# 实验六：聚类分析和主成分分析

沈雨萱 3180104691

## 目录



## 1 实验概况

一. 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。

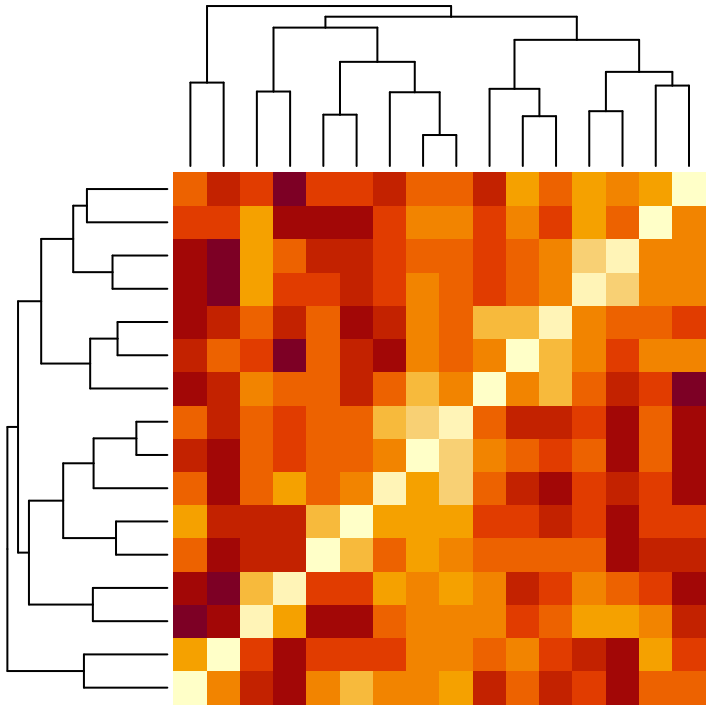
二. 实验内容（1）现有 16 种饮料的热量、咖啡因含量、钠含量和价格的数据（见 ex4.2), 根据这 4 个变量对 16 种饮料进行聚类。（2）中国 31 个城市 2011 年的空气质量数据（见 ex4.3）, 根据这个数据对 31 个城市进行聚类分析。（3）某市工业部门 13 个行业 8 项重要经济指标数据，其中 X1 为年末固定资产净值（单位：万元）；X2 为职工人数（单位：人），X3 为工业总产值（单位：万元）；X4 为全员劳动生产率（单位：元/人年）；X5 为百元固定资产原值实现产值（单位：元）；X6 为资金利税率（%）；X7 为标准燃料消费量（单位：吨）；X8 为能源利用效果（单位：万元/吨），数据见 case6.1。根据这些数据进行主成分分析。

## 2 实验结果

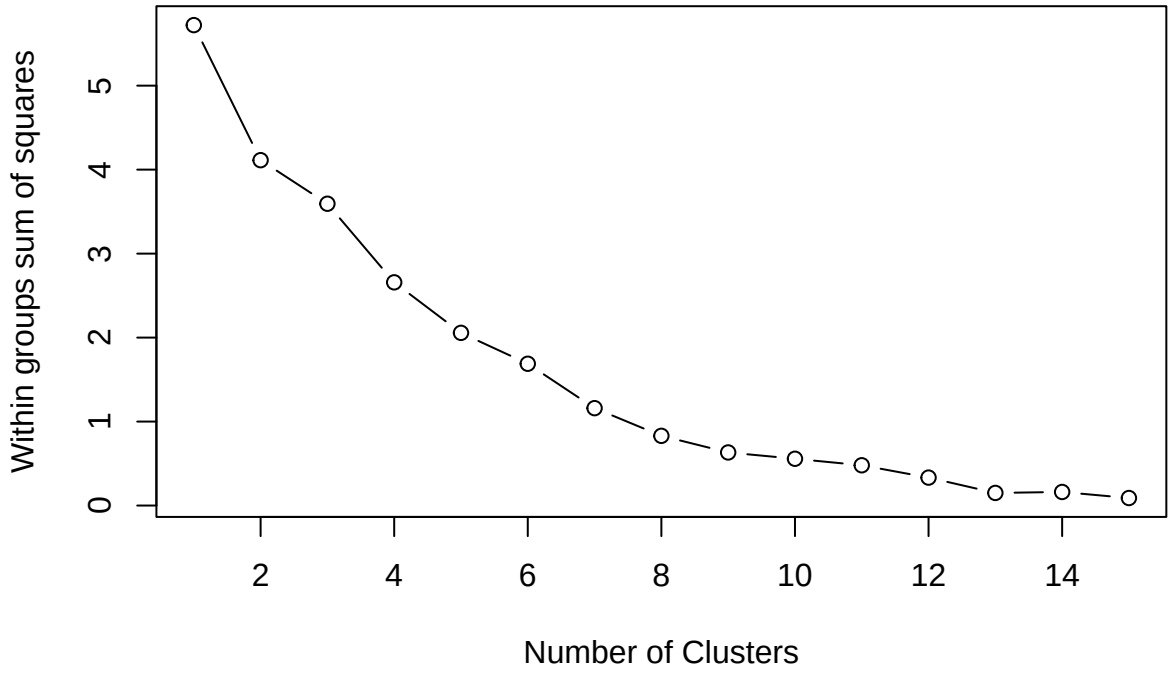
### 2.1 （1）

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。

```
data_0 <- read.csv("ex4.2.csv",encoding = "UTF-8",na.strings=c("","  ","NA"),header=T)
data_0 <- na.omit(data_0)
# 极差标准化
row.names(data_0) <- data_0[,1]
X<-data_0[,-1]
center<-sweep(X, 2, apply(X, 2, mean))# 按列中心化
R<-apply(X, 2, max)-apply(X, 2, min)# 计算列极差
X_star<-sweep(center, 2, R, "/")# 极差标准化，均值为 0，极差为 1
# 确定类数量
d<-dist(X_star,method = "euclidean")
heatmap(as.matrix(d),labRow = F, labCol = F)
```



```
wss <- (nrow(X_star)-1)*sum(apply(X_star,2,var))
for (i in 2:15)
wss[i] <- sum(kmeans(X_star,centers=i)$withinss)
### 这里的 wss(within-cluster sum of squares) 是组内平方和
plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within groups sum of squares")
```



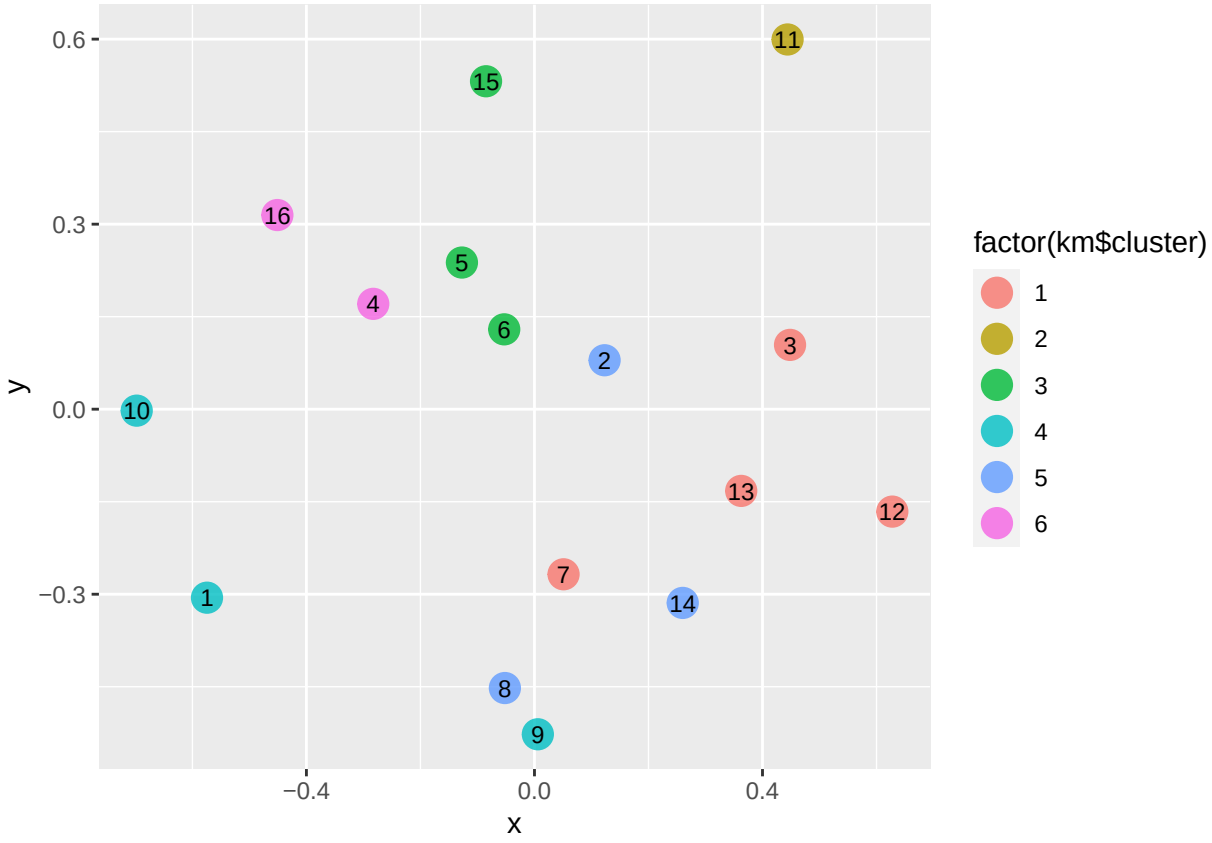

在分成 6 类时组内平方和下降存在一个拐点，结合热图利用 kmeans 将饮料分为 6 类。

```
km <- kmeans(X_star, 6, algorithm="MacQueen")
km
```

```
## K-means clustering with 6 clusters of sizes 4, 1, 3, 3, 3, 2
##
## Cluster means:
##         热量x1 咖啡因含量x2     钠含量x3      价格x4
## 1 -0.08976834   0.18125000 -0.20516304 -0.3211207
## 2 -0.18001931   0.55625000 -0.38722826  0.2909483
## 3  0.10987773   0.08958333 -0.10824275  0.2679598
## 4  0.38738739  -0.30625000  0.18161232 -0.1113506
## 5 -0.26850064   0.06875000  0.36277174 -0.1113506
## 6 -0.07360039  -0.41875000 -0.05027174  0.4288793
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
##  4  5  1  6  3  3  1  5  4  4  2  1  1  5  3  6
```

```
##
## Within cluster sum of squares by cluster:
## [1] 0.4652175 0.0000000 0.1160010 0.8980067 0.2553231 0.0897772
##  (between_SS / total_SS =  68.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

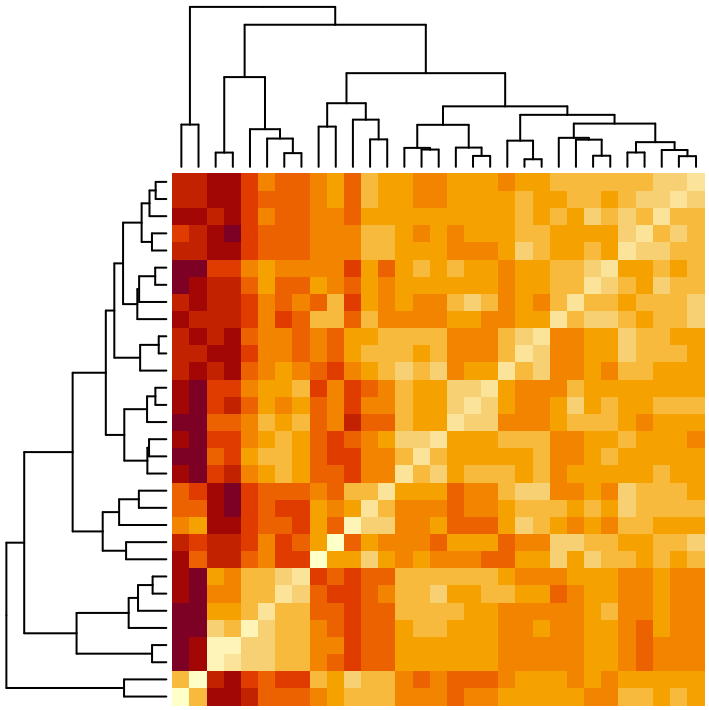
```
mds=cmdscale(d,k=2,eig=T)
x = mds$points[,1]
y = mds$points[,2]
p=ggplot(data.frame(x,y),aes(x,y))
p+geom_point(size=5,alpha=0.8,
             aes(colour=factor(km$cluster)))+geom_text(aes(x, y,label = rownames(X_star)),size=3)
```



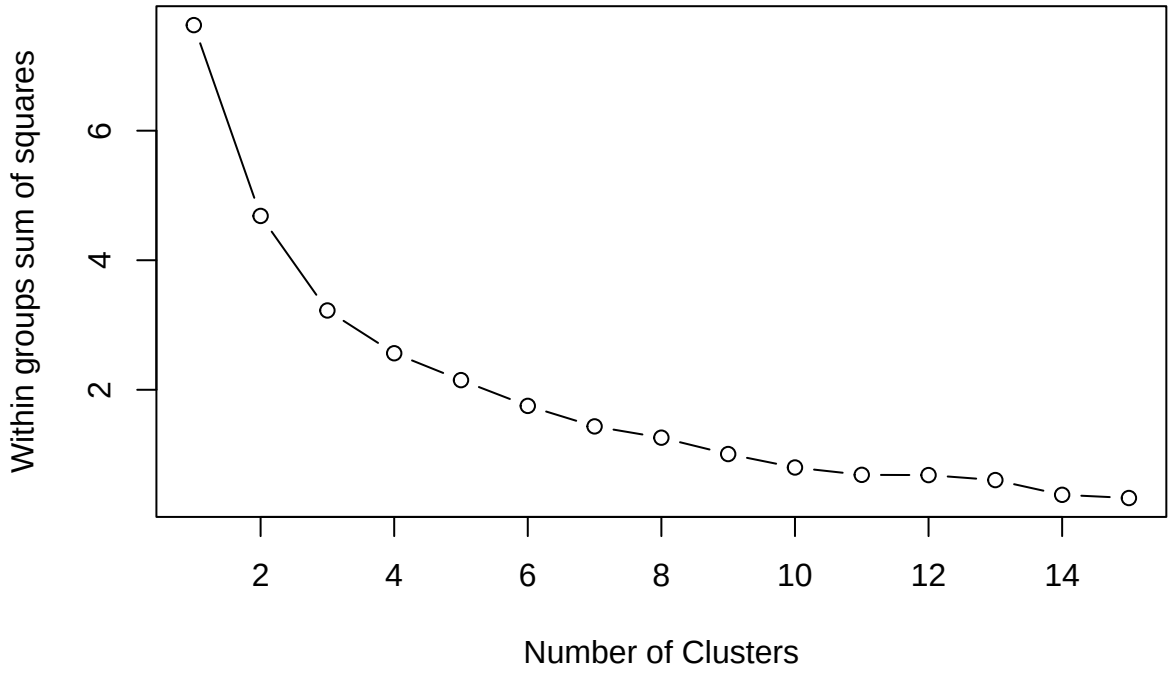
## 2.2 （2）

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

```
data_0 <- read.csv("ex4.3.csv",encoding = "UTF-8",na.strings=c("","  ","NA"),header=T,row.names = 1
X <- na.omit(data_0)
# 极差标准化
center<-sweep(X, 2, apply(X, 2, mean))# 按列中心化
R<-apply(X, 2, max)-apply(X, 2, min)# 计算列极差
X_star<-sweep(center, 2, R, "/")# 极差标准化，均值为 0，极差为 1
# 确定类数量
d<-dist(X_star,method = "euclidean")
heatmap(as.matrix(d),labRow = F, labCol = F)
```



```
wss <- (nrow(X_star)-1)*sum(apply(X_star,2,var))
for (i in 2:15)
wss[i] <- sum(kmeans(X_star,centers=i)$withinss)
### 这里的 wss(within-cluster sum of squares) 是组内平方和
plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within groups sum of squares")
```



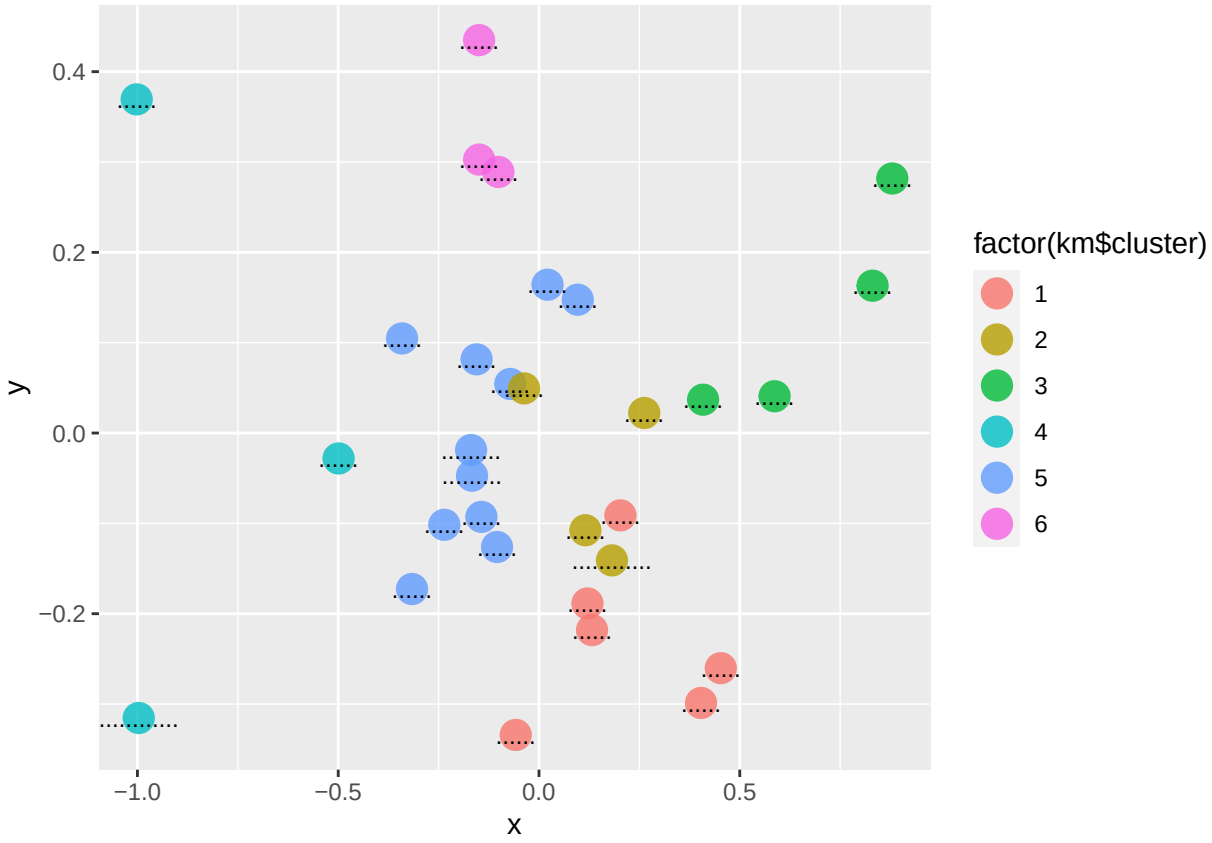

在分成 6 类时组内平方和下降存在一个拐点，结合热图利用 kmeans 城市分为 6 类

```
km <- kmeans(X_star, 6, algorithm="MacQueen")
km
```

```
## K-means clustering with 6 clusters of sizes 6, 4, 4, 3, 11, 3
##
## Cluster means:
##   可吸入颗粒物.PM10. 二氧化硫.SO2. 二氧化氮.NO2.
## 1        -0.11461898   -0.08188250    0.16311354
## 2        -0.06508153    0.21563756   -0.09498703
## 3        -0.36329941   -0.36709585   -0.27241063
## 4         0.37123878    0.17692405    0.29122359
## 5         0.09056443    0.03624216    0.04553808
## 6         0.09710420    0.05589742   -0.29456008
##   空气质量达到及好于二级的天数.天. 空气质量达到二级以上天数占全年比重...
## 1                      0.16551142                          0.16551142
## 2                      0.14002933                          0.14002933
## 3                      0.27432685                          0.27432685
```

```
## 4                    -0.48049409                          -0.48049409
## 5                    -0.06977533                          -0.06977533
## 6                    -0.14716076                          -0.14716076
##
## Clustering vector:
##     北京     天津   石家庄     太原 呼和浩特     沈阳     长春   哈尔滨
##        4        5        5        6        2        2        1        5
##     上海     南京     杭州     合肥     福州     南昌     济南     郑州
##        1        5        1        6        3        2        5        5
##     武汉     长沙     广州     南宁     海口     重庆     成都     贵阳
##        5        1        1        3        3        5        5        2
##     昆明     拉萨     西安     兰州     西宁     银川 乌鲁木齐
##        1        3        5        4        6        5        4
##
## Within cluster sum of squares by cluster:
## [1] 0.27764408 0.08399585 0.23374622 0.54809553 0.47680310 0.23070985
##  (between_SS / total_SS =  75.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
mds=cmdscale(d,k=2,eig=T)
x = mds$points[,1]
y = mds$points[,2]
p=ggplot(data.frame(x,y),aes(x,y))
p+geom_point(size=5,alpha=0.8,
             aes(colour=factor(km$cluster)))+geom_text(aes(x, y,label = rownames(X_star)),size=3)
```




## 2.3 （3）

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。

```
data_0 <- read.csv("case6.1.csv",encoding = "UTF-8",na.strings=c("","  ","NA"),header=T,row.names =
X <- na.omit(data_0)
# 极差标准化
center<-sweep(X, 2, apply(X, 2, mean))# 按列中心化
R<-apply(X, 2, max)-apply(X, 2, min)# 计算列极差
X_star<-sweep(center, 2, R, "/")# 极差标准化，均值为 0，极差为 1
X.pr <- princomp(X, cor = TRUE)
summary(X.pr, loadings=TRUE)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
```

```
## Standard deviation     1.7620762 1.7021873 0.9644768 0.80132532 0.55143824
## Proportion of Variance 0.3881141 0.3621802 0.1162769 0.08026528 0.03801052
## Cumulative Proportion  0.3881141 0.7502943 0.8665712 0.94683649 0.98484701
##                            Comp.6      Comp.7       Comp.8
## Standard deviation     0.29427497 0.179400062 0.0494143207
## Proportion of Variance 0.01082472 0.004023048 0.0003052219
## Cumulative Proportion  0.99567173 0.999694778 1.0000000000
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## X1  0.477  0.296  0.104         0.184         0.758  0.245
## X2  0.473  0.278  0.163 -0.174 -0.305        -0.518  0.527
## X3  0.424  0.378  0.156                      -0.174 -0.781
## X4 -0.213  0.451         0.516  0.539 -0.288 -0.249  0.220
## 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
```

```
predict(X.pr)
```

```
##          Comp.1      Comp.2      Comp.3      Comp.4      Comp.5       Comp.6
## 冶金  1.5354742  0.78961027  0.56001339  0.50981647  1.10179178 -0.002674682
## 电力  0.5185585 -2.69746855  0.23763437  0.88669141  0.16712505 -0.302963497
## 煤炭  1.0995810 -3.35723519  0.42612898  0.60624972 -0.96793634  0.061794018
## 化学  0.4786422  1.23197010 -1.03841942  1.66487001  0.01184091  0.077608546
## 机器  4.7133932  2.35482336  0.48674014 -0.78901797 -0.51657036  0.019902643
## 建材  0.3434470 -1.84603673  0.03241021 -0.97630012  0.38398448  0.214601348
## 森工 -1.1475233 -0.33091560  0.29333399 -0.71995334  0.09515880  0.315671049
## 食品 -2.2846030  2.33577406  1.14409872  0.57948492 -0.59525158  0.011742757
## 纺织 -0.8755175  0.93223117  0.36727669  0.13377155  0.54814203 -0.487867663
## 缝纫 -2.1148303  0.85885133  0.24048868 -0.53512434 -0.67391047 -0.185932496
## 皮革 -0.7424575 -0.78646014 -0.12755551 -1.15634344  0.24384184 -0.397822037
## 造纸 -1.2504626  0.03158169  0.29874009  0.08508599  0.38556365  0.668578329
## 文教 -0.2737020  0.48327422 -2.92089030 -0.28923086 -0.18377980  0.007361685
##             Comp.7        Comp.8
## 冶金  0.410987243  0.0045906628
## 电力 -0.132417759  0.0696050796
## 煤炭  0.085555594 -0.0249830548
```

```
## 化学 -0.008986494 -0.0540977524
## 机器 -0.126040107  0.0235021249
## 建材 -0.028389532 -0.0695329414
## 森工 -0.005296363 -0.0364517044
## 食品 -0.041535263 -0.0545827148
## 纺织 -0.299949326 -0.0009447066
## 缝纫  0.290797020  0.0756972450
## 皮革  0.018545326 -0.0307115193
## 造纸 -0.176242612  0.0818480991
## 文教  0.012972273  0.0160611822
```

由于前 4 个主成分已经达到 94%，因此就保留 4 个成分. 可以看到第一个主成分主要由 X1 为年末固定资产净值，X2 为职工人数，X3 为工业总产值决定，第二个主成分 X4 为全员劳动生产率，X5 为百元固定资产原值实现产值，X6 为资金利税率（%）占更多部分，第三个主成分主要由 X8 能源利用效果决定，第四个主成分主要由 X7 标准燃料消费量决定。