作业 **6** **PCA/FA/SVD/CCA**

1. 假设数据矩阵为 $X_{n\times p} = (\mathbf{x}_1,...,\mathbf{x}_n)^\top$, 且 X 已经中心化，假设样本协方差阵 $S = X^\top X/(n-1)$ 的谱分解为 $S = V\Lambda V^\top$, 其中 $V^\top V = VV^\top = I_p$, $\Lambda = \text{diag}(\lambda_1,...,\lambda_p)$, $\lambda_1 \geq ... \geq \lambda_p \geq 0$。主成分矩阵 $Y_{n\times p} = XV$ 的第 j 列为所有 n 个样本点的第 j 主成分，$j = 1,...,p$。

   (a) 证明 Y 是中心化的矩阵，其样本协方差矩阵为 $Y^\top Y/(n-1) = \Lambda$, Y 与 X 的样本协方差矩阵为 $Y^\top X/(n-1) = \Lambda V^\top$ (即第 k 个变量与第 j 个主成分之间的样本协方差为 $v_{kj}\lambda_j$，v_{kj} 是 V 的 (k,j) 元)。

   (b) 证明第 j 主成分和第 k 个变量的样本相关系数 $r_{jk} = v_{kj}\sqrt{\frac{\lambda_j}{s_{kk}}}$, 其中 s_{kk} 为第 k 个变量的样本方差, 即 S 的 (k,k) 元。

2. 假设 $\mathbf{x}_1,...,\mathbf{x}_n \in R^p$ 的样本方差-协方差矩阵具有如下形式

$$S = \sigma^2 I_p + \mathbf{1}_p\mathbf{1}_p^\top + \mathbf{e}\mathbf{e}^\top,$$

   其中 $\mathbf{1}_p = (1,1,...,1)^\top$, $\mathbf{e} = (1,-1,0,...,0)^\top$. 求 $\mathbf{x}_i, i = 1,...,n$ 的第一主成分和第二主成分，以及它们的方差在总方差中所占的累积比例。

3. Thurstone (1933) 关于 PMA(primary mental abilities) 的研究中，对 $n = 213$ 进行了 9 项与语言表达能力有关的测试。9 项测试分别是

   Sentences, Vocabulary, Sentence Completion, First Letters, Four Letter Words, Suffixes, Letter Series, Pedigrees, Letter group

   9 项测试的样本相关系数矩阵如下：

	Sentences	*Vocab*	*Sent.Comp*	*First.Letters*	*Four.Letters*	*Suffixes*	*Letter.Series*	*Pedigrees*	*Letter.Group*
Sentences	1.000	0.828	0.776	0.439	0.432	0.447	0.447	0.541	0.380
Vocab	0.828	1.000	0.779	0.493	0.464	0.489	0.432	0.537	0.358
Sent.Comp	0.776	0.779	1.000	0.460	0.425	0.443	0.401	0.534	0.359
First.Letters	0.439	0.493	0.460	1.000	0.674	0.590	0.381	0.350	0.424
Four.Letters	0.432	0.464	0.425	0.674	1.000	0.541	0.402	0.367	0.446
Suffixes	0.447	0.489	0.443	0.590	0.541	1.000	0.288	0.320	0.325
Letter.Series	0.447	0.432	0.401	0.381	0.402	0.288	1.000	0.555	0.598
Pedigrees	0.541	0.537	0.534	0.350	0.367	0.320	0.555	1.000	0.452
Letter.Group	0.380	0.358	0.359	0.424	0.446	0.325	0.598	0.452	1.000

   上述 9 项测试任务的目的分别是希望测试语言理解能力（Verbal comprehension,1-3）, 用词流利程度（Word fluency,4-6）和推理逻辑能力（Reasoning,7-9）。因为没有原始数据，只有相关系数矩阵，我们进行总体主成分分析（主因子分析），得到如下输出结果：

```
Call:
princomp(covmat = Thurstone)

Standard deviations:
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
2.202  1.044  1.019  0.690  0.669  0.612  0.567  0.484  0.409
```

前 3 个主成分的载荷：

	Sentences	Vocab	Sent.Comp	First.Letters	Four.Letter	Suffixes	Letter.Series	Pedigrees	Letter.Group
Comp.1	0.37	0.38	0.36	0.33	0.32	0.30	0.30	0.32	0.29
Comp.2	0.38	0.33	0.36	-0.44	-0.45	-0.33	-0.04	0.23	-0.24
Comp.3	0.15	0.21	0.19	0.21	0.13	0.36	-0.57	-0.33	-0.52

(a) 试计算前 3 个主成分的方差贡献率。

(b) 试根据成分载荷解释前三个主成分的含义，与 Thurstone 的目标是否一致？

4. 我们应用主成分分析方法研究欧洲 23 个国家的首都城市的气温数据集 temperature。23 个首都的信息如下表所示。

首都城市	国家	位置	首都城市	国家	位置	首都城市	国家	位置
Amsterdam (阿姆斯特丹)	荷兰	W	Athens (雅典)	希腊	S	Berlin (柏林)	德国	W
Brussels (布鲁塞尔)	比利时	W	Budapest (布达佩斯)	匈牙利	E	Copenhagen (哥本哈根)	丹麦	N
Dublin (都柏林)	爱尔兰	N	Helsinki (赫尔辛基)	芬兰	N	Kiev (基辅)	乌克兰	E
Krakow (克拉科夫)	波兰旧都	E	Lisbon (里斯本)	葡萄牙	S	London (伦敦)	英国	N
Madrid (马德里)	西班牙	S	Minsk (明斯克)	白俄罗斯	E	Moscow (莫斯科)	俄罗斯	E
Oslo (奥斯陆)	挪威	N	Paris (巴黎)	法国	W	Prague (布拉格)	捷克	E
Reykjavik (雷克雅未克)	冰岛	N	Rome (罗马)	意大利	S	Sarajevo (萨拉热窝)	波黑	S
Sofia (索菲亚)	保加利亚	E	Stockholm (斯德哥尔摩)	瑞典	N			

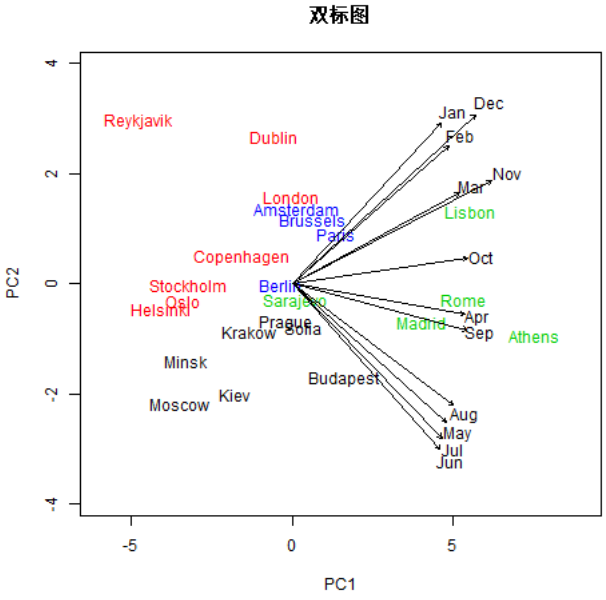
其中地理位置 W、E、S、N 分别表示欧洲西、东、南、北部，每个城市记录了历史上 12 个月份 (Jan-Dec) 的平均气温。对标准化后的数据应用主成分分析方法 (R: princomp)，部分输出结果即主成分的标准差 (Standard deviations) 如下所示（方框内）

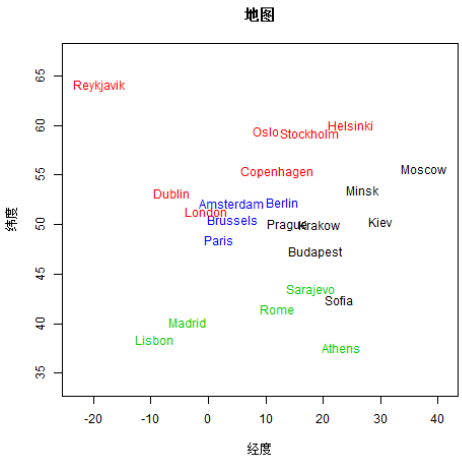
```
Call: princomp(x =  temperature, cor=T)
Standard deviations:
Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9 Comp.10 Comp.11 Comp.12
3.16    1.36    0.36    0.20    0.13    0.11    0.08    0.05    0.03    0.03    0.02    0.01
```

前两个主成分的载荷如下：

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Comp.1	0.27	0.28	0.30	0.31	0.28	0.26	0.27	0.29	0.31	0.31	0.30	0.28
Comp.2	0.39	0.34	0.21	-0.07	-0.34	-0.40	-0.37	-0.30	-0.11	0.06	0.21	0.35

左下图是主成分分析的双标图，作为参考，右下图给出了这 23 个城市的经纬度坐标图。





(a) 试计算第一、二主成分在总方差中的占比。根据载荷数据解释第一、二主成分的含义，它们与地理位置有什么关系？

(b) 四月份和哪个月份的气温相关系数最大？哪个或哪几个月份的温度最能反映南北差异？

(c) 哪个城市或地区春秋两季 (四月和九月）的温度相对较高？萨拉热窝 (Sarajevo) 是南欧城市，其气温与其它南欧城市（地图中的绿色标记城市）相差较远，该城市与其它南欧城市在地理环境上有什么差异？

5. 法国 Decastar 巡回赛是世界上最大的国际田联十项全能 (decathlon) 赛事，比赛次序如下：

第一天：100 米，跳远，铅球，跳高，400 米；

第二天：100 米栏，铁饼，撑杆跳，标枪，1500 米。

其中径赛跑步类以时间度量成绩（单位：秒，时间值越小越好），田赛包括跳跃和投掷项目，以高度或距离度量成绩（单位：米，值越大越好）。对 2004 年该赛事前 13 名运动员 10 项成绩的**相关系数**矩阵做主成分分析，前两个主成分 PC1 和 PC2 的载荷（即主成分方向）如下表所示

	100 米	跳远	铅球	跳高	400 米	100 米栏	铁饼	撑杆跳	标枪	1500 米
PC1	0.48	−0.37	−0.20	−0.19	0.34	0.50	−0.31	0.03	−0.01	−0.31
PC2	0.03	0.00	−0.43	0.27	−0.33	−0.20	−0.19	−0.53	−0.33	−0.40

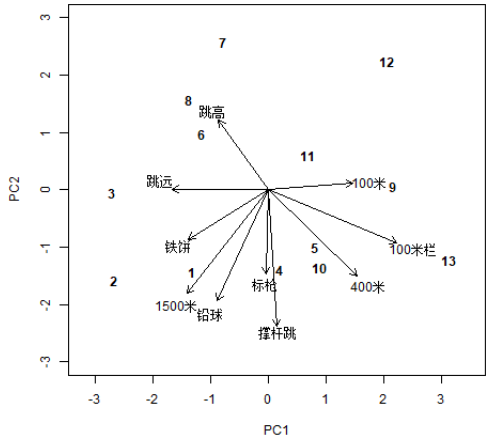
请回答如下问题:

(a) 根据上述载荷，解释第一主成分 PC1 的含义。

(b) 已知前两个主成分 PC1 和 PC2 的标准差分别为 1.76, 1.42, 计算它们的累计方差贡献率。

(c) 分析运动员的 PC 散点图（右图，数字代表运动员，大小代表名次），第一名（数字 1）的投掷类成绩如何（好、中、差）？他的 4 个径赛项目表现各如何（好、中、差）？最后一名（数字 13）有什么特点？

(d) 跳高和撑杆跳是正相关还是负相关？图中跳远和 100 米方向相反说明了什么？1500 米作为径赛项目与其它径赛项目类似吗？简单解释原因。



6. 假设随机向量 $\mathbf{x}=(x_1,x_2,x_3)^\top$ 的协方差矩阵（相关系数矩阵）为

$$\Sigma=\begin{pmatrix}1 & 0.63 & 0.45\\ 0.63 & 1 & 0.35\\ 0.45 & 0.35 & 1\end{pmatrix}$$

假设 $\mathbf{x}=(x_1,x_2,x_3)^\top$ 由如下单因子模型模型产生

$$\begin{aligned}x_1&=0.9F+\epsilon_1\\ x_2&=0.7F+\epsilon_2\\ x_3&=0.5F+\epsilon_3\end{aligned}$$

其中 $var(F)=1, cov(F,\boldsymbol{\epsilon})=\mathbf{0}$, 写出载荷矩阵 L 并求 $\Psi=var(\boldsymbol{\epsilon})$ 的估计。

7. 假设 $\mathbf{x}=(x_1,...,x_6)^\top$ 的相关系数矩阵如下

$$R=\begin{pmatrix}1.000 & & & & & \\ 0.505 & 1.000 & & & & \\ 0.569 & 0.422 & 1.000 & & & \\ 0.602 & 0.467 & 0.926 & 1.000 & & \\ 0.621 & 0.482 & 0.877 & 0.874 & 1.000 & \\ 0.603 & 0.450 & 0.878 & 0.894 & 0.937 & 1.000\end{pmatrix}.$$

对两因子模型，极大似然法得到如下载荷

变量	F_1	F_2
x_1	0.478	0.417
x_2	0.371	0.323
x_3	0.593	0.727
x_4	0.514	0.855
x_5	0.859	0.506
x_6	0.735	0.604

试计算

(a) 特殊方差.

(b) 公共方差.

(c) 每个因子解释的方差比例.

(d) 每个变量被 F_1,F_2 解释的方差的比例.

(e) 残差矩阵 $R-LL^\top-\Psi$.

8. 对第 3 题 Thurstone 数据进行 3 因子分析，得到如下载荷（空白处为 0 或接近 0）：

```
Call:
factanal(factors = 3, covmat = Thurstone, rotation = "promax")
Loadings:
                 Factor1 Factor2 Factor3
Sentences          0.919
Vocabulary         0.901
Sent.Completion    0.842
First.Letters              0.876
Four.Letter.Words          0.756
Suffixes           0.174   0.645  -0.103
Letter.Series                      0.893
Pedigrees          0.347           0.494
Letter.Group      -0.116   0.184   0.675
```

(a) 三个因子是否与第 3 题的主成分的含义相同或接近？

(b) 计算 3 个因子的累计方差贡献率，它一定小于主因子方法的贡献率，解释原因。

(c) 分别计算前 3 个、中间 3 个和最后 3 个变量的 3 因子方差解释比例。