主成份分析

王宁宁

目录


```
rm(list = ls(all = TRUE))
options(digits = 4  )
```

0.1 简介

- 主成分分析（Principal Component Analysis，PCA），将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。
- 主成分分析首先是由 K. 皮尔森对非随机变量引入的，尔后 H. 霍特林将此方法推广到随机向量的情形。信息的大小通常用离差平方和或方差来衡量。
- PCA 技术的一大好处是对数据进行降维的处理。我们可以对新求出的“主元”向量的重要性进行排序，根据需要取前面最重要的部分，将后面的维数省去，可以达到降维从而简化模型或是对数据进行压缩的效果。同时最大程度的保持了原有数据的信息。
- PCA 的思想是将 n 维特征映射到 k 维上（k<n），这 k 维是全新的正交特征。这 k 维特征称为主元，是重新构造出来的 k 维特征，而不是简单地从 n 维特征中去除其余 n-k 维特征。

0.2 PCA 的原理

PCA 的数学原理推荐看这里。

考虑一个 n 个观测 p 个变量的数据集：

$$X = [X_{ij}] = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{bmatrix} = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$$

X 的协方差阵 Σ，其特征值为：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

$$Y_i = \mathbf{a}_i'\mathbf{X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

$$Var(Y_i) = Var(\mathbf{a}_i'\mathbf{X}) = \mathbf{a}_i'\Sigma\mathbf{a}_i$$

- 第一主成份 Y_1 要满足：线性组合 $\mathbf{a}_1'\mathbf{X}$ 要使得

  $$Var(Y_1) = Var(\mathbf{a}_1'\mathbf{X}) = \mathbf{a}_1'\Sigma\mathbf{a}_1$$

  达到最大，且 $\mathbf{a}_1'\mathbf{a}_1 = 1$。

- 第二主成份 Y_2 要满足：线性组合 $\mathbf{a}_2'\mathbf{X}$ 要使得

  $$Var(Y_2) = Var(\mathbf{a}_2'\mathbf{X}) = \mathbf{a}_2'\Sigma\mathbf{a}_2$$

  达到次最大，且 $\mathbf{a}_2'\mathbf{a}_2 = 1$，且 $Cov(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_1'\mathbf{X}) = 0$。

- 第 k 主成份 Y_k 要满足：线性组合 $\mathbf{a}_k'\mathbf{X}$ 要使得

  $$Var(Y_k) = Var(\mathbf{a}_k'\mathbf{X}) = \mathbf{a}_k'\Sigma\mathbf{a}_k$$

  达到次最大，且 $\mathbf{a}_k'\mathbf{a}_k = 1$，且 $Cov(\mathbf{a}_k'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = 0 \quad \forall j < k$ 。

结论 1：考虑 Σ 的特征值和特种向量组合 $(\lambda_i, \mathbf{e}_i)$，$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 按特征值从大到小排列，则第 i 主成份是：

$$Y_i = e_{i1}X_1 + \cdots + e_{ip}X_p$$

满足：

$$Var(Y_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i = \lambda_i$$
$$Cov(Y_i, Y_j) = \mathbf{e}_i'\Sigma\mathbf{e}_j = 0 \quad \forall j \neq i$$

结论 2：上述的主成份具有下列的性质：

$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} \lambda_i$$

根据结论 2：第 i 个主成份的变异比例为：

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

0.3 实例

```
set.seed(1234)
#help(USJudgeRatings)
```

查看一下 R 自带的数据集：USJudgeRatings，美国律师对最高法院法官的评分。

变　量	描　述	变　量	描　述
CONT	律师与法官的接触次数	PREP	审理前的准备工作
INTG	法官正直程度	FAMI	对法律的熟稔程度
DMNR	风度	ORAL	口头裁决的可靠度
DILG	勤勉度	WRIT	书面裁决的可靠度
CFMG	案例流程管理水平	PHYS	体能
DECI	决策效率	RTEN	是否值得保留

图 1: 变量描述

```
library(graphics)
pairs(USJudgeRatings, main = "USJudgeRatings data")
```

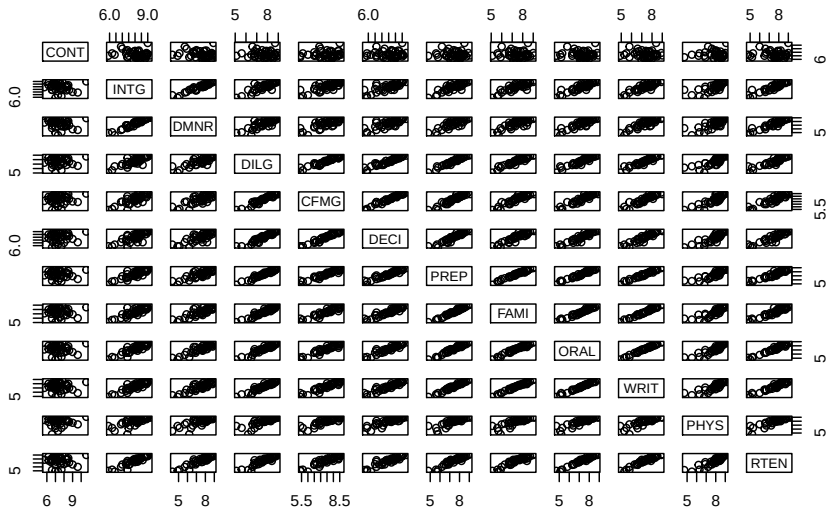

图 2:

```
summary(USJudgeRatings)
```

```
##       CONT            INTG           DMNR           DILG
##  Min.   : 5.70   Min.   :5.90   Min.   :4.30   Min.   :5.10
##  1st Qu.: 6.85   1st Qu.:7.55   1st Qu.:6.90   1st Qu.:7.15
##  Median : 7.30   Median :8.10   Median :7.70   Median :7.80
##  Mean   : 7.44   Mean   :8.02   Mean   :7.52   Mean   :7.69
##  3rd Qu.: 7.90   3rd Qu.:8.55   3rd Qu.:8.35   3rd Qu.:8.45
##  Max.   :10.60   Max.   :9.20   Max.   :9.00   Max.   :9.00
##       CFMG            DECI           PREP           FAMI
##  Min.   :5.40   Min.   :5.70   Min.   :4.80   Min.   :5.10
##  1st Qu.:7.00   1st Qu.:7.10   1st Qu.:6.90   1st Qu.:6.95
##  Median :7.60   Median :7.70   Median :7.70   Median :7.60
```

```
##  Mean   :7.48   Mean   :7.57   Mean   :7.47   Mean   :7.49
##  3rd Qu.:8.05   3rd Qu.:8.15   3rd Qu.:8.20   3rd Qu.:8.25
##  Max.   :8.70   Max.   :8.80   Max.   :9.10   Max.   :9.10
##       ORAL           WRIT           PHYS           RTEN
##  Min.   :4.70   Min.   :4.90   Min.   :4.70   Min.   :4.80
##  1st Qu.:6.85   1st Qu.:6.90   1st Qu.:7.70   1st Qu.:7.15
##  Median :7.50   Median :7.60   Median :8.10   Median :7.80
##  Mean   :7.29   Mean   :7.38   Mean   :7.93   Mean   :7.60
##  3rd Qu.:8.00   3rd Qu.:8.05   3rd Qu.:8.50   3rd Qu.:8.25
##  Max.   :8.90   Max.   :9.00   Max.   :9.10   Max.   :9.20
```

```
head(USJudgeRatings)
```

```
##            CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## AARONSON,L.H.  5.7  7.9  7.7  7.3  7.1  7.4  7.1  7.1  7.1  7.0  8.3  7.8
## ALEXANDER,J.M. 6.8  8.9  8.8  8.5  7.8  8.1  8.0  8.0  7.8  7.9  8.5  8.7
## ARMENTANO,A.J. 7.2  8.1  7.8  7.8  7.5  7.6  7.5  7.5  7.3  7.4  7.9  7.8
## BERDON,R.I.    6.8  8.8  8.5  8.8  8.3  8.5  8.7  8.7  8.4  8.5  8.8  8.7
## BRACKEN,J.J.   7.3  6.4  4.3  6.5  6.0  6.2  5.7  5.7  5.1  5.3  5.5  4.8
## BURNS,E.B.     6.2  8.8  8.7  8.5  7.9  8.0  8.1  8.0  8.0  8.0  8.6  8.6
```

```
boxplot(USJudgeRatings)
```

目的: 寻找变量的线性组合能够反映较好的反映对法官的评价

0.3.1 S 型分析

根据协方差矩阵的分析，称为 S 型分析。

```
pc1<-princomp(USJudgeRatings[,-1],cor = F)
screeplot(pc1,main = "scree plot1")
```

```
summary(pc1)
```

```
## Importance of components:
```

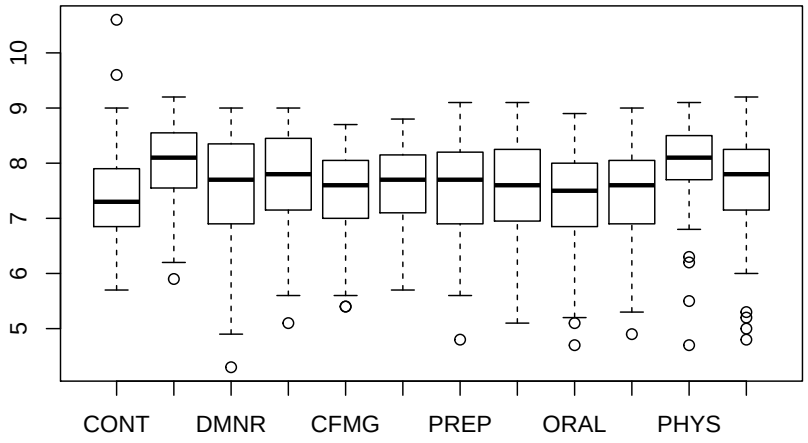
5
6
7
8
9
10
CONT
DMNR
CFMG
PREP
ORAL
PHYS

图 3:

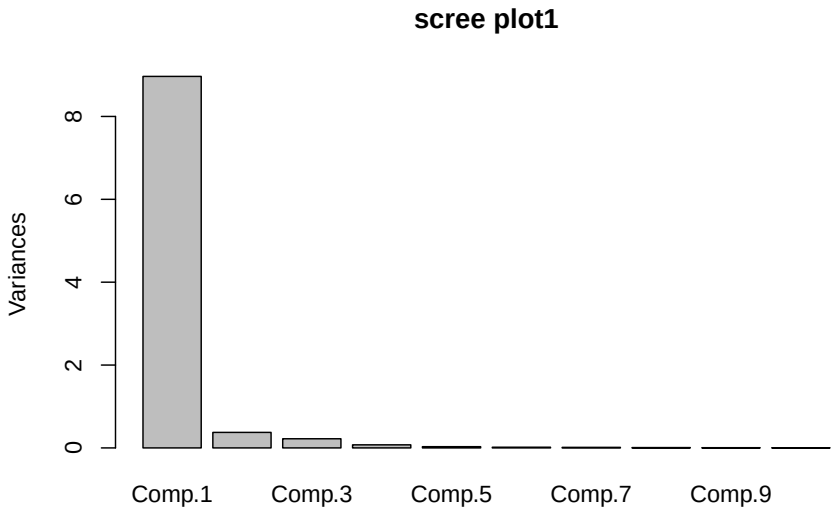
scree plot1
Variances
0
2
4
6
8
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9

图 4:

```
##                          Comp.1  Comp.2  Comp.3  Comp.4   Comp.5   Comp.6
## Standard deviation     2.994 0.61201 0.47007 0.27243 0.172116 0.128908
## Proportion of Variance  0.923 0.03856 0.02275 0.00764 0.003049 0.001711
## Cumulative Proportion   0.923 0.96158 0.98433 0.99197 0.995018 0.996728
##                           Comp.7    Comp.8    Comp.9  Comp.10   Comp.11
## Standard deviation     0.115737 0.0897567 0.0726041 0.056101 0.0437152
## Proportion of Variance 0.001379 0.0008293 0.0005426 0.000324 0.0001967
## Cumulative Proportion  0.998107 0.9989367 0.9994793 0.999803 1.0000000
```

```
pc1$loadings[,1]
```

```
##   INTG   DMNR   DILG   CFMG   DECI   PREP   FAMI   ORAL   WRIT
## -0.2347 -0.3476 -0.2868 -0.2721 -0.2534 -0.3091 -0.3051 -0.3320 -0.3140
##    PHYS    RTEN
## -0.2776 -0.3593
```

```
pc1
```

```
## Call:
## princomp(x = USJudgeRatings[, -1], cor = F)
##
## Standard deviations:
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
## 2.99443 0.61201 0.47007 0.27243 0.17212 0.12891 0.11574 0.08976 0.07260
## Comp.10 Comp.11
## 0.05610 0.04372
##
##  11  variables and  43 observations.
```

自己编程：

```
cov1<-cov(USJudgeRatings[,-1])
eg1<-eigen(cov1,symmetric=T)
eg1$vectors[,1]
```

```
## [1] -0.2347 -0.3476 -0.2868 -0.2721 -0.2534 -0.3091 -0.3051 -0.3320
##  [9] -0.3140 -0.2776 -0.3593
```

```
eg1$values[1]/sum(eg1$values)
```

```
## [1] 0.923
```

0.3.2 R 型分析

根据相关系数矩阵的分析，称为 R 型分析。

```
pc2<-princomp(USJudgeRatings[,-1],cor = T)
screeplot(pc2,main = "scree plot2")
```

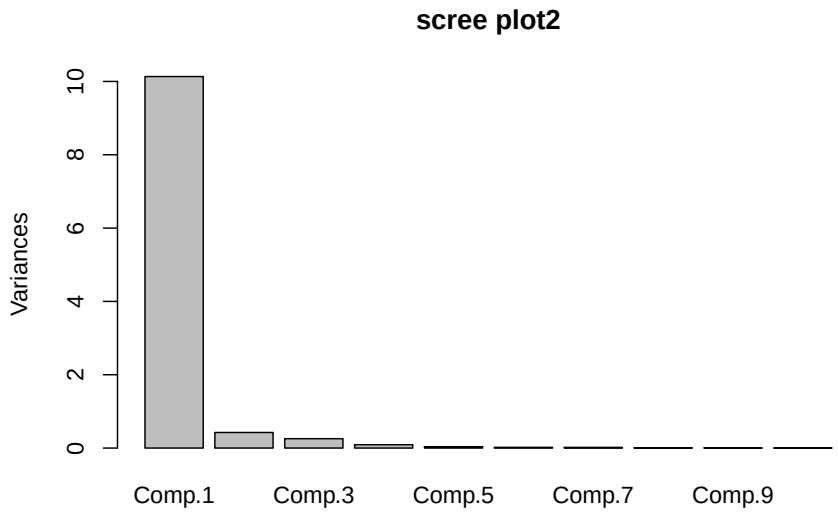

图 5:

```
summary(pc2)
```

```
## Importance of components:
```

```
##                          Comp.1 Comp.2  Comp.3 Comp.4   Comp.5   Comp.6
## Standard deviation     3.1833 0.6516 0.50525 0.3022 0.193133 0.140575
## Proportion of Variance 0.9212 0.0386 0.02321 0.0083 0.003391 0.001796
## Cumulative Proportion  0.9212 0.9598 0.98303 0.9913 0.994721 0.996517
##                           Comp.7    Comp.8    Comp.9   Comp.10   Comp.11
## Standard deviation     0.135921 0.0910572 0.0780508 0.0579730 0.0457250
## Proportion of Variance 0.001679 0.0007538 0.0005538 0.0003055 0.0001901
## Cumulative Proportion  0.998197 0.9989506 0.9995044 0.9998099 1.0000000
```

```
pc2$loadings[,1]
```

```
##   INTG   DMNR   DILG   CFMG   DECI   PREP   FAMI   ORAL   WRIT
## -0.2885 -0.2868 -0.3044 -0.3026 -0.3019 -0.3094 -0.3067 -0.3127 -0.3111
##    PHYS    RTEN
## -0.2807 -0.3098
```

```
pc2
```

```
## Call:
## princomp(x = USJudgeRatings[, -1], cor = T)
##
## Standard deviations:
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
## 3.18330 0.65164 0.50525 0.30216 0.19313 0.14058 0.13592 0.09106 0.07805
## Comp.10 Comp.11
## 0.05797 0.04573
##
##  11  variables and  43 observations.
```

自己编程：

```
cov2<-cor(USJudgeRatings[,-1])
eg2<-eigen(cov2,symmetric=T)
eg2
```

```
## $values
##  [1] 10.133417  0.424629  0.255280  0.091303  0.037300  0.019761  0.018474
##   [8]  0.008291  0.006092  0.003361  0.002091
##
## $vectors
##         [,1]     [,2]     [,3]    [,4]    [,5]    [,6]     [,7]
##  [1,] -0.2885  0.574468  0.117763  0.08381  0.37494 -0.50953 -0.229705
##  [2,] -0.2868  0.576357 -0.176987  0.23977 -0.39861  0.51408  0.167067
##  [3,] -0.3044 -0.138561  0.334740  0.26556  0.59149  0.29806  0.367529
##  [4,] -0.3026 -0.310012  0.019546  0.47774 -0.08203  0.10089 -0.722336
##  [5,] -0.3019 -0.336467  0.054444  0.38037 -0.39889 -0.44826  0.452352
##  [6,] -0.3094 -0.125254  0.229234 -0.20133  0.08470  0.33584 -0.006824
##  [7,] -0.3067 -0.122859  0.227526 -0.52405 -0.09944 -0.03819 -0.002373
##  [8,] -0.3127  0.005208 -0.005507 -0.22937 -0.14642  0.01946 -0.163556
##  [9,] -0.3111 -0.000300  0.148245 -0.31656 -0.23702 -0.07289 -0.060730
## [10,] -0.2807 -0.234798 -0.820161 -0.15475  0.29792  0.03755  0.042123
## [11,] -0.3098  0.152781 -0.201054  0.01114  0.03730 -0.23409  0.159968
##             [,8]       [,9]     [,10]       [,11]
##   [1,] -0.284904  0.145485 -0.10273 -0.0006869
##   [2,] -0.169286 -0.005467  0.10539 -0.0764810
##   [3,]  0.004789 -0.354686  0.02389 -0.0735830
##   [4,]  0.035844 -0.026425  0.20705 -0.0131127
##   [5,] -0.199577  0.150276 -0.13826 -0.0422633
##   [6,]  0.068955  0.717150 -0.25188  0.3049299
##   [7,] -0.222092  0.060538  0.54401 -0.4518560
##   [8,]  0.274475 -0.252450 -0.66685 -0.4660731
##   [9,] -0.099199 -0.492809 -0.01153  0.6804728
##  [10,] -0.272364 -0.001097 -0.03062  0.0487849
##  [11,]  0.797242  0.071825  0.33262  0.0835120
```

```
eg2$values[1]/sum(eg2$values)
```

```
## [1] 0.9212
```

0.4 思考和讨论

- 使用协方差阵与相关系数矩阵得到的主成份是否一致?

```
pc1$loadings[,1]
```

```
##   INTG   DMNR   DILG   CFMG   DECI   PREP   FAMI   ORAL   WRIT
## -0.2347 -0.3476 -0.2868 -0.2721 -0.2534 -0.3091 -0.3051 -0.3320 -0.3140
##    PHYS    RTEN
## -0.2776 -0.3593
```

```
pc2$loadings[,1]
```

```
##   INTG   DMNR   DILG   CFMG   DECI   PREP   FAMI   ORAL   WRIT
## -0.2885 -0.2868 -0.3044 -0.3026 -0.3019 -0.3094 -0.3067 -0.3127 -0.3111
##    PHYS    RTEN
## -0.2807 -0.3098
```

```
eg1$values
```

```
##  [1] 9.180118 0.383480 0.226229 0.075985 0.030329 0.017013 0.013714
##  [8] 0.008248 0.005397 0.003222 0.001957
```

```
eg2$values
```

```
##  [1] 10.133417  0.424629  0.255280  0.091303  0.037300  0.019761  0.018474
##  [8]  0.008291  0.006092  0.003361  0.002091
```

0.5 练习

尝试对 iris data 作主成份分析

```
#help(iris)
```