Chapter 5 - PCA

张笑竹 / 201618070114

2018年11月16日

第五章实验采用课本例5.1，例5.2和例5.3中给出的数据进行主成分分析。例5.1是Midwestern银行在1969-1971年之间雇员情况的数据；例5.2是我国部分省、直辖市、自治区独立核算的工业企业的经济效益指标数据；例5.3为全国重点水泥企业经济效益综合评价。
将本次的实验任务拆分如下：
1）调用函数princomp()，对例5.1进行主成分分析；
2）将函数进行封装，从而输出更加清晰的结果；
3）调用2）中的封装函数对例5.2进行主成分分析；
4）调用2）中的封装函数对例5.3进行主成分分析，并对排名情况进行分析。

1 例5.1的分析

1.1 选择主成分个数

首先，在R中读入例5.1的数据集，并对数据集的“行名”等进行处理。

```
X=read.csv("eg5_1.csv",header=T)
rownames(X)=X[,1]
X=X[,-1]
X1=X[,c(3,5:8)]
p=ncol(X1)
head(X1)
```

```
##   educ salary salbegin jobtime prevexp
## 1   15  57000    27000      98     144
## 2   16  40200    18750      98      36
## 3   12  21450    12000      98     381
## 4    8  21900    13200      98     190
## 5   15  45000    21000      98     138
## 6   15  32100    13500      98      67
```

上面输出了数据集的前6行，可以看出，经过处理后，数据集内的各个变量全部为连续变量。
下面，调用princomp()函数，进行主成分分析。需要注意的是，由于数据集的各个变量量纲不同，且差异较大，所以此处利用样本相关阵进行分析。

Step1: 主成分方差
求主成分的方差，就是求样本协方差矩阵Σ(此处指相关阵R)的特征值$\lambda_1,\lambda_2,...,\lambda_p$.

```
PCA.X=princomp(X1,cor=TRUE)
summary(PCA.X)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.5738588 1.0258172 1.0017379 0.60410805 0.31975329
## Proportion of Variance 0.4954063 0.2104602 0.2006958 0.07298931 0.02044843
## Cumulative Proportion  0.4954063 0.7058665 0.9065623 0.97955157 1.00000000
```

```
lambda=PCA.X$sdev^2
lambda
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 2.4770314 1.0523010 1.0034789 0.3649465 0.1022422
```

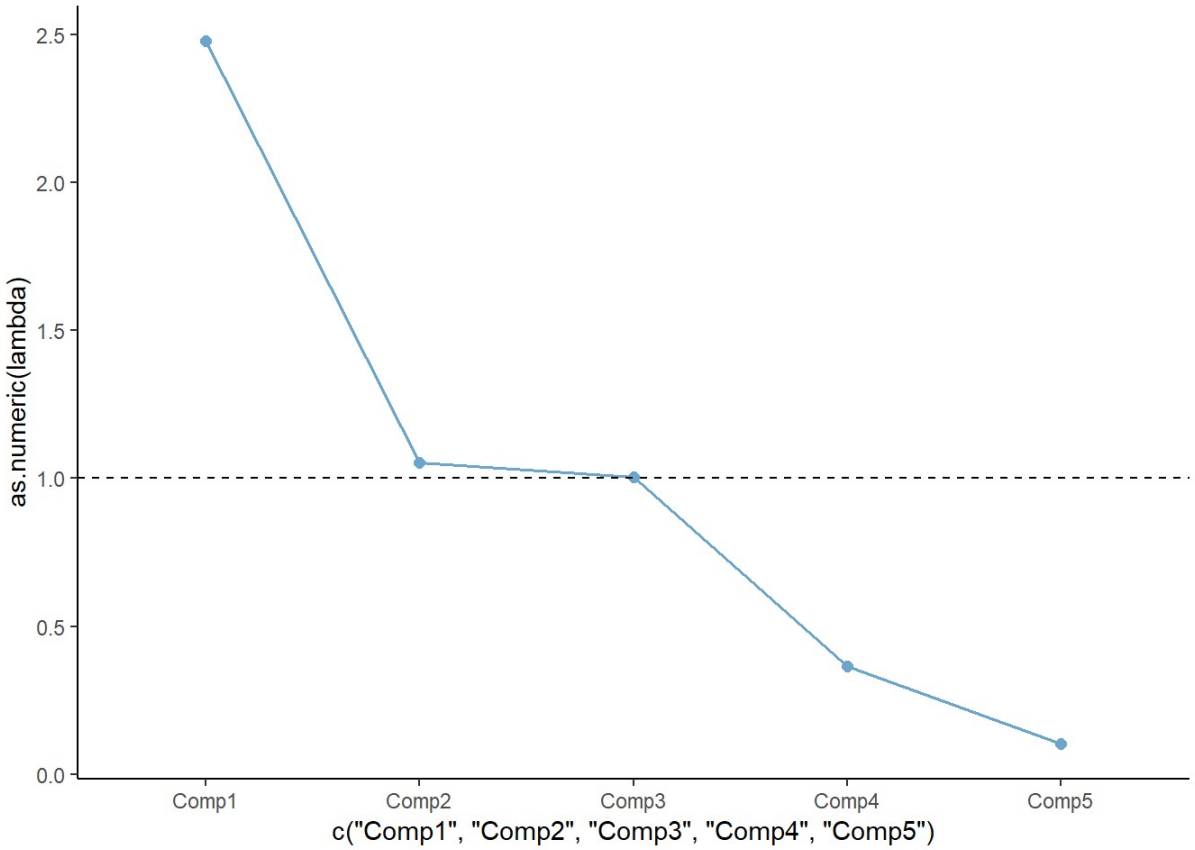
输出的结果中，“Standard deviation”指主成分的标准差，“Proportion of Variance”指方差贡献率，“Cumulative Proportion”指方差的累计贡献率。最后输出的“lambda”，是主成分的方差，也就是λ_i. 选择主成分个数m时，一个常用的原则，是取m使得

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geq 85\%$$

也就是取m使得其对应的“Cumulative Proportion”大于等于85%. 按照这种原则，应该选择前三个主成分。

Step2: 碎石图
事实上，在实际应用中，有些研究工作者习惯保留特征根大于1的那些主成分。因此，为了更加直观地反映特征根的大小与变化情况，可以绘制碎石图进行分析。

```
library(ggplot2)
ggplot(as.data.frame(lambda),aes(x=c("Comp1","Comp2","Comp3","Comp4","Comp5"), y=as.numeric(lambda), group=1)) +
  geom_line(colour="skyblue3", size=0.7)+geom_point(size=2,colour="skyblue3")+
  geom_hline(yintercept = 1, linetype = 2)+
  theme(panel.background = element_blank(),axis.line = element_line(colour = "black"))
```



图中，前3个特征根在虚线1之上。因此，按照这种原则，应该选择前三个主成分，这与第一个原则的结果是一

致的。

1.2 系数矩阵，因子负荷量，和主成分对原始变量的方差贡献率

Step1: 系数矩阵

求系数矩阵，就是求特征值$\lambda_1,\lambda_2,...,\lambda_p$所对应的标准化特征向量$\gamma_1,\gamma_2,...,\gamma_p$，以及它们所组成的矩阵.

```
k=3
#系数矩阵(特征向量)
eigV=PCA.X$loadings[1:p,1:k]
eigV
```

```
##               Comp.1      Comp.2      Comp.3
## educ      0.53764985  0.18897993 -0.01396416
## salary    0.59745724 -0.10183394  0.02852329
## salbegin  0.58245045 -0.25695231 -0.07676654
## jobtime   0.04324322  0.05093131  0.99415942
## prevexp  -0.11339743 -0.94090273  0.06888661
```

根据输出的系数矩阵，若记educ, salary, salbegin, jobtime, prevexp分别为X_1,X_2,X_3,X_4,X_5,则三个主成分可表示为：

$$Y_1 = 0.538X_1 + 0.597X_2 + 0.582X_3 + 0.043X_4 - 0.113X_5$$

$$Y_2 = 0.189X_1 - 0.102X_2 - 0.257X_3 + 0.051X_4 - 0.941X_5$$

$$Y_3 = -0.014X_1 + 0.029X_2 - 0.077X_3 + 0.994X_4 + 0.069X_5$$

Step2: 因子负荷量

因子负荷量的的绝对值大小刻画了该主成分的主要意义及其成因，反映了主成分和原始变量的相关程度。

第k个主成分Y_k与原始变量X_i的因子负荷量可表示为：

$$\rho(Y_k,X_i) = \frac{\gamma_{ik}\sqrt{\lambda_k}}{\sigma_{ii}} = \gamma_{ik}\sqrt{\lambda_k}$$

```
#Component Matrix(因子负荷量)
ComM=c()
for (i in 1:k){
  ComM=cbind(ComM,eigV[,i]*sqrt(lambda[i]))
}
ComM
```

```
##                [,1]        [,2]        [,3]
## educ      0.84618493  0.19385887 -0.01398843
## salary    0.94031332 -0.10446301  0.02857286
## salbegin  0.91669475 -0.26358611 -0.07689996
## jobtime   0.06805872  0.05224622  0.99588719
## prevexp  -0.17847154 -0.96519423  0.06900633
```

通过结果不难看出，第一主成分与变量educ, salary, salbegin联系密切，可主要被它们所解释；第二主成分与变量prevexp联系密切；而第三主成分则与变量jobtime联系密切。

Step3：主成分对原始变量的方差贡献率
主成分对原始变量的方差贡献率，可以大体刻画从每个原始变量中提取的信息量，由X_i与前m个主成分$Y_1, Y_2, \ldots, Y_m$的相关系数平方和计算得到：

$$v_i = \sum_{k=1}^{m} \rho^2(Y_k, X_i) = \frac{1}{\sigma_{ii}} \sum_{k=1}^{m} \lambda_k \gamma_{ik}^2$$

```
#Communalities(提取信息量)
Communal=c()
for (i in 1:p){
  ext=sum(ComM[i,]^2)
  Communal=rbind(Communal,ext)
}
rownames(Communal)=colnames(X1)
Communal
```

```
##               [,1]
## educ     0.7538059
## salary   0.8959181
## salbegin 0.9157205
## jobtime  0.9991530
## prevexp  0.9682139
```

得到的结果显示，除受教育程度信息损失较大外，主成分几乎包含了各个原始变量至少90%的信息。

1.3 计算得分

最后，可以计算出各个样品的主成分值。对于第j个样品的第k个主成分，计算公式如下：

$$\hat{Y}_{jk} = \sum_{i=1}^{p} \gamma_{ik} x_{ji}^*$$

其中，x_{ji}^*为样本资料阵中，标准化了的、第j行第i列的观测值。

```
#计算各个样品的主成分值
PCA.X.Score=as.matrix(scale(X1))%*%as.matrix(eigV)
head(PCA.X.Score)
```

```
##        Comp.1     Comp.2   Comp.3
## 1  1.83043178 -0.7093704 1.633739
## 2  0.93556441  0.6972791 1.610167
## 3 -1.33954290 -2.3363343 1.891282
## 4 -1.77338442 -0.9219074 1.773888
## 5  0.97305372 -0.3879458 1.668263
## 6  0.04365743  0.5725844 1.673102
```

结果展示了前6个样品的主成分值。
此外，以各个主成分的方差为权重，可以计算出各个样品的总得分，计算公式为：

$$score_j = \frac{\sum_{k=1}^{m} \lambda_k \hat{Y}_{jk}}{\sum_{k=1}^{m} \lambda_k}$$

```
#计算得分
scores=c()
for (i in 1:nrow(X1)){
  scores1=0
  for (j in 1:k){
    w=lambda[j]/sum(lambda[1:k])
    scores1=scores1+PCA.X.Score[i,j]*w
  }
  scores=c(scores,scores1)
}
X.score=cbind(X1,scores)
head(X.score)
```

```
##   educ salary salbegin jobtime prevexp     scores
## 1   15  57000    27000      98     144  1.1972676
## 2   16  40200    18750      98      36  1.0295903
## 3   12  21450    12000      98     381 -0.8557062
## 4    8  21900    13200      98     190 -0.7904132
## 5   15  45000    21000      98     138  0.8110012
## 6   15  32100    13500      98      67  0.5271770
```

输出的结果为新数据集的前6个样品。新数据集的最后一列，就是计算得到的得分。

2 例5.2的分析

2.1 函数的封装

将第1节中各个函数的输出进行整理，可以得到更为清晰的结果。函数封装如下：

```
Pca <- function(X,cor,z0){
  p=ncol(X)
  library(psych)
  PCA.X=princomp(X,cor) #cor=TRUE or cor=FALSE

  #主成分方差(特征根)
  lambda=PCA.X$sdev^2
  #碎石图
  library(ggplot2)
  pl<- ggplot(as.data.frame(lambda),aes(x=as.character(c(1:p)), y=as.numeric(lambda), group=1))
+
    geom_line(colour="skyblue3", size=0.7)+geom_point(size=2,colour="skyblue3")+
    geom_hline(yintercept = z0, linetype = 2)+
    theme(panel.background = element_blank(),axis.line = element_line(colour = "black"))

  #选择主成分个数
  k=0
  for(i in 1:p){
    if(lambda[i]>z0){
      k=k+1
    }
  }

  #系数矩阵(特征向量)
  eigV=PCA.X$loadings[1:p,1:k]

  #Component Matrix(因子负荷量)
  ComM=c()
  for (i in 1:k){
    ComM=cbind(ComM,eigV[,i]*sqrt(lambda[i]))
  }

  #Communalities(提取信息量)
  Communal=c()
  for (i in 1:p){
    ext=sum(ComM[i,]^2)
    Communal=rbind(Communal,ext)
  }
  rownames(Communal)=colnames(X)

  #计算得分
  PCA.X.Score=as.matrix(scale(X))%*%as.matrix(eigV)
  scores=c()
  for (i in 1:nrow(X)){
    scores1=0
    for (j in 1:k){
      w=lambda[j]/sum(lambda[1:k])
      scores1=scores1+PCA.X.Score[i,j]*w
    }
    scores=c(scores,scores1)
  }

  X.score=cbind(PCA.X.Score,scores)
```

```
  return(list("Total Variance Explained"=summary(PCA.X),"Variance of Component"=lambda,"Number o
f Chosen Components"=k,
       "Component Matrix"=ComM,"Communalities"=Communal,"Coefficient of Components"=eigV,
       "scores"=X.score,"Scree Plot"=pl))
}
```

函数的输入为：待分析的数据集、是否用相关阵(T&F)、临界特征值；这里选择主成分个数，采用的是上述第二种原则，即根据特征根的大小选取主成分。
函数的输出为：总方差解释、各个主成分方差、选择的主成分个数、因子负荷矩阵、信息提取量矩阵、系数矩阵、得分以及碎石图。

2.2 调用封装函数Pca()分析例5.2

首先，在R中读入例5.2的数据集，并对数据集的“行名”等进行处理。

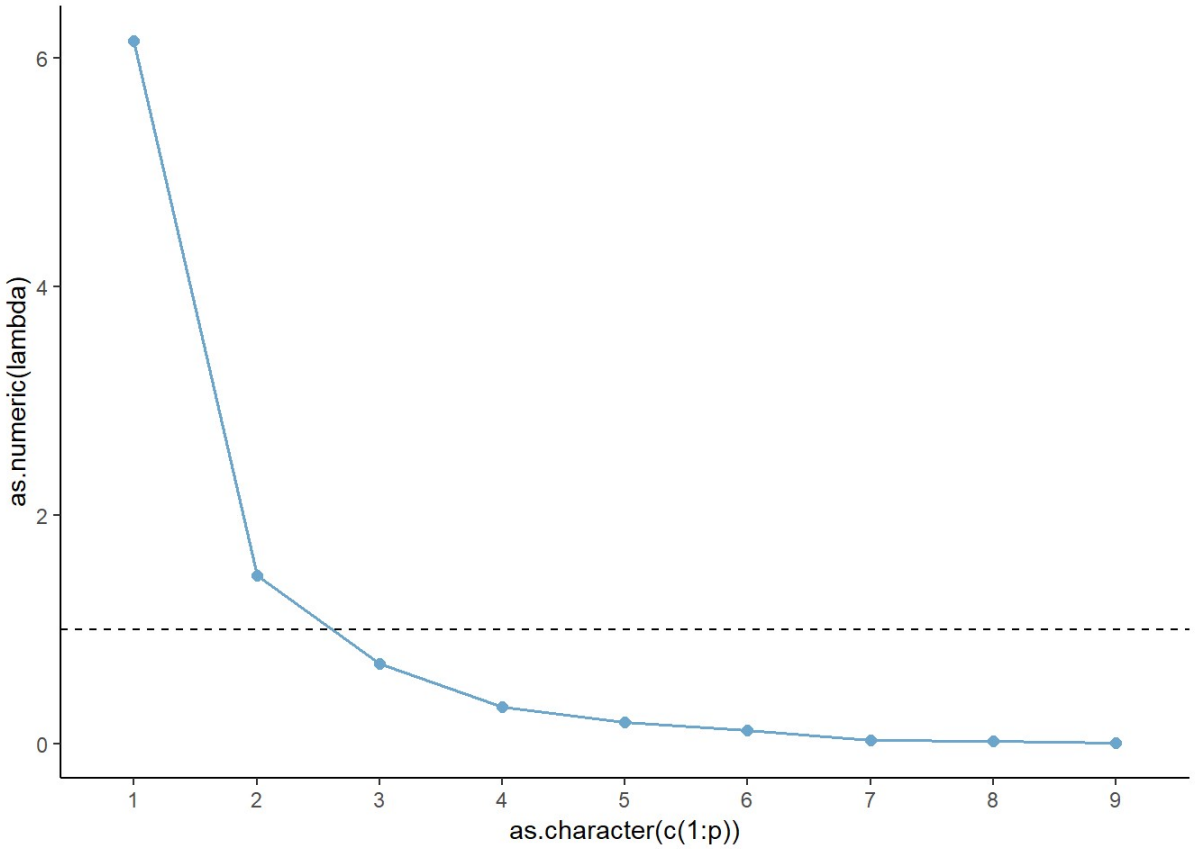
```
Y=read.csv("eg5-2.csv",header=T)
rownames(Y)=Y[,1]
Y=Y[,-1]
```

然后，调用Pca()进行分析。其中，从相关阵出发进行分析，保留特征根大于1的主成分。

```
Y.Pca<-Pca(Y,cor=T,1)
Y.Pca
```

```
## $`Total Variance Explained`
## Importance of components:
##                            Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     2.4798961 1.2136277 0.83511660 0.56375940
## Proportion of Variance 0.6833205 0.1636547 0.07749108 0.03531385
## Cumulative Proportion  0.6833205 0.8469752 0.92446628 0.95978013
##                            Comp.5     Comp.6      Comp.7      Comp.8
## Standard deviation     0.43594009 0.34065358 0.170638604 0.155962299
## Proportion of Variance 0.02111597 0.01289387 0.003235281 0.002702693
## Cumulative Proportion  0.98089611 0.99378998 0.997025262 0.999727955
##                              Comp.9
## Standard deviation     0.0494813355
## Proportion of Variance 0.0002720447
## Cumulative Proportion  1.0000000000
##
## $`Variance of Component`
##      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6
## 6.149884563 1.472892257 0.697419729 0.317824658 0.190043758 0.116044860
##      Comp.7      Comp.8      Comp.9
## 0.029117533 0.024324239 0.002448403
##
## $`Number of Chosen Components`
## [1] 2
##
## $`Component Matrix`
##                                      [,1]        [,2]
## 固定资产原值实现值...          -0.9313463  0.31495049
## X100元固定资产原值实现利税... -0.9755807 -0.16308056
## X100元资金实现利税...          -0.9305950 -0.32196827
## X100元工业总产值实现利税...    -0.2319553 -0.86328878
## X100元销售收入实现利税...      -0.4328881 -0.59628861
## 每吨标准煤实现工业产值.元.     -0.9228920  0.20019458
## 每千瓦时电力实现工业产值.元.   -0.8967712  0.27355496
## 全员劳动生产率.元.人.年.       -0.8712274  0.06355066
## X100元流动资金实现产值.元.     -0.8991959  0.15382823
##
## $Communalities
##                                     [,1]
## 固定资产原值实现值...          0.9665997
## X100元固定资产原值实现利税... 0.9783530
## X100元资金实现利税...          0.9696705
## X100元工业总产值实现利税...    0.7990708
## X100元销售收入实现利税...      0.5429522
## 每吨标准煤实现工业产值.元.     0.8918074
## 每千瓦时电力实现工业产值.元.   0.8790310
## 全员劳动生产率.元.人.年.       0.7630758
## X100元流动资金实现产值.元.     0.8322163
##
## $`Coefficient of Components`
##                                     Comp.1      Comp.2
## 固定资产原值实现值...          -0.37555860  0.25951162
## X100元固定资产原值实现利税... -0.39339581 -0.13437445
## X100元资金实现利税...          -0.37525562 -0.26529409
## X100元工业总产值实现利税...    -0.09353429 -0.71132915
## X100元销售收入实现利税...      -0.17455896 -0.49132744
## 每吨标准煤实现工业产值.元.     -0.37214945  0.16495551
```

```
## 每千瓦时电力实现工业产值.元.  -0.36161646  0.22540269
## 全员劳动生产率.元.人.年.       -0.35131608  0.05236421
## X100元流动资金实现产值.元.     -0.36259417  0.12675075
##
## $scores
##                   Comp.1      Comp.2      scores
## 北京 (1)    -2.81623780 -2.42574174 -2.74078516
## 天津 (2)    -3.73583159 -0.53648324 -3.11764539
## 河北 (3)     0.48683149  0.09774736  0.41165167
## 山西 (4)     2.02199676 -0.38456978  1.55699387
## 内蒙 (5)     2.97629761  0.73249650  2.54274467
## 辽宁 (6)     0.41801309 -1.26764724  0.09230553
## 吉林 (7)     1.61355311  0.80963271  1.45821757
## 黑龙江 (8)   1.04154171 -0.69226892  0.70653043
## 上海 (9)    -7.03772211 -1.37858319 -5.94424894
## 江苏 (10)   -3.94427676  2.80673714 -2.63982875
## 浙江 (11)   -4.36843328  1.82137825 -3.17242221
## 安徽 (12)   -0.07225380  0.59831730  0.05731564
## 福建 (13)   -0.51241108  0.52311393 -0.31232431
## 江西 (14)    1.18984730  1.25290514  1.20203150
## 山东 (15)   -0.98438511  0.36278512 -0.72408147
## 河南 (16)    1.02754308 -0.09775293  0.81011053
## 湖北 (17)   -0.35484983  0.30026057 -0.22826774
## 湖南 (18)    0.04353538 -0.38765552 -0.03978042
## 广东 (19)   -1.81893265  1.31473567 -1.21343731
## 广西 (20)   -0.13749809 -0.71890547 -0.24983910
## 四川 (21)    1.40993169  0.61027398  1.25541980
## 贵州 (22)    2.24418493 -1.31849457  1.55579497
## 云南 (23)    0.02115330 -2.40020473 -0.44670764
## 陕西 (24)    1.62418967  0.40611796  1.38883077
## 甘肃 (25)    1.64507169 -1.74861046  0.98933583
## 青海 (26)    3.40824628  0.59253114  2.86418666
## 宁夏 (27)    3.06812657  0.84212305  2.63801253
## 新疆 (28)    1.54276844  0.28576198  1.29988647
##
## $`Scree Plot`
```



根据输出的结果，选取前2个主成分，它们的特征根都大于1，累计方差贡献率为84.7%.
将9个变量分别记为$X_1, X_2, \ldots, X_9$,则第一主成分与$X_1, X_2, X_3, X_6, X_7, X_8, X_9$关系密切，可主要由它们解释；而第二主成分则与$X_4, X_5$关系密切。此外，除了变量“百元销售收入实现利税”，选取的主成分包含了各个原始变量的绝大部分信息。
由“主成分系数”(Coefficient of Components)得到主成分Y_1, Y_2的线性组合为：

$$Y_1 = -0.376X_1 - 0.393X_2 - 0.375X_3 - 0.094X_4 - 0.175X_5 - 0.372X_6 - 0.362X_7 - 0.351X_8 - 0.363X_9$$

$$Y_2 = 0.295X_1 - 0.134X_2 - 0.265X_3 - 0.711X_4 - 0.491X_5 + 0.165X_6 + 0.225X_7 + 0.052X_8 + 0.127X_9$$

2.3 可视化：散点图

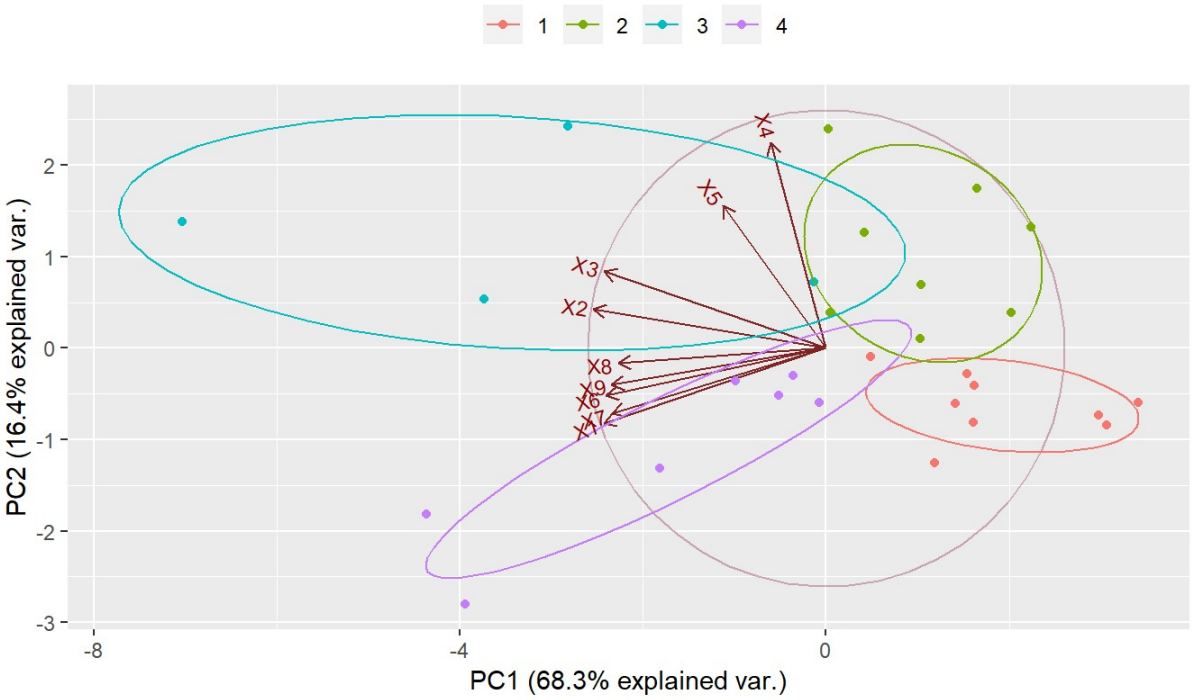
最后，将各个样品的两个主成分得分在二维空间中进行描绘。首先，按照四个象限将样品进行分类。

```
#先分类
type=c()
for(i in 1:nrow(Y)){
  if(Y.Pca$scores[i,1]>0 && Y.Pca$scores[i,2]>0){
    type=c(type,1)
  }
  if(Y.Pca$scores[i,1]>0 && Y.Pca$scores[i,2]<0){
    type=c(type,2)
  }
  if(Y.Pca$scores[i,1]<0 && Y.Pca$scores[i,2]<0){
    type=c(type,3)
  }
  if(Y.Pca$scores[i,1]<0 && Y.Pca$scores[i,2]>0){
    type=c(type,4)
  }
}
type
```

```
##  [1] 3 3 1 2 1 2 1 2 3 4 4 4 4 1 4 2 4 2 4 3 1 2 2 1 2 1 1 1
```

然后，以Y_1为横轴，Y_2为纵轴，绘制各个样品的散点图。

```
Y=cbind(Y,type)
colnames(Y)=c("X1","X2","X3","X4","X5","X6","X7","X8","X9","type")
y.pca <- prcomp(Y[,1:9], scale. = TRUE)
library(ggbiplot)
ggbiplot(y.pca, obs.scale = 1, var.scale = 1,
         groups = as.factor(type), ellipse = TRUE, circle = TRUE,size=2) +
  scale_color_discrete(name = '') +
  theme(legend.direction = 'horizontal', legend.position = 'top')
```



该散点图不仅对样品的4个类型做了区分，还能够表示9个主成分的方向。结合实际情况，可以看出，不同象限中各个省市的经济效益的大小关系为：第3象限>第4象限>第2象限>第1象限.

3 例5.3的分析

3.1 调用封装函数Pca()分析例5.3

首先，在R中读入例5.3的数据集，并对数据集的“行名”等进行处理。此外，由于指标“流动资金周转天数”和“万元产值能耗”为“成本型”变量”，“越小越好”，而其他变量为“收益型”变量，“越大越好”；因此，需要将这2个成本型变量取倒数。

```
Z=read.csv("eg5-3.csv",header=T)
attach(Z)
rownames(Z)=Z[,1]
Z=Z[,-1]
Z[,6]=1/Z[,6]
Z[,7]=1/Z[,7]
head(Z)
```

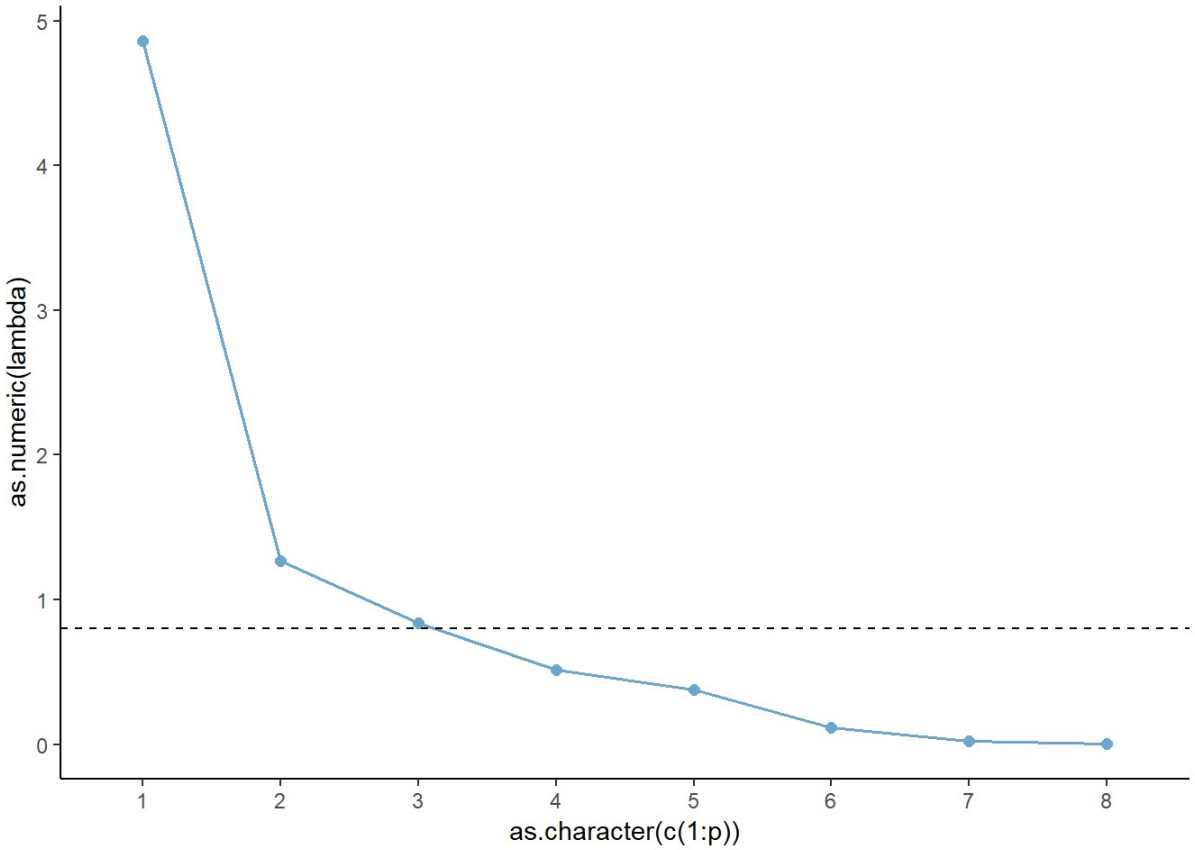
```
##          固定资产利税率 资金利税率 销售收入利税率 资金利润率 固定资产产值率
## 琉璃河           16.68       26.75           31.84       18.40           53.25
## 邯郸             19.70       27.56           32.94       19.20           59.82
## 大同             15.20       23.40           32.98       16.24           46.78
## 哈尔滨            7.29        8.97           21.30        4.76           34.39
## 华新             29.45       56.49           40.74       43.68           75.32
## 湘乡             32.93       42.78           47.98       33.87           66.46
##          流动资金周转天数 万元产值能耗 全员劳动生产率
## 琉璃河       0.01818182   0.03468609           1.75
## 邯郸         0.01818182   0.03037667           2.87
## 大同         0.01538462   0.02398657           1.53
## 哈尔滨       0.01612903   0.02545825           1.63
## 华新         0.01449275   0.03748126           2.14
## 湘乡         0.02000000   0.03042288           2.60
```

然后，调用前述封装函数Pca(),对该数据集进行主成分分析。

```
Z.Pca<-Pca(Z,cor=T,0.8)
Z.Pca
```

```
## $`Total Variance Explained`
## Importance of components:
##                            Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.2046848 1.1265763 0.9148794 0.71909488 0.61495939
## Proportion of Variance 0.6075794 0.1586468 0.1046255 0.06463718 0.04727188
## Cumulative Proportion  0.6075794 0.7662262 0.8708517 0.93548888 0.98276076
##                             Comp.6      Comp.7      Comp.8
## Standard deviation     0.33980626 0.145353408 0.036303945
## Proportion of Variance 0.01443354 0.002640952 0.000164747
## Cumulative Proportion  0.99719430 0.999835253 1.000000000
##
## $`Variance of Component`
##      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6
## 4.860635113 1.269174208 0.837004302 0.517097444 0.378175048 0.115468294
##      Comp.7      Comp.8
## 0.021127613 0.001317976
##
## $`Number of Chosen Components`
## [1] 3
##
## $`Component Matrix`
##                          [,1]          [,2]        [,3]
## 固定资产利税率    -0.9569816  0.0185733910 -0.23930221
## 资金利税率         -0.8989949  0.3955672959  0.03720129
## 销售收入利税率    -0.8618345 -0.0813848424 -0.33812955
## 资金利润率         -0.9275701  0.3504529503 -0.03762397
## 固定资产产值率    -0.7867504 -0.0002493985  0.18197264
## 流动资金周转天数 -0.4224862 -0.7733416983  0.34535298
## 万元产值能耗       -0.6404981  0.0778927259  0.64218500
## 全员劳动生产率    -0.5707680 -0.6154593954 -0.31276728
##
## $Communalities
##                         [,1]
## 固定资产利税率    0.9734242
## 资金利税率         0.9660492
## 销售收入利税率    0.8637137
## 资金利润率         0.9846192
## 固定资产产值率    0.6520902
## 流动资金周转天数 0.8958206
## 万元产值能耗       0.8287066
## 全员劳动生产率    0.8023898
##
## $`Coefficient of Components`
##                         Comp.1        Comp.2      Comp.3
## 固定资产利税率    -0.4340673  0.0164865803 -0.26156695
## 资金利税率         -0.4077657  0.3511233890  0.04066250
## 销售收入利税率    -0.3909105 -0.0722408601 -0.36958920
## 资金利润率         -0.4207269  0.3110778592 -0.04112451
## 固定资产产值率    -0.3568539 -0.0002213774  0.19890342
## 流动资金周转天数 -0.1916311 -0.6864530024  0.37748471
## 万元产值能耗       -0.2905168  0.0691410998  0.70193405
## 全员劳动生产率    -0.2588887 -0.5463095430 -0.34186722
##
## $scores
##              Comp.1      Comp.2      Comp.3      scores
## 琉璃河 -0.04944916 -0.80688332  1.68006726  0.02035273
```

```
## 邯郸   -0.84046336 -2.22495495 -0.08804192 -1.00228498
## 大同    1.56943785  0.16905735 -0.80699111  1.02881652
## 哈尔滨  3.73936076 -0.86260748  0.36170405  2.49520338
## 华新   -3.95706517  1.72335609  0.36765037 -2.40266021
## 湘乡   -3.88945540 -1.84328506 -0.78135299 -3.14328334
## 柳州   -1.61104477 -0.32266644 -1.02899882 -1.30640744
## 峨嵋    2.80457471  0.71743217 -0.87616874  1.98214054
## 耀县    0.47423958  0.60491878 -1.08213162  0.31106100
## 永登   -0.66308953  0.31178986 -0.48247014 -0.46379168
## 工源    0.26952753 -0.08333228  1.54448517  0.35842103
## 抚顺    0.75209630  0.05208774  1.08605054  0.66469455
## 大连    1.06653413  1.15139415 -0.31879162  0.91555816
## 江南   -1.72341594  1.60407382  0.64421588 -0.83278178
## 江油    2.05821249 -0.19038044 -0.21922631  1.37496151
##
## $`Scree Plot`
```



根据输出的结果，选取前3个主成分，它们的特征根都大于0.8，这是在输入Pca()函数的参数时进行特殊设定造成的结果。这样做的目的，是为了提高主成分的累计方差贡献率；事实上，若设定的特征根临界值过高，就会造成选入的主成分过少，从而累计方差贡献率较低。此处的累计方差贡献率为84.7%.

将8个变量分别记为$X_1, X_2, \ldots, X_8$,则第一主成分与X_1, X_2, X_3, X_4, X_5关系密切，可主要由它们解释；而第二主成分则与X_6, X_8关系密切；第三主成分与X_7关系密切。此外，除了变量“固定资产产值率”，主成分包含了各个原始变量的绝大部分信息。

由“主成分系数”(Coefficient of Components)得到主成分Y_1, Y_2, Y_3的线性组合为：

$$Y_1 = -0.434X_1 - 0.408X_2 - 0.391X_3 - 0.421X_4 - 0.357X_5 - 0.192X_6 - 0.291X_7 - 0.259X_8$$

$$Y_2 = 0.016X_1 + 0.351X_2 - 0.072X_3 + 0.311X_4 - 0.000X_5 - 0.686X_6 + 0.069X_7 - 0.546X_8$$

$$Y_3 = -0.262X_1 + 0.041X_2 - 0.370X_3 - 0.041X_4 + 0.199X_5 + 0.377X_6 + 0.702X_7 - 0.342X_8$$

3.2 得分排名

最后，用样本主成分得分进行排序。事实上，依据每个主成分Y_k的方差贡献率α_k作为权数构造的综合评价函数存在较大的争议。因此，只用第一主成分作评价指数，并且最终得到了排名结果。

```
y1=as.numeric((-1)*Z.Pca$scores[,1])
rk=nrow(Z)-rank(y1)+1
demo.rk=data.frame(y1,rk)
colnames(demo.rk)=c("y1","名次")
rownames(demo.rk)=rownames(Z)
demo.rk
```

```
##                  y1 名次
## 琉璃河  0.04944916    7
## 邯郸    0.84046336    5
## 大同   -1.56943785   12
## 哈尔滨 -3.73936076   15
## 华新    3.95706517    1
## 湘乡    3.88945540    2
## 柳州    1.61104477    4
## 峨嵋   -2.80457471   14
## 耀县   -0.47423958    9
## 永登    0.66308953    6
## 工源   -0.26952753    8
## 抚顺   -0.75209630   10
## 大连   -1.06653413   11
## 江南    1.72341594    3
## 江油   -2.05821249   13
```

可见，华新水泥厂的综合经济效益最好，湘乡水泥厂其次，而哈尔滨水泥厂的综合效益最差。