重复例 1-3，**提交练习** 1-6

1 主成分分析

1.1 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**.

1.2 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。

2 因子分析

因子分析的 R 函数为 factanal (极大似然法), 以及 princomp/prcomp (主成分方法或主因子方法)。我们仅考虑前者。主要函数包括因子分析函数 factanal，因子旋转函数 varimax，promax。调用方式如下

```
##总体因子分析（基于协方差矩阵）
> myFA = factanal(covmat=my.cov, factors=2,n.obs=..)
#covmat: covariance/correlation matrix
#factors: number of factors
#n.obs: 若给出样本量，则基于协方差/相关系数矩阵可以计算拟合优度检验。
#若不给n.obs，则不能计算拟合优度检验（如例1）

##样本因子分析（基于原始数据）
> myFA = factanal(x=my.data, factors=2, scores=...)
#use data.frame #scores: 因子的预测
> varimax(L)
#L: loadings obtained from factor analysis (princomp or factanal)
> promax(L) #non-orthogonal rotation
```

下面两个例子分别演示总体因子分析（即只有协方差或相关系数矩阵的情形）和样本因子分析（即有原始数据的情形）.

例 **2.** (总体因子分析，Johnson and Wichern 课本例 9.3) 一项问卷调查中，消费者对一种新的食品产品的 5 个属性（"Taste", "Good buy", "Flavor", "Suitabl for snack", "Lots of energy"）进行了评价打分（1-7，1 分最低，7 分最高），5 个属性的样本相关系数如下：

Attribute (*Variable*)		1	2	3	4	5
Taste	1	1.00	.02	.96	.42	.01
Good buy for money	2	.02	1.00	.13	.71	.85
Flavor	3	.96	.13	1.00	.50	.11
Suitable for snack	4	.42	.71	.50	1.00	.79
Provides lots of energy	5	.01	.85	.11	.79	1.00

下面对该相关系数矩阵进行因子分析（因子个数取为 2）。两因子模型将 5 个变量的打分 $\mathbf{x} = (x_1, x_2, ..., x_5)^\top$ 表示为

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{LF} + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \mathbf{l}_1 F_1 + \mathbf{l}_2 F_2 + \boldsymbol{\epsilon},$$

其中 $\boldsymbol{\mu} = E(\mathbf{x})$, $\mathbf{F} = (F_1, F_2)^\top \sim N_2(\mathbf{0}, I_2)$ 代表两个独立的因子，$\mathbf{l}_1, \mathbf{l}_2$ 分别是 F_1, F_2 的载荷，$L = (\mathbf{l}_1, \mathbf{l}_2)$ 为 5×2 载荷矩阵, $\boldsymbol{\epsilon} \sim N_5(\mathbf{0}, \Psi)$ 为特殊因子 (误差), $\Psi = diag(\psi_1, ..., \psi_5)$ 对角元为 5 个特殊方差。需要估计的参数为 $\boldsymbol{\mu}$, L, Ψ, 但因这里我们仅有相关系数矩阵，所以 $\boldsymbol{\mu}$ 不能估计，这并不影响 L,Ψ 的估计以及因子模型的解释。数据分析结果如下：

```
> R=matrix(c(1,0.02,0.96,0.42,0.01,0.02,1,0.13,0.71,0.85,0.96,0.13,
      1,0.5,0.11,0.42,0.71,0.5,1,0.79,0.01,0.85,0.11,0.79,1),5,5)
> name=c("Taste","Good.buy","Flavor","Suitable.for.snack","Lots.of.energy")
> dimnames(R)=list(name,name)

> (myFA = factanal(covmat=R, factors=2)) # 两因子的因子分析模型
      # 本例我们不知道样本量，故不指定 n.obs，因而分析结果不给出拟合优度检验。
      Call:
      factanal(factors = 2, covmat = R)

      Uniquenesses:
      Taste    Good.buy    Flavor Suitable.for.snack   Lots.of.energy
      0.028       0.237     0.040             0.168            0.052

      Loadings:
      Factor1  Factor2
      Taste                    0.985
      Good.buy          0.873
      Flavor            0.131  0.971
      Suitable.for.snack 0.817  0.405
      Lots.of.energy    0.973

      Factor1 Factor2
      SS loadings      2.396   2.078
      Proportion Var   0.479   0.416
      Cumulative Var   0.479   0.895

      The degrees of freedom for the model is 1 and the fit was 0.0233
```

输出结果主要是 Ψ, L 的估计，以及由载荷计算得到的方差贡献率，细节如下：

- Uniqueness（特殊方差）Ψ

  5 个变量的个特殊方差（uniqueness, specific variance）分别为 0.028, 0.237, 0.040, 0.168, 0.052, 其中 “good buy” 和”Suitable for snack” 的 uniquenss 值远大于另外三个变量，这说明两个公共因子不能很好地解释这两个变量。从另外一个角度讲，“Good buy” 和 “ Suitable for snack” 两个变量作为问卷问题让消费者打分可能不是十分恰当，这两个问题以及 “Lots of energy” 的目的可能

是希望得到消费者对新产品的 “实用性” 方面的感受，但显然它们在测量 “ 实用性” 的方面不如 “Lots of energy” 精准。从这个角度看，该问卷设计中的问题有改进余地。为了更好地拟合得到更小的 uniqueness, 我们可以增加因子个数（注意载荷旋转的目的是得到更好解释的因子，并不改变 uniqueness），也可尝试用主因子方法。

- 载荷 L 和因子的解释

  拟合得到的两因子模型为：

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix} + \begin{pmatrix} 0 \\ 0.873 \\ 0.131 \\ 0.817 \\ 0.973 \end{pmatrix} F_1 + \begin{pmatrix} 0.985 \\ 0 \\ 0.971 \\ 0.405 \\ 0 \end{pmatrix} F_2 + error$$

  其中变量 2,4,5 即 Goog buy, Suitable for snack, Lots of energy 在 F_1 上的载荷值分别是 0.873, 0.817, 0.973, 而 Taste 和 Flavor 的载荷接近于 0 (分别为 0, 0.131$\approx$ 0)，即 F_1 与 Taste，flavor 不相关，与其他变量正相关。第一因子载荷说明 Goog buy, Suitable for snack, Lots of energy, 甚至包括 Flavor 可能是同一类变量，F_1 的含义应该是这三或四个变量的某种共同含义，我们可称之为 “实用性”，因为 Lots of energy 的载荷最大 (0.973), 这说明该变量是消费者在实用性考量上最为看重的一个方面，其次是 Good buy(花钱合算) 和 “是不是适合当作小吃”, 而 Flavor 在实用性考虑中最不重要 (0.131).

  第二个因子 F_2 在 Taste, Flavor, Suitable for snack 上的载荷分别是 0.985, 0.971, 0.405 这三个变量都与口味有关，这说明 F_2 可能是口味因子。载荷的取值 Taste 最大，Flavor 次之，这说明对于口味的考量方面，Taste 最重要，Flavor 也很重要，而 Suitable for snack 显得不那么重要。

- 共性方差、方差贡献

  Communality(共性方差)：输出结果中没有列出各变量的共性方差 $h_i^2 = 1 - \text{uniquess}_i, i = 1,...,5$。由所给的特殊方差可计算得到：$(h_1^2, h_2^2, h_3^2, h_4^2, h_5^2) = (0.972, 0.763, 0.960, 0.832, 0.948)$。两个因子能解释变量 1，3，5 的方差的 90% 以上。对变量 2，4 的解释能力稍差，分别是 76.3%,83.2%.

  另外，总方差为 $tr(R) = 5$。输出结果的倒数第 4 行给出了因子载荷的第 1，2 列的平方和（SS）分别是 2.396，2.078，所以它们解释总方差的比率分别为 2.396/5=0.479, 2.078/5=0.416（输出结果的倒数第 3 行），两个因子的累积方差解释比例为 0.895（倒数第 2 行）。

- 因子旋转/载荷旋转 （注意载荷旋转是为了得到更好的解释，并不能改进拟合效果）

  函数 factanal 的缺省旋转方法为 rotation = ”varimax”(旋转后使得 varimax 最大)，所以上述结果中的载荷是载荷的原始极大似然估计通过正交旋转得到的。下面尝试 promax 旋转，得到如下载荷：

```
> promax(loadings(myFA))
Loadings:
Factor1 Factor2
Taste                    1.004
Good_buy           0.892 -0.103
Flavor                   0.975
Suitable_for_snack 0.786  0.313
Lots_of_energy     0.997 -0.137
```

  与 varimax 旋转的结果并没有多大差异。

例 **3.** （样本因子分析）数据集 geopol 给出了 41 个国家和地区的 10 项指标/变量数据，描述如下

变量	描述
pop	population
GDP	Gross Internal Product per habitant
pop.inc	rate of increase of the population
urban	rate of urban population
illiter	rate of illiteracy in the population
student	rate of students in the population
life	expected lifetime of people
nutri	rate of nutritional needs realized
magz	number of newspapers and magazines per 1000 habitants
tv	number of television per 1000 habitants

41 个国家和地区的名称缩写如下：

AFS	South Africa	DAN	Denmark	MAR	Marocco
ALG	Algeria	EGY	Egypt	MEX	Mexico
BRD	Germany	ESP	Spain	NOR	Norway
GBR	Great Britain	FRA	France	PER	Peru
ARS	Saudi Arabia	GAB	Gabun	POL	Poland
ARG	Argentine	GRE	Greece	POR	Portugal
AUS	Australia	HOK	Hong Kong	SUE	Sweden
AUT	Austria	HON	Hungary	SUI	Switzerland
BEL	Belgium	IND	India	THA	Tailand
CAM	Cameroon	IDO	Indonesia	URS	USSR
CAN	Canada	ISR	Israel	USA	USA
CHL	Chile	ITA	Italia	VEN	Venezuela
CHN	China	JAP	Japan	YOU	Yugoslavia
CUB	Cuba	KEN	Kenia		

你可以从原始数据 geopol 计算得到 10 个指标的协方差矩阵或相关系数矩阵，然后应用总体因子分析（并指定样本量 n.obs），也可基于原始数据直接进行因子分析，两者结果相同。

```
geopol= read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/geopol.txt",head=T)
R=cor(geopol)
#R=cov(geopol)
myfa =factanal(covmat=R, factors=2, n.obs=41) # 总体因子分析
myfa2=factanal(x = geopol, factors = 2) # 样本因子分析
> myfa2
        Call:
        factanal(x = geopol, factors = 2)

        Uniquenesses:
        pop     GDP pop.inc   urban illiter student    life   nutri    magz      tv
        0.956   0.005   0.243   0.589   0.169   0.530   0.111   0.525   0.325   0.215

        Loadings:
        Factor1 Factor2
        pop     -0.122  -0.170
        GDP      0.276   0.959
        pop.inc -0.770  -0.406
        urban    0.493   0.409
        illiter -0.829  -0.379
        student  0.466   0.503
        life     0.807   0.488
        nutri    0.649   0.231
        magz     0.391   0.722
        tv       0.483   0.743

        Factor1 Factor2
        SS loadings      3.289   3.041
```

```
Proportion Var   0.329   0.304
Cumulative Var   0.329   0.633

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 45.9 on 26 degrees of freedom.
The p-value is 0.00937
```

输出结果最后三行计算拟合优度统计量（chi-square statistic）

$$W = n_B\left\{\log(|LL^\top + \Psi|/|R|)\right\},$$

其中 n_B 是 Bartlett 校正的样本量 $n_B = (n - 1 - (2p + 4m + 5)/6) \approx n$，及其 p 值

$$pvalue = P(\chi^2_{[(p-m)^2-p-m]/2} \geq W)$$

p 值越大，拟合效果越好。本例的拟合优度 p 值较小，说明拟合不太好。拟合 3 因子模型：

```
myfa3=factanal(x = geopol, factors = 3) # 样本因子分析
        Call:
        factanal(x = geopol, factors = 3)

        Uniquenesses:
        pop     GDP pop.inc   urban illiter student    life   nutri    magz      tv
        0.819   0.005   0.005   0.482   0.174   0.351   0.141   0.489   0.311   0.190

        Loadings:
        Factor1 Factor2 Factor3
        pop                    -0.420
        GDP      0.271   0.908   0.311
        pop.inc -0.910  -0.406
        urban    0.383   0.282   0.541
        illiter -0.755  -0.282  -0.419
        student  0.344   0.349   0.639
        life     0.753   0.418   0.343
        nutri    0.664   0.174   0.201
        magz     0.413   0.698   0.175
        tv       0.444   0.646   0.442

        Factor1 Factor2 Factor3
        SS loadings      3.112   2.385   1.537
        Proportion Var   0.311   0.239   0.154
        Cumulative Var   0.311   0.550   0.703

        Test of the hypothesis that 3 factors are sufficient.
        The chi square statistic is 24.94 on 18 degrees of freedom.
        The p-value is 0.127
```

拟合有较大改进，p 值 0.127。因子 1 和 2 的载荷类似, 人口增长率 pop.inc 和文盲率 illiter 的载荷都是负数，因此因子 1，2 都和这两个指标负相关，大概都代表了国家发达程度（一个代表物质发达程度，一个代表文明发达程度）。我们可画出载荷散点图进行考察, 并结合每个国家的因子得分以及国家背景知识了解两个因子的含义。以两因子模型为例，因子载荷和因子得分散点图：

```
loadings (myfa2)->L
plot(L,type="n")
text(L,  rownames(L))

        # 因子得分 (score):
myfa2 = factanal(x=geopol, factors=2,scores ="Bartlett" )  #
s = myfa2$score
plot (s,type="n")
text(s, rownames(geopol))
```

练习 5. 下面给出了 112 个孩子的 6 项测试（general: 综合, picture: 绘画,blocks: 积木,maze: 迷宫,reading: 阅读,vocab: 词汇量）成绩的协方差矩阵 (R 数据集 ability.cov):

	general	picture	blocks	maze	reading	vocab
general	24.641	5.991	33.520	6.023	20.755	29.701
picture	5.991	6.700	18.137	1.782	4.936	7.204
blocks	33.520	18.137	149.831	19.424	31.430	50.753
maze	6.023	1.782	19.424	12.711	4.757	9.075
reading	20.755	4.936	31.430	4.757	52.604	66.762
vocab	29.701	7.204	50.753	9.075	66.762	135.292

试对该数据做因子分析，两因子模型是否能很好地拟合数据? （提示：首先标准化，把该矩阵转换为相关系数矩阵）。

练习 6. （Johnson and Wichern 课本 Table 9.12）一家公司对其销售员工的销售业绩进行了评估，同时还希望能找到一种或几种能揭示员工销售潜力的测试或考试。公司随机抽取了 50 名销售人员，并对他们的三种销售表现做了评估：销售额增长速度，销售获利成绩，对新客户的销售能力。这三种表现的度量都以 100 作为平均业绩，分数越高表现越好。另外，这 50 人接受了 4 项测试，分别用于测试他们的创造力，物理推理能力 (mechanical reasoning)，抽象推理能力，和数学能力。

```
>sales= read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/T9-12.DAT")
>dimnames(sales)[[2]]=c("Sale.growth", "Sales. profitabiity", "NewAccount.sales",
   "Creativity", "Manchanicl.reasoning.test", "Abstract.reasoning.test", "Math.test")
>sales
```

1. 首先将数据标准化（标准化函数 scale），求 $m = 2$ 和 $m = 3$ 个公共因子的因子分析的极大似然解。
2. 由 (a) 的解，求 $m = 2$ 和 $m = 3$ 的旋转载荷，比较两组旋转载荷，解释因子的含义。
3. 对 $m = 2, 3$, 给出公共方差、个性方差的估计值, 以及残差矩阵 $R - (LL^\top + \Psi)$，你倾向于选择 $m = 2$ 还是 $m = 3$?
4. 计算每个人的因子得分 (factor score), 并画出因子得分的散点图（因子得分计算：函数 factanal 中指定 scores="regression" 或 "Bartlett"）。

提示: 公司的主要目的是找出与销售业绩有关的测试。