第十二周作业

主成分分析与因子分析

9.1 主成分分析城市工业主体结构

确定主成分并对主成分进行解释

```
we91 <- data.frame(
  X1=c(90342,4903,6735,49454,139190,12215,2372,11062,17111,1206,2150,5251,14341),
  X2=c(52455,1973,21139,36241,203505,16219,6572,23078,23907,3930,5704,6155,13203),
  X3=c(101091,2035,3767,81557,215898,10351,8103,54935,52108,6126,6200,10383,19396),
  X4=c(19272,10313,1780,22504,10609,6382,12329,23804,21796,15586,10870,16875,14691),
  X5=c(82.0,34.2,36.1,98.1,93.2,62.5,184.4,370.4,221.5,330.4,184.2,146.4,94.6),
  X6=c(16.1,7.1,8.2,25.9,12.6,8.7,22.2,41.0,21.5,29.5,12.0,27.5,17.8),
  X7=c(197435,592077,726396,348226,139572,145818,20921,65486,63806,1840,8913,78796,6354),
  X8=c(0.172,0.003,0.003,0.985,0.628,0.066,0.152,0.263,0.276,0.437,0.274,0.151,1.574)
  );
we91.pr <- princomp(we91, cor = TRUE);
summary(we91.pr, loadings = TRUE);
```

```
> we91.pr
Call:
princomp(x = we91, cor = TRUE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
1.76207620 1.70218731 0.96447683 0.80132532 0.55143824 0.29427497 0.17940006  0.04941432
> summary(we91.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 #前n项贡献率
                           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:                                                      #新的变量和旧变量的关系
   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
```

由于前四个主成分的累积贡献率已达94.68%，为此可用该4个主成分来代替8个指标以达到降维的目的；
解释：

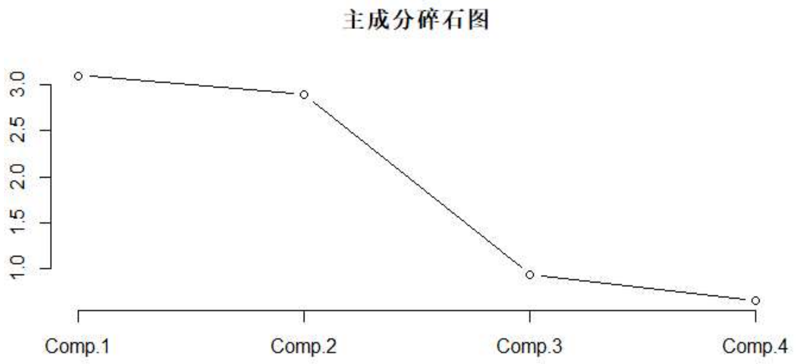
- 第1个主成分可称为行业规模因子，因为其对应系数较大的是前3个指标，即年末固定资产净值、职工人数和工业总产值，该3个指标都可反映行业的生产规模；由于系数为负，为此该因子越小，则行业规模越大；反之亦然；
- 第2个主成分可称为行业效率因子，因为其对应系数较大的是第4个指标，即全员劳动生产率。该因子越小，行业生产效率越高；
- 第3个主成分可称为行业利能因子，因为其对应系数较大的是第8个指标，即能源利用效果。该因子越大，行业能源利用效果越明显；
- 第4个主成分可称为行业耗能因子，因为其对应系数较大的是第7个指标，即标准燃料消费量。该因子越大，行业能源消费量越高。

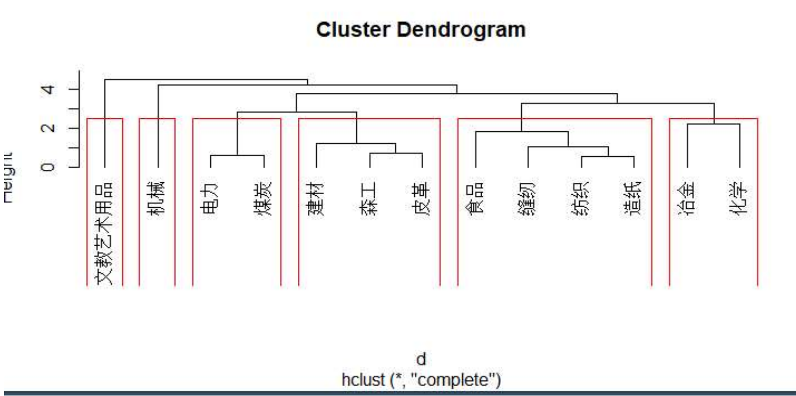
利用主成分得分对行业进行排序和分类

```
# 计算各样本的主成分得分
we91.prdt <- predict(we91.pr); we91.prdt
rownames(we91.prdt) <- c("冶金", "电力", "煤炭", "化学", "机械", "建材", "森工",
                          "食品", "纺织", "缝纫", "皮革", "造纸", "文教艺术用品");
# 基于主成分得分对行业进行排序(排序越靠前越好)
rownames(we91.prdt)[order(we91.prdt[,1])]
rownames(we91.prdt)[order(we91.prdt[,2])]
rownames(we91.prdt)[order(we91.prdt[,3])]
rownames(we91.prdt)[order(we91.prdt[,4])]

#得出主成分的碎石图
screeplot(we91.pr,npcs=4,type="lines",main="主成分碎石图")

d = dist(scale(we91.prdt[,1:4]), method = "euclidean");
hc1<-hclust(d,"complete"); #最长距离法聚类
plclust(hc1,hang=-1);rect.hclust(hc1, k=6,border="red"); #分类数可根据自己的理解设定;
```





9.2 消费品销售量回归方程

```
we92 = data.frame(
  X1=c(82.9,88.0,99.9,105.3,117.7,131.0,148.2,161.8,174.2,184.7),
  X2=c(92,93,96,94,100,101,105,112,112,112),
  X3=c(17.1,21.3,25.1,29.0,34.0,40.0,44.0,49.0,51.0,53.0),
  X4=c(94,96,97,97,100,101,104,109,111,111),
  Y=c(8.4,9.6,10.4,11.4,12.2,14.2,15.8,17.9,19.6,20.8));
lm.sol = lm(Y~X1+X2+X3+X4,data = we92);
> summary(lm.sol)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = we92)

Residuals:
        1         2         3         4         5         6         7         8 
 0.024803  0.079476  0.012381 -0.007025 -0.288345  0.216090 -0.142085  0.158360 
        9        10 
-0.135964  0.082310 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) -17.66768    5.94360  -2.973  0.03107 * 
X1            0.09006    0.02095   4.298  0.00773 **
X2           -0.23132    0.07132  -3.243  0.02287 * 
X3            0.01806    0.03907   0.462  0.66328   
X4            0.42075    0.11847   3.552  0.01636 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.2037 on 5 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9978
F-statistic:  1021 on 4 and 5 DF,  p-value: 1.827e-07
```

从结果来看，回归方程效果不太好，X3回归系数未通过显著性检验

```
we92.pr = princomp(~X1+X2+X3+X4, data = we92, cor = TRUE);
summary(we92.pr,loadings = TRUE)
Importance of components:
                          Comp.1      Comp.2     Comp.3       Comp.4
Standard deviation     1.9859037 0.199906992 0.11218966 0.0603085506
Proportion of Variance 0.9859534 0.009990701 0.00314663 0.0009092803
Cumulative Proportion  0.9859534 0.995944090 0.99909072 1.0000000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
X1 -0.502 -0.237  0.579  0.598
X2 -0.500  0.493 -0.610  0.367
X3 -0.498 -0.707 -0.368 -0.342
X4 -0.501  0.449  0.396 -0.626
```

由于前两个主成分的累积贡献率已达到99%，因此舍去其他主成分，达到降维的目的。

```
## 预测样本主成分，并作主成分分析
pre = predict(we92.pr);
we92$z1 = pre[,1]; we92$z2 = pre[,2];
lm.sol = lm(Y~z1+z2, data = we92);
summary(lm.sol)


Call:
lm(formula = Y ~ z1 + z2, data = we92)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74323 -0.29223  0.01746  0.30807  0.80849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.03000    0.17125  81.927 1.06e-11 ***
z1          -2.06119    0.08623 -23.903 5.70e-08 ***
z2          -0.62409    0.85665  -0.729     0.49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5415 on 7 degrees of freedom
Multiple R-squared:  0.9879,    Adjusted R-squared:  0.9845
F-statistic: 285.9 on 2 and 7 DF,  p-value: 1.945e-07
```

回归系数和回归方程均通过检验，效果显著。
Y = 14.03000 - 2.06119Z1 - 0.62409Z2

```
## 做变换，得到原坐标下的关系表达式
beta = coef(lm.sol);
A = loadings(we92.pr);
x.bar = we92.pr$center; x.sd = we92.pr$scale;
coef = (beta[2]*A[,1]+beta[3]*A[,2])/x.sd;
beta0= beta[1]-sum(x.bar*coef);
print(c(beta0,coef))
```

回归方程为：Y = -16.8846 + 0.03421X1 + 0.09376X2 + 0.11955X3 + 0.12360X4

9.3 女中学生的体型指标

```
x <- c(1.000, 0.846, 0.805, 0.859, 0.473, 0.398, 0.301,0.382,
    0.846, 1.000, 0.881, 0.826, 0.376,0.326, 0.277, 0.277,
    0.805, 0.881, 1.000, 0.801, 0.380,0.319, 0.237, 0.345,
    0.859, 0.826, 0.801, 1.000, 0.436, 0.329,0.327, 0.365,
    0.473, 0.376, 0.380, 0.436, 1.000,0.762, 0.730, 0.629,
    0.398, 0.326, 0.319, 0.329, 0.762,1.000, 0.583, 0.577,
    0.301, 0.277, 0.237, 0.327, 0.730,0.583, 1.000, 0.539,
    0.382, 0.415, 0.345, 0.365, 0.629,0.577, 0.539, 1.000);
names=c("身高 x1","手臂长 x2", "上肢长 x3","下肢长 x4", "体重 x5",
        "颈围 x6", "胸围 x7","胸宽 x8");
r <- matrix(x, nrow=8, dimnames=list(names, names));
source("factor.analy1.R");
fa <- factor.analy1(r, m = 2); fa # 选取2个因子

$method
[1] "Principal Component Method"

$loadings
      Factor1    Factor2
X1 -0.8624962 -0.3785039
X2 -0.8444116 -0.4447482
X3 -0.8162445 -0.4631786
X4 -0.8426517 -0.4011731
X5 -0.7580163  0.5136264
X6 -0.6740489  0.5229470
X7 -0.6168803  0.5693859
X8 -0.6429662  0.4554198

$var
      common    spcific
X1 0.8871649 0.11283507
X2 0.9108319 0.08916808
X3 0.8807894 0.11921058
X4 0.8710016 0.12899837
X5 0.8384007 0.16159927
X6 0.7278156 0.27218442
X7 0.7047417 0.29525833
X8 0.6208127 0.37918733

$B
                 Factor1   Factor2
SS loadings    4.6561250 1.7854336
Proportion Var 0.5820156 0.2231792
Cumulative Var 0.5820156 0.8051948

> vm1 <- varimax(fa$loadings, normalize = F); vm1
$loadings

Loadings:
   Factor1 Factor2
X1 -0.913   0.232
X2 -0.939   0.168
X3 -0.929   0.136
X4 -0.911   0.201
X5 -0.282   0.871
X6 -0.210   0.827
X7 -0.137   0.828
X8 -0.227   0.754

               Factor1 Factor2
SS loadings      3.603   2.839
Proportion Var   0.450   0.355
Cumulative Var   0.450   0.805

$rotmat
          [,1]       [,2]
[1,] 0.7888429 -0.6145949
[2,] 0.6145949  0.7888429
```

结论：在计算结果中，因子Factor1前几个变量（X1,X2,X3,X4）的载荷因子接近1，可称Factor1是长度因子。而因子Factor2后几个变量（X5,X6,X7,X8）的载荷因子接近1，可称Factor2是宽度因子。

9.4 学生5门课成绩的公共因子

```
we94 = data.frame(
  x1 = c(99,99,100,93,100,90,75,93,87,95,76,85),  # 政治
  x2 = c(94,88,98,88,91,78,73,84,73,82,72,75),    # 语文
  x3 = c(93,96,81,88,72,82,88,83,60,90,43,50),    # 外语
  x4 = c(100,99,96,99,96,75,97,68,76,62,67,34),   # 数学
  x5 = c(100,97,100,96,78,97,89,88,84,39,78,37)); # 物理
r94 = cor(we94);
fa94 <- factor.analy1(r94, m = 3); fa94 # 选取3个因子
vm94 <- varimax(fa94$loadings, normalize = F); vm94
```

- f1（政治语文）可称为文科因子;
- f2（数学物理）可称为理工因子;
- f3（外语）可称为外语因子