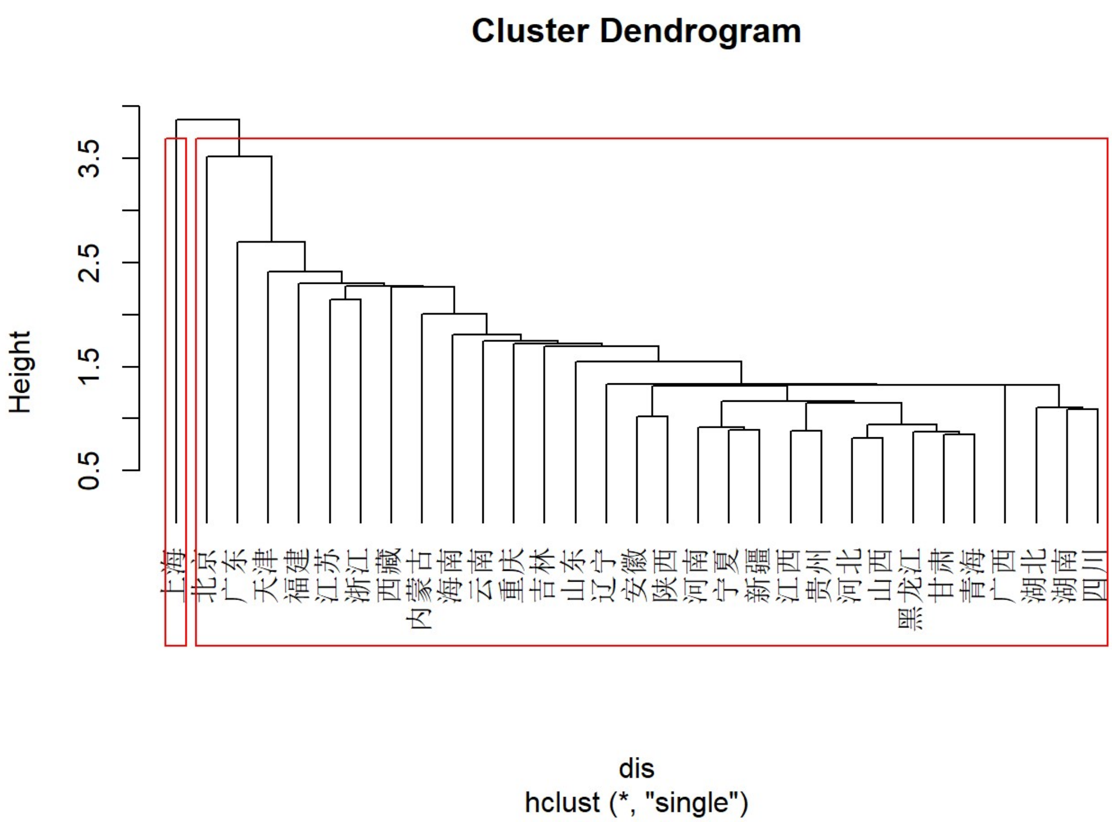



```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in
Hubert
##                 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 D
index
##                 second differences plot) that corresponds to a significant increase of the va
lue of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 7 proposed 2 as the best number of clusters
## * 6 proposed 3 as the best number of clusters
## * 5 proposed 4 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 3 proposed 6 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```


Clusters Chosen
Number of Criteria
0
1
2
3
4
5
6
7
0
2
3
4
5
6
7
Number of Clusters


Cluster Dendrogram
Height
0.5
1.5
2.5
3.5
上海
北京
广东
天津
福建
江苏
浙江
西藏
内蒙古
海南
云南
重庆
吉林
山东
辽宁
安徽
陕西
河南
宁夏
新疆
江西
贵州
河北
山西
黑龙江
甘肃
青海
广西
湖北
湖南
四川
dis
hclust (*, "single")

```
## $clusters
## clusters
##  1  2
## 30  1
##
## $median
##   clusters      x1      x2       x3      x4      x5      x6      x7
## 1        1 5492.93 1771.65 1374.205  982.54 1924.98 1606.74 1047.78
## 2        2 9655.60 2111.17 1790.480 1906.49 4563.80 3723.74 1016.65
##          x8
## 1  571.245
## 2 1485.530
```

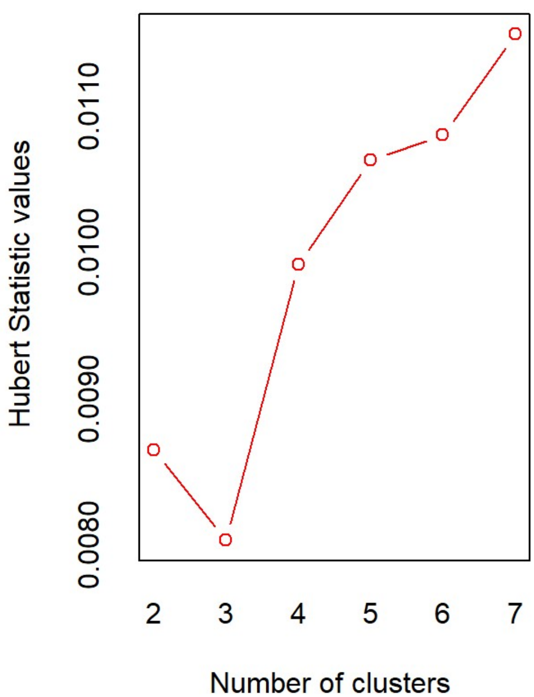
根据结果不难看出，在最短距离法下，最优聚类数目为2类，上海单独为一类，其他所有省份为另外一类。聚类图呈现明显的链式形状。

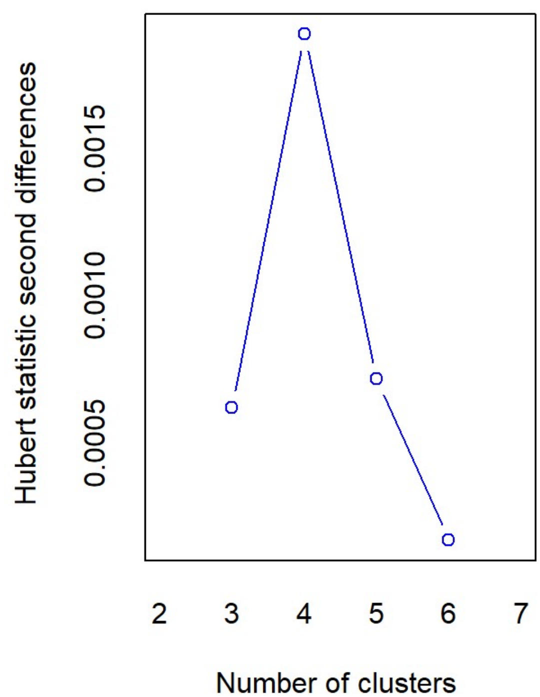
### 1.3.2 最长距离法

同理，调用H.clust()函数，选择最长距离法，对样品进行聚类。

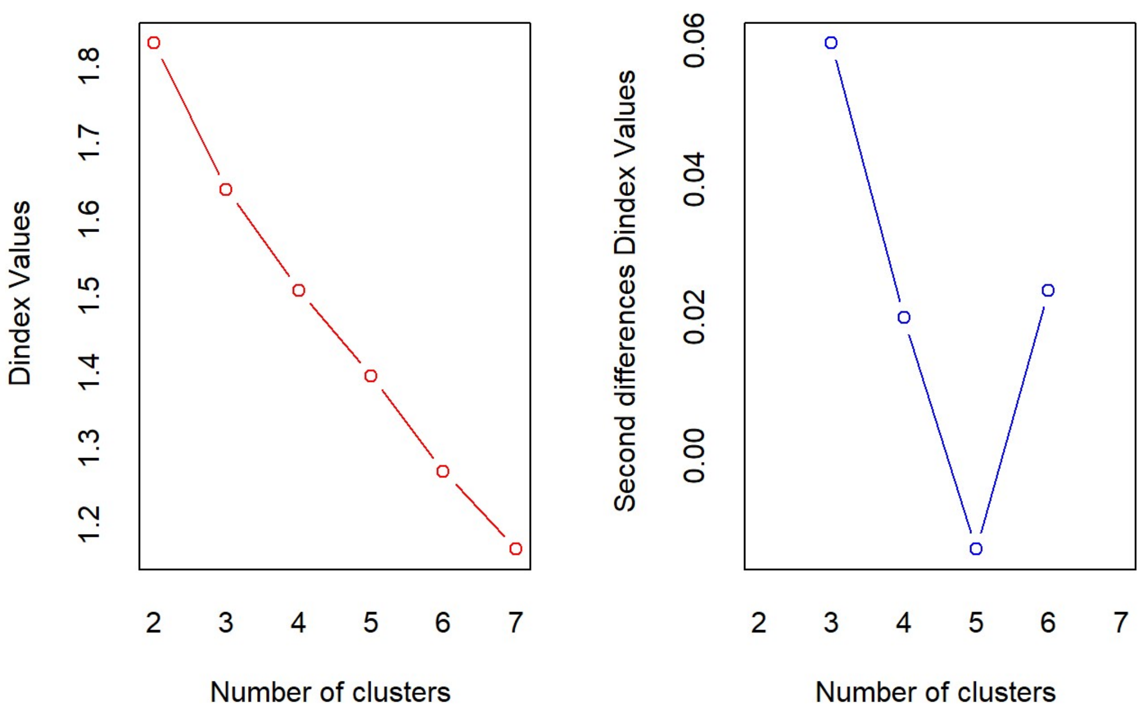
```
H.clust(X,"euclidean","complete",2,7,"all")
```

```
## Warning in pf(beale, pp, df2): 产生了NaNs
```



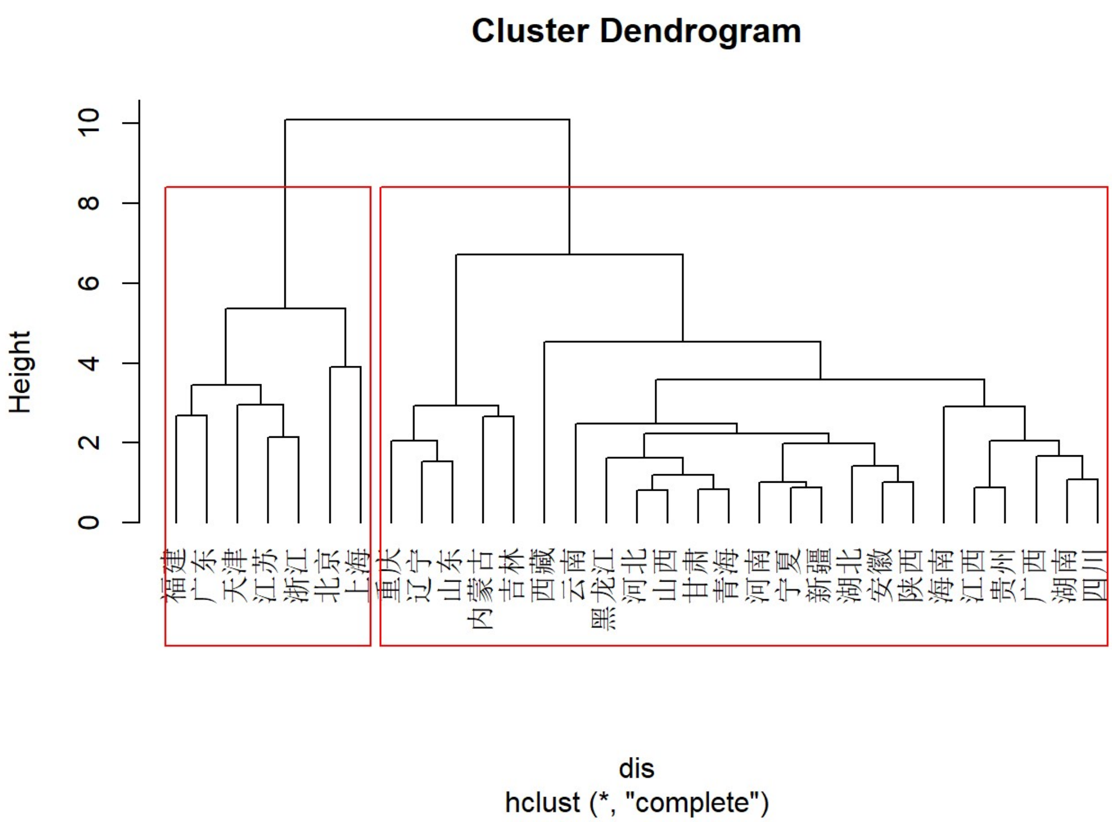



```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                In the plot of Hubert index, we seek a significant knee that corresponds to a
##                significant increase of the value of the measure i.e the significant peak in
Hubert
##                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 D
index
##                second differences plot) that corresponds to a significant increase of the va
lue of
##                the measure.
##
## *******************************************************************
## * Among all indices:
## * 11 proposed 2 as the best number of clusters
## * 2 proposed 3 as the best number of clusters
## * 6 proposed 4 as the best number of clusters
## * 4 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
##
##                   ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```


Clusters Chosen
Number of Criteria
0
2
4
6
8
10
0
2
3
4
6
7
Number of Clusters


Cluster Dendrogram
Height
0
2
4
6
8
10
福建
广东
天津
江苏
浙江
北京
上海
重庆
辽宁
山东
内蒙古
吉林
西藏
云南
黑龙江
河北
山西
甘肃
青海
河南
宁夏
新疆
湖北
安徽
陕西
海南
江西
贵州
广西
湖南
四川
dis
hclust (*, "complete")

```
## $clusters
## clusters
##  1  2
##  7 24
##
## $median
##   clusters      x1       x2      x3      x4      x5       x6       x7
## 1        1 7535.29 1915.970 1790.48 1288.42 3781.51 2996.590 1058.110
## 2        2 5340.26 1705.515 1311.91  939.59 1799.05 1511.055 1028.575
##         x8
## 1 871.300
## 2 543.375
```

这里，样品同样被分为2类，其中北京、上海、天津、江苏、浙江、广东和福建被聚为一类，其他省份被聚为另外一类。聚类图分类特征较为明显，结果较好。

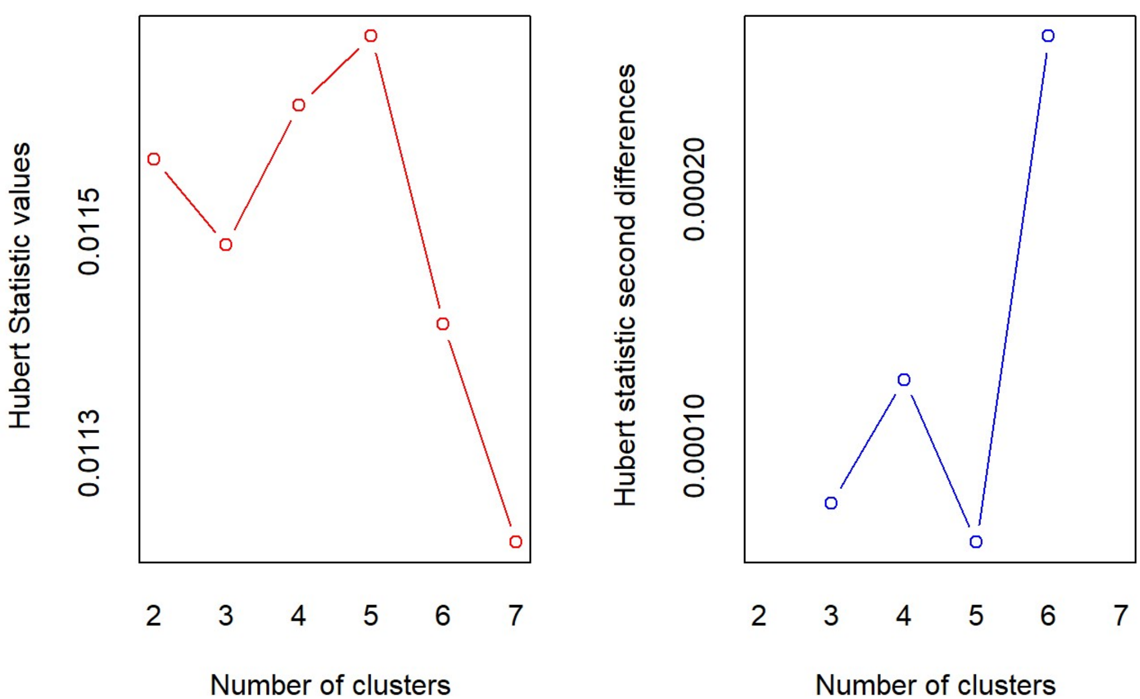
### 1.3.3 重心法

调用H.clust()函数，选择重心法，对样品进行聚类。

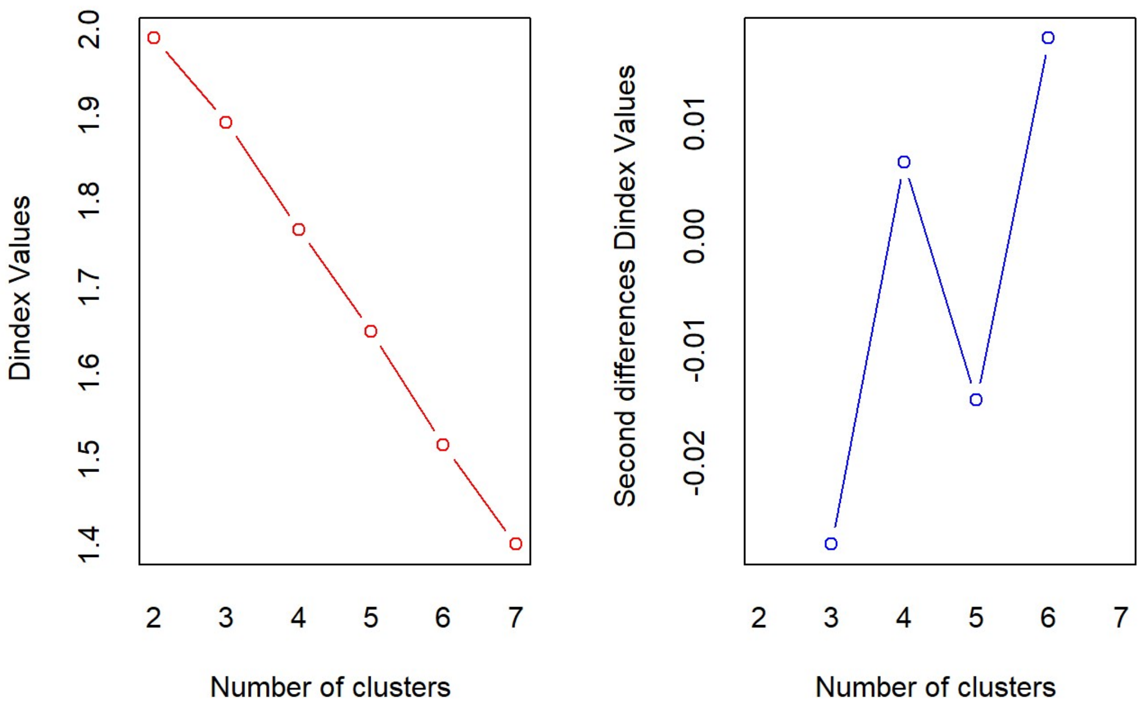
```
H.clust(X,"euclidean","centroid",2,7,"all")
```

```
## Warning in pf(beale, pp, df2): 产生了NaNs
```

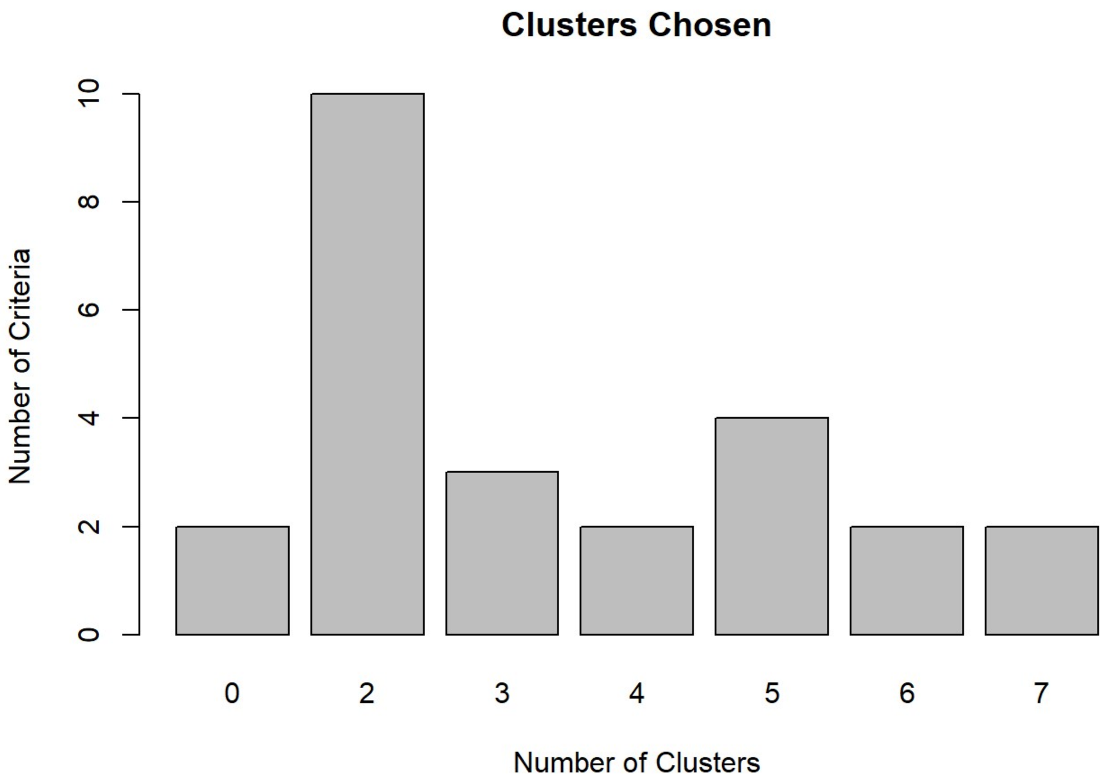
```
## [1] "Frey index : No clustering structure in this data set"
```

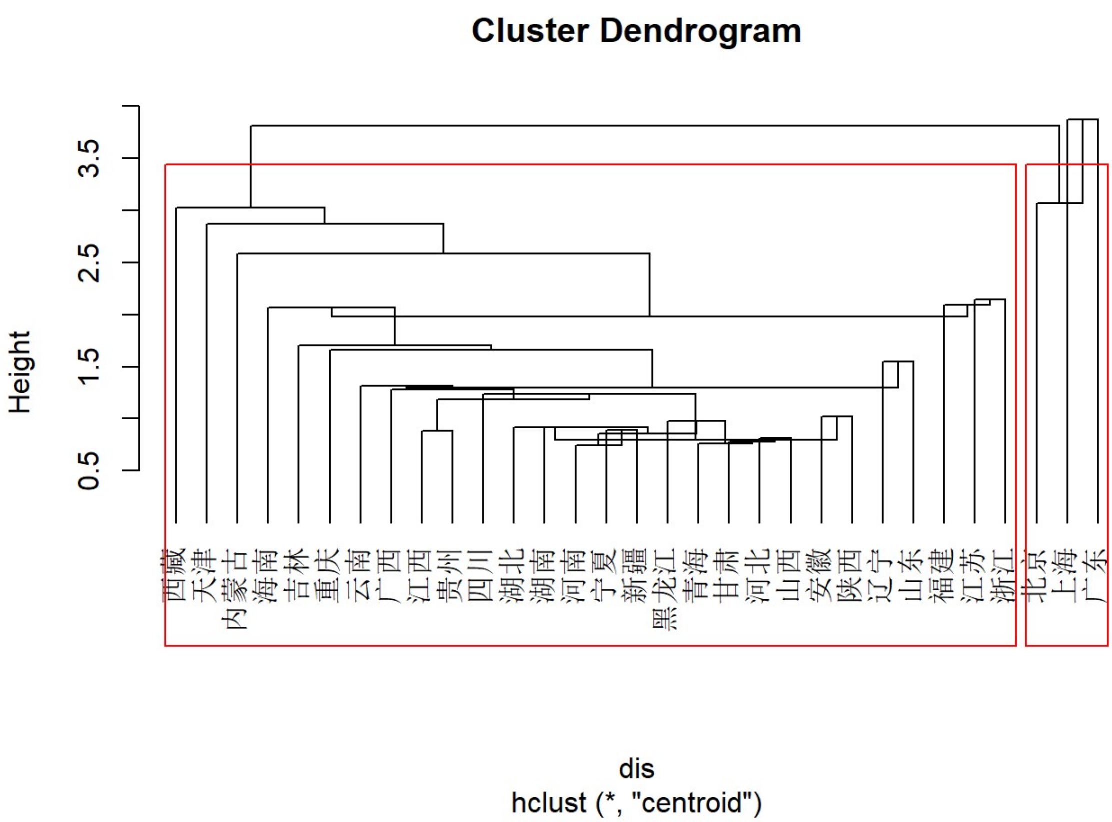


```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in
Hubert
##                 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 D
index
##                 second differences plot) that corresponds to a significant increase of the va
lue of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 10 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 2 proposed 4 as the best number of clusters
## * 4 proposed 5 as the best number of clusters
## * 2 proposed 6 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```


Clusters Chosen
Number of Criteria
0
2
4
6
8
10
0
2
3
4
5
6
7
Number of Clusters


Cluster Dendrogram
Height
0.5
1.5
2.5
3.5
西藏
天津
内蒙古
海南
吉林
重庆
云南
广西
江西
贵州
四川
湖北
湖南
河南
宁夏
新疆
黑龙江
青海
甘肃
河北
山西
安徽
陕西
辽宁
山东
福建
江苏
浙江
北京
上海
广东
dis
hclust (*, "centroid")

```
## $clusters
## clusters
##  1  2
##  3 28
##
## $median
##   clusters       x1      x2      x3       x4       x5      x6       x7
## 1        1 8258.440 2111.17 1970.94 1610.700 4176.660 3695.98 1048.280
## 2        2 5465.675 1771.65 1354.00  972.245 1897.135 1556.38 1038.415
##         x8
## 1 1154.180
## 2  562.285
```

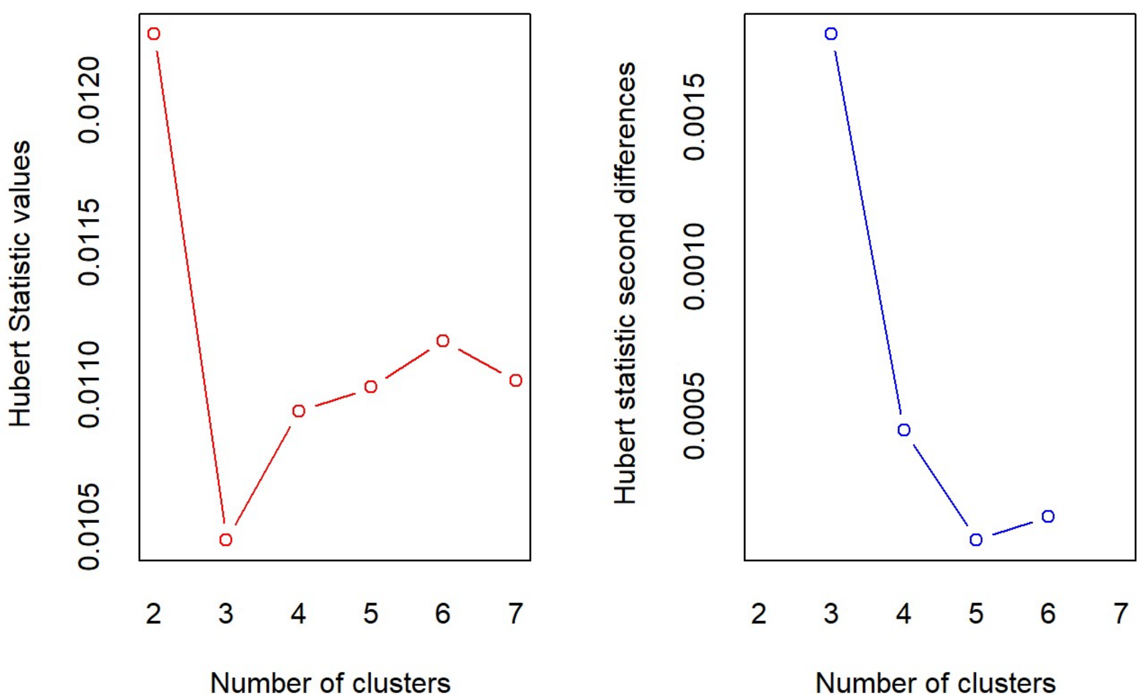
从输出的结果不难看出，在重心法下，最优聚类数目为2类，北京、上海、广东为一类，其他所有省份为另外一类；聚类结果不是非常明显。

### 1.3.4 类平均法

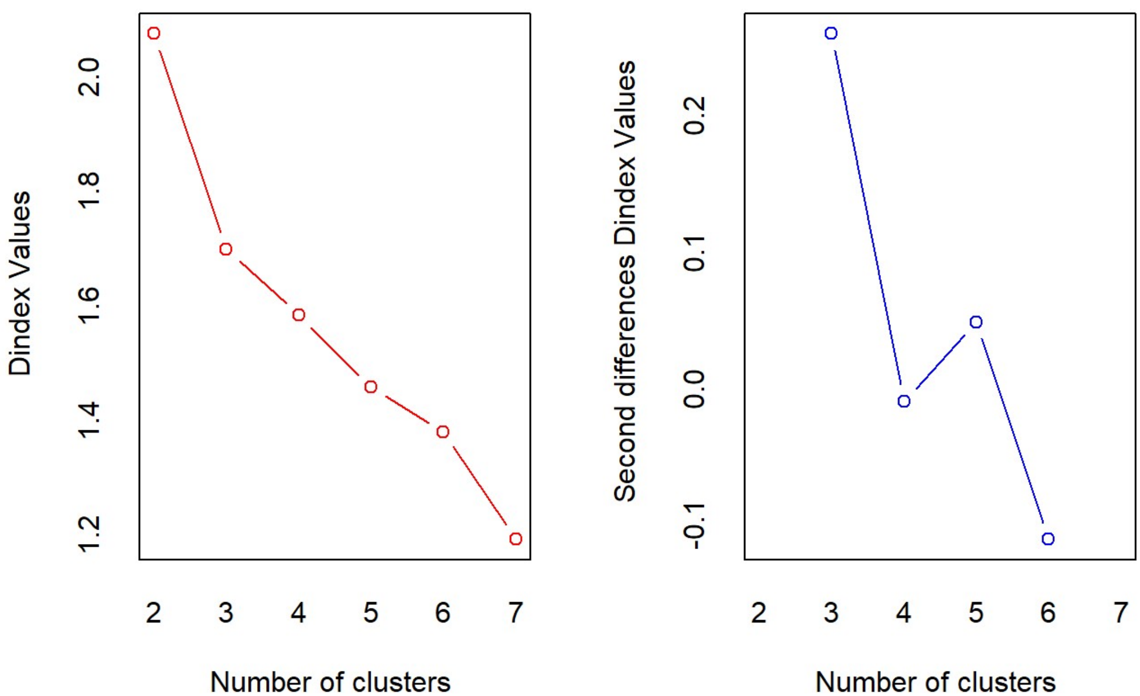
下面，调用H.clust()函数，利用类平均法进行聚类。

```
H.clust(X,"euclidean","average",2,7,"all")
```

```
## Warning in pf(beale, pp, df2): 产生了NaNs
```

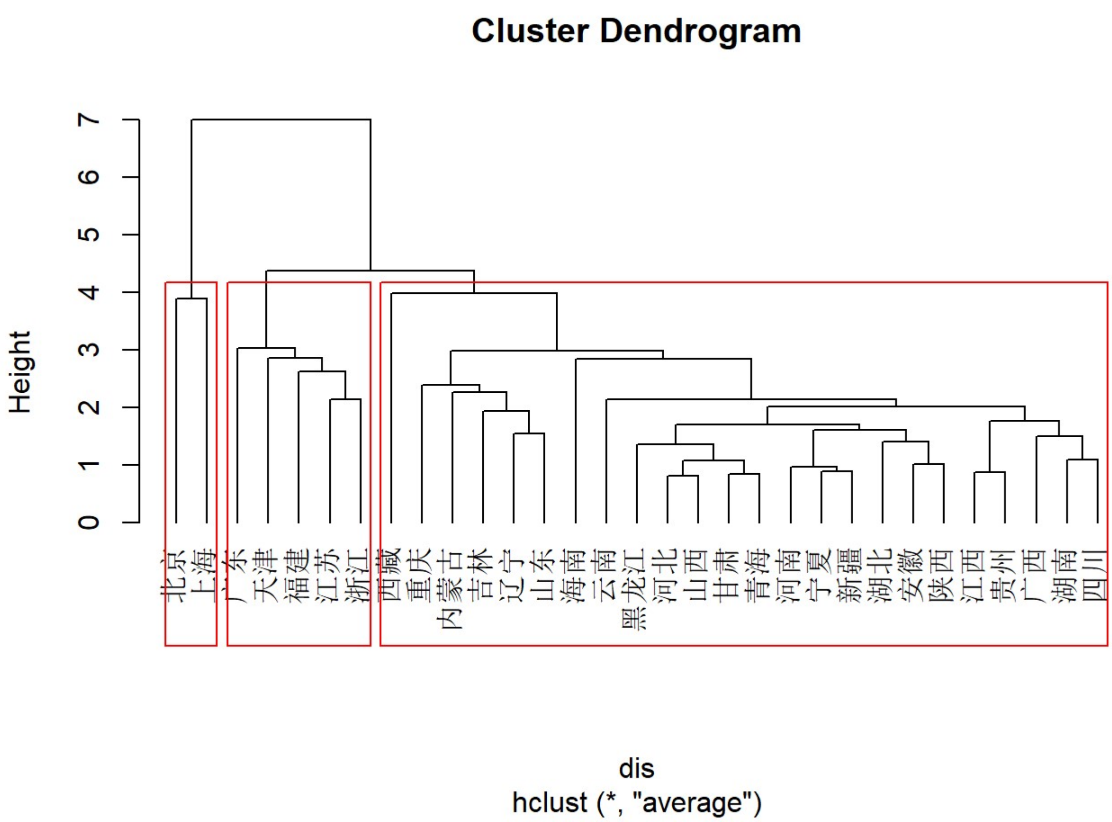


```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                In the plot of Hubert index, we seek a significant knee that corresponds to a
##                significant increase of the value of the measure i.e the significant peak in
Hubert
##                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 D
index
##                second differences plot) that corresponds to a significant increase of the va
lue of
##                the measure.
##
## *******************************************************************
## * Among all indices:
## * 3 proposed 2 as the best number of clusters
## * 16 proposed 3 as the best number of clusters
## * 3 proposed 5 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
##
##                   ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```


Clusters Chosen
Number of Criteria
0
5
10
15
0
2
3
5
7
Number of Clusters


Cluster Dendrogram
Height
0
1
2
3
4
5
6
7
北京
上海
广东
天津
福建
江苏
浙江
西藏
重庆
内蒙古
吉林
辽宁
山东
海南
云南
黑龙江
河北
山西
甘肃
青海
河南
宁夏
新疆
湖北
安徽
陕西
江西
贵州
广西
湖南
四川
dis
hclust (*, "average")

```
## $clusters
## clusters
##  1  2  3
##  2  5 24
##
## $median
##   clusters       x1       x2      x3       x4       x5       x6       x7
## 1        1 8595.445 2375.035 1880.71 1758.595 4172.655 3709.860 1337.510
## 2        2 7343.640 1881.430 1753.86 1254.710 3083.370 2954.130 1058.110
## 3        3 5340.260 1705.515 1311.91  939.590 1799.050 1511.055 1028.575
##         x8
## 1 1319.855
## 2  812.390
## 3  543.375
```

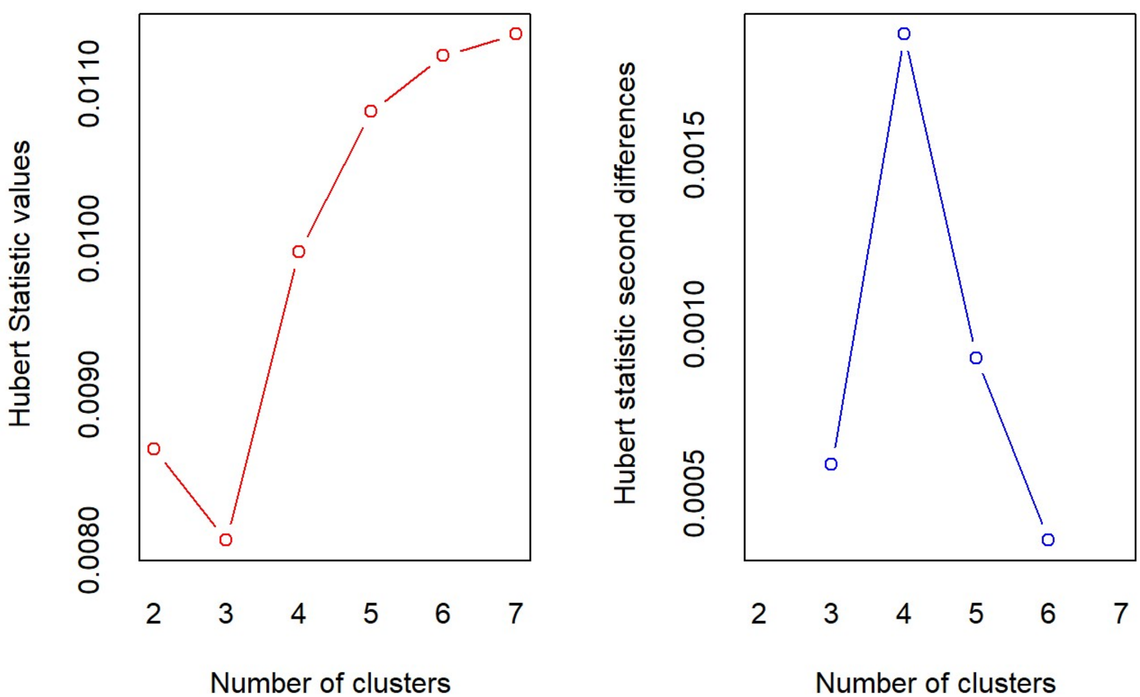
可见，在类平均法下，最优聚类数目为3类，北京、上海为一类，广东、天津、福建、江苏、浙江为一类，其他所有省份为另外一类；聚类效果十分明显。

### 1.3.5 Ward法

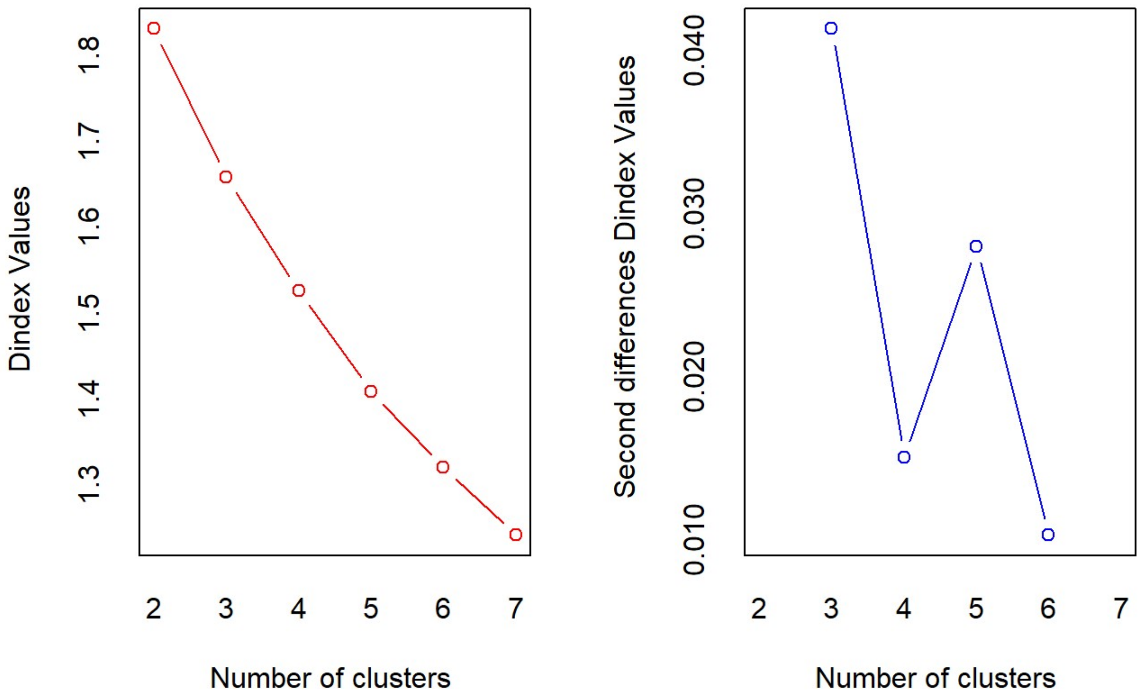
最后，调用H.clust()函数，采用Ward法进行聚类。

```
H.clust(X,"euclidean","ward.D",2,7,"all")
```

```
## Warning in pf(beale, pp, df2): 产生了NaNs
```

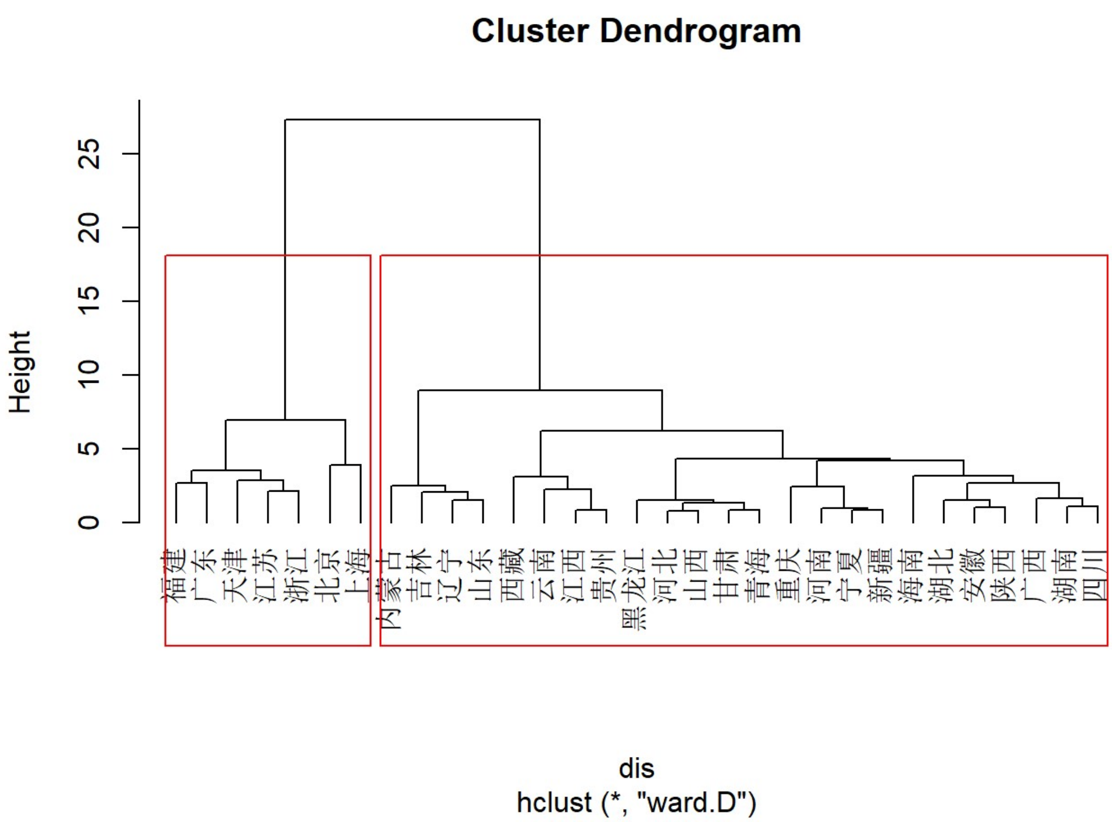


```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in
Hubert
##                 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 D
index
##                 second differences plot) that corresponds to a significant increase of the va
lue of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 13 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 3 proposed 4 as the best number of clusters
## * 3 proposed 5 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```


Clusters Chosen
Number of Criteria
0
2
4
6
8
10
12
0
2
3
4
5
7
Number of Clusters


Cluster Dendrogram
Height
0
5
10
15
20
25
福建
广东
天津
江苏
浙江
北京
上海
内蒙古
吉林
辽宁
山东
西藏
云南
江西
贵州
黑龙江
河北
山西
甘肃
青海
重庆
河南
宁夏
新疆
海南
湖北
安徽
陕西
广西
湖南
四川
dis
hclust (*, "ward.D")

```
## $clusters
## clusters
##  1  2
##  7 24
##
## $median
##   clusters      x1       x2      x3      x4      x5       x6       x7
## 1        1 7535.29 1915.970 1790.48 1288.42 3781.51 2996.590 1058.110
## 2        2 5340.26 1705.515 1311.91  939.59 1799.05 1511.055 1028.575
##        x8
## 1 871.300
## 2 543.375
```

在Ward法下，最优聚类数目为2类，北京、上海、广东、天津、江苏、浙江、福建为一类，其他所有省份为另外一类；聚类效果较为明显。

### 1.3.6 总结

将上述5种系统聚类法的结果整理如下：

| 聚类方法 | 最优聚类数 | 类别是否明显 | 各类别样品数量 |
|---|---|---|---|
| 最短距离法 | 2 | 否，呈现链式 | 1,30 |
| 最长距离法 | 2 | 是 | 7,24 |
| 重心法 | 2 | 否 | 3,28 |
| 类平均法 | 3 | 是 | 2,5,24 |
| Ward法 | 2 | 是 | 7,24 |

综上所述，最短距离法和重心法聚类效果较差，而最长距离法、类平均法和Ward法聚类效果较好。 但是，类与类之间的样品数量差异过大，这或许是由最优聚类数的选取原则造成的。

# 2 K-means聚类法

## 2.1 确定聚类数目

与系统聚类法不同的是，K-means聚类法需要预先确定聚类的数目。利用factoextra包中的函数，画出“聚合”系数随分类数变化曲线，从而判定最佳聚类数。

```
#K-means聚类法
library(ggplot2)
```

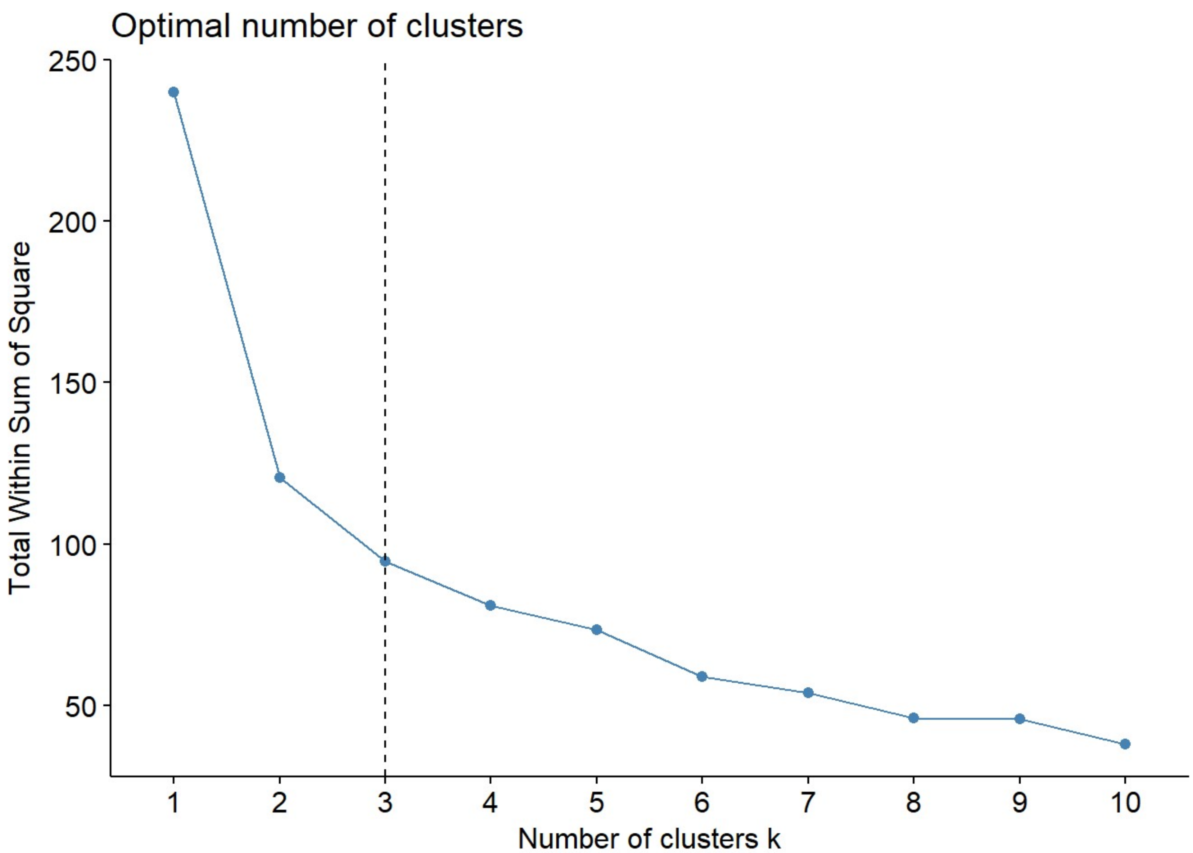
```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.4.4
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
df <- scale(X)
fviz_nbclust(df, kmeans, method = "wss") + geom_vline(xintercept = 3, linetype = 2)
```




不难看出，当分类数为3或4时，曲线变得比较平滑。因此，决定将所有省份分为3类。

## 2.2 进行聚类

最后，利用kmeans()函数和fviz_cluster()函数，输出聚类中心、迭代次数、聚类频数和聚类图。

```
set.seed(123)
km <- kmeans(df,3)
km$centers
```

```
##           x1         x2         x3         x4         x5        x6
## 1  1.8598918  0.8029115  1.5610874  1.7442512  2.2093299  2.037769
## 2 -0.5139319 -0.5239240 -0.5483537 -0.5455692 -0.5638815 -0.571046
## 3  0.2012452  0.6909984  0.4028909  0.3159158  0.1458385  0.236417
##           x7         x8
## 1  0.7456442  1.8804830
## 2 -0.4679504 -0.5932886
## 3  0.6045034  0.3508069
```

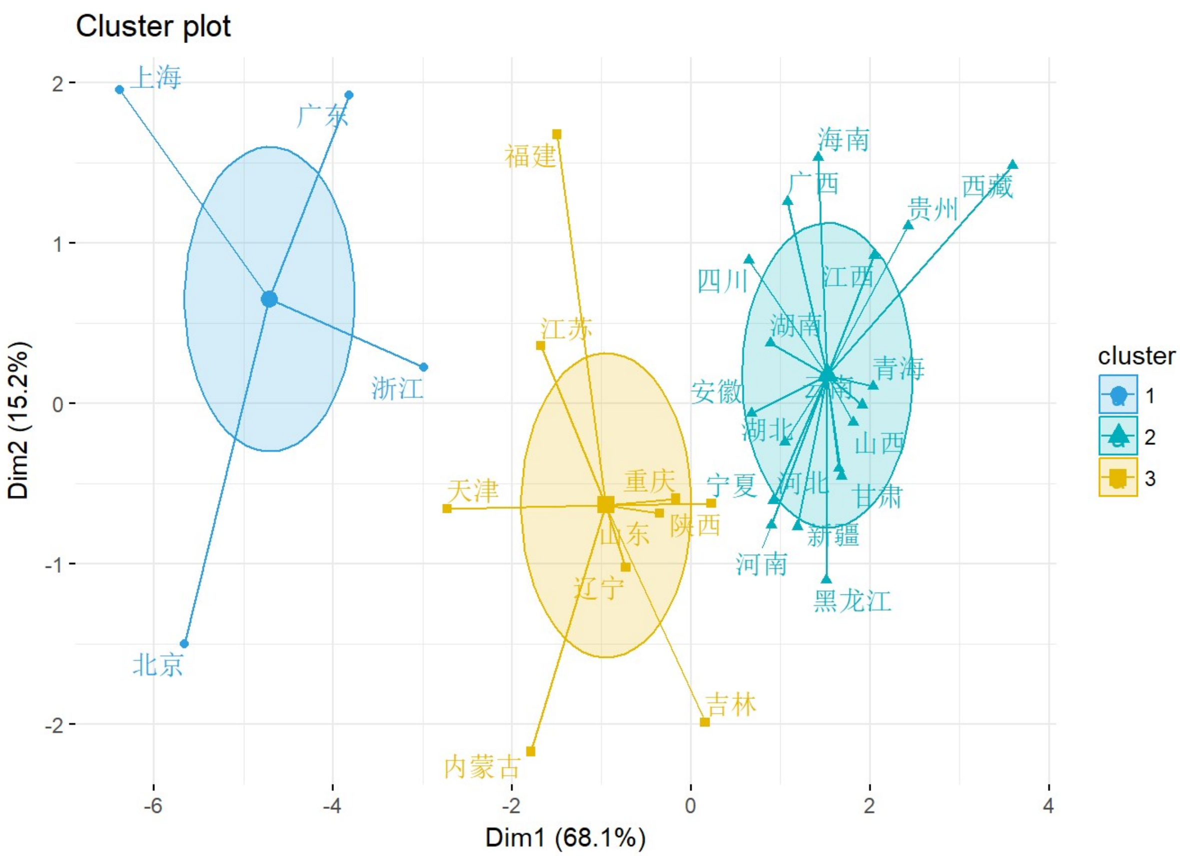
```
km$iter
```

```
## [1] 2
```

```
dd <- cbind(X, cluster = km$cluster)
table(dd$cluster)
```

```
##
##  1  2  3
##  4 18  9
```

```
fviz_cluster(km, data = df,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)
```




经过2次迭代即完成聚类，3个类别的样品数分别为(4,8,19)。从某种程度上，K-means聚类法克服了系统聚类法“类与类之间包含的样品数目差别过大”的缺陷，将经济发展水平相近的省份化作一类，结果较为合理。

## 3 模糊聚类法

本节中利用模糊聚类法，对例3.7进行聚类分析。

相比起前面的“硬聚类”，FCM方法会计算每个样本对所有类的隶属度，这给了我们一个参考该样本分类结果可靠性的计算方法。若某样本对某类的隶属度在所有类的隶属度中具有绝对优势，则该样本分到这个类是一个十分保险的做法，反之若该样本在所有类的隶属度相对平均，则我们需要其他辅助手段来进行分类。

首先，读入例3.7的数据集，并对数据集结构进行微调，将各个公司名称标记为行名。

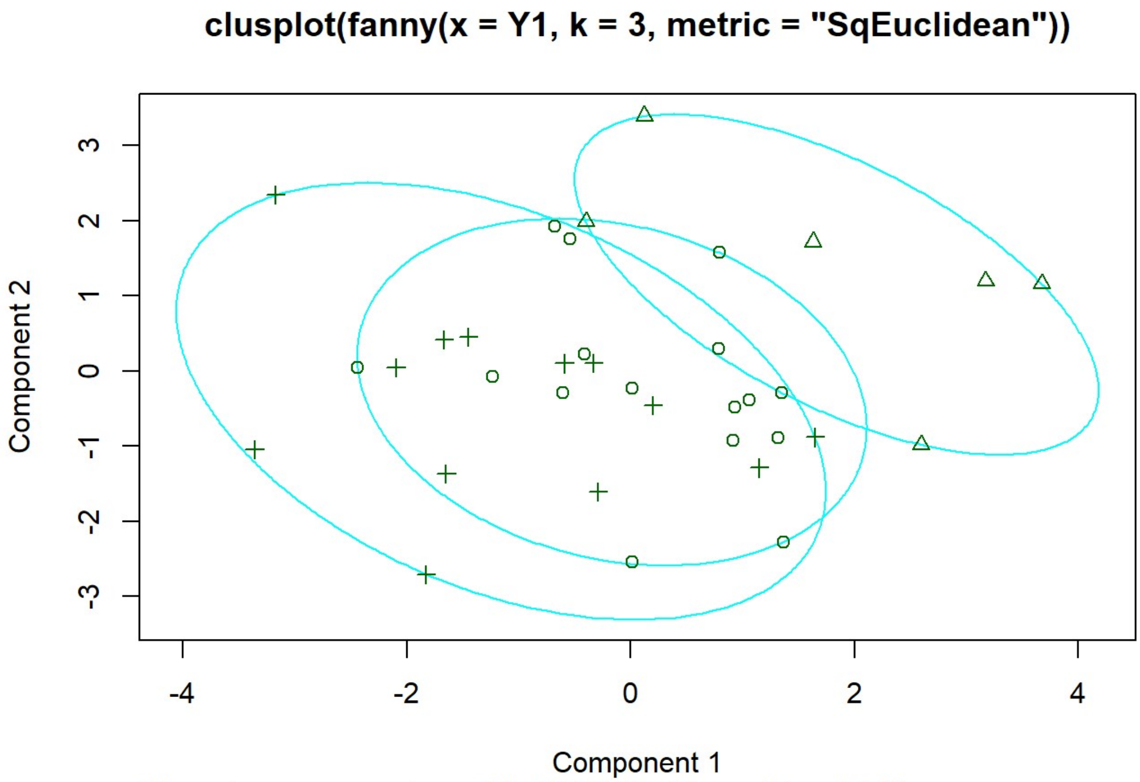
```
Y=read.csv("eg3-7.csv")
Y1=Y
rownames(Y1)=Y1[,1]
Y1=Y1[,-1]
Y1=Y1[,-1]
```

然后，利用fanny()函数，将样品分为3类。能够得到分类分布、样本隶属度以及聚类图。

```
library(cluster)
fannyz=fanny(Y1,3,metric="SqEuclidean")
list("分类分布"=fannyz$clustering,"样本隶属度"=fannyz$membership)
```

```
## $分类分布
## 深圳能源 深南电Ａ 富龙热电 穗恒运Ａ 粤电力Ａ 韶能股份    ST惠天  城投控股
##        1        1        2        2        1        3        1        3
## 大连热电 华电能源 国电电力 长春经开 大龙地产 金丰投资   新黄浦 浦东金桥
##        1        3        1        1        2        2        1        3
##   外高桥 中华企业 渝开发Ａ 莱茵置业 粤宏远Ａ 中国国贸    万科A  三木集团
##        3        3        1        1        2        1        3        1
## 国兴地产   中关村 中兴通讯 长城电脑 南天信息 同方股份 永鼎股份 宏图高科
##        2        1        3        1        3        1        3        3
##   新大陆 方正科技 复旦复华
##        3        1        3
##
## $样本隶属度
##               [,1]       [,2]       [,3]
## 深圳能源 0.4627634 0.12736060 0.40987601
## 深南电Ａ 0.7345217 0.16278392 0.10269436
## 富龙热电 0.1051439 0.84951603 0.04534005
## 穗恒运Ａ 0.2759819 0.61552614 0.10849196
## 粤电力Ａ 0.5399324 0.07394796 0.38611962
## 韶能股份 0.4165089 0.06369778 0.51979333
## ST惠天   0.6117684 0.05954193 0.32868963
## 城投控股 0.3403638 0.14991098 0.50972524
## 大连热电 0.9122160 0.03022853 0.05755548
## 华电能源 0.4130828 0.09566793 0.49124932
## 国电电力 0.6881510 0.11404219 0.19780681
## 长春经开 0.5026950 0.15780149 0.33950355
## 大龙地产 0.1182967 0.82206364 0.05963965
## 金丰投资 0.1052065 0.84333634 0.05145713
## 新黄浦   0.4292140 0.41029852 0.16048745
## 浦东金桥 0.2472810 0.07509445 0.67762454
## 外高桥   0.2349805 0.05236040 0.71265906
## 中华企业 0.1273929 0.03223870 0.84036841
## 渝开发Ａ 0.4722216 0.35175332 0.17602508
## 莱茵置业 0.4560396 0.22828453 0.31567590
## 粤宏远Ａ 0.2398206 0.67411705 0.08606233
## 中国国贸 0.7950557 0.05062774 0.15431651
## 万科A    0.1700771 0.02752104 0.80240190
## 三木集团 0.5018610 0.25848080 0.23965818
## 国兴地产 0.2106219 0.65936591 0.13001216
## 中关村   0.5430426 0.18176835 0.27518905
## 中兴通讯 0.1392618 0.03291750 0.82782070
## 长城电脑 0.4506898 0.31274785 0.23656240
## 南天信息 0.3529242 0.25921507 0.38786075
## 同方股份 0.5827573 0.14226081 0.27498185
## 永鼎股份 0.1791461 0.03999553 0.78085840
## 宏图高科 0.1163364 0.02708346 0.85658013
## 新大陆   0.4257938 0.12898591 0.44522027
## 方正科技 0.6370565 0.23662056 0.12632292
## 复旦复华 0.4420571 0.04421677 0.51372617
```

```
clusplot(fannyz)
```




These two components explain 56.94 % of the point variability.

由于聚类图并不十分明显，因此，画出3D聚类图以提高视觉效果。

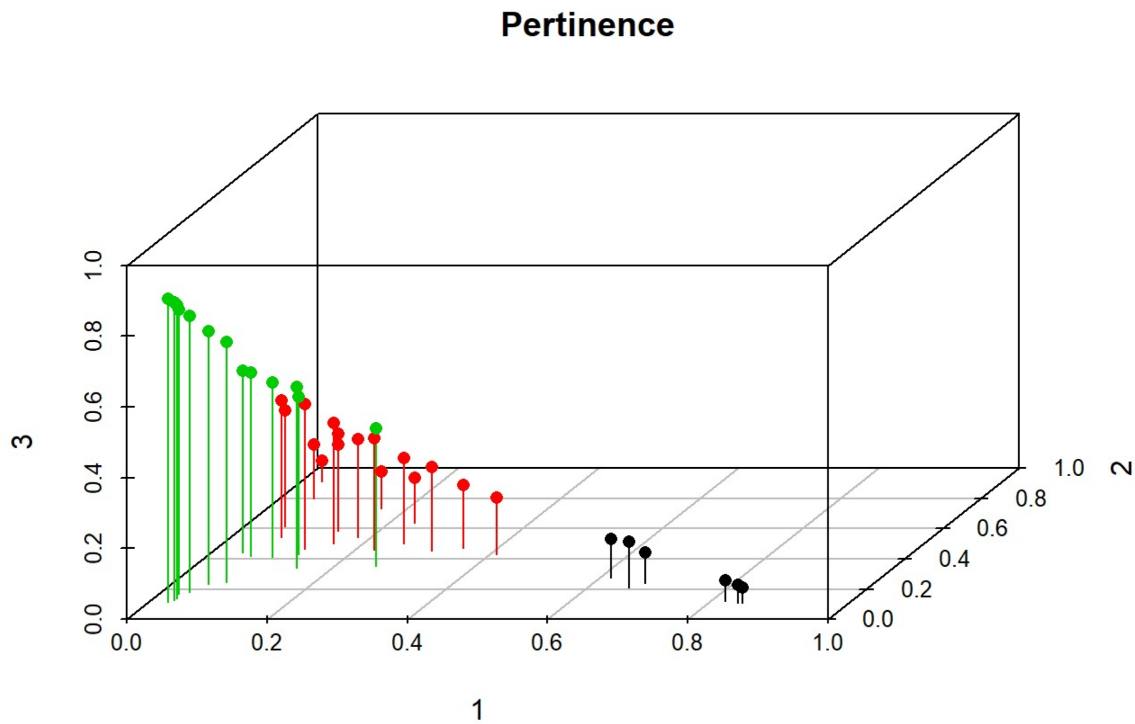
```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.4.4
```

```
result1<-cmeans(Y1,3,50)
library(scatterplot3d)
```

```
## Warning: package 'scatterplot3d' was built under R version 3.4.4
```

```
scatterplot3d(result1$membership, color=result1$cluster, type="h",
                     angle=55, scale.y=0.7, pch=16, main="Pertinence")
```




根据3D聚类图可以看出，类别之间的差距较为明显。事实上，该数据集中的公司来自于3大行业；为了验证聚类的合理性，将原行业与聚类后的类别进行对比，计算聚类的正确率。

```
t=table(Y[,2],fannyz$clustering)
sum(diag(prop.table(t)))
```

```
## [1] 0.4571429
```

遗憾的是，最终的正确率仅有45.71%。鉴于这种情况，或许模糊聚类的效率与准确性仍待商榷。