Lab7: 主成分分析

1. 内容: 练习主成分方法的使用
2. 作业提交: 完成后面的作业, 现场演示给助教并解释结果.

1 主成分方法的推断

```
# Pricipal Components Analysis
# entering raw data and extracting PCs
# from the correlation matrix
mydata<-iris[,1:4]
fit1 <- princomp(mydata, cor=TRUE)  #variances are computed with the divisor N
fit2 <- prcomp(mydata,cor=TRUE) #variances are computed with the usual divisor N-1
summary(fit1) # print variance accounted for
loadings(fit1) # pc loadings
plot(fit1,type="lines") # scree plot
fit1$scores # the principal components
biplot(fit1)
```

练习 1. 完成课本*8.28*题.

2 主成分方法的应用

1. 维数缩减方面的应用

```
install.packages("plsgenomics")
library(plsgenomics)
data(leukemia)
?leukemia
names(leukemia)
```

练习 2. 使用*leukemia*数据, 该数据集为38×3051维, 变量个数远远大于样本量。试使用PCA降低维数并在低维空间图示化表达原始数据, 两类癌症病人是否明显分开? 使用SVD方法计算矩阵特征向量和特征根, 以提高计算效率.

2. 使用主成分进行预测

```
###pca - 使用前4列
data(iris)
dat <- as.matrix(iris[,-5])
pca <- prcomp(dat, retx=TRUE, center=TRUE, scale=TRUE)

### 假设多元正态成立，计算各组参数
setosa.mean <- apply(iris[iris$Species=="setosa",-5], 2, mean)
    setosa.cov <- cov(iris[iris$Species=="setosa",-5])

versicolor.mean <- apply(iris[iris$Species=="versicolor",-5], 2, mean)
    versicolor.cov <- cov(iris[iris$Species=="versicolor",-5])

virginica.mean <- apply(iris[iris$Species=="virginica",-5], 2, mean)
    virginica.cov <- cov(iris[iris$Species=="virginica",-5])

#模拟产生一组新的观测（使用上面假设的多元正态分布）

require(MASS)
set.seed(1)
n <- 30
new.setosa <- mvrnorm(n, setosa.mean, setosa.cov)
new.versicolor <- mvrnorm(n, versicolor.mean, versicolor.cov)
new.virginica <- mvrnorm(n, virginica.mean, virginica.cov)

### 使用 predict.prcomp 函数预测新数据的主成分得分

pred.setosa <- predict(pca, new.setosa)
pred.versicolor <- predict(pca, new.versicolor)
pred.virginica <- predict(pca, new.virginica)

### 图示化结果
SPP <- iris$Species
COLOR <- c(2:4)
PCH <- c(1,16)

pc <- c(1,2)
plot(pca$x[,pc[1]], pca$x[,pc[2]], col=COLOR[SPP], cex=PCH[1],
  xlab=paste0("PC ", pc[1], " (", round(pca$sdev[pc[1]]/sum(pca$sdev)*100,0), "%)"),
  ylab=paste0("PC ", pc[2], " (", round(pca$sdev[pc[2]]/sum(pca$sdev)*100,0), "%)"))
```

```
points(pred.setosa[,pc[1]], pred.setosa[,pc[2]],
   col=COLOR[levels(SPP)=="setosa"], pch=PCH[2])
points(pred.versicolor[,pc[1]], pred.versicolor[,pc[2]],
   col=COLOR[levels(SPP)=="versicolor"], pch=PCH[2])
points(pred.virginica[,pc[1]], pred.virginica[,pc[2]],
   col=COLOR[levels(SPP)=="virginica"], pch=PCH[2])
legend("topright", legend=levels(iris$Species),
   col=COLOR, pch=17)
legend("topleft", legend=c("Original data", "New data"),
   col=1, pch=PCH)
```

练习 3. 使用*plsgenomics*包里的*SRBCT*数据, 将数据随机划分为训练集*(80%)*和检验集*(20%)*，类似上述过程，在低维空间图示化检验集的预测类别结果.