《多元统计分析》课后作业

姓名：李倩倩

学号：2024017349

班级：统计 24-1 班

中国石油大学（北京）克拉玛依校区文理学院数学与统计系

Friday 8th May, 2026

作业要求

1. 可以和其他同学讨论作业当中的问题，但应当自己独立完成作业
2. 计算、证明等要有过程，要有主要步骤的说明
3. 请将计算、绘图所用的 R 代码以及生成的结果和图像一并添加在作业文件当中
4. 请使用 LaTeX编辑并生成 PDF 格式的文件，第 X 周作业文件命名方式：学号-姓名-X.pdf
5. 评分标准：每一问得分 $\in \{2, 1, 0\}$
    - 2: 按时完成并上交作业，且答案基本正确
    - 1: 按时完成并上交作业，且答案部分正确
    - 0: 答案完全错误，或者迟交作业(规定时间72小时之后)
6. 请将完成的 PDF 格式的作业文件发送至邮箱：xiaolei@cup.edu.cn
7. 每位同学可以有一次迟交作业的机会，但不得晚于规定时间三日之后
8. 第 8 周作业截止时间：2026年5月8日24:00

目录

Chapter 1

第 8 周作业

第 8 周作业截止时间： 2026年5月8日24:00

第 8 周作业完成时间： Friday 8th May, 2026 23:08

1. 利用数据矩阵的因子分解方法，简要分析瑞士银行钞票数据集 (mclust 包中的 banknote 数据集).

   (a) [2 分] 利用 R 中的 scale() 函数对数据进行标准化，将标准化之后的数据集记为 $\mathcal{X}$.

   【解】

   加载 mclust 包并读取 banknote 数据集，提取 6 个数值变量（Length, Left, Right, Bottom, Top, Diagonal），利用 `scale()` 对每列进行标准化（各列均值减为 0，样本标准差化为 1），所得矩阵记为 $\mathcal{X}$，维度为 200×6，即 $n = 200$，$p = 6$.

```r
rm(list = ls(all = TRUE))
library(mclust)
data(banknote)

n <- nrow(banknote)        # n = 200
p <- ncol(banknote) - 1    # p = 6

# Extract 6 numeric variables (drop the first column: Status)
X_data <- banknote[, -1]

# Standardize: subtract column mean, divide by sample std (ddof = n-1)
X <- scale(X_data)

# Verify: column means = 0, column sds = 1
round(colMeans(X), 6)
round(apply(X, 2, sd), 6)
```

标准化后，$\mathcal{X}$ 各列均值均为 0，各列样本标准差均为 1，即

$$\mathcal{X}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, \ldots, 200,\ j = 1, \ldots, 6.$$

其中 $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ 为第 j 个变量的样本标准差.

(b) [2 分] 为处理更为规范，我们将 $\mathcal{X}$ 的所有元素除以 $\sqrt{n-1}$，其中 n 为样本容量，得到的数据矩阵记为 $\mathcal{Y}$，现在 $\mathcal{Y}$ 即是我们要分析的数据矩阵.

【解】

将标准化数据矩阵 $\mathcal{X}$ 的每个元素除以 $\sqrt{n-1} = \sqrt{199} \approx 14.1067$，得到分析数据矩阵 $\mathcal{Y}$：

```
Y <- X / sqrt(n - 1)   # sqrt(199) = 14.10674 (the analysis data matrix)
```

由此，$\mathcal{Y}_{ij} = \dfrac{x_{ij} - \bar{x}_j}{s_j\sqrt{n-1}}$，维度仍为 200×6. 经此变换后，$\mathcal{R} = \mathcal{Y}^{\mathrm{T}}\mathcal{Y}$ 恰好等于原始数据的样本相关矩阵，详见下一小题.

(c) [2 分] 求矩阵 $\mathcal{R} = \mathcal{Y}^{\mathrm{T}}\mathcal{Y}$ 的特征值及其对应的单位特征向量. 矩阵 $\mathcal{R} = \mathcal{Y}^{\mathrm{T}}\mathcal{Y}$ 是原始数据的相关矩阵.

【解】

计算 $\mathcal{R} = \mathcal{Y}^{\mathrm{T}}\mathcal{Y}$ 并对其进行谱分解，求得特征值与对应单位特征向量.

```
# Compute correlation matrix R = Y^T Y
R <- t(Y) %*% Y

# Verify: R equals the sample correlation matrix of the raw data
round(R, 4)
round(cor(X_data), 4)   # should be identical to R above

# Spectral decomposition: eigen() returns eigenvalues in descending order
eig    <- eigen(R)
lambda <- eig$values    # eigenvalue vector
U      <- eig$vectors   # unit eigenvector matrix (columns are eigenvectors)
round(lambda, 4)
round(U, 4)
```

相关矩阵 $\mathcal{R}$：

$$\mathcal{R} = \begin{pmatrix} 1.0000 & 0.2313 & 0.1518 & -0.1898 & -0.0613 & 0.1943 \\ 0.2313 & 1.0000 & 0.7433 & 0.4138 & 0.3623 & -0.5032 \\ 0.1518 & 0.7433 & 1.0000 & 0.4868 & 0.4007 & -0.5165 \\ -0.1898 & 0.4138 & 0.4868 & 1.0000 & 0.1419 & -0.6230 \\ -0.0613 & 0.3623 & 0.4007 & 0.1419 & 1.0000 & -0.5940 \\ 0.1943 & -0.5032 & -0.5165 & -0.6230 & -0.5940 & 1.0000 \end{pmatrix}$$

（行列对应顺序：Length, Left, Right, Bottom, Top, Diagonal）

特征值（降序排列）：

$$\lambda_1 = 2.9456, \quad \lambda_2 = 1.2781, \quad \lambda_3 = 0.8690, \quad \lambda_4 = 0.4498, \quad \lambda_5 = 0.2687, \quad \lambda_6 = 0.1889$$

验证：$\sum_{k=1}^{6} \lambda_k = 6.0000 = p$，与相关矩阵迹 $\mathrm{tr}(\mathcal{R}) = 6$ 一致.

对应的单位特征向量（列向量 $\boldsymbol{u}_k$）：

变量	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	$\boldsymbol{u}_3$	$\boldsymbol{u}_4$	$\boldsymbol{u}_5$	$\boldsymbol{u}_6$
Length	−0.0070	−0.8155	−0.0177	−0.5746	−0.0588	−0.0311
Left	0.4678	−0.3420	0.1034	0.3949	0.6395	0.2977
Right	0.4867	−0.2525	0.1235	0.4303	−0.6141	−0.3492
Bottom	0.4068	0.2662	0.5835	−0.4037	−0.2155	0.4624
Top	0.3679	0.0915	−0.7876	−0.1102	−0.2198	0.4190
Diagonal	−0.4935	−0.2739	0.1139	0.3919	−0.3402	0.6318

各 $\boldsymbol{u}_k$ 均满足 $\|\boldsymbol{u}_k\| = 1$ 且两两正交.

(d) [2 分] 因子变量是数据矩阵 $\mathcal{Y}$ 的六个变量 $Y_1 \sim Y_6$ 的线性组合，写出前两个因子变量 Z_1 和 Z_2 的表达式.

【解】

由讲义第 10 章，第 k 个因子变量 $Z_k = \mathcal{Y}\boldsymbol{u}_k$，其元素为初始变量 $Y_1, \ldots, Y_6$ 的线性组合，线性组合系数由 $\boldsymbol{u}_k$ 的各分量给出（Y_1 =Length，Y_2 =Left，Y_3 =Right，Y_4 =Bottom，Y_5 =Top，Y_6 =Diagonal）.

```
# Direction vectors of the first two factor axes
u1 <- U[, 1]   # corresponds to the largest eigenvalue lambda_1
u2 <- U[, 2]   # corresponds to the second largest eigenvalue lambda_2
```

将 $\boldsymbol{u}_1$ 和 $\boldsymbol{u}_2$ 的分量代入，得前两个因子变量的表达式：

$$Z_1 = -0.0070\,Y_1 + 0.4678\,Y_2 + 0.4867\,Y_3 + 0.4068\,Y_4 + 0.3679\,Y_5 - 0.4935\,Y_6$$

$$Z_2 = -0.8155\,Y_1 - 0.3420\,Y_2 - 0.2525\,Y_3 + 0.2662\,Y_4 + 0.0915\,Y_5 - 0.2739\,Y_6$$

说明： Z_1 的系数显示 Left、Right、Bottom、Top 对其有正向贡献，Diagonal 有较强负向贡献，Length 贡献几乎为零；Z_2 主要由 Length（系数最大）主导.

(e) [2 分] 计算前两个因子变量 Z_1 和 Z_2 对应的特征值之和占所有特征值之和的比例.

【解】

由讲义定义，该比例 τ_2 反映了二维因子表示所能解释的总惯量（方差）的比例：

```
tau2 <- sum(lambda[1:2]) / sum(lambda)
cat("tau_2 =", round(tau2, 4), "\n")
```

```
# output: tau_2 = 0.7039
```

$$\tau_2 = \frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{6} \lambda_k} = \frac{2.9456 + 1.2781}{6.0000} = \frac{4.2237}{6.0000} \approx \boxed{0.7039}$$

前两个因子变量共解释了总方差的约 **70.39**%，二维图形可以较好地展现数据的主要结构.

(f) [2 分] 计算观测数据 (行) 在前两个因子变量 Z_1 和 Z_2 上的坐标值.

【解】

n 个观测点在第 k 个因子轴 $\boldsymbol{u}_k$ 上的坐标为 $\boldsymbol{z}_k = \mathcal{Y}\boldsymbol{u}_k$，全部观测的坐标矩阵为 $Z = \mathcal{Y}U_{[:,1:2]}$，维度为 200×2.

```
# Compute coordinates of 200 observations along Z1 and Z2
Z <- Y %*% U[, 1:2]
colnames(Z) <- c("Z1", "Z2")

# Preview first 6 rows (genuine) and rows 101-106 (counterfeit)
round(Z[1:6, ], 5)
round(Z[101:106, ], 5)
```

前 6 个观测（真钞）与第 101–106 个观测（伪钞）的坐标如下：

观测序号	Z_1	Z_2	类别
1	0.12356	−0.11673	genuine
2	−0.16082	0.03810	genuine
3	−0.16104	0.00761	genuine
4	−0.16147	0.00620	genuine
5	−0.18612	−0.00277	genuine
6	0.05363	−0.21841	genuine
101	0.07817	−0.05174	counterfeit
102	0.10553	−0.01264	counterfeit
103	0.09100	0.02020	counterfeit
104	0.11840	−0.03290	counterfeit
105	0.13066	0.01028	counterfeit
106	0.08716	−0.07267	counterfeit

真钞的 Z_1 均值约为 -0.1059，伪钞的 Z_1 均值约为 $+0.1059$，两类在 Z_1 方向上有明显分离.

(g) [2 分] 作观测数据在前两个因子变量 Z_1 和 Z_2 上的散点图，将真钞与伪钞的数据点分别用不同的颜色表示，你能看到什么现象.

【解】

```
# Get banknote class labels
status <- banknote$Status
```

```r
# Scatter plot on Z1-Z2: genuine = blue circle, counterfeit = red triangle
plot(Z[, 1], Z[, 2],
     col = ifelse(status == "genuine", "steelblue", "firebrick"),
     pch = ifelse(status == "genuine", 16, 17),
     xlab = expression(Z[1]),
     ylab = expression(Z[2]),
     main = "Banknote: Observation Scatter Plot on Z1-Z2")
abline(h = 0, v = 0, lty = 2, col = "gray50")
legend("topright",
       legend = c("Genuine (genuine)", "Counterfeit (counterfeit)"),
       col    = c("steelblue", "firebrick"),
       pch    = c(16, 17),
       bty    = "n")
```

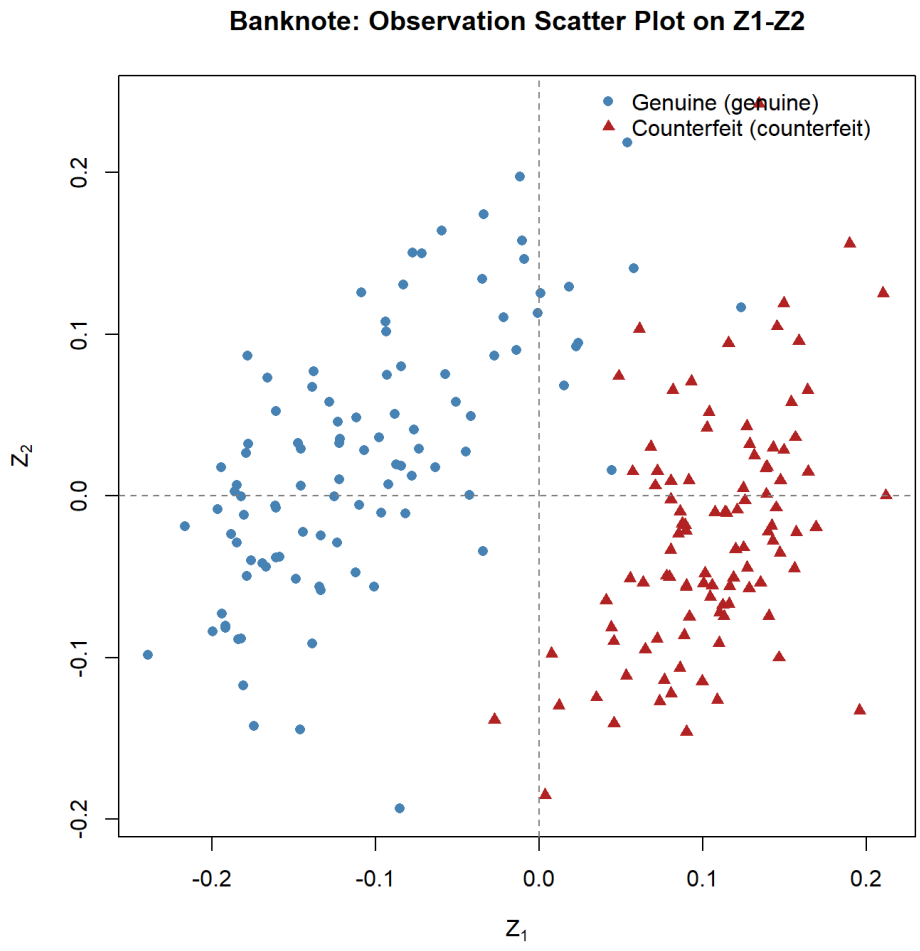

图 1.1: 200 个观测在因子变量 Z_1-Z_2 上的散点图（蓝色圆点为真钞，红色三角为伪钞）

现象分析：

- 沿 Z_1 轴（第一因子变量），真钞（蓝色）集中于负半轴，伪钞（红色）集中于正半轴，两组有明显的分离.

- 沿 Z_2 轴（第二因子变量），两组的分离程度远不如 Z_1 轴明显，两组在 Z_2 方向上有较大重叠.
- 这说明第一因子变量 Z_1 是区分真伪钞票的主要判别方向，该因子主要反映 Left、Right、Bottom、Top（对角线两侧尺寸）与 Diagonal（对角线长度）之间的相对大小关系.

(h) [2 分] 计算变量数据 (列) 在前两个因子变量 W_1 和 W_2 上的坐标值.

【解】

由讲义定理 10.4，p 个变量点在第 k 个因子轴（G_k 方向，对应 $\mathcal{Y}\mathcal{Y}^{\mathrm{T}}$ 的特征向量 $\boldsymbol{v}_k$）上的坐标为

$$\boldsymbol{w}_k = \mathcal{Y}^{\mathrm{T}}\boldsymbol{v}_k = \sqrt{\lambda_k}\,\boldsymbol{u}_k,$$

即 W_k 是对应特征值的平方根与单位特征向量的乘积，不需要显式计算 $\boldsymbol{v}_k$.

```
# Variable coordinates: W_k = sqrt(lambda_k) * u_k
# Matrix form: W = U[,1:2] %*% diag(sqrt(lambda[1:2]))
W <- U[, 1:2] %*% diag(sqrt(lambda[1:2]))
rownames(W) <- colnames(X_data)
colnames(W) <- c("W1", "W2")
round(W, 4)
```

6 个变量在 W_1-W_2 平面上的坐标为：

变量	$W_1 = \sqrt{\lambda_1}\,u_{j1}$	$W_2 = \sqrt{\lambda_2}\,u_{j2}$
Length	−0.0120	−0.9219
Left	0.8028	−0.3866
Right	0.8353	−0.2854
Bottom	0.6981	0.3010
Top	0.6314	0.1034
Diagonal	−0.8469	−0.3097

注意 $\boldsymbol{w}_k$ 即是变量 Y_j 与第 k 个因子变量 Z_k 的协方差（相关系数），因此 $|W_{jk}|$ 越大表明该变量与第 k 个因子的关联越密切.

(i) [2 分] 作变量点在前两个因子变量 W_1 和 W_2 上的散点图，你能看到什么现象.

【解】

```
# Scatter plot of variable points on W1-W2
plot(W[, 1], W[, 2],
     type = "n",
     xlim = c(-1.1, 1.1), ylim = c(-1.1, 1.1),
     xlab = expression(W[1]),
     ylab = expression(W[2]),
     main = "Banknote: Variable Scatter Plot on W1-W2",
     asp  = 1)
text(W[, 1], W[, 2], rownames(W), cex = 1.0, col = "darkgreen")
```

```r
abline(h = 0, v = 0, lty = 2, col = "gray50")
# Draw unit circle as reference
theta <- seq(0, 2 * pi, length.out = 200)
lines(cos(theta), sin(theta), col = "gray70", lty = 3)
```

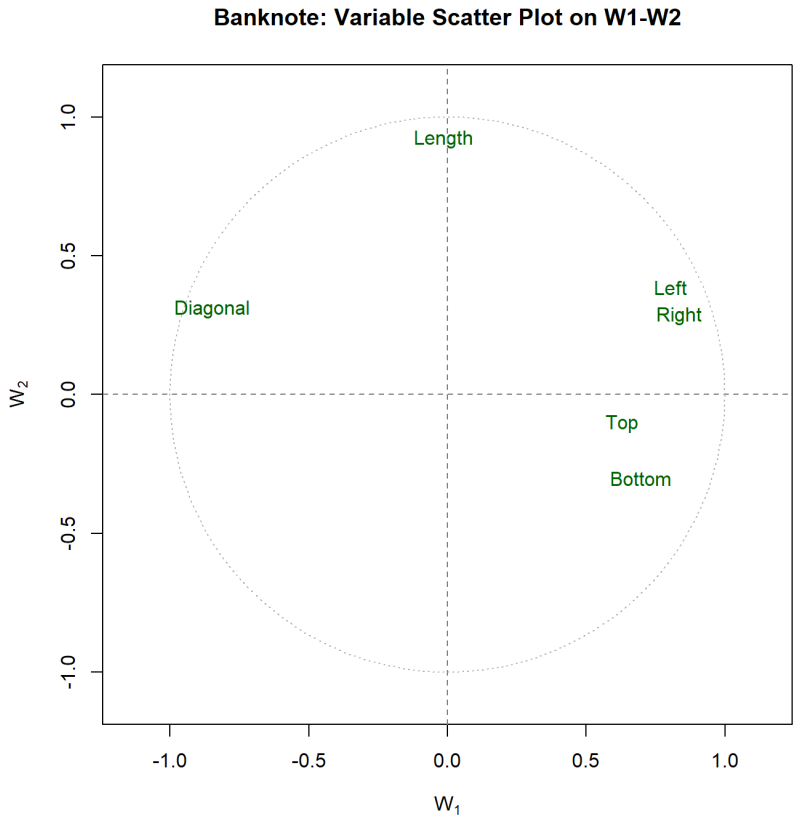

图 1.2: 6 个变量在因子变量 W_1-W_2 上的散点图

现象分析：

- 在 W_1 方向（第一因子轴）上，Left、Right、Bottom、Top 的坐标均为正值（约在 0.63 ∼ 0.84），而 Diagonal 的坐标为负值（−0.85），Length 接近零. 这说明第一因子 Z_1 主要刻画了钞票宽度/高度尺寸（Left、Right、Bottom、Top）与对角线长度（Diagonal）之间的对立关系.
- 在 W_2 方向（第二因子轴）上，Length 的坐标绝对值最大（−0.9219），独立于其他变量，说明第二因子 Z_2 主要反映钞票长度（Length）的信息.
- Left 与 Right 靠近，Bottom 与 Top 靠近，说明这两对变量高度相关（与相关矩阵中 Left-Right 相关系数 0.74，Bottom-Top 为 0.14 相呼应）.
- 从变量散点图可知：真钞与伪钞在钞票尺寸上的系统差异主要体现在 Diagonal 与 Left/Right/Bottom/T 的比例关系上，这与第一因子变量 Z_1 能够有效区分真伪钞票的结论一致.