多元统计作业 2

交作业时间：**2025 年 11 月 19 日 23:59（周三晚）**

请在规定时间前提交到大夏学堂，超时得分将有折扣。
计算、证明题提交 pdf 电子版（可手写拍照），编程题提交 R 或 Python 代码、结果及必要的解释。

1. 设$\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2)'$的协方差阵为

$$\Sigma = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix},$$

试从 Σ 和相关阵 R 出发分别求出总体主成分，并加以比较.

2. 对 Fisher iris 数据, 忽略 species 变量, 进行主成分分析, 计算主成分得分，画出碎石图, 主成分得分之间的散点图矩阵.

3. 设标准化变量 $X = (X_1, X_2)'$, $Y = (Y_1, Y_2)'$. 已知 $Z = (X', Y')'$ 的相关矩阵为

$$\mathrm{R} = \begin{pmatrix} 1.0 & 0.5 & 0.7 & 0.7 \\ 0.5 & 1.0 & 0.7 & 0.7 \\ 0.7 & 0.7 & 1.0 & 0.6 \\ 0.7 & 0.7 & 0.6 & 1.0 \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}.$$

求X和Y的典型相关变量和典型相关系数.

4. (Johnson & Wichern Exercise 9.24) 对表 8.5 (T8-5.dat) 的人口普查数据作因子分析. 从相关系数矩阵 R 着手, 求极大似然和主成分解. 评论一下你选择的 m 值. 你的分析需包含因子旋转和因子得分的计算. 数据表的变量名称如下.

Table 8.5 Census-tract Data

Tract	Total population (thousands)	Professional degree (percent)	Employed age over 16 (percent)	Government employment (percent)	Median home value ($100,000)
1	2.67	5.71	69.02	30.3	1.48
2	2.25	4.37	72.98	43.3	1.44
3	3.12	10.27	64.94	32.0	2.11
4	5.14	7.44	71.29	24.5	1.85
5	5.54	9.25	74.94	31.0	2.23
6	5.04	4.84	53.61	48.2	1.60
7	3.14	4.82	67.00	37.6	1.52
8	2.43	2.40	67.20	36.8	1.40
9	5.38	4.30	83.03	19.7	2.07
10	7.34	2.73	72.60	24.5	1.42
⋮	⋮	⋮	⋮	⋮	⋮
52	7.25	1.16	78.52	23.6	1.50
53	5.44	2.93	73.59	22.3	1.65
54	5.83	4.47	77.33	26.2	2.16
55	3.74	2.26	79.70	20.2	1.58
56	9.21	2.36	74.58	21.8	1.72
57	2.14	6.30	86.54	17.4	2.80
58	6.62	4.79	78.84	20.0	2.33
59	4.24	5.82	71.39	27.1	1.69
60	4.72	4.71	78.01	20.6	1.55
61	6.48	4.93	74.23	20.9	1.98

Note: Observations from adjacent census tracts are likely to be correlated. That is, these 61 observations may not constitute a random sample. Complete data set available at www.prenhall.com/statistics.

5. 数据文件 digits.txt 为 USPS 的手写数字数据的一部分, 其为 3300*256 的一个矩阵,每一行为一个 16*16 分辨率的手写数据, 1-1100 行为数字“1”, 1101-2200 行为数字“2”, 2201-3300 行为数字 “3”.
   (1) 将数据随机划分为训练集和测试集.
   (2) 使用训练集, 选择三种不同的判别分类方法, 对分类器进行训练.
   (3) 使用测试集计算错误分类率,对比这些方法(错误分类率和时间成本).
   (4) 由于每个手写数字为 256 维, 使用主成分方法进行降维, 重复上面 1-3 过程, 对比两类做法.

   示例代码:

```
digits<-read.table("digits.txt")
digits<-as.matrix(digits)
library(RColorBrewer)
showMatrix <- function(x, ...)
image(t(x[nrow(x):1,]), xaxt = ’none’, yaxt = ’none’,
col = rev(colorRampPalette(brewer.pal(9, ’GnBu’))(100)), ...)
par(mfrow=c(1,3),mar=rep(0,4))
showMatrix(matrix(digits[1,],16,16))
showMatrix(matrix(digits[1101,],16,16))
showMatrix(matrix(digits[2201,],16,16))
#let’s start with lda classifier.
library(MASS)
digits<-data.frame(cl=rep(1:3,each=1100),digits)
#take 75% samples as training set, you can change this
idx<-c(1:825,1100+1:825,2200+1:825)
#you should take the training set randomly.
train<-digits[idx,]
## record the running time
ptm <- proc.time()
z<-lda(cl~.,data=train)
proc.time() – ptm
#Now it’s your turn to complete...
```