多元统计 8

罗震林 17306071

May 2021

1 7.1

1.1 加载相关包

```
library(ggplot2) # 画图
library(MASS)  # 生成随机数据
library(ggpubr)
# library(ggbiplot)
```

1. ggpubr 是一款基于 ggplot2 的可视化包 ggpubr。主要是使用其中的 ggarange 实现一页多图
2. ggbiplot 是一款 PCA 分析结果可视化的 R 包工具，可以直接采用 ggplot2 来可视化 R 中基础函数 prcomp() PCA 分析的结果。为了更好地理解主成分分析，在这里不使用该包，直接通过 ggplot2 实现

1.2 生成随机数据

```
set.seed(1)
n<-100
sigma <- matrix(c(10,0,0,0,1,0,0,0,1),3,3)
data1 <- as.data.frame(mvrnorm(n,mu=c(0,0,0),sigma))
colnames(data1) <- c("X1","X2","X3")
```

生成样本量为 100 的 3 维正态分布数据样本，其总体可表示为

$$\mathbf{X}=\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}\right)$$

三个随机变量的均值都为 0，其中 X_1 的方差为 10，明显大于另外两个方差为 1 的变量

1.3 使用协方差阵进行主成分分析

```
d1_pca <- princomp(data1)
load1 <- loadings(d1_pca)
```

```
d1_pca$sdev^2    # 特征值
```

```
     Comp.1      Comp.2      Comp.3
8.180742e+00 1.596308e+00 3.330669e-16
```

```
load1[,]      # 对应的特征向量
```

```
      Comp.1     Comp.2       Comp.3
X1 0.9855061  0.1696398  1.871701e-17
X2 0.1199535 -0.6968581  7.071068e-01
X3 0.1199535 -0.6968581 -7.071068e-01
```

1. 特征值计算中，主成分所计算得到为标准差，而特征值为方差，所以需要进行平方计算
2. 对应的特征向量即结果中各主成分所对应的列

1.3.1 碎石图

```
p_scre <- data.frame(PC = paste0("Comp.", 1:3) , Variances = d1_pca$sdev^2)
ggplot(p_scre,aes(x=PC, y=Variances, group=1)) + geom_line() + geom_point(size=2)
```

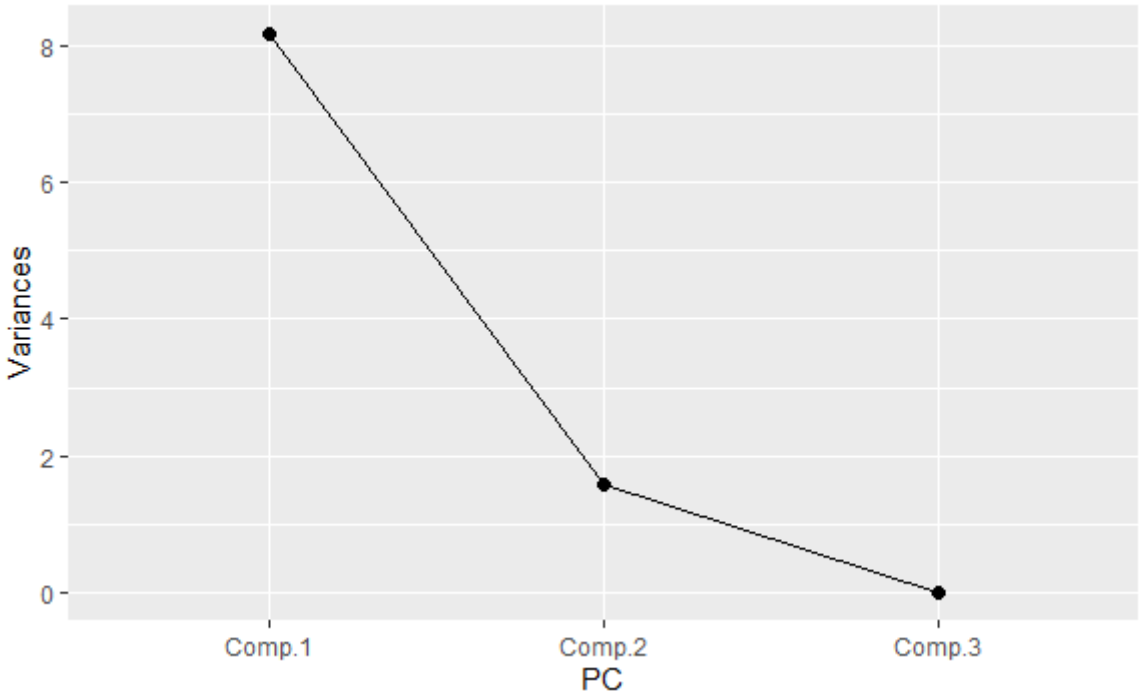

图 1: 碎石图（由协方差阵得到）

图 1.3.2中横坐标为对应三个主成分，纵坐标为主成分的方差，可以看出第 1 主成分方差明显大于另外两个主成分

1.3.2 PC scores

```
pc_score <- as.data.frame(d1_pca$scores)
rownames(pc_score) <- 1:100
head(pc_score)
```

	Comp.1 <dbl>	Comp.2 <dbl>	Comp.3 <dbl>
1	-2.3529734	-0.7735555	-1.110223e-16
2	0.2391921	0.1060821	1.387779e-17
3	-3.0222087	-1.1590962	-5.551115e-17
4	4.7569299	0.2622079	1.110223e-16
5	0.7065224	-0.8172471	-1.665335e-16
6	-2.9756272	2.3913089	4.440892e-16

PC scores 可以直接由主成分得到的结果中提取出来，并且将 100 个样本直接以数字 1 至 100 来命名，这里只展示前 6 个样本的 PC scores

- 作图

```
library(scales) # 格式化坐标轴文本的包
pc_score <- as.data.frame(d1_pca$scores)
rownames(pc_score) <- 1:100
head(pc_score)
# ggbiplot可以直接实现

# 第1和第2主成分
p1 <- ggplot(pc_score, aes(x=Comp.1,y=Comp.2))+geom_point()

# 第1和第3主成分
p2 <- ggplot(pc_score, aes(x=Comp.1,y=Comp.3))+geom_point() +
    scale_y_continuous(labels = scientific)

# 第2和第3主成分
p3 <- ggplot(pc_score, aes(x=Comp.2,y=Comp.3))+geom_point() +
    scale_y_continuous(labels = scientific)

ggarrange(p1, p2, p3,
          ncol = 3, nrow = 1,align = "hv")
```

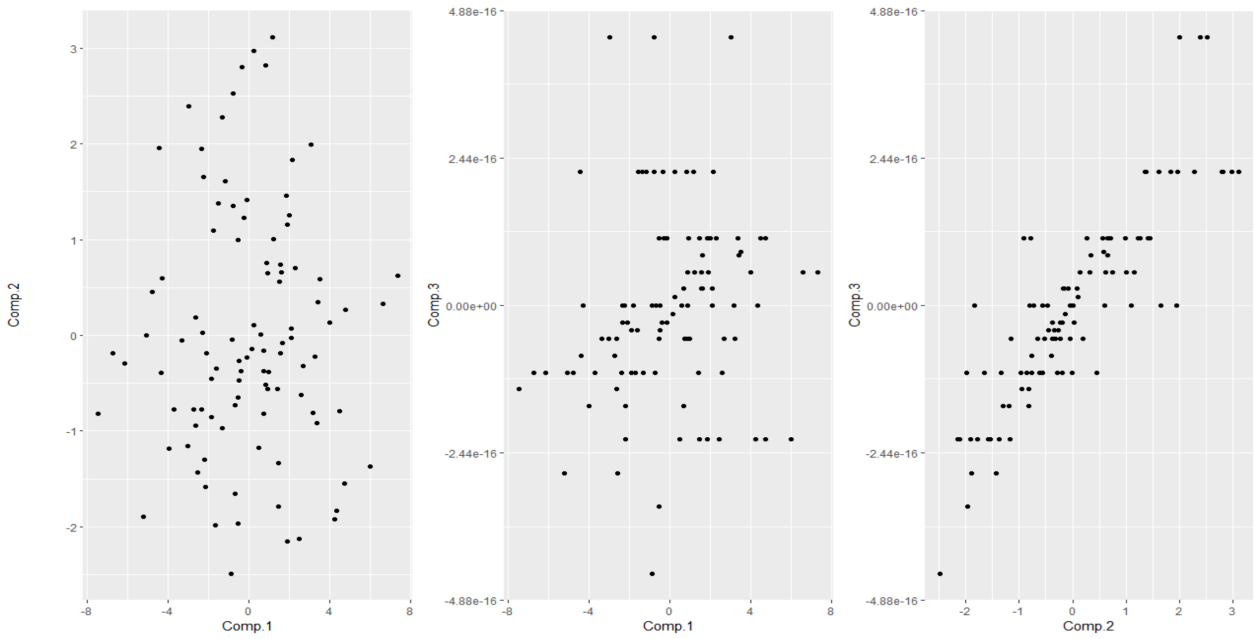

图 2: 各对主成分下的 PC scores

1.4 使用相关矩阵进行主成分分析

```
d2_pca <- princomp(data1,cor = TRUE)
load2 <- loadings(d2_pca)
# 特征值和特征向量
d2_pca$sdev^2    # 特征值
```

```
   Comp.1    Comp.2    Comp.3
1.0531676 0.9993515 0.9474810
```

```
load2[,]      # 对应的特征向量
```

```
       Comp.1      Comp.2     Comp.3
X1  0.2560712  0.93749561  0.2356470
X2  0.7048632 -0.01426598 -0.7091998
X3 -0.6615100  0.34770457 -0.6644592
```

1.4.1 碎石图

```
p2_scre <- data.frame(PC = paste0("Comp.", 1:3) , Variances = d2_pca$sdev^2)
ggplot(p2_scre,aes(x=PC, y=Variances, group=1)) + geom_line() + geom_point(size=2) + ylim(0,2)
```

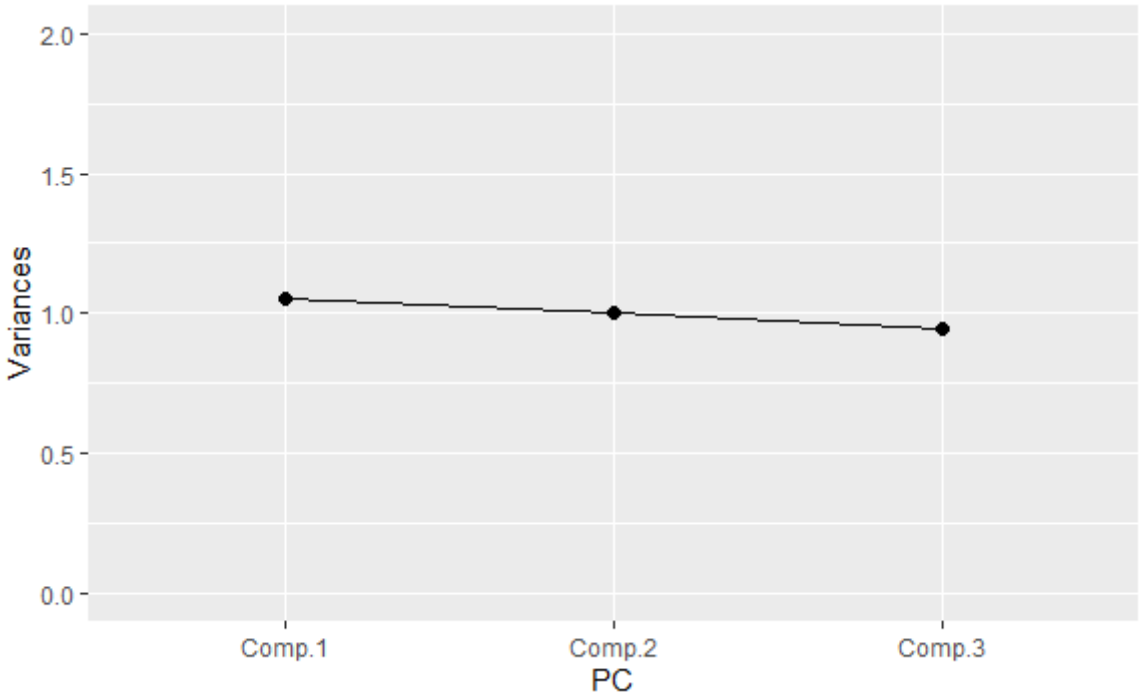

图 3: 碎石图（由相关矩阵得到）

1.4.2 PC scores

```
pc2_score <- as.data.frame(d2_pca$scores)
rownames(pc2_score) <- 1:100
head(pc2_score)
```

	Comp.1 <dbl>	Comp.2 <dbl>	Comp.3 <dbl>
1	0.4537362	-0.98917380	-0.04944156
2	1.1024056	0.08457649	-1.17946065
3	1.4017762	-1.33092015	-0.71329715
4	0.0429925	1.63568342	0.50389286
5	-1.0942782	0.03850960	2.08351073
6	0.1712690	-0.35057500	-3.20434614

```
# 第1和第2主成分
pp1 <- ggplot(pc2_score, aes(x=Comp.1,y=Comp.2))+geom_point()
# 第1和第3主成分
pp2 <- ggplot(pc2_score, aes(x=Comp.1,y=Comp.3))+geom_point()
# 第2和第3主成分
pp3 <- ggplot(pc2_score, aes(x=Comp.2,y=Comp.3))+geom_point()
ggarrange(pp1, pp2, pp3, ncol = 3, nrow = 1,align = "hv")
```

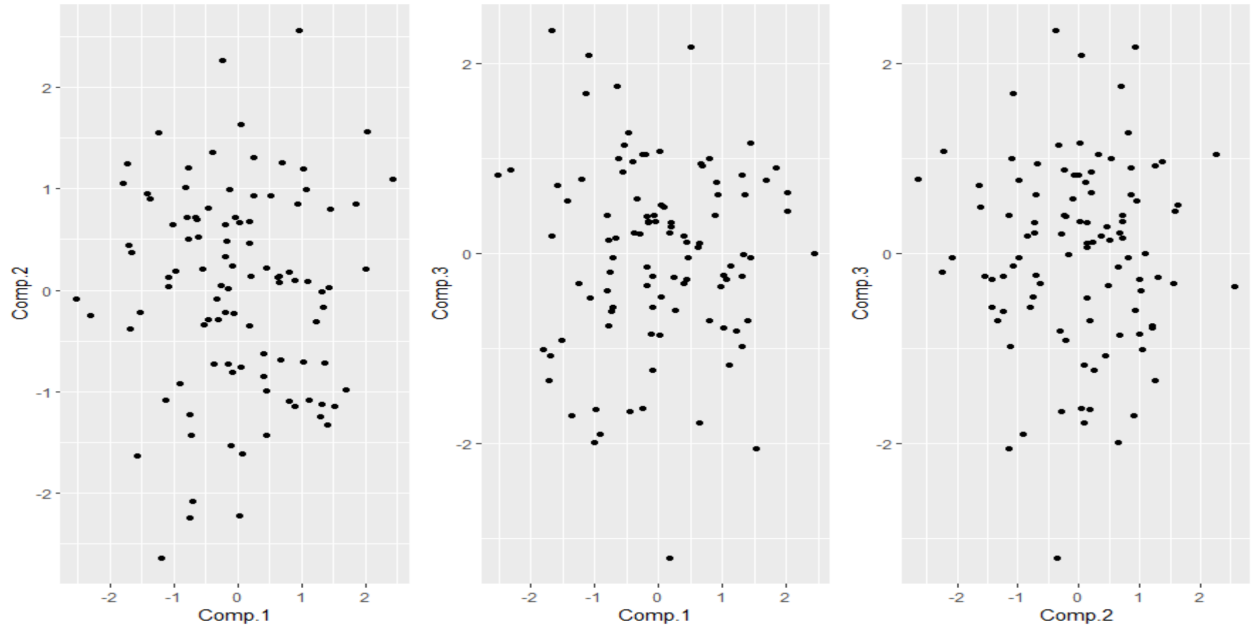

图 4: PC scores

1.5 总结

1. 使用协方差阵进行主成分分析得到的主成分中，第 1 主成分的贡献率远大于第 2 和第 3 主成分，且在方差最大的变量 X_1 上的载荷最大，说明变量 X_1 占了原始变量的绝大部分信息
2. 从图 2中可以看出，第 3 主成分的贡献率基本上可以忽略不计，所以在第 3 主成分作纵坐标的对比图像中，会呈现一定的相关性
3. 使用相关矩阵进行主成分分析得到的主成分中，3 个主成分的贡献率都很接近。主成分 1 在变量 X_2 和 X_3 上的载荷较大，主成分 2 在变量 X_1 上的载荷较大，主成分 3 在变量 X_2 和 X_3 上的载荷较大，说明降维效果并不好
4. 从图 4中很难看出一定规律，因为各主成分贡献率相近，所以 PC scores 显得较为随机

2 7.4

2.1 读取数据

```
rm(list = ls())
library(ggbiplot)
pendigit <- read.table("pendigits.tes.txt",sep = ",")
```

pendigit 数据集中，前 16 列是我们需要进行分析的数据，第 17 列表示实际中该样本的数字

2.2 计算方差

```
#apply(pendigit[, 1:16], 2, var) 该列为计算方差
apply(pendigit[, 1:16], 2, sd) # 计算标准差
```

```
      V1       V2       V3       V4       V5       V6       V7
35.99529 14.73708 26.48636 17.96088 32.13472 26.13041 30.35197
      V8       V9      V10      V11      V12      V13      V14
28.22755 35.08918 26.78593 37.55554 26.18340 21.74573 32.93756
     V15      V16
42.23408 35.70502
```

该 16 个变量的标准差很接近，大都处于 20 40 之间，但是平方计算方差后，差距就会显得很大，所以这里只展示计算的标准差结果，可以说明这 16 个变量的方差很近似

2.3 主成分分析

```
pca1 <- princomp(pendigit[,1:16]) # 默认使用协方差阵进行计算
summary(pca1)
ggscreeplot(pca1) # 碎石图，ggbiplot包中的函数
```

```
Importance of components:
                           Comp.1     Comp.2     Comp.3
Standard deviation     64.7632306 61.2259782 47.5402576
Proportion of Variance  0.2867429  0.2562755  0.1545108
Cumulative Proportion   0.2867429  0.5430184  0.6975292
                           Comp.4     Comp.5      Comp.6
Standard deviation     35.74096212 30.7869010 24.92736551
Proportion of Variance  0.08733094  0.0647989  0.04248038
Cumulative Proportion   0.78486012  0.8496590  0.89213941
                           Comp.7      Comp.8     Comp.9
Standard deviation     21.19110867 18.50226023 16.0831517
Proportion of Variance  0.03070033  0.02340374  0.0176839
Cumulative Proportion   0.92283973  0.94624347  0.9639274
```

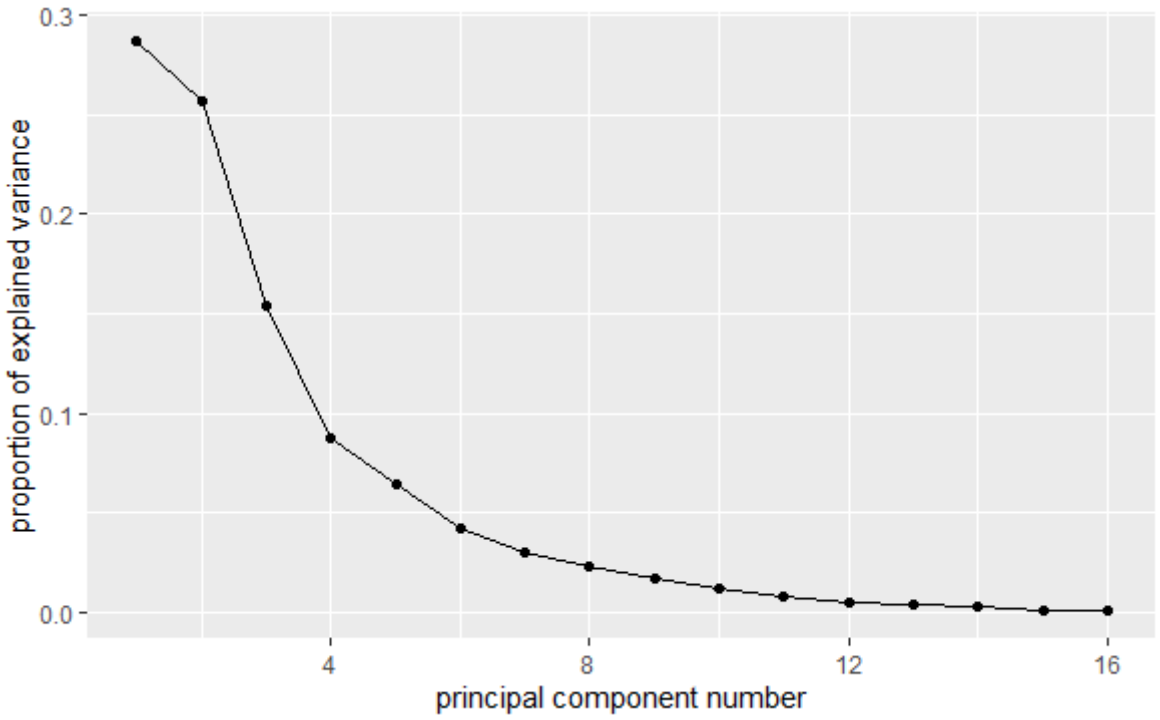

图 5: 碎石图

从输出结果和碎石图中可以知道：5 个主成分能够解释 80% 的总方差；7 个主成分能够解释 90% 的总方差。(注：这里使用 ggscreeplot 可以从纵坐标直接看出各主成分的方差贡献率)

2.4 前 3 个主成分散点图

```
# 数据整合
pc_score <- data.frame(pca1$scores[,1:3], factor(pendigit$V17))
rownames(pc_score) <- 1:3498
colnames(pc_score)[4] <- "number"
head(pc_score)

# 三个主成分的三维情况
library(deSolve)
library(gg3D)   # 画3D图需要的包
theta=-30 #方位角的度数
phi=0 # 渐近线
p0 <- ggplot(pc_score, aes(x=Comp.1, y=Comp.2, z=Comp.3 , colour= number)) +
        axes_3D(theta=theta, phi=phi) +
        stat_3D(theta=theta, phi=phi) +
        labs_3D(theta=theta, phi=phi,
                hjust=c(2,0,1),vjust=c(3,3.5,0),
                size=5,
                labs = c("Comp.1", "Comp.2", "Comp.3")) +
  theme_void()
```

```
# 第1和第2主成分
p1 <- ggplot(pc_score, aes(x=Comp.1, y=Comp.2, colour= number))+geom_point()

# 第1和第3主成分
p2 <- ggplot(pc_score, aes(x=Comp.1, y=Comp.3, colour= number))+geom_point()

# 第2和第3主成分
p3 <- ggplot(pc_score, aes(x=Comp.2, y=Comp.3, colour= number))+geom_point()

library(ggpubr)
ggarrange(p0, p1, p2, p3,
          ncol = 2, nrow =2,align = "hv",
          common.legend = TRUE,
          legend="bottom")
```

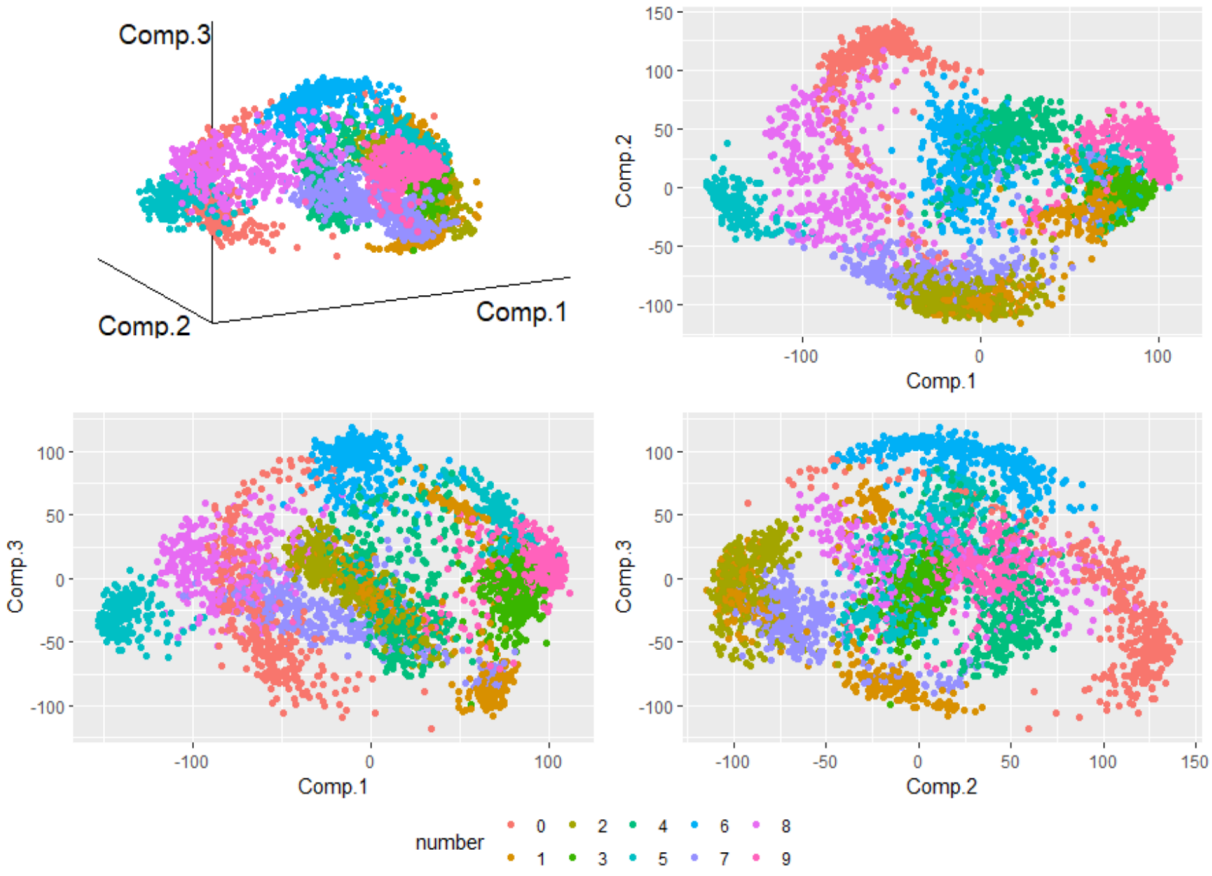

图 6: 各维度下 PC scores

图 6中第 1 个三维图为前三个主成分的散点图，其余三个图分别为前三个主成分两两比较的散点图。可以看出不同数字各自聚集在一定的范围内，且与使用相关矩阵进行主成分分析的结果类似。因为在方差近似的情况下，各变量标准化后的差异不会太明显，所以两种方法得到的结果类似

3 7.5

```
p_scre <- data.frame(PC = c(1:16) , Variances = pca1$sdev^2)
ggplot(p_scre,aes(x=PC, y=Variances, group=1)) + geom_line() + geom_point(size=2)
```

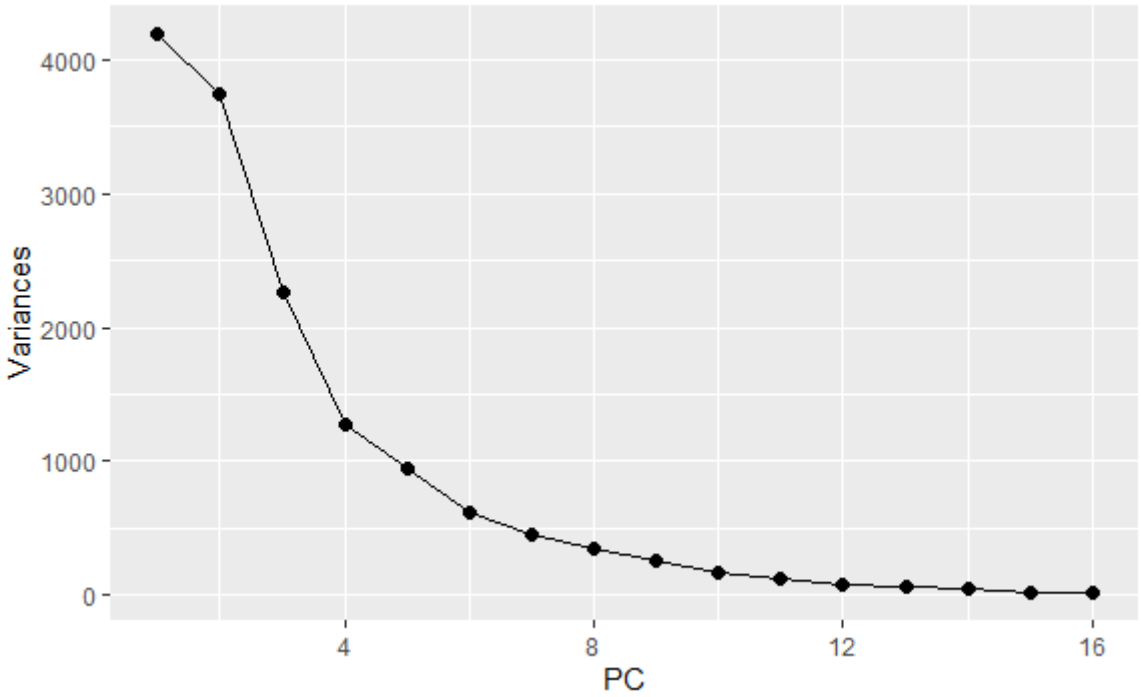

图 7: 碎石图

由碎石图可以看到，曲线在第 4 主成分时存在一个较大的转折，所以选择到其之后的第 1 个即前 5 个主成分较为合适，且前 5 个主成分能够解释 85% 的总方差，已经可以有较好的解释结果

4 7.9

```
library(ggplot2)
library(ggpubr)

iris_pca <- princomp(iris[,1:4])
pc_score <- data.frame(iris_pca$scores,factor(iris$Species))
rownames(pc_score) <- 1:150
colnames(pc_score)[5] <- "Species"
head(pc_score)

# 第1和第2主成分
p1 <- ggplot(pc_score, aes(x=Comp.1, y=Comp.2, colour = Species))+geom_point()
```

```
# 第1和第3主成分
p2 <- ggplot(pc_score, aes(x=Comp.1, y=Comp.3, colour = Species)) + geom_point()

# 第2和第3主成分
p3 <- ggplot(pc_score, aes(x=Comp.2, y=Comp.3, colour = Species))+geom_point()

ggarrange(p1, p2, p3,
          ncol = 3, nrow = 1, align = "hv",
          common.legend = TRUE,
          legend="bottom")
```

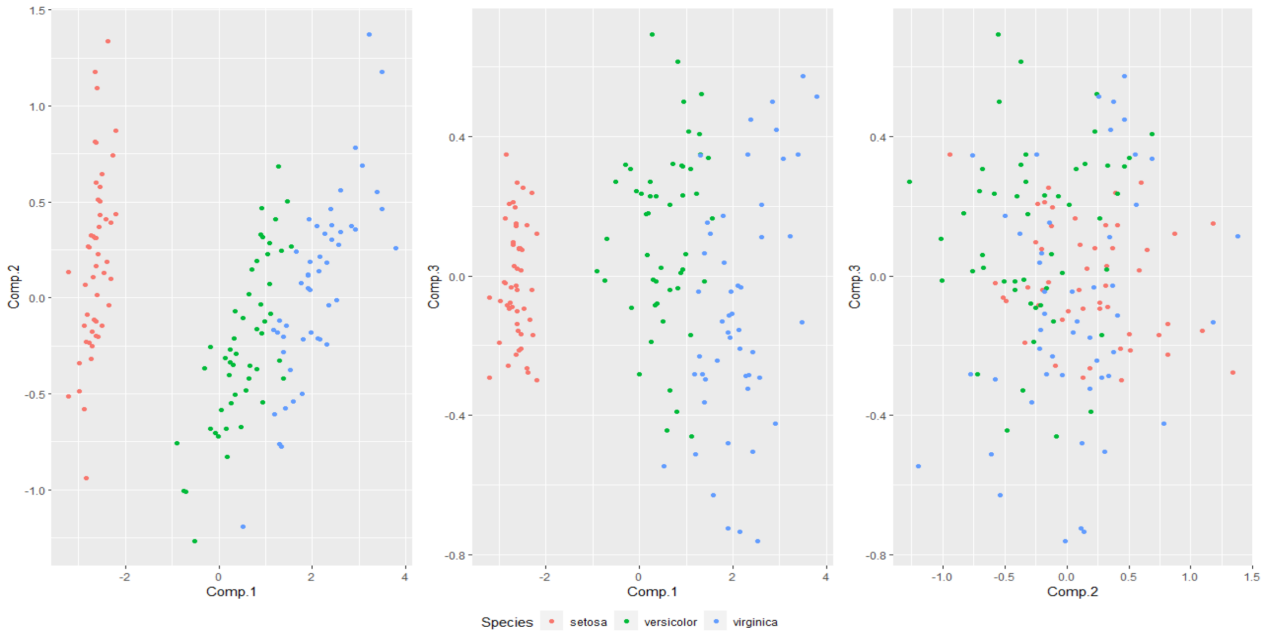

图 8: PC scores

从图 8中可以看出，通过第 1 主成分和第 2 主成分，第 1 和第 3 主成分都能够很好地将三种花形成不同的聚集，从而有较好的分类结果，且前者效果优于后者；而使用第 2 主成分和第 3 主成分的效果则很差，无法进行有效分类

5 7.12

由已知, 可以计算得到

$$\hat{\Sigma}_{XX}^{-1} = \hat{\Sigma}_{YY}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

$$\hat{\Sigma}_{XX}^{-\frac{1}{2}} = \hat{\Sigma}_{YY}^{-\frac{1}{2}} = \frac{1}{2}\begin{pmatrix} \frac{1}{\sqrt{1+\rho}} + \frac{1}{\sqrt{1-\rho}} & \frac{1}{\sqrt{1+\rho}} - \frac{1}{\sqrt{1-\rho}} \\ \frac{1}{\sqrt{1+\rho}} - \frac{1}{\sqrt{1-\rho}} & \frac{1}{\sqrt{1+\rho}} + \frac{1}{\sqrt{1-\rho}} \end{pmatrix}$$

由于该题中 $\Sigma_{XX}=\Sigma_{YY}$，所以 $\hat{R}=\hat{R}^*$, 可以计算得到

$$\hat{R}=\hat{R}^*=\frac{2\rho^2}{(1+\rho)^2}\begin{pmatrix}1&1\\1&1\end{pmatrix}$$

其特征值和对应得特征向量分别为

$$\hat{\lambda}_1^2=\frac{4\rho^2}{(1+\rho)^2},\quad \widehat{\mathbf{v}}_1^T=\frac{1}{\sqrt{2}}(1,1)$$
$$\hat{\lambda}_2^2=0,\quad \widehat{\mathbf{v}}_2^T=\frac{1}{\sqrt{2}}(-1,1)$$

从而可以计算得到

$$\begin{aligned}\hat{G}^{(2)}&=\begin{pmatrix}\hat{\lambda}_1\hat{\lambda}_1^{-1}\widehat{\mathbf{v}}_1^T\\ \hat{\lambda}_2\hat{\lambda}_2^{-1}\widehat{\mathbf{v}}_2^T\end{pmatrix}\hat{\Sigma}_{XX}^{-\frac{1}{2}}\\&=\frac{1}{\sqrt{2}}\begin{pmatrix}1&1\\0&0\end{pmatrix}\frac{1}{2}\begin{pmatrix}\frac{1}{\sqrt{1+\rho}}+\frac{1}{\sqrt{1-\rho}}&\frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}}\\ \frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}}&\frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}}\end{pmatrix}\\&=\frac{1}{\sqrt{2(1+\rho)}}\begin{pmatrix}1&1\\0&0\end{pmatrix}\\ \hat{H}^{(2)}&=\frac{1}{\sqrt{2}}\begin{pmatrix}\widehat{\mathbf{v}}_{\mathbf{1}}^T\\ \widehat{\mathbf{v}}_{\mathbf{2}}^T\end{pmatrix}\hat{\Sigma}_{YY}^{-\frac{1}{2}}\\&=\frac{1}{\sqrt{2}}\begin{pmatrix}1&1\\-1&1\end{pmatrix}\frac{1}{2}\begin{pmatrix}\frac{1}{\sqrt{1+\rho}}+\frac{1}{\sqrt{1-\rho}}&\frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}}\\ \frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}}&\frac{1}{\sqrt{1+\rho}}-\frac{1}{\sqrt{1-\rho}}\end{pmatrix}\\&=\frac{1}{\sqrt{2}}\begin{pmatrix}\frac{1}{\sqrt{1+\rho}}&\frac{1}{\sqrt{1+\rho}}\\ -\frac{1}{\sqrt{1-\rho}}&\frac{1}{\sqrt{1-\rho}}\end{pmatrix}\end{aligned}$$

容易可以看到第 2 对典型相关变量包含零向量，所以只考虑第 1 对典型相关变量。记 $\hat{G}^{(2)},\hat{H}^{(2)}$ 的第 1 行分别为 $\widehat{\mathbf{g}}_1^T,\widehat{\mathbf{h}}_1^T$, 则第 1 对典型相关变量为

$$\hat{\xi}_1=\widehat{\mathbf{g}}_1^T\mathbf{X}=\frac{1}{\sqrt{2(1+\rho)}}(X_1+X_2)$$
$$\hat{\omega}_1=\widehat{\mathbf{h}}_1^T\mathbf{Y}=\frac{1}{\sqrt{2(1+\rho)}}(Y_1+Y_2)$$

以及它们的典型相关系数为

$$\begin{aligned}\hat{\rho}_1&=\frac{\widehat{\mathbf{g}}_1^T\hat{\Sigma}_{XY}\widehat{\mathbf{h}}_1}{(\widehat{\mathbf{g}}_1^T\hat{\Sigma}_{XX}\widehat{\mathbf{g}}_1)^{\frac{1}{2}}(\widehat{\mathbf{h}}_1^T\hat{\Sigma}_{YY}\widehat{\mathbf{h}}_1)^{\frac{1}{2}}}\\&=\frac{2\rho}{1+\rho}\end{aligned}$$