实验六：聚类分析和主成分分析

林泽钦 3160104013 统计 1601

目录

1 实验目的和要求

1.1 实验目的

通过本试验项目，能够理解并掌握如下内容：

- 熟练利用 R 对数据进行聚类分析；
- 利用主成分分析方法进行变量降维。

1.2 实验内容

1.2.1 饮料数据聚类

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。

1.2.2 城市空气质量聚类

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。

1.2.3 行业数据主成分分析

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。

1.2.4 消费数据主成分分析

我国 2010 各地区城镇居民家庭平均每人全年消费数据如 ex6.7 所示，这些数据指标分别从食品，衣着，居住，医疗，交通，通信，教育，家政和耐用消费品来描述消费。试对该数据进行主成分分析。

1.3 实验环境

- R-3.5.1
- RStudio

2 实验过程与结果

以下为实验过程中需要用到的程序包。

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

2.1 饮料数据聚类

2.1.1 数据读入与处理

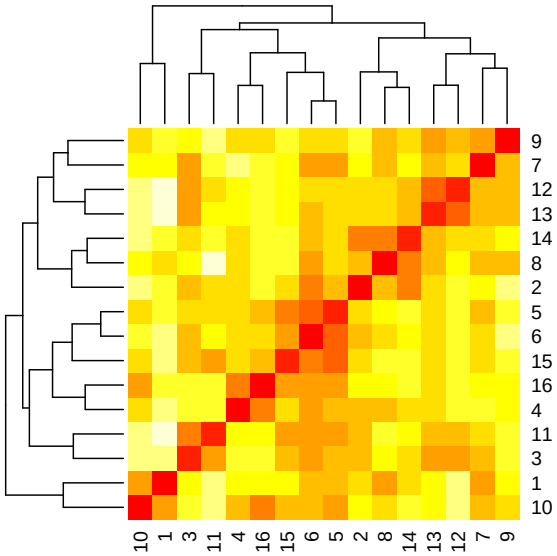
首先将实验提供的 ex4.2.xls 转化为容易读取的 csv 格式，然后读入数据，丢弃第一列 (饮料序号)。由于各个变量量纲不一致，所以将数据标准化。

```
data1 <- read.csv("Pro6Data1.csv", encoding = "UTF-8")
data1 <- scale(data1[, 2:5])
```

2.1.2 类数目的确定

我们需要确定聚类的数目，为此可以用距离矩阵的热图来发现类。先计算距离矩阵，然后使用 heatmap 函数即可绘出距离矩阵的热图，heatmap 使用的默认聚类方法为 hclust 函数默认的最大距离法。

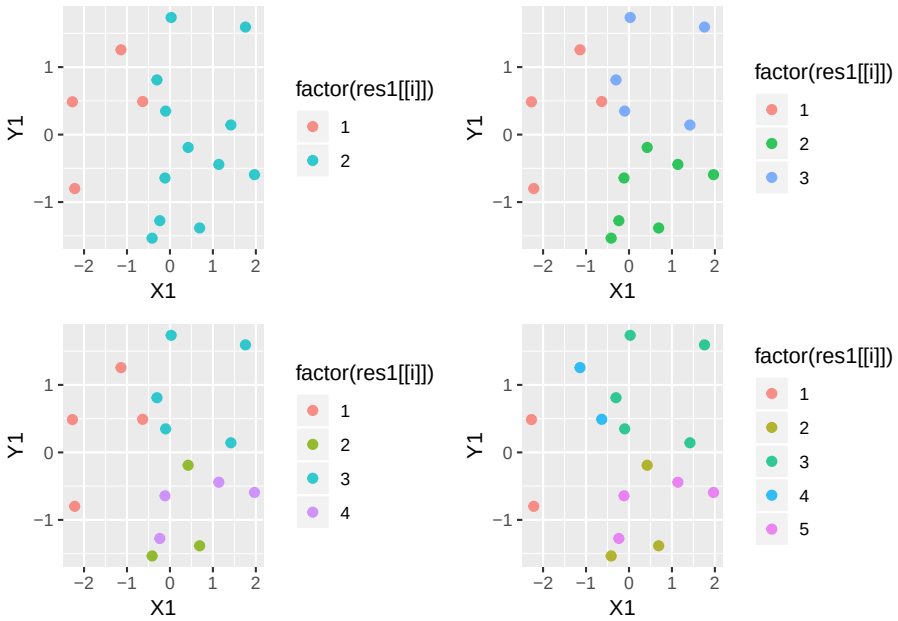
```
dist1 <- dist(data1, method = "euclidean")
heatmap(as.matrix(dist1))
```



上述图像很难看出应该分几个类，因此尝试通过主成分的方法来确定类的数目。为方便可视化，我们取前两个主成分。这么做的前提是前两个主成

分的贡献率足够高，考虑到样本只有四个变量，这种做法是可行的。下面的代码中，我们首先使用 Ward 离差平方和方法进行系统聚类，然后用 cutree 函数将样本分别切分为 2, 3, 4, 5 个类。然后对每一次切分都绘制出数据前两个主成分的平面图。

```
model1 <- hclust(dist1, method = "ward.D")
mds1 <- cmdscale(dist1, k = 2, eig = TRUE)
X1 <- mds1$points[, 1]
Y1 <- mds1$points[, 2]
res1 <- lapply(2:5, function(i) {
  cutree(model1, k = i)
})
pies1 <- lapply(1:4, function(i) {
  ggplot(data.frame(X1, Y1), aes(X1, Y1)) +
    geom_point(size = 2, alpha = 0.8, aes(colour = factor(res1[[i]])))
})
grid.arrange(pies1[[1]], pies1[[2]], pies1[[3]], pies1[[4]],
             ncol = 2, nrow = 2)
```



如上所示，当聚类数目大于 3 时，样本开始有交错的情况，因此，选择

将样本聚合成 3 个类。

2.1.3 系统聚类结果

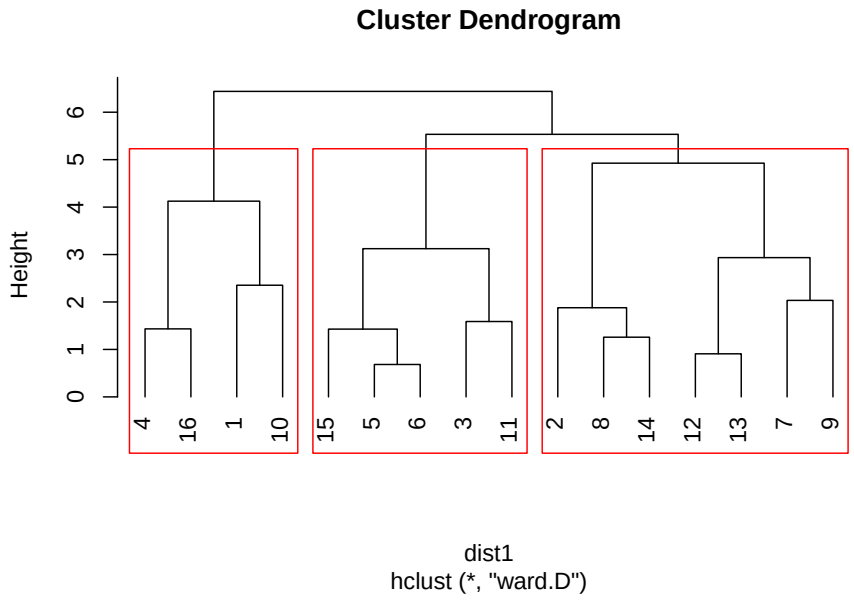
使用 Ward 样本离差阵法的系统聚类的结果为：

```
res1[[2]]
```

```
##  [1] 1 2 3 1 3 3 2 2 2 1 3 2 2 2 3 1
```

相应的图像为：

```
plot(model1, hang = -1)
rect.hclust(model1, k = 3, border = "red")
```



2.1.4 Kmeans 聚类结果

可以将上述聚类结果作为 Kmeans 的初始输入。具体来说，就是计算上述各类的中心，将其作为迭代的初始中心点，如下所示：

```r
center.list1 <- lapply(1:3, function(i) {
  colMeans(subset(data1, sapply(1:nrow(data1), function(j){
    res1[[2]][j] == i}))[])})
centers1 <- rbind(center.list1[[1]],
                  center.list1[[2]],
                  center.list1[[3]])
kmeans(data1, centers1, algorithm = "MacQueen")
```

```
## K-means clustering with 3 clusters of sizes 4, 7, 5
##
## Cluster means:
##            x1           x2         x3         x4
## 1  0.82704526 -1.053319677  0.5004200  0.7980331
## 2 -0.51035816  0.008598528  0.1449447 -0.8592063
## 3  0.05286523  0.830617803 -0.6032586  0.5644624
##
## Clustering vector:
##  [1] 1 2 3 1 3 3 2 2 2 1 3 2 2 2 3 1
##
## Within cluster sum of squares by cluster:
## [1] 11.547689 17.381343  6.333132
##  (between_SS / total_SS =  41.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

可以看到，kmeans 迭代得到的聚类结果与之前系统聚类的结果一致。

2.2 城市空气质量聚类

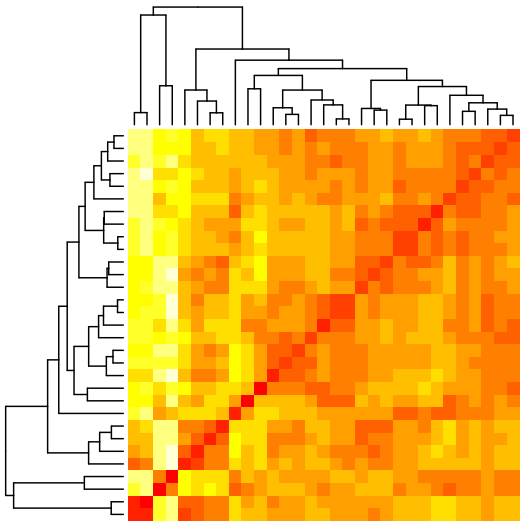
2.2.1 数据读入与处理

同样将提供的 ex4.3.xls 转化为容易读取的 csv 格式，然后读入数据，将第一列城市名作为各个观测的标签。由于最后一列 (空气质量达到二级以上天数占全年比重) 与倒数第二列 (空气质量达到及好于二级的天数) 成比例，所以删去最后一列。另外，各个变量量纲不一致，所以将数据标准化。

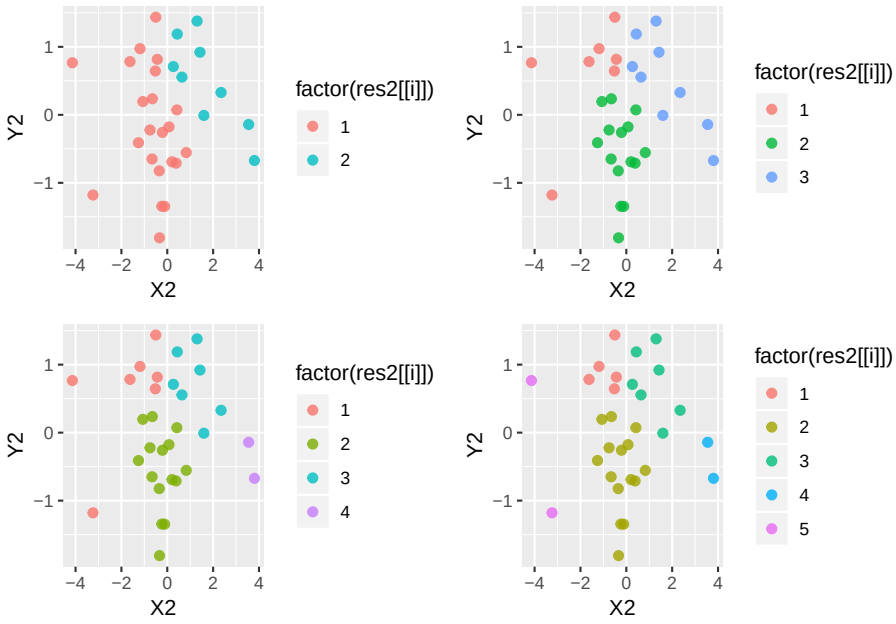
```
data2 <- read.csv("Pro6Data2.csv", fileEncoding = "UTF-8")
data2 <- scale(data2[, 2:5])
```

2.2.2 类数目的确定

步骤与前一部分基本一致。首先是矩阵热图，可以看到，依旧没什么用，可能是由于观测数目还不够多。

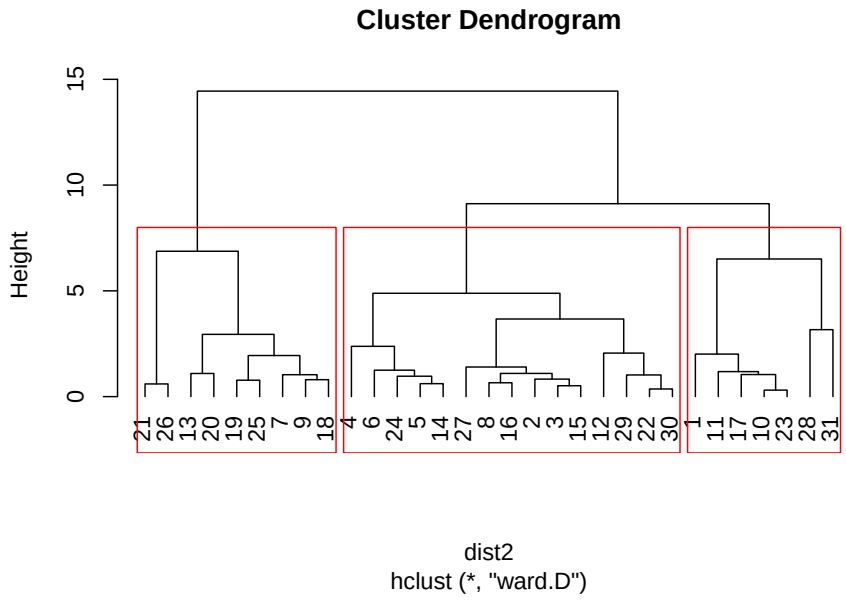


然后是主成分方法，如下图所示。可以看到各个分类都有一定的合理性。这里为了避免某个类的样本数目太少，我们将样本划分为 3 个类。



2.2.3 系统聚类结果

```
plot(model2, hang = -1)
rect.hclust(model2, k = 3, border = "red")
```



2.3 行业数据主成分分析

2.3.1 数据读取

与之前相同，读取数据并进行标准化。

```
data3 <- read.csv("Pro6Data3.csv", fileEncoding = "UTF-8")
industry <- data3$Industry
data3 <- scale(data3[2:9])
rownames(data3) <- industry
```

2.3.2 主成分分析

下面给出了样本主成分的贡献率以及载荷。可以看到，当取前四个主成分的时候，累计贡献率就高达 94.7%。

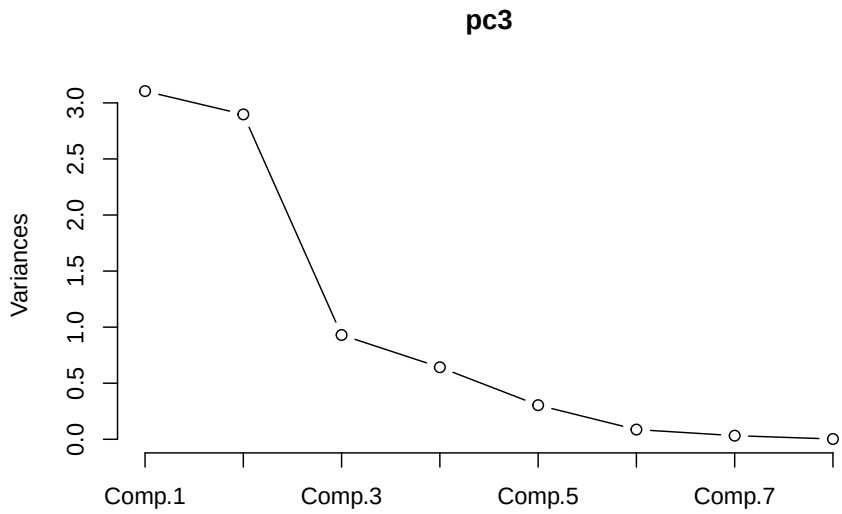
```
pc3 <- princomp(data3, cor = TRUE)
summary(pc3, loadings = TRUE)
```

```
## 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
```

下面是对应的碎石图。

```
screeplot(pc3, type = "lines")
```



各个样本的主成分得分如下：

```
predict(pc3)
```

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

```
## 冶金 -0.002674682  0.410987243  0.0045906628
## 电力 -0.302963497 -0.132417759  0.0696050796
## 煤炭  0.061794018  0.085555594 -0.0249830548
## 化学  0.077608546 -0.008986494 -0.0540977524
## 机器  0.019902643 -0.126040107  0.0235021249
## 建材  0.214601348 -0.028389532 -0.0695329414
## 森工  0.315671049 -0.005296363 -0.0364517044
## 食品  0.011742757 -0.041535263 -0.0545827148
## 纺织 -0.487867663 -0.299949326 -0.0009447066
## 缝纫 -0.185932496  0.290797020  0.0756972450
## 皮革 -0.397822037  0.018545326 -0.0307115193
## 造纸  0.668578329 -0.176242612  0.0818480991
## 文教  0.007361685  0.012972273  0.0160611822
```

2.4 消费数据主成分分析

2.4.1 数据读取

读取数据并进行标准化。

```
data4 <- read.csv("Pro6Data4.csv", fileEncoding = "UTF-8")
area <- data4$Area
data4 <- scale(data4[2:9])
rownames(data4) <- area
```

2.4.2 主成分分析

下面给出了样本主成分的贡献率以及载荷。可以看到，当取前四个主成分的时候，累计贡献率就高达 93.1%。

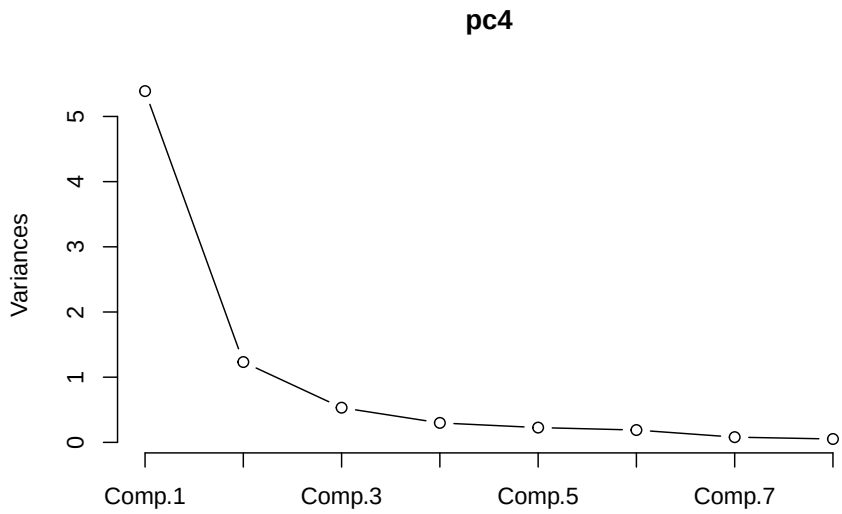
```
pc4 <- princomp(data4, cor = TRUE)
summary(pc4, loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4     Comp.5
## Standard deviation     2.3213318 1.1100881 0.72943408 0.5464987 0.47643779
```

```
## Proportion of Variance 0.6735727 0.1540369 0.06650926 0.0373326 0.02837412
## Cumulative Proportion  0.6735727 0.8276096 0.89411886 0.9314515 0.95982558
##                            Comp.6     Comp.7      Comp.8
## Standard deviation     0.4351019 0.28334611 0.227588880
## Proportion of Variance 0.0236642 0.01003563 0.006474587
## Cumulative Proportion  0.9834898 0.99352541 1.000000000
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## x1  0.358  0.396  0.158  0.288  0.503         0.282  0.522
## x2  0.257 -0.536  0.703        -0.130 -0.336         0.135
## x3  0.374        -0.412 -0.570 -0.112 -0.512  0.224  0.198
## x4  0.275 -0.599 -0.336         0.600  0.148 -0.248
## x5  0.393  0.292  0.137  0.120  0.166 -0.233  0.114 -0.795
## x6  0.386         0.195 -0.466 -0.178  0.729  0.168
## x7  0.396  0.264               -0.211        -0.837  0.152
## x8  0.361 -0.205 -0.373  0.599 -0.503  0.114  0.251
```

下面是对应的碎石图。

```
screeplot(pc4, type = "lines")
```



各个样本的主成分得分如下：

```
predict(pc4)
```

```
##             Comp.1      Comp.2       Comp.3       Comp.4      Comp.5
## 北京     4.8582289 -2.09242507 -0.359297546  1.254130515  1.08515552
## 天津     2.6825196 -0.97401186 -1.619890965  0.046361189  0.56263875
## 河北    -0.8677972 -0.78473178 -0.564157705  0.096880480 -0.27401520
## 山西    -0.5814968 -0.65025281 -0.462745156 -0.989863757 -0.55555864
## 内蒙古   0.2413737 -1.81467595  1.172934581  0.055893034 -0.28434548
## 辽宁     0.4870452 -0.36076668 -0.063224086 -0.886972143  0.69619790
## 吉林    -0.4088039 -1.18771573 -0.150746501 -1.231757095  0.23102101
## 黑龙江  -1.3983441 -1.18531922  0.306830704 -0.656247279  0.28011277
## 上海     7.0190860  1.85171579  1.018604590  0.403132469 -0.69681244
## 江苏     1.2367107  0.17453478  0.117134796  0.297373869 -0.37838868
## 浙江     3.9735319  0.07669694  1.624062309 -1.072005590  0.32292246
## 安徽    -0.8809803  0.33759733  0.008779653 -0.350298974 -0.21180207
## 福建     1.5189274  1.58234582  0.186643097 -0.134008568 -0.48731714
## 江西    -2.0072831  0.66564153  0.099106764  0.280933763 -0.35130939
```

```
## 山东    0.8937090 -1.12558708 -0.079310718  0.297027365 -0.89796779
## 河南   -1.0350837 -0.96749855 -0.388743034  0.308970111 -0.50611620
## 湖北   -0.9304652 -0.02309539  0.224015366  0.067408257 -0.07965707
## 湖南   -0.6464029 -0.16110037 -0.005220805 -0.008692239  0.22000481
## 广东    3.9297232  2.22419898 -1.042362598 -0.560538152  0.36766168
## 广西   -1.1820194  1.34072074 -0.577714511  0.573584432 -0.16757131
## 海南   -1.4200826  2.05221455 -1.663039483 -0.074564340  0.21904386
## 重庆    0.7299521 -0.82846680 -0.170110946  0.762160460 -0.25114153
## 四川   -1.4469078  0.71957115  0.511780963  0.426432287  0.27775825
## 贵州   -2.2836358  1.01849219 -0.027313885 -0.071021530 -0.14413516
## 云南   -2.1152790  0.63226283  0.599422940  0.581963877  0.95504756
## 西藏   -3.4779063  1.43895003  1.414873548  0.255938208  0.88176261
## 陕西   -0.2725945 -0.45319366 -0.352218459 -0.386721895 -0.09490918
## 甘肃   -1.9684347 -0.20859500 -0.072649099  0.096163122 -0.20213721
## 青海   -2.3225907  0.03089682 -0.221335827  0.314177057 -0.11419906
## 宁夏   -0.6272406 -0.80743068 -0.284424777  0.042503368 -0.20142956
## 新疆   -1.6974592 -0.52097286  0.820316791  0.261657698 -0.20051407
##              Comp.6       Comp.7       Comp.8
## 北京    0.225362094 -0.357122361 -0.234458565
## 天津    0.014004832  0.618092965  0.320831332
## 河北   -0.602605653 -0.159176117 -0.214158672
## 山西   -0.302827505 -0.205917478 -0.048148962
## 内蒙古 -0.511183047  0.085616644 -0.017085896
## 辽宁   -0.051238946  0.320643016  0.166176733
## 吉林   -0.456349966  0.087013566  0.058575124
## 黑龙江  0.216991027 -0.189214279 -0.055866606
## 上海   -0.075388354  0.252461369 -0.082678600
## 江苏    0.713184936 -0.459127021  0.252070181
## 浙江    0.761555981  0.252882463 -0.323769635
## 安徽    0.584586397  0.092557409  0.221585700
## 福建   -0.283132848 -0.067043550  0.297024546
## 江西   -0.056563752 -0.109454470  0.127996287
## 山东   -0.573657472  0.590681245 -0.119547637
```

```
## 河南    0.004112821 -0.026811230 -0.198471232
## 湖北    0.447563434  0.106898663  0.299523500
## 湖南    0.289053140 -0.302364812  0.217899553
## 广东   -0.966911492 -0.684612212 -0.045106742
## 广西    0.412285234  0.227667950 -0.399551873
## 海南    0.317808556  0.405541313 -0.255015842
## 重庆   -0.091092103  0.009881314  0.644253947
## 四川   -0.011292832 -0.044488005  0.121089205
## 贵州    0.256858016 -0.143298717 -0.004449726
## 云南   -0.323522372  0.128604886 -0.144565696
## 西藏   -0.618290799  0.169671703  0.172229228
## 陕西    0.857636178 -0.300993787 -0.046153947
## 甘肃    0.140164305 -0.141721777 -0.145356328
## 青海    0.206451631  0.124341818 -0.125641474
## 宁夏   -0.329441216 -0.032456010 -0.181896521
## 新疆   -0.194120226 -0.248754498 -0.257331382
```