《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

Thursday 14$^{\text{th}}$ 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 9 周作业截止时间：2026年5月15日24:00

目录

Chapter 1

第 9 周作业

第 9 周作业截止时间： 2026年5月15日24:00

第 9 周作业完成时间： Thursday 14th May, 2026 17:18

1. 美国犯罪数据集 (uscrime.csv) 的主成分分析.

   该数据集由 11 个变量的 50 个观测值组成，它提供了 1985 年美国 50 个州报告的犯罪数量及其它一些信息，我们利用其中 7 个变量 $(X_3, \ldots, X_9)$ 的数据来作主成分分析. 数据集当中各个变量的定义如表 **??** 所示.

表 1.1: 美国犯罪数据集的变量含义.

X_1	land area	占地面积
X_2	popu 1985	1985年的人口数量
X_3	murder	凶杀
X_4	rape	强奸
X_5	robbery	抢劫
X_6	assault	人身袭击
X_7	burglary	入室盗窃
X_8	larceny	偷盗
X_9	autotheft	汽车盗窃
X_{10}	region	美国各州所处地区
X_{11}	division	美国各州所属分部

   (a) [2 分] 读入原始数据，提取拟分析的数据集 $\mathscr{X}$.

   【解】

读入数据集，提取第 3 至第 9 列（即 $X_3, \ldots, X_9$，对应 murder、rape、robbery、assault、burglary、larceny、autotheft）构成数据集 $\mathscr{X}$，共 50 行 7 列.

```
crime <- read.csv("uscrime.csv", row.names = 1)
dim(crime)      # 50 x 11
names(crime)    # check all variable names

# extract X3-X9
X <- crime[, c("murder","rape","robbery","assault",
               "burglary","larceny","autotheft")]
dim(X)          # 50 x 7
head(X)
```

输出结果：

```
> dim(crime)
[1] 50 11
> dim(X)
[1] 50  7
> head(X)
   murder rape robbery assault burglary larceny autotheft
ME    1.5  7.0    12.6      62      562    1055       146
NH    2.0  6.0    12.1      36      566     929       172
VT    1.3 10.3     7.6      55      731     969       124
MA    3.5 12.0    99.5      88     1134    1531       878
RI    3.2  3.6    78.3     120     1019    2186       859
CT    3.5  9.1    70.4      87     1084    1751       484
```

(b) [2 分] 由于数据集 $\mathscr{X}$ 中各变量的数据尺度有较大差异，我们先对其作标准化处理，将标准化后的数据集记为 $\mathscr{Y}$.

【解】

使用 `scale()` 函数对 $\mathscr{X}$ 进行标准化（减去均值、除以标准差），得到标准化数据集 $\mathscr{Y}$，其中每列均值为 0、方差为 1.

```
Y <- scale(X)    # standardize: mean = 0, sd = 1
round(head(Y), 4)
```

输出结果：

```
> round(head(Y), 4)
   murder    rape robbery assault burglary larceny autotheft
ME -1.3924 -1.1725 -0.9750 -1.0770  -1.0215 -1.2519   -1.1115
```

```
NH -1.2625 -1.3086 -0.9804 -1.4584  -1.0104 -1.4294   -0.9812
VT -1.4444 -0.7234 -1.0298 -1.1797  -0.5534 -1.3731   -1.2217
MA -0.8727 -0.4921 -0.0220 -0.6956   0.5628 -0.5813    2.5557
RI -0.9506 -1.6352 -0.2545 -0.2262   0.2443  0.3414    2.4605
CT -0.8727 -0.8867 -0.3411 -0.7103   0.4243 -0.2714    0.5818
```

(c) [2 分] 计算数据集 $\mathscr{Y}$ 的样本协方差矩阵 $\mathscr{S}$，并与数据集 $\mathscr{X}$ 的相关矩阵 $\mathscr{R}$ 进行比较.

【解】

对标准化数据 $\mathscr{Y}$ 计算样本协方差矩阵 $\mathscr{S}$，并计算原始数据 $\mathscr{X}$ 的相关矩阵 $\mathscr{R}$，两者完全相等（误差在机器精度范围内），这是因为标准化变量的协方差就等于原始变量的相关系数.

```r
S <- cov(Y)           # sample covariance matrix of Y
R <- cor(X)           # correlation matrix of X
round(S, 6)
round(R, 6)
all.equal(S, R)       # verify S and R are equal
```

输出结果（$\mathscr{S}$ 与 $\mathscr{R}$ 完全一致）：

```
> round(R, 6)
          murder    rape robbery  assault burglary  larceny autotheft
murder   1.000000 0.519868 0.341058 0.812557 0.276724 0.064783  0.109829
rape     0.519868 1.000000 0.551439 0.695932 0.680154 0.600606  0.440703
robbery  0.341058 0.551439 1.000000 0.563203 0.622192 0.436181  0.617053
assault  0.812557 0.695932 0.563203 1.000000 0.520720 0.316700  0.330380
burglary 0.276724 0.680154 0.622192 0.520720 1.000000 0.801101  0.700100
larceny  0.064783 0.600606 0.436181 0.316700 0.801101 1.000000  0.554779
autotheft 0.109829 0.440703 0.617053 0.330380 0.700100 0.554779  1.000000
> all.equal(S, R)
[1] TRUE
```

由输出可知，$\mathscr{S} = \mathscr{R}$，两者数值完全相同，最大绝对误差为 5.55×10^{-16}（机器精度范围内）.

(d) [2 分] 对相关矩阵 $\mathscr{R}$ 作谱分解 $\mathscr{R} = \Gamma \Lambda \Gamma^{\mathrm{T}}$，给出谱分解的结果并作验证运算.

【解】

对相关矩阵 $\mathscr{R}$ 调用 `eigen()` 函数进行谱分解，得到特征值矩阵 $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_7)$ 和特征向量矩阵 Γ，其中 Γ 为正交矩阵，各列为对应特征值的单位特征向量，特征值由大到小排列.

```r
eig   <- eigen(R)
Lambda <- diag(eig$values)      # diagonal matrix of eigenvalues
Gamma  <- eig$vectors           # matrix of eigenvectors
```

```r
# flip sign of PC1 if all loadings are negative
if (all(Gamma[, 1] < 0)) Gamma[, 1] <- -Gamma[, 1]

round(eig$values, 6)            # eigenvalues
round(Gamma, 6)                 # eigenvectors

# verify: R_check should equal R
R_check <- Gamma %*% Lambda %*% t(Gamma)
max(abs(R_check - R))
```

谱分解结果：

特征值（由大到小）：

$$\lambda_1 = 4.0768, \quad \lambda_2 = 1.4316, \quad \lambda_3 = 0.6312, \quad \lambda_4 = 0.3401, \quad \lambda_5 = 0.2484, \quad \lambda_6 = 0.1397, \quad \lambda_7 = 0.1322$$

特征向量矩阵 Γ（各列对应 $\lambda_1, \ldots, \lambda_7$）：

```
> round(eig$values, 6)
[1] 4.076779 1.431642 0.631169 0.340146 0.248395 0.139670 0.132198

> round(Gamma, 6)
               PC1       PC2       PC3       PC4       PC5       PC6       PC7
murder    0.276102 -0.644479 -0.009512 -0.328660 -0.202982  0.100249 -0.590809
rape      0.421366 -0.116320 -0.360072  0.295788  0.759031 -0.065174 -0.106780
robbery   0.387459  0.045959  0.604141  0.645146 -0.189568  0.068970 -0.160844
assault   0.388088 -0.456244  0.010739 -0.066876 -0.135881  0.100406  0.779805
burglary  0.436358  0.257153 -0.155373 -0.143573 -0.291594 -0.782918 -0.026886
larceny   0.360361  0.400882 -0.508163  0.048057 -0.359894  0.560949 -0.068614
autotheft 0.353856  0.366076  0.472002 -0.600736  0.337223  0.207943  0.012045

> max(abs(R_check - R))
[1] 8.88e-16
```

验证结果：$\Gamma\Lambda\Gamma^{\mathrm{T}}$ 与 $\mathscr{R}$ 的最大绝对误差为 8.88×10^{-16}，在机器精度范围内，谱分解正确.

(e) [2 分] 画碎石图，问各个主成分的贡献率是多少？前 2 个主成分的累积贡献率是多少，前 3 个主成分的累积贡献率又是多少.

【解】

各主成分的贡献率 $= \lambda_k / \sum_{i=1}^{7} \lambda_i$，累积贡献率为各贡献率之和.

```r
lambda    <- eig$values
contrib   <- lambda / sum(lambda)
```

```r
cum_contr <- cumsum(contrib)
result    <- data.frame(
  PC       = paste0("PC", 1:7),
  lambda   = round(lambda, 4),
  rate     = round(contrib * 100, 2),
  cum_rate = round(cum_contr * 100, 2)
)
print(result)

# scree plot
par(mfrow = c(1, 2))
plot(1:7, lambda, type = "b", pch = 16, col = "steelblue",
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot (US Crime Data)")
abline(h = 1, col = "red", lty = 2)

barplot(contrib * 100, names.arg = paste0("PC", 1:7),
        col = "steelblue", ylim = c(0, 70),
        xlab = "Principal Component", ylab = "Contribution Rate (%)",
        main = "Contribution Rate")
lines(seq(0.7, by = 1.2, length.out = 7), cum_contr * 100,
      type = "b", col = "red", pch = 16)
abline(h = 80, col = "gray", lty = 2)
```

各主成分贡献率：

```
   PC   lambda   rate cum_rate
1 PC1   4.0768  58.24    58.24
2 PC2   1.4316  20.45    78.69
3 PC3   0.6312   9.02    87.71
4 PC4   0.3401   4.86    92.57
5 PC5   0.2484   3.55    96.12
6 PC6   0.1397   2.00    98.11
7 PC7   0.1322   1.89   100.00
```

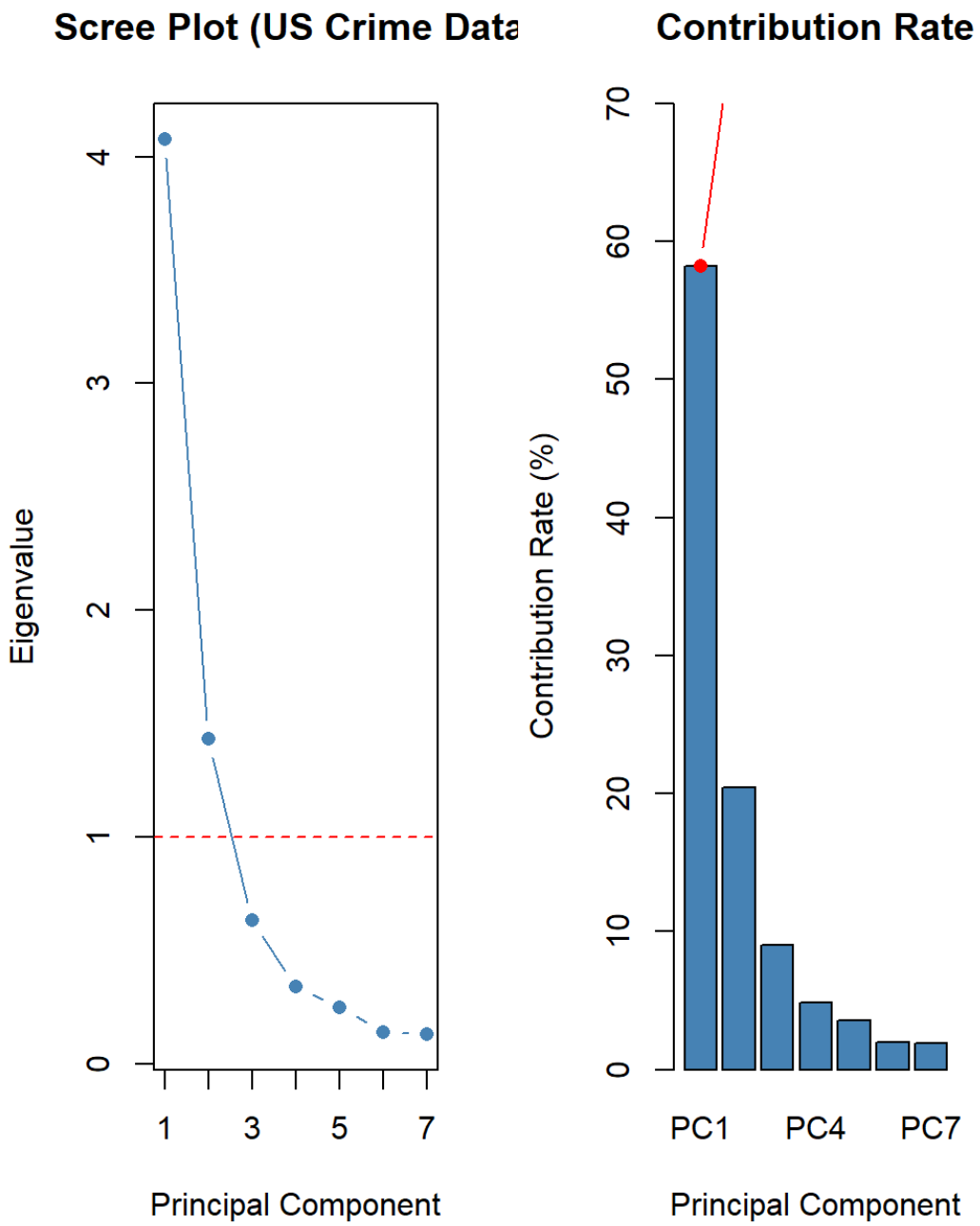

图 1.1: 美国犯罪数据集的碎石图（左）与贡献率图（右）.

由图和表可知：

- 第 1 主成分贡献率为 58.24%，第 2 主成分贡献率为 20.45%，第 3 主成分贡献率为 9.02%，后四个主成分合计贡献不足 13%.
- **前 2 个主成分的累积贡献率为** 78.69%.
- **前 3 个主成分的累积贡献率为** 87.71%.
- 从碎石图观察，$\lambda_1 = 4.08 > \lambda_2 = 1.43 > 1 > \lambda_3 = 0.63$，在 PC2 之后出现明显”折弯”，说明前 2 个主成分已经捕捉了数据的主要信息.

(f) [2 分] 计算各变量可以用前 2 个主成分解释的比例.

【解】

设 γ_{ij} 为特征向量矩阵 Γ 的第 i 行第 j 列元素，λ_j 为第 j 个特征值. 则变量 X_i 与第 j 个主成分的相关系数（载荷）为

$$\ell_{ij} = \gamma_{ij}\sqrt{\lambda_j}$$

变量 X_i 可由前 2 个主成分解释的比例（公因子方差/共同度）为

$$h_i^2 = \sum_{j=1}^{2} \ell_{ij}^2 = \gamma_{i1}^2\lambda_1 + \gamma_{i2}^2\lambda_2$$

```
# proportion of each variable explained by first 2 PCs (communality)
h2 <- Gamma[, 1]^2 * lambda[1] + Gamma[, 2]^2 * lambda[2]
names(h2) <- colnames(X)
round(h2, 4)
```

输出结果：

```
> round(h2, 4)
   murder      rape   robbery   assault  burglary   larceny autotheft
   0.9054    0.7432    0.6150    0.9120    0.8709    0.7595    0.7023
```

各变量可由前 2 个主成分解释的比例（共同度）如下表：

变量	murder	rape	robbery	assault	burglary	larceny	autotheft
h_i^2	0.9054	0.7432	0.6150	0.9120	0.8709	0.7595	0.7023
百分比	90.54%	74.32%	61.50%	91.20%	87.09%	75.95%	70.23%

可以看出，assault（91.20%）、murder（90.54%）和 burglary（87.09%）被前 2 个主成分解释的比例最高，robbery（61.50%）和 autotheft（70.23%）相对最低. 整体来看，前 2 个主成分对所有变量的解释比例均超过 60%，解释效果较好.

(g) [2 分] 作变量在前 2 个主成分平面上的散点图，对结果进行解释.

【解】

以各变量与前 2 个主成分的相关系数（载荷）(ℓ_{i1}, ℓ_{i2}) 为坐标，作变量在 PC1–PC2 平面上的散点图（相关系数圆图）.

```
# compute correlations between variables and first 2 PCs (loadings)
L <- matrix(NA, 7, 2)
for (j in 1:2) L[, j] <- Gamma[, j] * sqrt(lambda[j])
rownames(L) <- colnames(X)
colnames(L) <- c("PC1", "PC2")
round(L, 4)
```

```r
# correlation circle plot (variable scatter on PC1-PC2 plane)
plot(L[, 1], L[, 2],
     xlim = c(-1.2, 1.2), ylim = c(-1.2, 1.2), asp = 1,
     xlab = sprintf("PC1 (%.1f%%)", contrib[1]*100),
     ylab = sprintf("PC2 (%.1f%%)", contrib[2]*100),
     main = "Variable Biplot on PC1-PC2 Plane")
arrows(0, 0, L[, 1], L[, 2], length = 0.1, col = 1:7, lwd = 2)
text(L, rownames(L), col = 1:7, pos = 4, cex = 0.9)
theta <- seq(0, 2*pi, length.out=200)
lines(cos(theta), sin(theta), col="gray")
abline(h = 0, v = 0, lty = 2)
```

载荷（各变量与前 2 个主成分的相关系数）：

```
> round(L, 4)
             PC1     PC2
murder    0.5575 -0.7711
rape      0.8508 -0.1392
robbery   0.7823  0.0550
assault   0.7836 -0.5459
burglary  0.8811  0.3077
larceny   0.7276  0.4797
autotheft 0.7145  0.4380
```

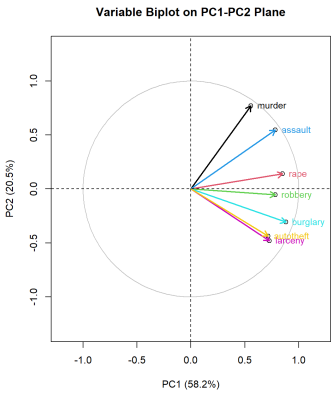

图 1.2: 美国犯罪数据各变量在 PC1–PC2 平面上的散点图.

解释：

- **第一主成分（PC1）**：所有变量在 PC1 上的载荷均为正（0.56 ∼ 0.88），说明 PC1 代表犯罪的整体水平，可称为”综合犯罪指数”，PC1 越大表示整体犯罪率越高.

- **第二主成分（PC2）**：burglary、larceny、autotheft 在 PC2 上的载荷为正（财产型犯罪），而 murder、assault 在 PC2 上的载荷为负（暴力型犯罪）. 因此 PC2 反映了"财产犯罪"与"暴力犯罪"之间的对比关系.

- murder 和 assault 两个向量方向相近（均位于第四象限），说明两者高度正相关；burglary、larceny、autotheft 三者方向相近（位于第一象限），说明财产型犯罪之间高度相关.

(h) [2 分] 作每个州的观测数据在前 2 个主成分平面上的散点图，能否看出美国四个地区存在不同? 各州所在地区由变量 X_{10} 提供.

【解】

计算各州的主成分得分，以 PC1、PC2 为坐标，按各州所属地区（X_{10}）用不同颜色标记，作散点图.

```
# compute principal component scores
scores <- Y %*% Gamma     # Y is standardized; Gamma is the eigenvector matrix
# note: if Gamma[,1] was flipped, scores[,1] is flipped accordingly

region <- crime$region
col_map <- c(Midwest = "blue", Northeast = "green",
             South = "red", West = "purple")
pch_map <- c(Midwest = 16, Northeast = 15, South = 17, West = 18)

plot(scores[, 1], scores[, 2],
     col = col_map[region], pch = pch_map[region],
     xlab = sprintf("PC1 (%.1f%%)", contrib[1]*100),
     ylab = sprintf("PC2 (%.1f%%)", contrib[2]*100),
     main = "Observations on PC1-PC2 Plane (by Region)")
text(scores[, 1], scores[, 2], rownames(crime),
     col = col_map[region], cex = 0.6, pos = 4)
legend("topright", legend = names(col_map),
       col = col_map, pch = c(16,15,17,18), cex = 0.9)
abline(h = 0, v = 0, lty = 2)
```

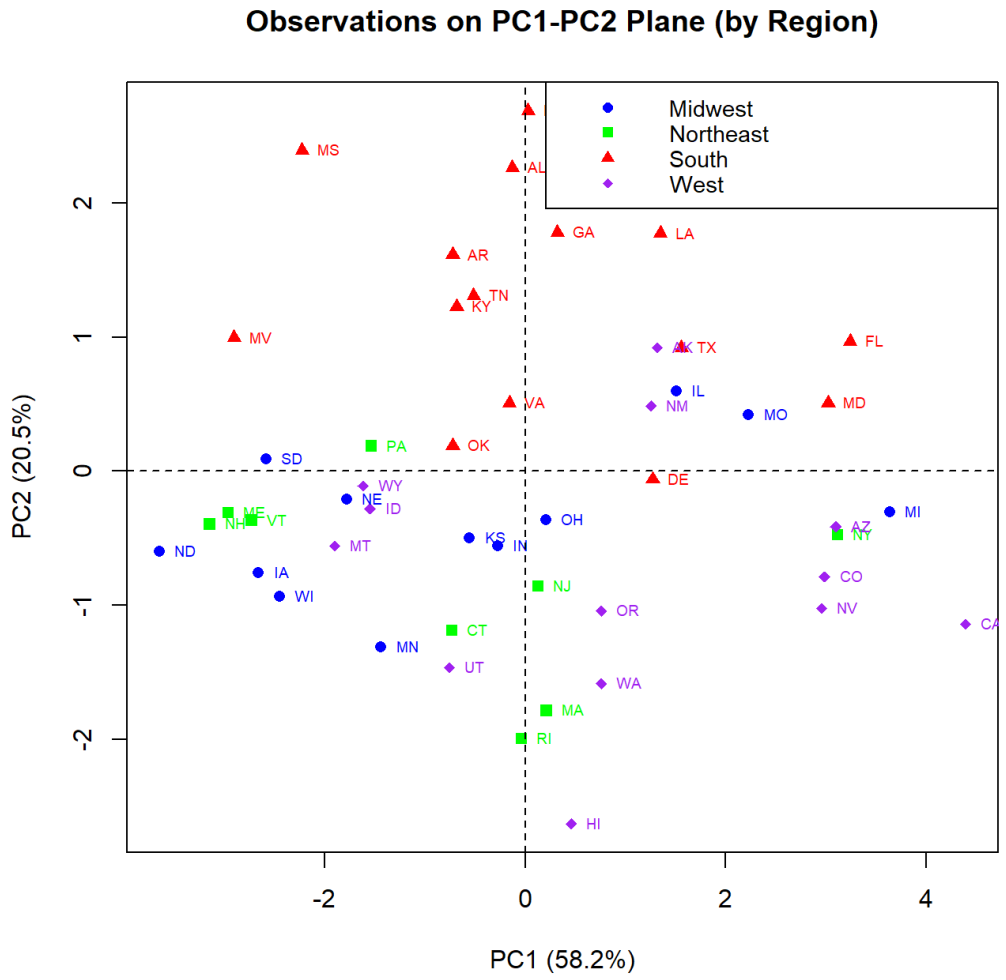

图 1.3: 美国各州在 PC1–PC2 平面上的散点图（按地区着色）.

结论：

- 从散点图可以看出四个地区存在明显差异：
  - **Northeast（东北部，绿色）**：集中在 PC1 较小、PC2 较大的区域，说明东北部整体犯罪率居中，但财产型犯罪相对偏高（如 MA、RI 的 autotheft 较高）.
  - **Midwest（中西部，蓝色）**：多分布在 PC1 较大（即犯罪率整体偏低）的区域，但内部有所分化，MO 例外偏向犯罪率较高的方向.
  - **South（南部，红色）**：多集中在 PC1 较小（犯罪率高）且 PC2 较低（暴力犯罪比例高）的区域，如 FL、GA、SC 等州，表明南部整体犯罪率高，且暴力犯罪较突出.
  - **West（西部，紫色）**：在 PC1–PC2 平面上分布较分散，AK、CA、NV 等地犯罪率较高，MT、ID 等州犯罪率较低.
- 四个地区在 PC1–PC2 平面上的分布存在一定规律性的聚集，说明可以通过前 2 个主成分区分出部分地区差异，尤其是 South 与 Northeast 的差异最为明显.

(i) [2 分] 是否有必要考虑第 3 个主成分?

【解】

```
# eigenvalue and contribution rate of the 3rd PC
cat(sprintf("lambda_3 = %.4f\n", lambda[3]))
cat(sprintf("PC3 contribution rate: %.2f%%\n", contrib[3]*100))
cat(sprintf("Cumulative contribution of first 3 PCs: %.2f%%\n", cum_contr[3]*100))
```

```
lambda_3 = 0.6312
PC3 contribution rate: 9.02%
Cumulative contribution of first 3 PCs: 87.71%
```

结论：从以下几个角度分析是否需要考虑第 3 个主成分：

- **特征值准则**：$\lambda_3 = 0.6312 < 1$，按 Kaiser 准则应保留特征值大于 1 的主成分，故 PC3 不必保留.
- **累积贡献率准则**：前 2 个主成分的累积贡献率已达 78.69%，接近常用的 80% 阈值；加入 PC3 后累积贡献率为 87.71%，提升 9.02%.
- **碎石图准则**：从碎石图来看，在 PC2 之后曲线出现明显的"拐点"，PC3 之后的下降趋势相对平缓，说明 PC2 已经是信息量递减的分界点.
- **综合判断**：若分析目的是保留 80% 以上的信息量，则有必要考虑第 3 个主成分；若对 78.69% 的解释率满意，且更注重简洁性，则前 2 个主成分已足够. 在实际应用中，由于 PC3 的贡献率仅为 9.02%，且特征值小于 1，**通常不必考虑第 3 个主成分**.

2. 乳腺癌数据集 (Breast Cancer Wisconsin Data.csv) 的主成分分析.

数据集 (Breast Cancer Wisconsin Data.csv) 来自 Wisconsin 大学附属医院，由 William H. Wolberg 博士提供. 数据集由 11 个变量的 699 个观测值组成，数据集当中各个变量的定义见表 **??**. 我们用其中的 9 个变量 $(X_2, \ldots, X_{10})$ 的数据来作主成分分析.

(a) [2 分] 读入数据集"Breast Cancer Wisconsin Data.csv"，根据上述定义对每一个变量进行命名 (建议用英文单词或字母缩写). 检查所有变量的类型，最后一个变量 (X_{11}) 是分类变量，将它的属性转变为因子. 其余变量均为数值型，若读入的数据集中有变量非数值型，将它转变为数值型.

【解】

原始 CSV 文件无表头，读入后手动命名. BareNuclei 列因含有"?"字符而被读为字符型，其余数值列均为整型. 将 Class 转为因子（2=Benign，4=Malignant），BareNuclei 暂时保留字符型（含缺失值，在下一问处理后再转换）.

```
cancer <- read.csv("Breast Cancer Wisconsin Data.csv",
                   header = FALSE, stringsAsFactors = FALSE)
colnames(cancer) <- c("SampleCode","ClumpThickness",
```

表 1.2: 乳腺癌数据集的变量含义.

变量	Definition	中文含义
X_1	Sample code number	样本代码编号
X_2	Clump Thickness	肿块厚度
X_3	Uniformity of Cell Size	细胞大小的一致性
X_4	Uniformity of Cell Shape	细胞形状的一致性
X_5	Marginal Adhesion	边缘黏附(用于描述细胞边缘与周围组织的黏附程度)
X_6	Single Epithelial Cell Size	单个上皮细胞大小
X_7	Bare Nuclei	裸露的细胞核
X_8	Bland Chromatin	良性染色质
X_9	Normal Nucleoli	正常核仁
X_{10}	Mitoses	有丝分裂
X_{11}	Class	分类 (2 表示良性，4 表示恶性)

```
                    "CellSizeUniform","CellShapeUniform",
                    "MarginalAdhesion","SingleEpiCellSize",
                    "BareNuclei","BlandChromatin",
                    "NormalNucleoli","Mitoses","Class")
# check variable types
sapply(cancer, class)

# convert Class to factor
cancer$Class <- factor(cancer$Class, levels = c(2, 4),
                       labels = c("Benign", "Malignant"))
table(cancer$Class)
```

输出结果：

```
> sapply(cancer, class)
    SampleCode ClumpThickness CellSizeUniform CellShapeUniform
      "integer"      "integer"       "integer"        "integer"
MarginalAdhesion SingleEpiCellSize    BareNuclei BlandChromatin
      "integer"         "integer"    "character"      "integer"
  NormalNucleoli          Mitoses           Class
      "integer"        "integer"       "integer"
```

```
> table(cancer$Class)

   Benign Malignant
      458       241
Note: Class is now a factor; statistics include missing rows
```

共 699 个观测，其中良性（Benign）458 例，恶性（Malignant）241 例，另有 16 个含缺失值的观测（在下一问中处理）. BareNuclei 因含"?"被读为字符型，其余数值变量均为整型.

(b) [2 分] 数据集当中有 16 个数据含有单一缺失值，这些缺失值在原数据集中用 "?" 来表示. 找到含有缺失数据的观测值，将它们从数据集当中剔除. 从剔除缺失数据的数据集当中提取变量 $(X_2, \ldots, X_{10})$ 的数据子集，它就是我们要作主成分分析的对象.

【解】

缺失值以"?"表示，用 `apply()` 检查每行是否含有"?"，找到 16 行并将其剔除，剩余 683 行. 之后提取 X_2–X_{10} 并将 BareNuclei 转为数值型.

```
# find rows with missing values (containing "?")
missing_rows <- which(apply(cancer, 1, function(x) any(x == "?")))
cat("Rows with missing values:", missing_rows, "\n")
cat("Number of missing rows:", length(missing_rows), "\n")

# remove rows with missing values
cancer_clean <- cancer[-missing_rows, ]
dim(cancer_clean)    # 683 x 11

# extract X2-X10 and convert to numeric
X2 <- cancer_clean[, c("ClumpThickness","CellSizeUniform",
                       "CellShapeUniform","MarginalAdhesion",
                       "SingleEpiCellSize","BareNuclei",
                       "BlandChromatin","NormalNucleoli","Mitoses")]
X2$BareNuclei <- as.numeric(X2$BareNuclei)
dim(X2)    # 683 x 9
head(X2)
```

输出结果：

```
> cat("Rows with missing values:", missing_rows, "\n")
Rows with missing values: 24 41 140 146 159 165 236 250 276 293 295 298 316 322 412 618

> cat("Number of missing rows:", length(missing_rows), "\n")
Number of missing rows: 16

> dim(cancer_clean)
```

```
[1] 683  11

> dim(X2)
[1] 683   9

> head(X2)
  ClumpThickness CellSizeUniform CellShapeUniform MarginalAdhesion
1              5               1                1                1
2              5               4                4                5
3              3               1                1                1
4              6               8                8                1
5              4               1                1                3
6              8              10               10                8
  SingleEpiCellSize BareNuclei BlandChromatin NormalNucleoli Mitoses
1                 2          1              3              1       1
2                 7         10              3              2       1
3                 2          2              3              1       1
4                 3          4              3              7       1
5                 2          1              3              1       1
6                 7         10              9              7       1
```

共找到 16 行含缺失值（第 24、41、140 等行），剔除后剩余 683 个完整观测，提取的数据子集 $\mathscr{X}$ 维度为 683×9.

(c) [2 分] 将得到的数据集进行标准化，计算相关矩阵并给出结果.

【解】

对数据集 $\mathscr{X}$（683 行 9 列）用 `scale()` 进行标准化，再计算相关矩阵（等价于标准化数据的协方差矩阵）.

```
Y2 <- scale(X2)           # standardize
R2 <- cor(X2)             # correlation matrix
round(R2, 4)
```

输出结果（相关矩阵 $\mathscr{R}$）：

```
> round(R2, 4)
                 ClumpTh CellSize CellShape MargAdh EpiCell BareNuc BlandChr NormNuc
    Mitoses
ClumpThickness    1.0000   0.6425    0.6535  0.4878  0.5236  0.5931   0.5537  0.5341
    0.3510
CellSizeUniform   0.6425   1.0000    0.9072  0.7070  0.7535  0.6917   0.7556  0.7193
    0.4608
```

```
CellShapeUniform   0.6535   0.9072    1.0000   0.6859   0.7225   0.7139   0.7353   0.7180
    0.4413
MarginalAdhesion   0.4878   0.7070    0.6859   1.0000   0.5945   0.6706   0.6686   0.6031
    0.4189
SingleEpiCellSize  0.5236   0.7535    0.7225   0.5945   1.0000   0.5857   0.6181   0.6289
    0.4806
BareNuclei         0.5931   0.6917    0.7139   0.6706   0.5857   1.0000   0.6806   0.5843
    0.3392
BlandChromatin     0.5537   0.7556    0.7353   0.6686   0.6181   0.6806   1.0000   0.6656
    0.3460
NormalNucleoli     0.5341   0.7193    0.7180   0.6031   0.6289   0.5843   0.6656   1.0000
    0.4338
Mitoses            0.3510   0.4608    0.4413   0.4189   0.4806   0.3392   0.3460   0.4338
    1.0000
```

由相关矩阵可见，CellSizeUniform 与 CellShapeUniform 的相关系数高达 0.9072，表明细胞大小与形状的一致性高度相关. 大多数变量之间的相关系数在 0.4–0.8 之间，整体呈较强的正相关，说明这 9 个变量具有较高的共性，适合进行主成分分析.

(d) [2 分] 作相关矩阵的谱分解，给出谱分解的结果并作验证运算.

【解】

对相关矩阵 $\mathscr{R}_2$ 作谱分解 $\mathscr{R}_2 = \Gamma\Lambda\Gamma^{\mathrm{T}}$，其中 Λ 为特征值对角矩阵，Γ 为正交特征向量矩阵.

```r
eig2   <- eigen(R2)
Lambda2 <- diag(eig2$values)
Gamma2  <- eig2$vectors
rownames(Gamma2) <- colnames(X2)
colnames(Gamma2) <- paste0("PC", 1:9)

round(eig2$values, 6)    # eigenvalues
round(Gamma2, 6)         # eigenvectors

# verify
R2_check <- Gamma2 %*% Lambda2 %*% t(Gamma2)
max(abs(R2_check - R2))
```

谱分解结果：

特征值（由大到小）：

$$\lambda_1 = 5.8995,\ \ \lambda_2 = 0.7759,\ \ \lambda_3 = 0.5393,\ \ \lambda_4 = 0.4596,\ \ \lambda_5 = 0.3803$$

$$\lambda_6 = 0.3019,\ \ \lambda_7 = 0.2944,\ \ \lambda_8 = 0.2607,\ \ \lambda_9 = 0.0884$$

```
> round(eig2$values, 6)
[1] 5.899499 0.775947 0.539252 0.459627 0.380276 0.301876 0.294403 0.260736 0.088383

> round(Gamma2, 6)
                     PC1       PC2       PC3       PC4       PC5       PC6       PC7
      PC8       PC9
ClumpThickness   0.302063  0.140801  0.866372  0.107828 -0.080321  0.242518 -0.008516
    0.247707 -0.002747
CellSizeUniform  0.380793  0.046640 -0.019938 -0.204255  0.145653  0.139032 -0.205434
    -0.436300 -0.733211
CellShapeUniform 0.377583  0.082422  0.033511 -0.175866  0.108392  0.074527 -0.127209
    -0.582727  0.667481
MarginalAdhesion 0.332724  0.052094 -0.412647  0.493173  0.019569  0.654629  0.123830
    0.163434  0.046019
SingleEpiCell    0.336234 -0.164404 -0.087743 -0.427384  0.636693 -0.069309  0.211018
    0.458669  0.066891
BareNuclei       0.335068  0.261261  0.000691  0.498618  0.124773 -0.609221  0.402790
    -0.126653 -0.076510
BlandChromatin   0.345747  0.228077 -0.213072  0.013047 -0.227666 -0.298897 -0.700417
    0.383719  0.062241
NormalNucleoli   0.335591 -0.033966 -0.134248 -0.417113 -0.690210 -0.021518  0.459783
    0.074012 -0.022079
Mitoses          0.230206 -0.905557  0.080492  0.258988 -0.105042 -0.148345 -0.132117
    -0.053537  0.007496

> max(abs(R2_check - R2))
[1] 2.66e-15
```

验证结果：$\Gamma\Lambda\Gamma^{\mathrm{T}}$ 与 $\mathscr{R}_2$ 的最大绝对误差为 2.66×10^{-15}，在机器精度范围内，谱分解正确.

(e) [2 分] 画碎石图，问各个主成分的贡献率是多少？前 2 个主成分的累积贡献率是多少，前 3 个主成分的累积贡献率又是多少.

【解】

```r
lambda2    <- eig2$values
contrib2   <- lambda2 / sum(lambda2)
cum_contr2 <- cumsum(contrib2)
result2    <- data.frame(
  PC       = paste0("PC", 1:9),
  lambda   = round(lambda2, 4),
  rate     = round(contrib2 * 100, 2),
  cum_rate = round(cum_contr2 * 100, 2)
)
```

```r
print(result2)

# scree plot
par(mfrow = c(1, 2))
plot(1:9, lambda2, type = "b", pch = 16, col = "coral",
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot (Breast Cancer Data)")
abline(h = 1, col = "red", lty = 2)

barplot(contrib2 * 100, names.arg = paste0("PC", 1:9),
        col = "coral", ylim = c(0, 75),
        xlab = "Principal Component", ylab = "Contribution Rate (%)",
        main = "Contribution Rate")
lines(seq(0.7, by = 1.2, length.out = 9), cum_contr2 * 100,
      type = "b", col = "red", pch = 16)
abline(h = 80, col = "gray", lty = 2)
```

各主成分贡献率：

```
   PC   lambda   rate cum_rate
1 PC1   5.8995  65.55    65.55
2 PC2   0.7759   8.62    74.17
3 PC3   0.5393   5.99    80.16
4 PC4   0.4596   5.11    85.27
5 PC5   0.3803   4.23    89.50
6 PC6   0.3019   3.35    92.85
7 PC7   0.2944   3.27    96.12
8 PC8   0.2607   2.90    99.02
9 PC9   0.0884   0.98   100.00
```

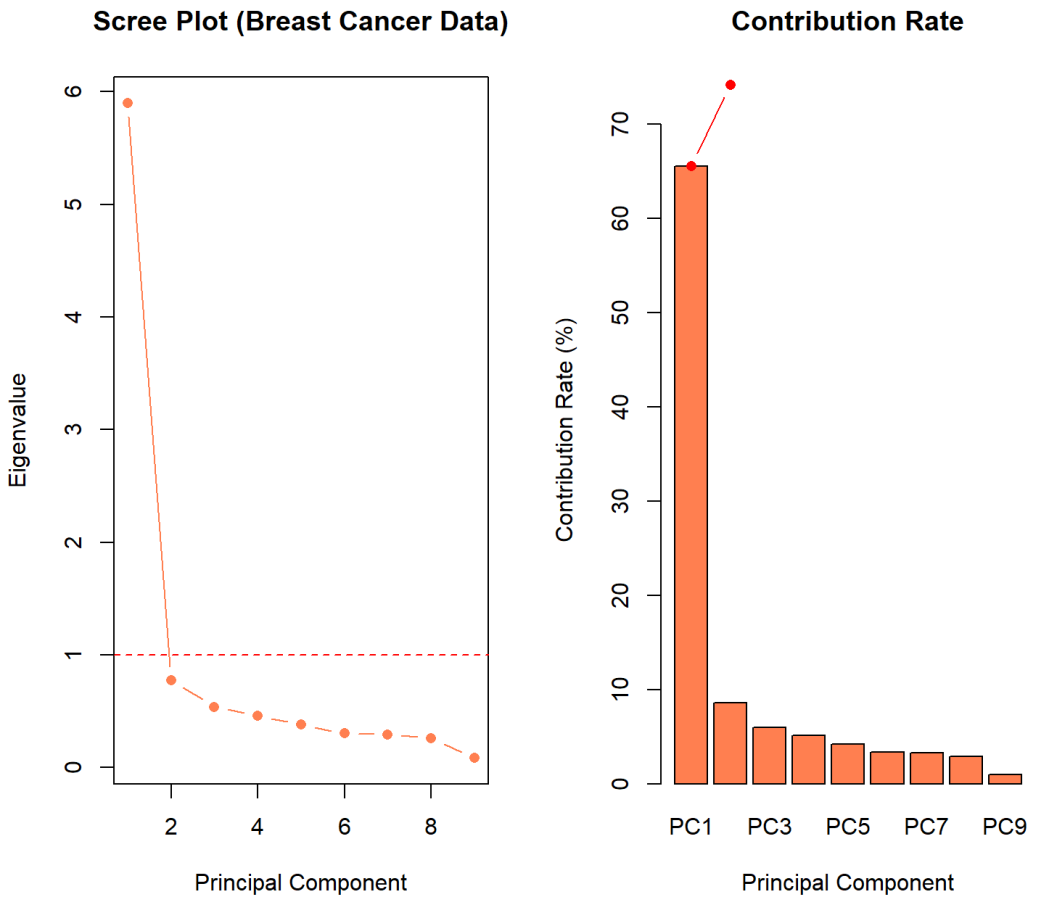

图 1.4: 乳腺癌数据集的碎石图（左）与贡献率图（右）.

由图和表可知：

- 第 1 主成分贡献率高达 65.55%，远超其他主成分，占据主导地位；第 2 主成分贡献率为 8.62%，第 3 主成分贡献率为 5.99%，之后各主成分贡献率均低于 6%.
- **前 2 个主成分的累积贡献率为** 74.17%.
- **前 3 个主成分的累积贡献率为** 80.16%.
- 碎石图在 PC1 之后出现急剧下降，PC1 远大于其他特征值（$\lambda_1 = 5.90 \gg \lambda_2 = 0.78$），说明数据的主要信息集中在第 1 主成分上.

(f) [2 分] 计算初始变量与前 2 个主成分的相关系数并给出结果. 前 2 个主成分对每个变量解释的比例是多少?

【解】

变量 X_i 与第 j 个主成分的相关系数（载荷）为 $\ell_{ij} = \gamma_{ij}\sqrt{\lambda_j}$，前 2 个主成分对变量 X_i 的解释比例（共同度）为 $h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 = \gamma_{i1}^2\lambda_1 + \gamma_{i2}^2\lambda_2$.

```
# loading matrix (correlations between variables and first 2 PCs)
L2 <- matrix(NA, 9, 2)
for (j in 1:2) L2[, j] <- Gamma2[, j] * sqrt(lambda2[j])
```

```r
rownames(L2) <- colnames(X2)
colnames(L2) <- c("PC1", "PC2")
round(L2, 4)

# proportion of each variable explained by first 2 PCs (communality)
h2_2 <- Gamma2[, 1]^2 * lambda2[1] + Gamma2[, 2]^2 * lambda2[2]
names(h2_2) <- colnames(X2)
round(h2_2, 4)
```

载荷（各变量与前 **2** 个主成分的相关系数）：

```
> round(L2, 4)
                    PC1     PC2
ClumpThickness   0.7337  0.1240
CellSizeUniform  0.9249  0.0411
CellShapeUniform 0.9171  0.0726
MarginalAdhesion 0.8081  0.0459
SingleEpiCell    0.8167 -0.1448
BareNuclei       0.8138  0.2301
BlandChromatin   0.8398  0.2009
NormalNucleoli   0.8151 -0.0299
Mitoses          0.5591 -0.7977
```

前 **2** 个主成分对每个变量的解释比例（共同度）：

```
> round(h2_2, 4)
ClumpThickness CellSizeUniform CellShapeUniform MarginalAdhesion SingleEpiCell
        0.5537          0.8571           0.8464           0.6552        0.6879
    BareNuclei  BlandChromatin  NormalNucleoli         Mitoses
        0.7153          0.7456          0.6653         0.9489
```

变量	ℓ_{i1}（PC1 载荷）	ℓ_{i2}（PC2 载荷）	h_i^2（解释比例）
ClumpThickness	0.7337	0.1240	55.37%
CellSizeUniform	0.9249	0.0411	85.71%
CellShapeUniform	0.9171	0.0726	84.64%
MarginalAdhesion	0.8081	0.0459	65.52%
SingleEpiCell	0.8167	−0.1448	68.79%
BareNuclei	0.8138	0.2301	71.53%
BlandChromatin	0.8398	0.2009	74.56%
NormalNucleoli	0.8151	−0.0299	66.53%
Mitoses	0.5591	−0.7977	94.89%

Mitoses（有丝分裂）由前 2 个主成分解释的比例最高（94.89%），其中 PC2 的贡献最为突出. CellSizeUniform 和 CellShapeUniform 由 PC1 解释的比例最高（¿92%），而 ClumpThickness 的共同度最低（55.37%），说明其与前 2 个主成分的关联相对较弱.

(g) [2 分] 作变量在前 2 个主成分平面上的散点图，对结果进行解释.

【解】

以各变量与前 2 个主成分的相关系数（载荷）(ℓ_{i1}, ℓ_{i2}) 为坐标，作变量在 PC1–PC2 平面上的相关系数圆图.

```
var_labels <- c("ClumpTh","CellSize","CellShape","MargAdh",
                "EpiCell","BareNuc","BlandChr","NormNuc","Mitoses")
plot(L2[, 1], L2[, 2],
     xlim = c(-1.2, 1.2), ylim = c(-1.2, 1.2), asp = 1,
     xlab = sprintf("PC1 (%.1f%%)", contrib2[1]*100),
     ylab = sprintf("PC2 (%.1f%%)", contrib2[2]*100),
     main = "Variable Biplot on PC1-PC2 Plane\n(Breast Cancer Data)")
arrows(0, 0, L2[, 1], L2[, 2], length = 0.1, col = 1:9, lwd = 2)
text(L2, var_labels, col = 1:9, pos = 4, cex = 0.85)
theta <- seq(0, 2*pi, length.out = 200)
lines(cos(theta), sin(theta), col = "gray")
abline(h = 0, v = 0, lty = 2)
```

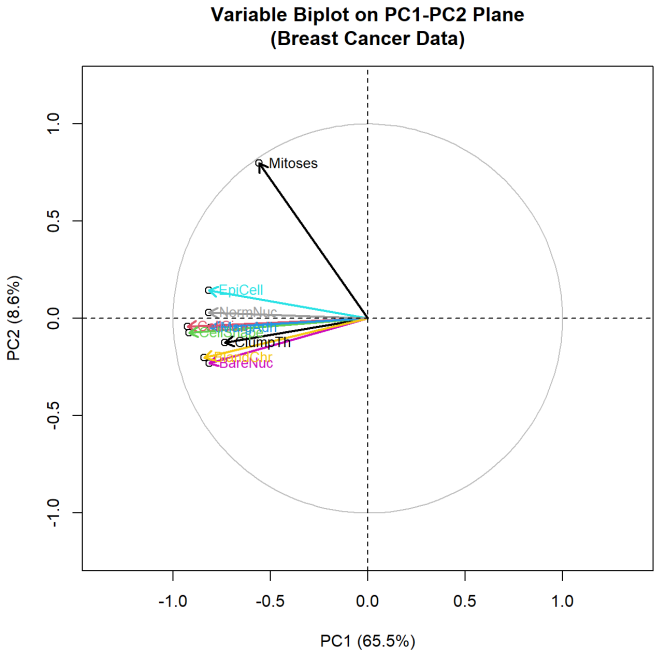

图 1.5: 乳腺癌数据各变量在 PC1–PC2 平面上的圆图.

解释：

- **第一主成分（PC1）**：所有 9 个变量在 PC1 上的载荷均为正（0.56 ∼ 0.92），说明 PC1 代表总体"恶性程度"指标，PC1 得分越高意味着各指标的异常程度越高.
- **第二主成分（PC2）**：Mitoses（有丝分裂）在 PC2 上的载荷为 −0.80，而 BareNuclei 和 BlandChromatin 在 PC2 上载荷为正. 因此 PC2 主要反映了有丝分裂相对于核特征的对比，代表"细胞分裂活跃度"的特殊维度.
- 除 Mitoses 外，其余 8 个变量的箭头方向几乎一致（均指向右侧，PC1 方向），说明它们共同代表细胞异常的综合特征. Mitoses 箭头方向与其余变量存在明显差异，说明有丝分裂具有独特的生物学意义.
- CellSizeUniform 和 CellShapeUniform 的箭头几乎重叠（载荷分别为 0.9249 和 0.9171），进一步验证了两者之间极高的相关性（$r = 0.9072$）.

(h) [2 分] 变量 Class (X_{11}) 是分类变量，分别对应于良性与恶性. 作观测数据在前 2 个主成分平面上的散点图，能否看出良性与恶性的表现存在不同?

【解】

计算各样本的主成分得分，以 PC1、PC2 为坐标，按 Class（良性/恶性）用不同颜色和形状标记，作散点图.

```
# compute principal component scores
scores2 <- Y2 %*% Gamma2


# color by class
class_vec <- cancer_clean$Class
```

```r
col_class <- ifelse(class_vec == "Benign", "steelblue", "red")
pch_class <- ifelse(class_vec == "Benign", 1, 2)

plot(scores2[, 1], scores2[, 2],
     col = col_class, pch = pch_class, cex = 0.7,
     xlab = sprintf("PC1 (%.1f%%)", contrib2[1]*100),
     ylab = sprintf("PC2 (%.1f%%)", contrib2[2]*100),
     main = "Observations on PC1-PC2 Plane\n(Breast Cancer Data, by Class)")
legend("topright",
       legend = c("Benign (2)", "Malignant (4)"),
       col = c("steelblue", "red"), pch = c(1, 2), cex = 0.9)
abline(h = 0, v = 0, lty = 2)
```

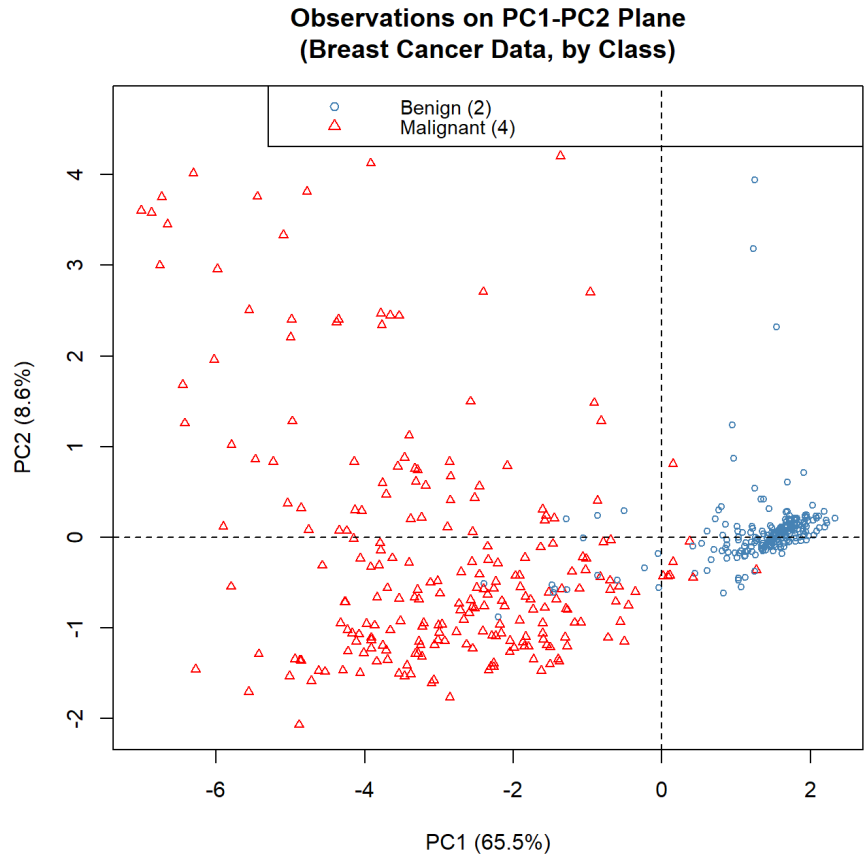

图 1.6: 乳腺癌各样本在 PC1–PC2 平面上的散点图（按分类着色）.

结论：

- 从散点图可以清楚地看出，良性（Benign，蓝色圆点）与恶性（Malignant，红色三角）样本在 PC1–PC2 平面上存在**明显的分离**：
  - 良性样本主要集中在 **PC1 较大（得分较高）**的区域右侧，对应 PC1 轴正方向的聚集区域. 注意 PC1 正方向表示各细胞特征异常程度较低（良性肿瘤特征）.

  - 恶性样本则聚集在 **PC1 较小（得分较低）**的区域左侧，表现出更高的异常指标水平.
- 两类样本在 PC1 方向上的区分最为明显，PC2 方向的区分能力相对较弱.
- 这一结果表明，前 2 个主成分（尤其是 PC1）能够有效捕捉良恶性肿瘤之间的差异，具有良好的判别能力，可为后续的分类分析（如判别分析）提供基础.

(i) [2 分] 是否有必要考虑第 3 个主成分?

【解】

```
cat(sprintf("lambda_3 = %.4f\n", lambda2[3]))
cat(sprintf("PC3 contribution rate: %.2f%%\n", contrib2[3]*100))
cat(sprintf("Cumulative contribution of first 3 PCs: %.2f%%\n", cum_contr2[3]*100))
```

```
lambda_3 = 0.5393
PC3 contribution rate: 5.99%
Cumulative contribution of first 3 PCs: 80.16%
```

结论：从以下几个角度综合判断：

- **特征值准则**：$\lambda_1 = 5.90 \gg 1$，而 $\lambda_2 = 0.78 < 1$，按 Kaiser 准则仅保留 PC1 即可. 若以特征值大于 1 为标准，甚至连 PC2 也不必保留，更无需考虑 $\lambda_3 = 0.54 < 1$ 的 PC3.
- **累积贡献率准则**：前 2 个主成分的累积贡献率已达 74.17%；加入 PC3 后为 80.16%，可达到常用的 80% 阈值. 若以 80% 为目标，则**有必要**考虑第 3 个主成分.
- **碎石图准则**：碎石图在 PC1 处最为陡峭，PC2 之后趋于平坦，”拐点”位于 PC1 和 PC2 之间，进一步说明 PC2 和 PC3 对整体方差的解释贡献有限.
- **应用目的**：从第（8）问的散点图可以看出，PC1 已经能较好地区分良性与恶性，PC2 主要反映有丝分裂的特殊维度（贡献率 8.62%）. 若以分类为主要目标，前 2 个主成分已足够；若追求 80% 的解释率，可考虑加入 PC3（ClumpThickness 方向，PC3 的载荷 $\ell_{3,1} = 0.87$ 最大）.
- **综合判断**：在大多数情况下，由于 $\lambda_3 < 1$ 且 PC3 仅额外解释 5.99% 的方差，**通常不必考虑第 3 个主成分**，以前 2 个主成分进行分析即可满足需求. 但若严格要求累积贡献率达到 80%，则可视情况保留 PC3.