实验11、聚类分析

年级： 15级　　专业：生信　　学号：1513401013　　姓名：郑磊

编号　　一　　二　　三　　四　　总分　　评阅人

得分

软硬件平台：

1. 硬件平台：（硬件配置）i5，2.9HZ处理器，16G内存，64位操作系统

2. 系统平台：（操作系统及其版本号）Windows10 企业版

3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供URL地址）R3.4.1 ，Rstudio

一、目的要求：

1、加深对主成分分析（PCA）的理解；

2、加深对聚类分析的理解；

3、熟悉并掌握主成分分析（PCA）、聚类分析所涉及的R语言函数和脚本。

二、实验内容：

1、学生成绩的主成分分析：

利用教师提供的学生多门课程成绩表，开展主成分分析实践，看看能够有效地提取反应不同课程成绩分布特征信息的主成分，以及这些主成分能够有效地区分不同学生的学习成绩特点。

1.1、 读取学生成绩数据：

```
dir="D:/RFile/实验十一"
setwd(dir)
file="15生信成绩.txt"
scores<-read.table(file,head=T,sep="\t")
```

```
colnames(scores)

ncol(scores)

#[1] 25 =》24门课程

nrow(scores)

#[1] 30

#创建数据框

data<-data.frame(scores[,2:25])

colnames(data)<-paste("x",1:24, sep="")

rownames(data)<-scores[,1]

data #查看数据
```

1.2、 分析各门课程成绩之间的相关性：

```
library(psych)

corr.test(data)
```

1.4、数据的标准化处理前后的对比：

```
#原数据：

png("boxplot1.png")

boxplot(data,las=2)

dev.off()

#数据中心化，使其均值变为零【原点】

data2<-scale(data, center=T,scale=F)

data2

png("lboxplot2.png")
```

```
boxplot(data2,las=2)

dev.off()

#数据围绕0附近波动，但是方差变异很大

#数据标准化，除以方差

data3<-scale(data, center=T, scale=T)

data3

png("boxplot3.png")

boxplot(data3,las=2)

dev.off()
```

1.6、标准化数据协方差矩阵的计算：

```
mc<-cov(data3)

mc
```

1.7、主成分分析（PCA）：

```
#cor：逻辑变量，若为cor=T表示用样本的相关矩阵R作主成分分析，cor=F，表

示用样本

的协方差矩阵s作为主成分分析

pca<-princomp(data,cor=T)

pca2<-princomp(data2,cor=T)

pca3<-princomp(data3,cor=T)

#以上几个结果相同，princomp自动进行上述中心化和标准化处理

Pca
```

1.8、观察主成分分析的摘要信息：

summary(pca)

pca[]#查看详细信息

pca$sdev #Standard deviation

pca$loadings #loading系数矩阵

pca$center #每一门课程均值=》数据中心化

pca$scale #每一门课程方差=》数据标准化

pca$scores #每个样本每个组分的得分

pca$loadings #查看loadins信息

pca$loadings[] #查看loadings全部数值

#计算得到各个样本主成分的数据=》等价于pca$scores

pca_data <- predict(pca)

1.9、绘图查看主成分的变异贡献度：

#针对princomp()对象的plot方法#

#该方法可以绘制展示每个主成分与其自身方差贡献度相关性的悬崖碎石图。

png("lec12_bar-stone_plot1.png",width=600*3,height=3*300,res=72*3)

par(mfrow=c(1,2),las=2)

#条形图

plot(pca)

abline(h=1,type="2",col="red")

#主成分的碎石图

screeplot(pca, type="lines")

abline(h=1,type="2",col="red")

```
dev.off()
```

1.10、绘制得分（scores）图：

```
#=》主成分分布更为离散=》把30个样本区分的更好
#得分图（Score plot）
png("lec12_15scores_scores_plot6.png",width=600*3,height=3*400,res=72*3)
par(mfrow=c(2,3))
#主成分分析之后的前两个主成分得分绘图
plot(pca$scores[,1], pca$scores[,2],type="n")
text(pca$scores[,1],pca$scores[,2],labels=rownames(pca$scores),cex=0.8)
plot(pca$scores[,1], pca$scores[,3],type="n")
text(pca$scores[,1],pca$scores[,3],labels=rownames(pca$scores),cex=0.8)
plot(pca$scores[,1], pca$scores[,4],type="n")
text(pca$scores[,1],pca$scores[,4],labels=rownames(pca$scores),cex=0.8)
plot(pca$scores[,2], pca$scores[,3],type="n")
text(pca$scores[,2],pca$scores[,3],labels=rownames(pca$scores),cex=0.8)
plot(pca$scores[,2], pca$scores[,4],type="n")
text(pca$scores[,2],pca$scores[,4],labels=rownames(pca$scores),cex=0.8)
plot(pca$scores[,3], pca$scores[,4],type="n")
text(pca$scores[,3],pca$scores[,4],labels=rownames(pca$scores),cex=0.8)
dev.off()
```

1.11、绘制载荷（loadings）图：

```
png("lec12_15scores_loadings_plot6.png",width=600*3,height=3*400,res=72*3)
```

```
par(mfrow=c(2,3))
#主成分分析之后的前两个主成分得分绘图
plot(pca$loadings[,1], pca$loadings[,2],type="n")
text(pca$loadings[,1],pca$loadings[,2],labels=rownames(pca$loadings),cex=0.8)
plot(pca$loadings[,1], pca$loadings[,3],type="n")
text(pca$loadings[,1],pca$loadings[,3],labels=rownames(pca$loadings),cex=0.8)
plot(pca$loadings[,1], pca$loadings[,4],type="n")
text(pca$loadings[,1],pca$loadings[,4],labels=rownames(pca$loadings),cex=0.8)
plot(pca$loadings[,2], pca$loadings[,3],type="n")
text(pca$loadings[,2],pca$loadings[,3],labels=rownames(pca$loadings),cex=0.8)
plot(pca$loadings[,2], pca$loadings[,4],type="n")
text(pca$loadings[,2],pca$loadings[,4],labels=rownames(pca$loadings),cex=0.8)
plot(pca$loadings[,3], pca$loadings[,4],type="n")
text(pca$loadings[,3],pca$loadings[,4],labels=rownames(pca$loadings),cex=0.8)
dev.off()
```

2、基于肿瘤组和对照组基因表达谱的聚类分析：

该环节需要大家提前准备好一个基因表达谱数据，如果没有，则有授课教师提供（gds4794）。

2.1、读取数据：

在R语言环境中，加载GEOquery包，读取 gds4794数据集；其中包含了肺癌和正常对照样本两种病理类型数据；

2.2、差异表达基因筛选：

按照基因（探针）表达水平采用t检验进行统计分析，计算差异显著性p值，以及差异表达倍数（肺癌/正常）；

设定好筛选的p值和上、下调倍数【注意：每个同学按照学号顺序，上调倍数从起始11倍开始设定，下调倍数是该数值的倒数】；

按照设定阈值，筛选出区别这两组样本的主要基因（探针）；

提取这些差异表达基因的表达水平数据。

```
#变量初始化，用来存放计算结果中的p.value和fold change值
p=NULL
fold.change=NULL
#R用Sys.time()可以查看当前系统时间
#程序开始时记录：
timestart<-Sys.time()
#基因表达谱遍历
for(i in 1:nrow(data))
{
  a <- unlist(data[i,3:25])
  b <- unlist(data[i,26:67])
  fold.change<-c(fold.change,mean(a,na.rm=TRUE)/mean(b,na.rm=TRUE))
  x<-t.test(a,b)
  p<-c(p,x$p.value)
}
#程序临结束时记录：
```

```
timeend<-Sys.time()

#程序运行时间：

timeend-timestart

#Time difference of 51.29762 secs

#data第一列探针名IDs作为p和fold.change的名称

names(p)<-data[,1]

names(fold.change)<-data[,1]

#设定阈值进行筛选

p_value = 0.05

up = 10 #lung cancer  上调2倍

down = 0.1 #lung cancer  下调2倍

#筛选

p2 <- p[p<p_value] #p值筛选

fc.up <- fold.change[fold.change>up] #上调基因

fc.down <- fold.change[fold.change<down] #下调基因

length(p2); length(fc.up); length(fc.down) #查看筛选结果

#交集计算

probes.up<-intersect(names(p2),names(fc.up)) #符合统计学显著性的上调基因

length(probes.up)

probes.down<-intersect(names(p2),names(fc.down)) #符合统计学显著性的下调
基因

length(probes.down)
```

probes<-union(probes.up,probes.down) #合并合统计学显著性的上调和下调基因

#上述过程合并进行

#probes <- intersect(names(p2),union(names(fc.up),names(fc.down)))

length(probes)

subdata<-log(data[probes,3:67]) #从原始基因表达谱数据表中提取筛选出来的基因数据

rownames(subdata)<-probes #设定探针IDs为行标题

nrow(subdata)

2.3、数据标准化前后的对比：

注意表达水平数据矩阵的行列转换，原数据矩阵列为样本，行为基因（探针），后续分析需要进行行列转置。

#数据标准化，除以方差

subdata2<-scale(t(subdata), center=T, scale=T)

rownames(subdata2)<-rep(1:65) #使用数据编号代替样本名称

#subdata2

png("lec12_gds4794_clustering_boxplot1.png",width=600*3,height=300*3,res=72*3)

par(mfrow=c(1,2),las=2)

boxplot(t(subdata))

boxplot(subdata2)

dev.off()

2.4、层次聚类

根据标准化的基因表达水平计算不同样本之间的距离，然后按照“最短距离”策略急性层次聚类分析。

```
d<-dist(subdata2, method = "euclidean")
#r语言中使用hclust(d, method = "complete", members=NULL) 来进行层次聚类。
hc<-hclust(d,"single")
png("lec12_gds4794_clustering_tree_plot.png",width=600,height=300)
plot(hc)
dev.off()
```

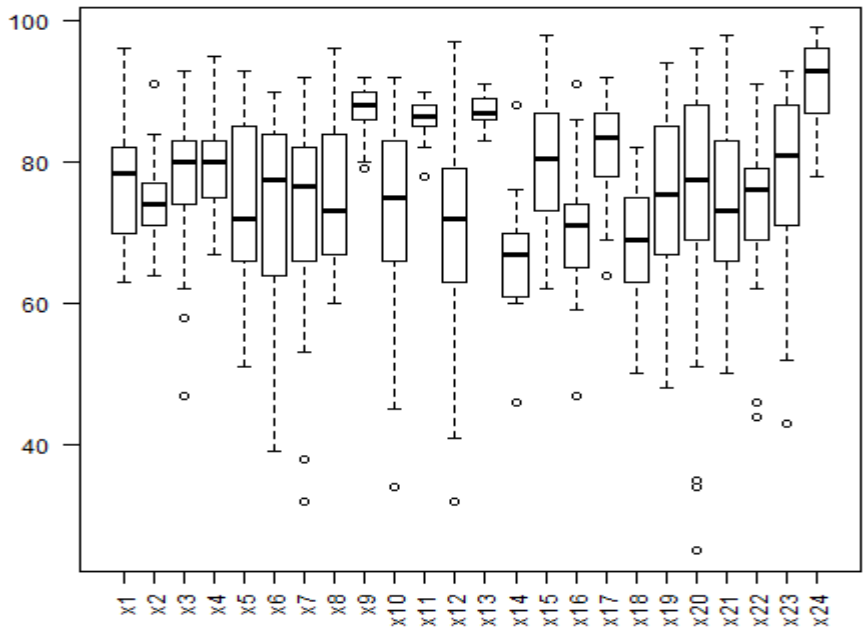
2.5、确定分类：

根据2.4步的绘图结果，自己选择合适的分类参数k来确定分类结果，并对分类结果加以探讨。

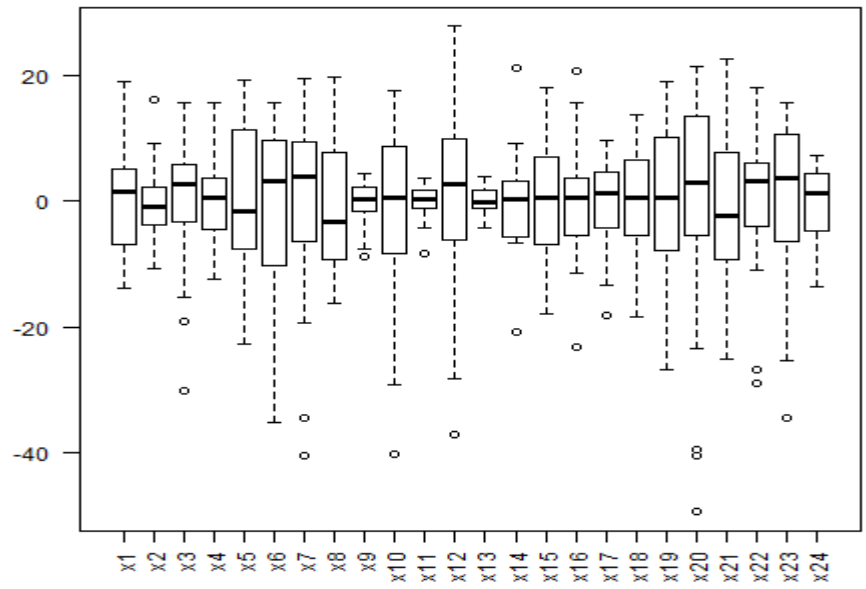
```
#使用rect.hclust(tree, k = NULL, which = NULL, x = NULL, h = NULL,border =2, cluster = NULL)来确定类的个数。 tree就是求出来的对象。k为分类的个数，h为类间距离的阈值。border是画出来的颜色，用来分类的
png("lec12_gds4794_clustering_tree_plot2.png", width=600,height=300)
plot(hc)
rect.hclust(hc,k=2)
dev.off()
result=cutree(hc,k=3) #该函数可以用来提取每个样本的所属类别
result
```

三、实验结果：

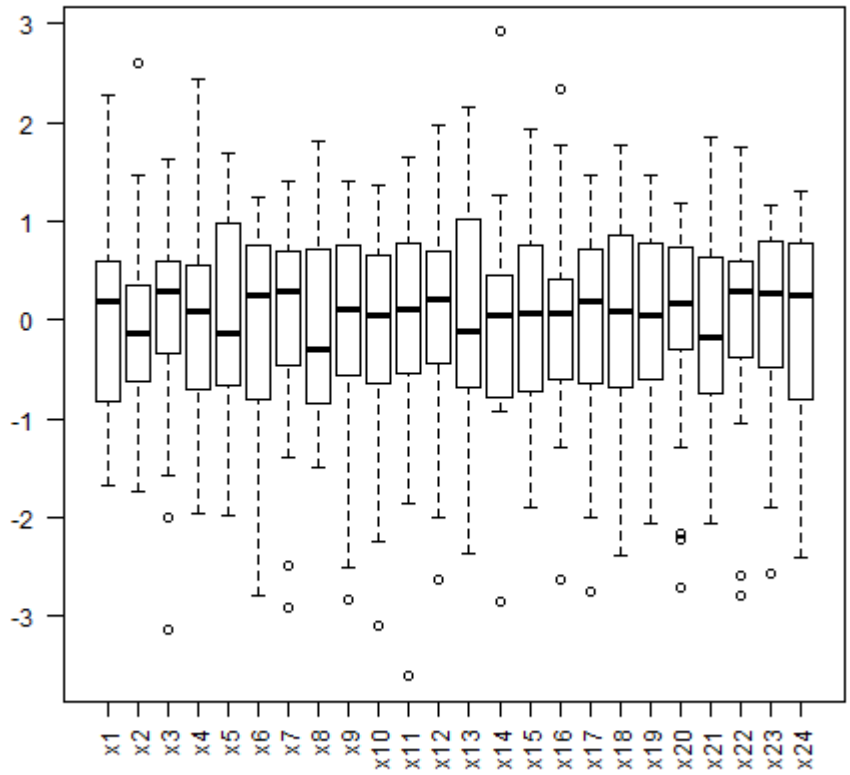
1.4



数据中心化：



数据标准化：



1.8

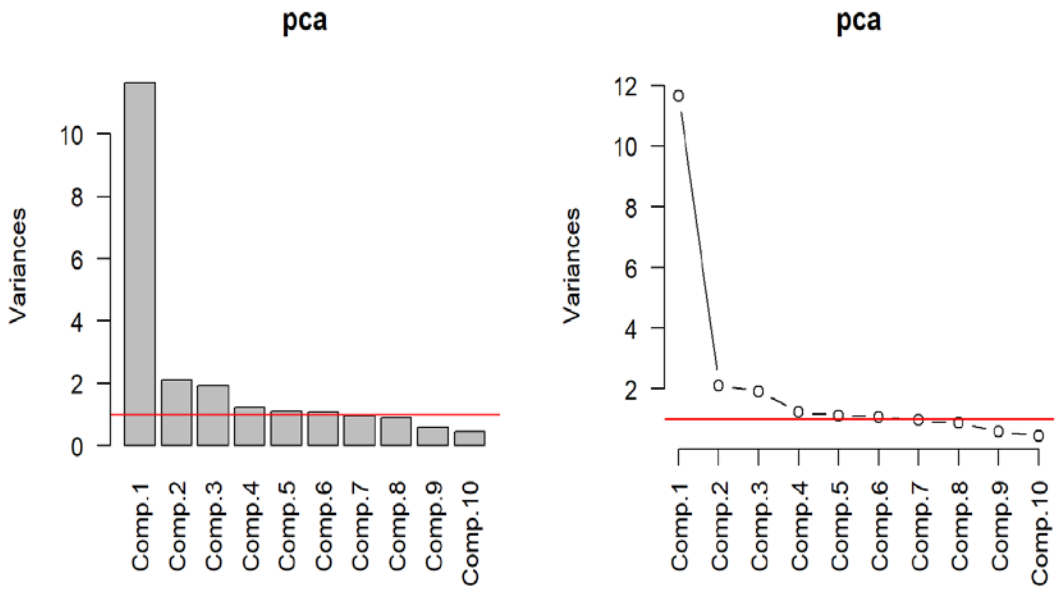
```
> summary(pca)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4
 Comp.5     Comp.6     Comp.7     Comp.8     Comp.9     Comp.10
Standard deviation     3.4137449 1.45259866 1.38680827 1.11067294 1.0
5631990 1.03609774 0.98067582 0.93999471 0.7737224 0.67347203
Proportion of Variance 0.4855689 0.08791845 0.08013488 0.05139977 0.0
4649216 0.04472911 0.04007188 0.03681625 0.0249436 0.01889852
Cumulative Proportion  0.4855689 0.57348738 0.65362227 0.70502203 0.7
5151419 0.79624329 0.83631517 0.87313142 0.8980750 0.91697355
                          Comp.11    Comp.12    Comp.13    Comp.14
  Comp.15    Comp.16    Comp.17    Comp.18    Comp.19
Standard deviation     0.61676815 0.56500613 0.5396866 0.459367126 0.
405782714 0.389534649 0.356694090 0.341719621 0.298026846
```

Proportion of Variance 0.01585012 0.01330133 0.0121359 0.008792423 0.006860817 0.006322385 0.005301278 0.004865512 0.003700833

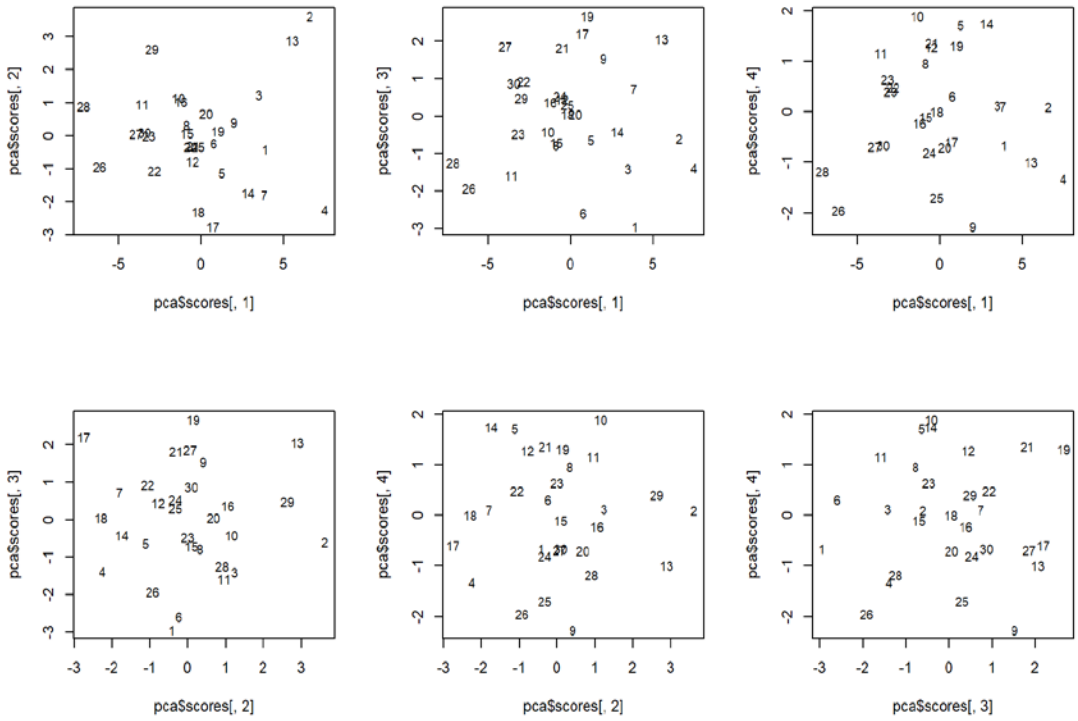
Cumulative Proportion 0.93282367 0.94612500 0.9582609 0.967053325 0.973914142 0.980236527 0.985537805 0.990403317 0.994104151

	Comp.20	Comp.21	Comp.22	Comp.23	Comp.24
Standard deviation	0.228854761	0.222882892	0.1419850169	0.1193752039	0.0709853054
Proportion of Variance	0.002182271	0.002069866	0.0008399894	0.0005937683	0.0002099547
Cumulative Proportion	0.996286422	0.998356288	0.9991962770	0.9997900453	1.0000000000

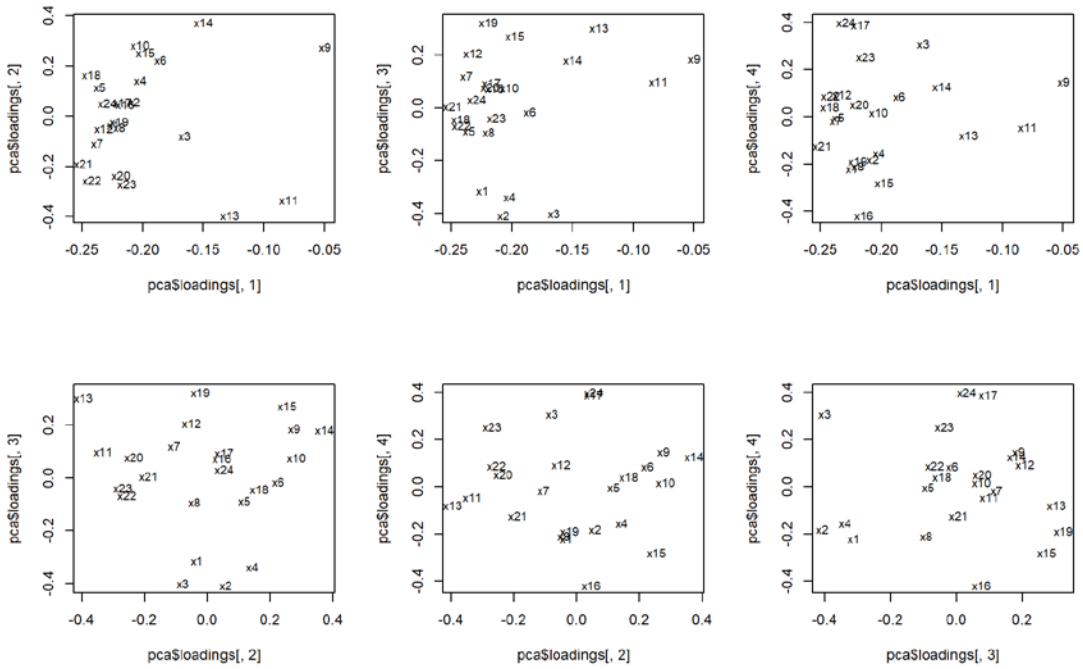
1.9



1.10



1.11

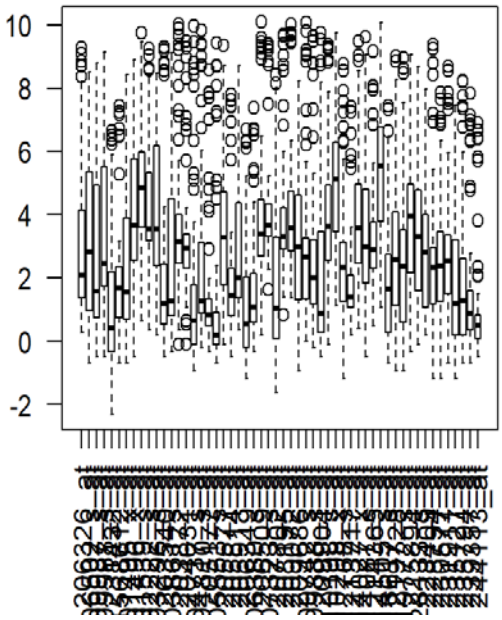


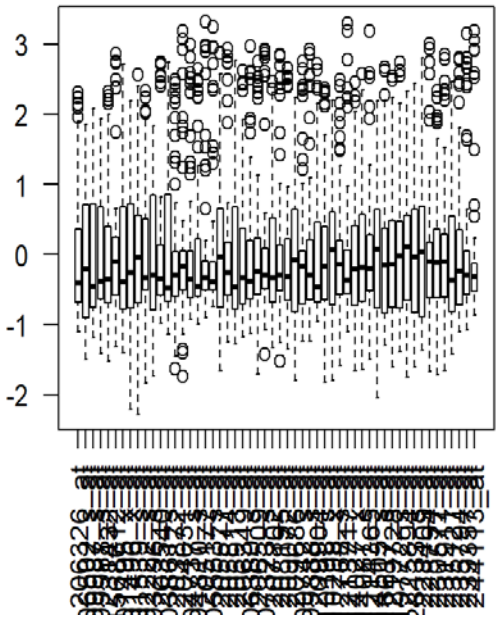
2.2

```
> timeend-timestart
Time difference of 1.956635 mins
```

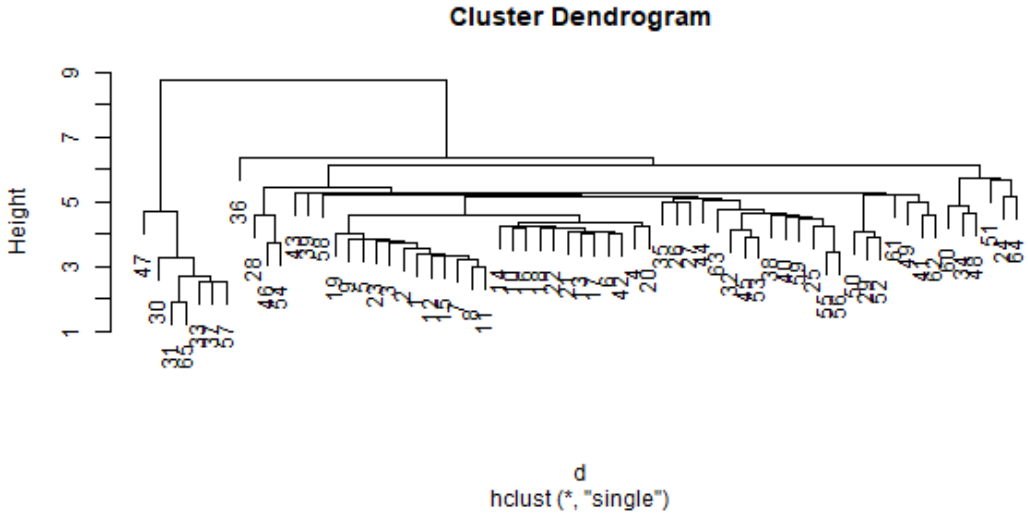
```
> length(p2); length(fc.up); length(fc.down) #查看筛选结果
[1] 25077
[1] 6
[1] 169
> length(probes.up)
[1] 5
> length(probes.down)
[1] 49
> length(probes)
[1] 54
> nrow(subdata)
[1] 54
```

2.3

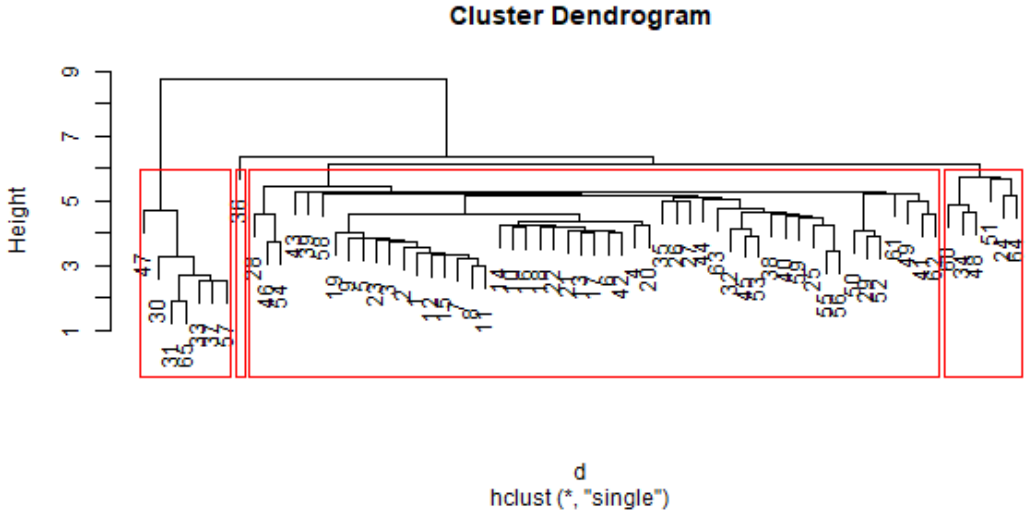




2.4



2.5




```
> result
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 2  1  1  1  1  1  3  3  1  3  2  1  4  3  1  1  1  1  1  1  1
45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65
 1  1  3  2  1  1  2  1  1  1  1  1  3  1  1  2  1  1  1  2  3
```

四、讨论

1.6

协方差与相关系数一致，只不过协方差结果保留的小数位较多。

1.10

图中的离散点较为分散，第四与第六附图没有重叠，区分度较好，其余的几乎也都可以较好的区分。

1.11

主成分一对原始变量几乎全是负相关，其余主成分正负皆有。

2.5

这次聚类不是很成功，共聚了4类：第一类是中间一大片，第二类是一些normal样品，第三类是大脑相关样品，第四类只有1个，心脏样品。