《多元统计分析》第六次上机作业
聚类分析和主成分分析

3180103000 许乐乐

- 实验目的与要求：

  通过本试验项目，使学生理解并掌握如下内容：

  1. 能够熟练利用R对数据进行聚类分析；
  2. 能够利用主成分分析方法进行变量降维。

- 实验内容：

习题一

现有16种饮料的热量、咖啡因含量、钠含量和价格的数据（见ex4.2),根据这4个变量对16种饮料进行聚类。

- 读取数据

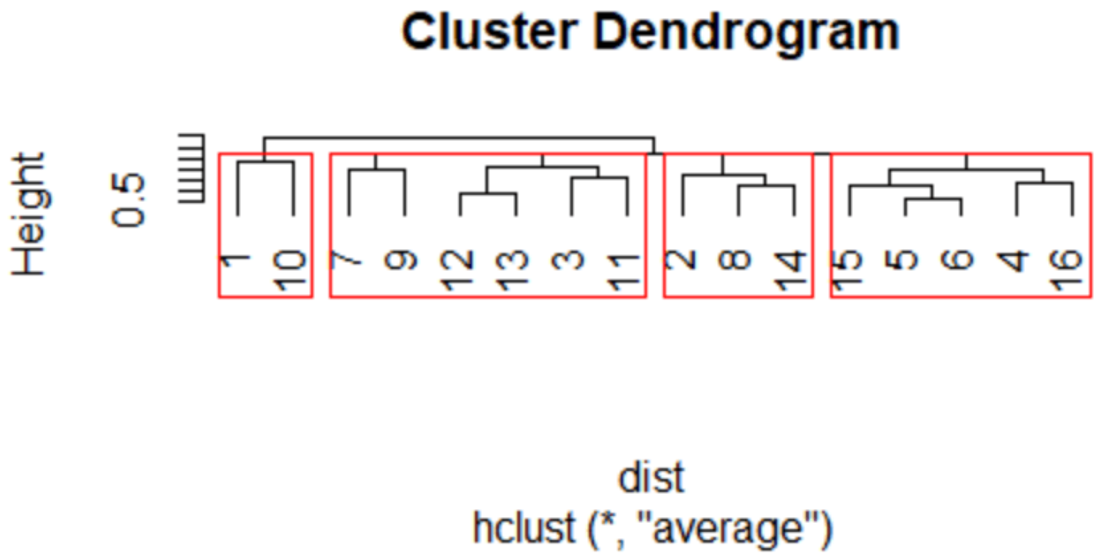
```
d<-read_csv("ex4.2.csv")
dd<-d[,2:5]
```

- 数据标准化

```
d.s<-scale(dd, center=T, scale=T)
```

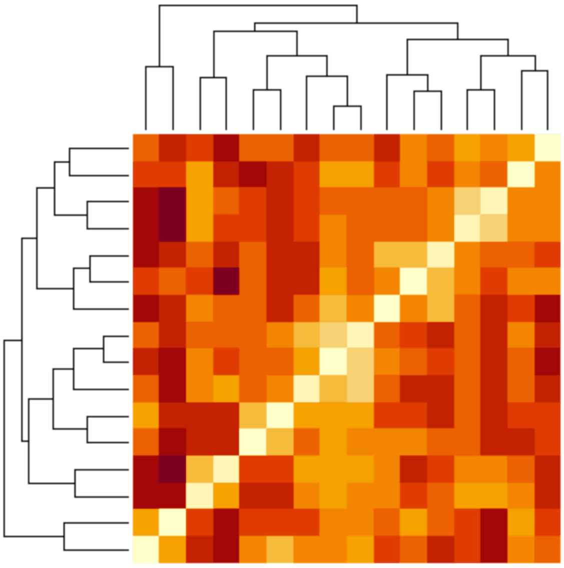
- 生成系统聚类

```
dist<-dist(d.s) # 计算类内距离
hc<-hclust(dist, "average") #计算类间距离（类平均法）
Opar<-par(mfrow=c(2,2))
plot(hc, hang=-1);re1<-rect.hclust(hc, k=4, border="red")
```



- 采用矩阵热图来发现类

```
dist.e<-dist(d.s, method="euclidean")
heatmap(as.matrix(dist.e), labRow=F, labCol=F)
```



```
# 从图中确定聚类的类的个数，然后进行聚类分析
model1=hclust(dist.e, method="ward.D")
re<-cutree(model1,k=4)
```

饮料	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
分类	1	2	3	1	3	3	4	2	4	1	3	4	4	2	3	1

- 动态聚类法

```
km<-kmeans(dist, 5, algorithm="MacQueen")
```

```
K-means clustering with 4 clusters of sizes 3, 6, 3, 4

Cluster means:
         1        2        3        4         5
1 2.012143 3.370107 3.749065 2.636555 2.2685085
2 3.977349 2.791789 1.681443 3.227834 2.3198112
3 3.103283 2.090545 1.985194 2.165289 0.5588516
4 3.303143 1.483124 2.721080 1.890449 2.3949882
          6        7        8        9       10
1 2.4146731 2.792278 2.977134 3.240309 1.410122
2 2.1846981 1.929909 2.771900 2.296561 3.625631
3 0.7242647 2.087205 2.649949 3.025186 2.419685
4 1.8670662 2.758240 1.414115 3.014008 3.372785
        11       12       13       14        15
1 4.183442 4.330209 3.531141 3.754148 2.7702258
2 2.369114 1.725434 1.486921 2.731340 2.6571207
3 2.115400 2.920142 2.190132 2.823880 0.8279874
4 3.345680 3.059485 2.385803 1.345966 3.0351927
        16
1 1.853340
2 3.310241
3 1.930439
4 2.877531
```

```
Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 
 1  4  2  4  3  3  2  4  2  1  2  2  2  4  3  1 

Within cluster sum of squares by cluster:
[1] 24.99658 56.25790  7.81510 24.99757
 (between_SS / total_SS =  52.0 %)

Available components:

[1] "cluster"      "centers"      "totss"       
[4] "withinss"     "tot.withinss" "betweenss"   
[7] "size"         "iter"         "ifault"      
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	4	2	4	3	3	2	4	2	1	2	2	2	4	3	1

- 分类结果：

  综上所述，动态聚类法与系统聚类法将饮料样品分为四类的结果几乎相同：

  第一类：1、10、16

  第二类：2、4、8、14

  第三类：3、7、9、11、12、13

  第四类：5、6、15

习题二

中国31个城市2011年的空气质量数据（见ex4.3）,根据这个数据对31个城市进行聚类分析。

- 读取数据

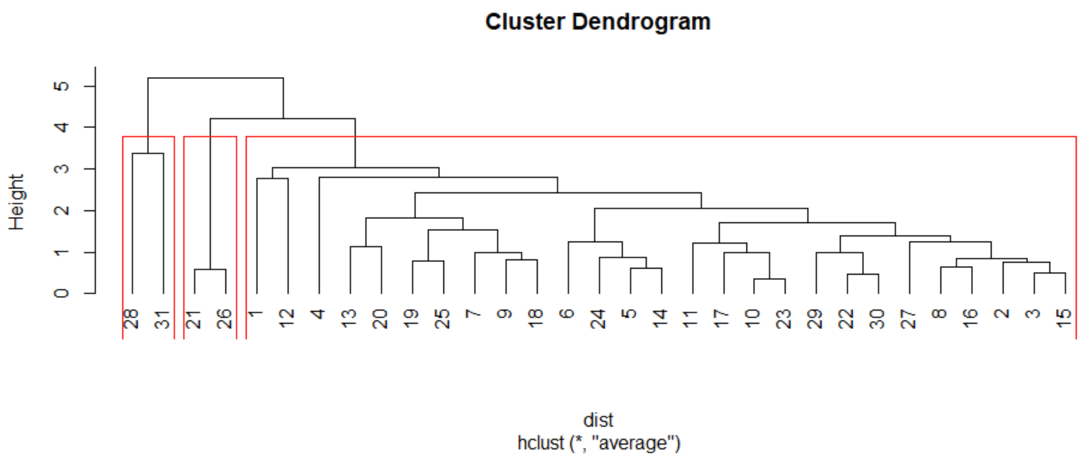
```
d<-read_csv("ex4.3.csv")
dd<-d[,2:6]
```

- 数据标准化

```
d.s<-scale(dd, center=T, scale=T)
```

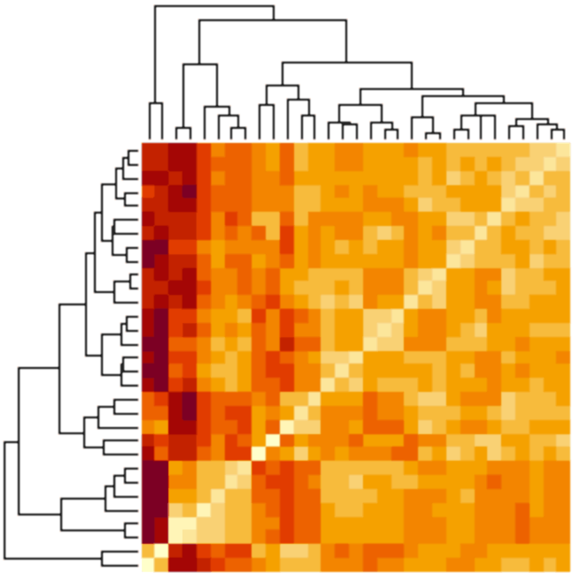
- 生成系统聚类

```
dist<-dist(d.s) # 计算类内距离
hc<-hclust(dist, "average") #计算类间距离（类平均法）
Opar<-par(mfrow=c(2,2))
plot(hc, hang=-1);re1<-rect.hclust(hc, k=3, border="red")
```



- 采用矩阵热图来发现类

```
dist.e<-dist(d.s, method="euclidean")
heatmap(as.matrix(dist.e), labRow=F, labCol=F)
```



```
# 从图中确定聚类的类的个数，然后进行聚类分析
model1=hclust(dist.e, method="ward.D")
re<-cutree(model1,k=3)
```

```
 [1] 1 1 1 1 1 1 2 1 2 1 1 1 2 1 1 1 1 2 2 2 2 1 1 1 2
[26] 2 1 3 1 1 3
```

- 动态聚类法

```
km<-kmeans(dist, 3, algorithm="MacQueen")
```

```
K-means clustering with 3 clusters of sizes 10, 4, 17

Cluster means:
         1        2        3        4        5
1 4.077880 2.068607 2.344343 3.347206 1.489169
2 4.728317 4.363805 4.485937 4.821016 4.755631
3 2.495061 1.262135 1.339044 2.405617 2.374726
         6        7        8        9       10
1 2.240890 1.392074 2.305871 1.533647 2.324112
2 4.692710 4.560649 4.418166 4.611182 4.468197
```

```
3 1.809185 2.178852 1.254563 2.190342 1.445615
        11        12        13        14        15
1 2.115916 3.504955 2.047131 1.639797 2.443600
2 4.784738 4.667572 4.539140 4.797090 4.490475
3 2.021111 2.296246 3.631636 2.202084 1.349731
        16        17        18        19        20
1 2.529389 3.017518 1.301587 1.475332 1.416214
2 4.515449 4.568941 4.582583 4.860646 4.388748
3 1.409654 1.794530 1.966825 3.141318 2.760272
        21        22        23        24        25
1 3.428177 2.044259 2.241061 1.532797 1.450213
2 4.445854 4.366202 4.540788 4.653525 4.876541
3 4.913014 1.475487 1.582126 2.431490 3.240618
        26        27        28        29        30
1 3.074468 3.183207 6.275272 2.746660 1.865403
2 4.322353 4.468197 4.982578 4.509333 4.433890
3 4.664231 1.573876 4.303717 1.652615 1.672586
        31
1 5.872377
2 5.171854
3 4.312857

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 
 3  3  3  3  1  3  1  3  1  3  3  3  1  1  3  3  3  1 
19 20 21 22 23 24 25 26 27 28 29 30 31 
 1  1  2  3  3  1  1  2  3  2  3  3  2 

Within cluster sum of squares by cluster:
[1] 132.8647 363.1814 265.9822
 (between_SS / total_SS =  53.0 %)

Available components:

[1] "cluster"      "centers"      "totss"       
[4] "withinss"     "tot.withinss" "betweenss"   
[7] "size"         "iter"         "ifault"      
```

习题三

某市工业部门13个行业8项重要经济指标数据，其中X1为年末固定资产净值（单位：万元）；X2为职工人数（单位：人），X3为工业总产值（单位：万元）；X4为全员劳动生产率（单位：元/人年）；X5为百元固定资产原值实现产值（单位：元）；X6为资金利税率（%）；X7为标准燃料消费量（单位：吨）；X8为能源利用效果（单位：万元/吨），数据见case6.1。根据这些数据进行主成分分析。

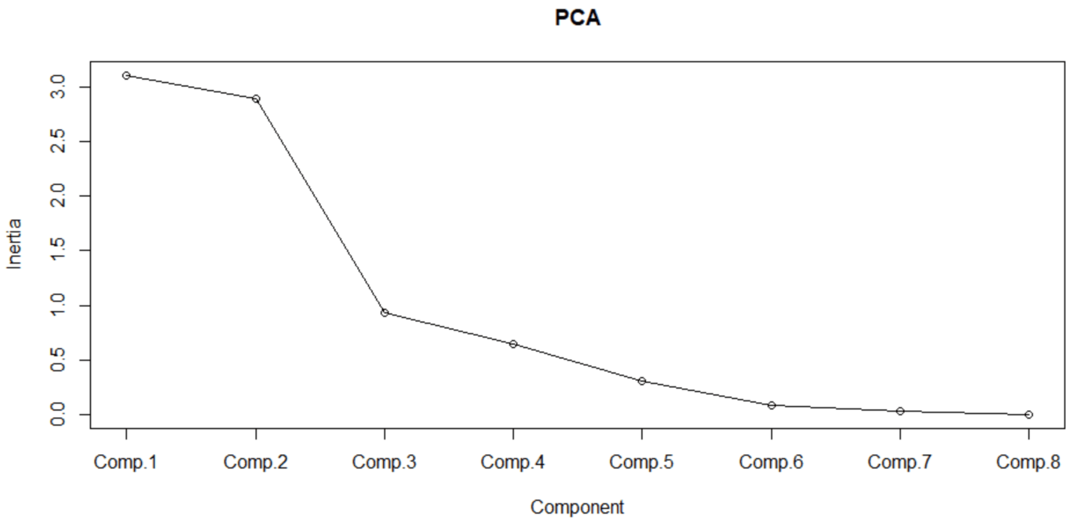
- 主成分分析

```
d<-read_csv("hw6_3_1.csv")
dd<-read_csv("hw6_3_2.csv")
PCA<-princomp(d, cor=T)
# 主成分的方差和累计占比
summary(PCA,loadings=TRUE)
screeplot(PCA, type="lines")
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     1.7620762 1.7021873 0.9644768
Proportion of Variance 0.3881141 0.3621802 0.1162769
Cumulative Proportion  0.3881141 0.7502943 0.8665712
```

```
                           Comp.4     Comp.5
Standard deviation     0.80132532 0.55143824
Proportion of Variance 0.08026528 0.03801052
Cumulative Proportion  0.94683649 0.98484701
                           Comp.6      Comp.7
Standard deviation     0.29427497 0.179400062
Proportion of Variance 0.01082472 0.004023048
Cumulative Proportion  0.99567173 0.999694778
                             Comp.8
Standard deviation     0.0494143207
Proportion of Variance 0.0003052219
Cumulative Proportion  1.0000000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
X1  0.477  0.296  0.104         0.184         0.758
X2  0.473  0.278  0.163 -0.174 -0.305        -0.518
X3  0.424  0.378  0.156                      -0.174
X4 -0.213  0.451         0.516  0.539 -0.288 -0.249
X5 -0.388  0.331  0.321 -0.199 -0.450 -0.582  0.233
X6 -0.352  0.403  0.145  0.279 -0.317  0.714
X7  0.215 -0.377  0.140  0.758 -0.418 -0.194
X8         0.273 -0.891        -0.322 -0.122
   Comp.8
X1  0.245
X2  0.527
X3 -0.781
X4  0.220
X5
X6
X7
X8
```



由Cumulative Proportion知，8项指标若综合为三个主成分，可解释原变量信息的86.66%；若综合为四个主成分，可解释原变量信息的94.68%

- 利用主成分得分进行排序

```
# 主成分得分（可以根据第一主成分得分进行排序，负数<正数）
PCAS<-PCA$scores
PCAS
```

```
        Comp.1      Comp.2      Comp.3      Comp.4
 [1,]  1.5354742  0.78961027  0.56001339  0.50981647
 [2,]  0.5185585 -2.69746855  0.23763437  0.88669141
 [3,]  1.0995810 -3.35723519  0.42612898  0.60624972
 [4,]  0.4786422  1.23197010 -1.03841942  1.66487001
 [5,]  4.7133932  2.35482336  0.48674014 -0.78901797
 [6,]  0.3434470 -1.84603673  0.03241021 -0.97630012
 [7,] -1.1475233 -0.33091560  0.29333399 -0.71995334
 [8,] -2.2846030  2.33577406  1.14409872  0.57948492
 [9,] -0.8755175  0.93223117  0.36727669  0.13377155
[10,] -2.1148303  0.85885133  0.24048868 -0.53512434
[11,] -0.7424575 -0.78646014 -0.12755551 -1.15634344
[12,] -1.2504626  0.03158169  0.29874009  0.08508599
[13,] -0.2737020  0.48327422 -2.92089030 -0.28923086
          Comp.5       Comp.6       Comp.7
 [1,]  1.10179178 -0.002674682  0.410987243
 [2,]  0.16712505 -0.302963497 -0.132417759
 [3,] -0.96793634  0.061794018  0.085555594
 [4,]  0.01184091  0.077608546 -0.008986494
 [5,] -0.51657036  0.019902643 -0.126040107
 [6,]  0.38398448  0.214601348 -0.028389532
 [7,]  0.09515880  0.315671049 -0.005296363
 [8,] -0.59525158  0.011742757 -0.041535263
 [9,]  0.54814203 -0.487867663 -0.299949326
[10,] -0.67391047 -0.185932496  0.290797020
[11,]  0.24384184 -0.397822037  0.018545326
[12,]  0.38556365  0.668578329 -0.176242612
[13,] -0.18377980  0.007361685  0.012972273
             Comp.8
 [1,]  0.0045906628
 [2,]  0.0696050796
 [3,] -0.0249830548
 [4,] -0.0540977524
 [5,]  0.0235021249
 [6,] -0.0695329414
 [7,] -0.0364517044
 [8,] -0.0545827148
 [9,] -0.0009447066
[10,]  0.0756972450
[11,] -0.0307115193
[12,]  0.0818480991
[13,]  0.0160611822
```

由第一主成分的得分知，由小到大对13个行业排的次序为：8，10，12，7，9，11，13，6，4，2，3，1，5。