1 总体主成分分析

总体主成分分析是总体的协方差矩阵或相关系数矩阵已知情况下的 PCA，也就是只有协方差或相关系数矩阵而没有原始数据的情形。总体 PCA 的 R 命令如下

```
princomp(covmat=Sigma) #covmat: covariance/correlation matrix
```

例 1 (课本例 8.4). 雄性乌龟外壳尺寸长、宽、高，分别记为 L,W,H, 它们的对数 $x_1 = \log(L), x_2 = \log(W), x_3 = \log(H)$ 的协方差矩阵为

$$S = 10^{-3}\begin{pmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.773 \end{pmatrix}.$$

```
> S = matrix(10^(-3)*c(11.072,8.019,8.160,8.019,6.417,6.005,8.160,6.005,6.773),3,3)
> mypca = princomp(covmat = S) # 指定方差-协方差矩阵为S
> summary(mypca,loading=T) #loading=T: 输出载荷(loadings),即S的特征向量
Importance of components:
Comp.1 Comp.2 Comp.3
Standard deviation 0.1526548 0.02446816 0.01896889
Proportion of Variance 0.9604934 0.02467606 0.01483055
Cumulative Proportion 0.9604934 0.98516945 1.00000000

Loadings:
Comp.1 Comp.2 Comp.3
[1,] 0.683 -0.158 0.713
[2,] 0.510 -0.595 -0.621
[3,] 0.523 0.788 -0.326
```

- 输出结果的 Importance of components 部分:

  第一行为三个主成分的标准差: $\sqrt{\lambda_1} = 0.1526548$, $\sqrt{\lambda_2} = 0.02446816$, $\sqrt{\lambda_3} = 0.01896889$. 第二、三行分别是方差在总和中的比例以及累积比例，比如 $\lambda_1/(\lambda_1+\lambda_2+\lambda_3) = 0.9604934$ 等等。

- 输出结果的 Loadings 部分:

  第一列是 $\mathbf{v}_1 = (0.683, 0.510, 0.523)^\top$, 所以对任何 $\mathbf{x} = (x_1, x_2, x_3)^\top$, 第一主成分为

  $$y_1 = \mathbf{v}_1^\top \mathbf{x} = 0.683x_1 + 0.51x_2 + 0.523x_3,$$

  三个分量/loading 符号相同，而且取值接近。近似地可以认为

  $$y_1 \approx (x_1+x_2+x_3)/\sqrt{3} = \frac{1}{\sqrt{3}}\log(L\times W\times H) = \frac{1}{\sqrt{3}}\log(\text{Volume})$$

  它能解释总方差的 96.05%, 所以 y_1 或体积几乎包含了原随机向量 $\mathbf{x} = (x_1, x_2, x_3)^\top$ 的所有信息, 即乌龟之间的体型差别主要体现在体积上。

练习 1. 重复第 11 讲例 4(儿童能力和智力测试)，数据为 R 自带数据集 ability.cov。

练习 2. (Johnson and Wichern 课本例 8.5). 为了考察 5 个公司：Allied Chemical, du Pont, Union Carbide, Exxon, Texaco 的股票价格之间的关系，根据 1975 年-1976 年的各周五收盘价数据计算得到股票周收益率（定义为: [本周五收盘价- 上周五收盘价]/上周五收盘价)，记 $x_1,...,x_5$ 分别是上述 5 种股票的周收益率。共得到 100 组这样的数据，由此计算得到 $\mathbf{x}=(x_1,...,x_5)^\top$ 的相关系数矩阵为：

$$R=\begin{pmatrix} 1 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1 & 0.599 & 0.389 & 0322 \\ 0.509 & 0.599 & 1 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1 \end{pmatrix}.$$

```
R=c(1,0.577,0.509,0.387,0.462,0.577,1,0.599,0.389,0.322,0.509,0.599,
       1,0.436,0.426,0.387,0.389,0.436,1,0.523,0.462,0.322,0.426,0.523,1)
R=matrix(R,5,5)
company=c("Allied Chem", "du Pont", "Union Carbide", "Exxon", "Texaco")
rownames(R)=colnames(R)=company
```

试基于 5 个股票的相关系数矩阵（R 函数: cor)，进行 PCA 分析，解释结果.

2 样本主成分分析

总体 PCA 因为没有原始数据，不能计算每个样本的主成分值（称为 pc 得分,score)，也不能画双标图。当有原始数据样本时，则能考察每个样本的 pc 得分以及样本与变量之间的关系（biplot)。

对于有原始数据的情形，R 中的样本 PCA 函数主要有两个：princomp, prcomp （后者能处理 $n<p$ 的情形)。样本 PCA 的 R 命令如下

```
mypca = princomp(x,cor=F, ...) #x: data matrix,  n>p;
mypca=prcomp(x,scale=F,...) #x: data matrix, n>p 或 n<p;
biplot(mypca) #双标图
```

对于样本 PCA，两个函数结果相同（但载荷可能相差一个符号)。princomp 要求 $n>p$。prcomp 不能处理总体 PCA，但能处理变量个数 p 大于样本个数 n 的情形（等价于奇异值分解 svd，关于用 svd 作主成分分析以后会提及，这里暂时忽略)。建议一般情形即 $n>p$ 时，使用 princomp 函数作总体 PCA 或样本 PCA，只有 $n<p$ 时才使用 prcomp 或者 svd。

princomp 输出的主成分和载荷分别称为 scores,loadings; prcomp 输出的主成分和载荷分别称为 x,rotations. 两个函数的缺省 (default) 情况都使用协方差矩阵。cor=T, scale=T 都是指定使用相关系数矩阵（标准化数据的协方差)。

练习 3. 重复第 12 讲例 2（欧洲气温数据)，数据集为

```
read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/temperature.csv",
 head=T, row.name=1, sep=",") #row.name=1指定csv文档第一列是rowname
#or
read.csv("http://staff.ustc.edu.cn/~ynyang/vector/data/temperature.csv", head=T, row.name=1)
```

练习 4. R 程序包 FactoMineR 的数据集 decathlon 包含 2004 年 8 月 23-24 雅典奥运会和同年 9 月 25-26 日举行的 Decastar 男子十项全能比赛的成绩（部分运动员同时参加了这两场赛事，后者是世界上最大的国际田联十项全能专业赛事，每年在法国塔朗斯举行）。

```
library(FactoMineR)
data(decathlon)
```

数据集的最后三列分别是 Rank（名次），Points（总分），Competition（赛事：OlympicG，Decastar,）；数据集的前十列为十个比赛项目的成绩，十个项目如下 (也是比赛次序)：

第一天：100m, long jump, shot put, high jump, 400m；

第二天：110m hurdles, discus, pole vault, javelin, 1500m.

试只使用奥运会的十项成绩进行主成分分析，考虑的问题包括但不限于：

1. 因为各个比赛项目成绩单位不同，需要标准化（即使用相关系数矩阵进行分析）；考察载荷，解释 PC1，PC2 的含义；求 PC1，PC2（princomp 输出的 score，或 prcomp 输出的 x）与运动员比赛成绩（Points）的相关系数，哪个主成分可认为代表了运动员的成绩？

2. 分析双标图，你认为哪些项目是取得好的总成绩的关键（考察前几名和最后几名的特点）？

3. 分析双标图, 说明跳远和短跑成绩是否有关，是否认为属于同一类比赛？跳高和投掷类项目（铁饼、铅球、标枪）成绩是否有关，是否可以认为属于同一类项目？1500m 作为径赛项目与哪个（些）项目关系密切？说明理由。