高级统计方法 第10次作业:

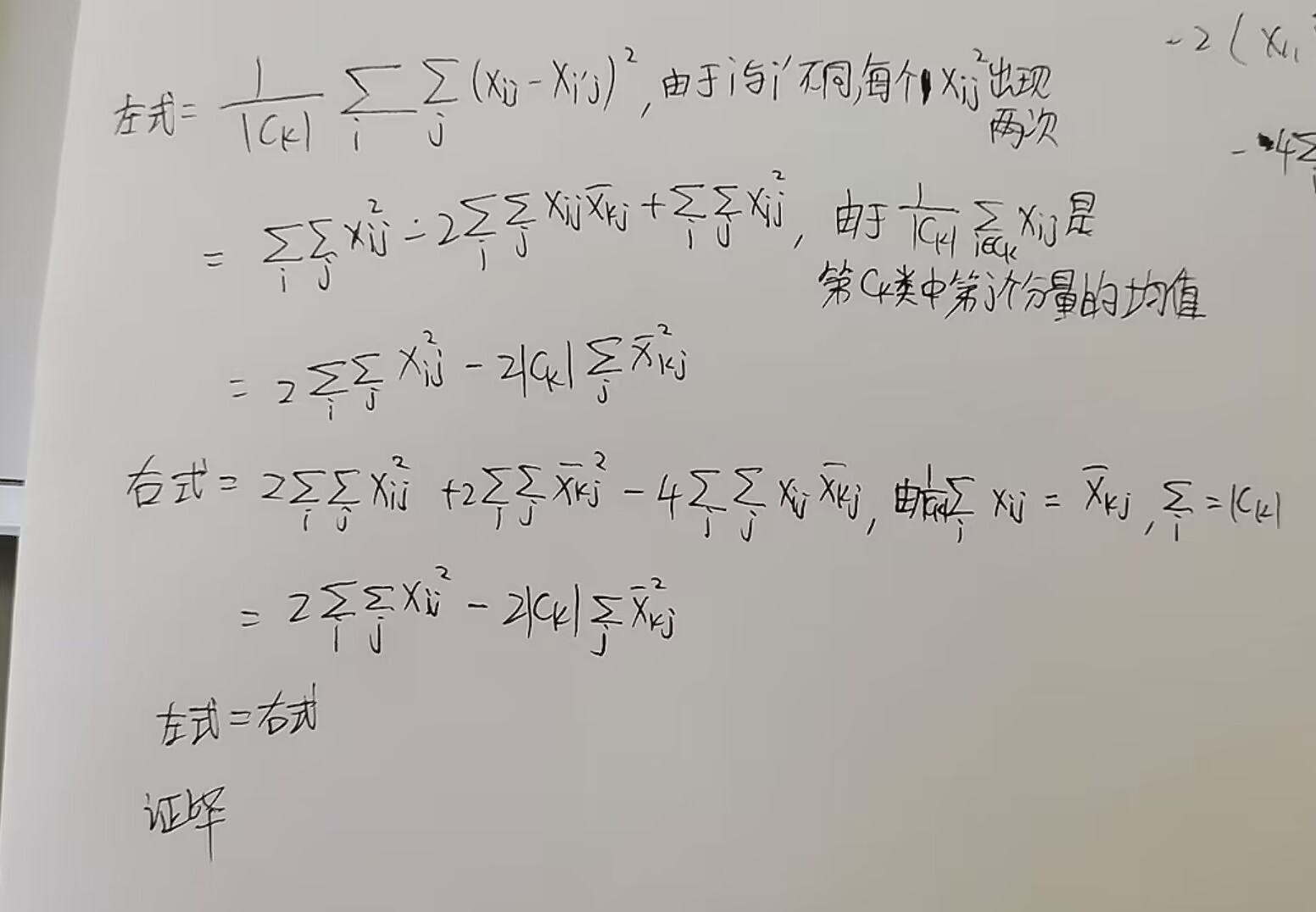
序号：1 姓名：杨程瑜 学号：20202241107 班级：网2001

在完成word文档时需要沿用课本上习题序号，因此第二次作业格式应为：

**概念**

1.问题（略）

（a）问题（略）

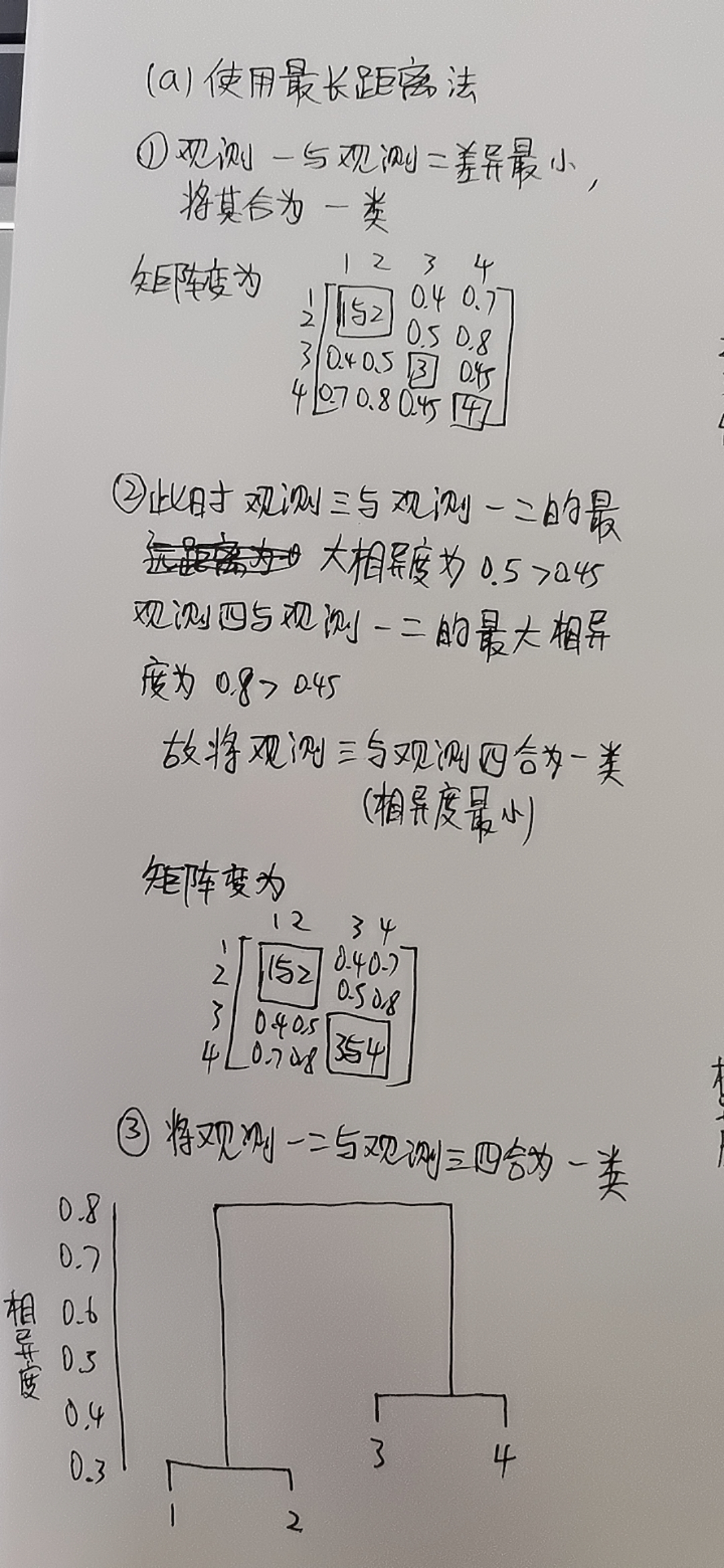


（b）问题（略）

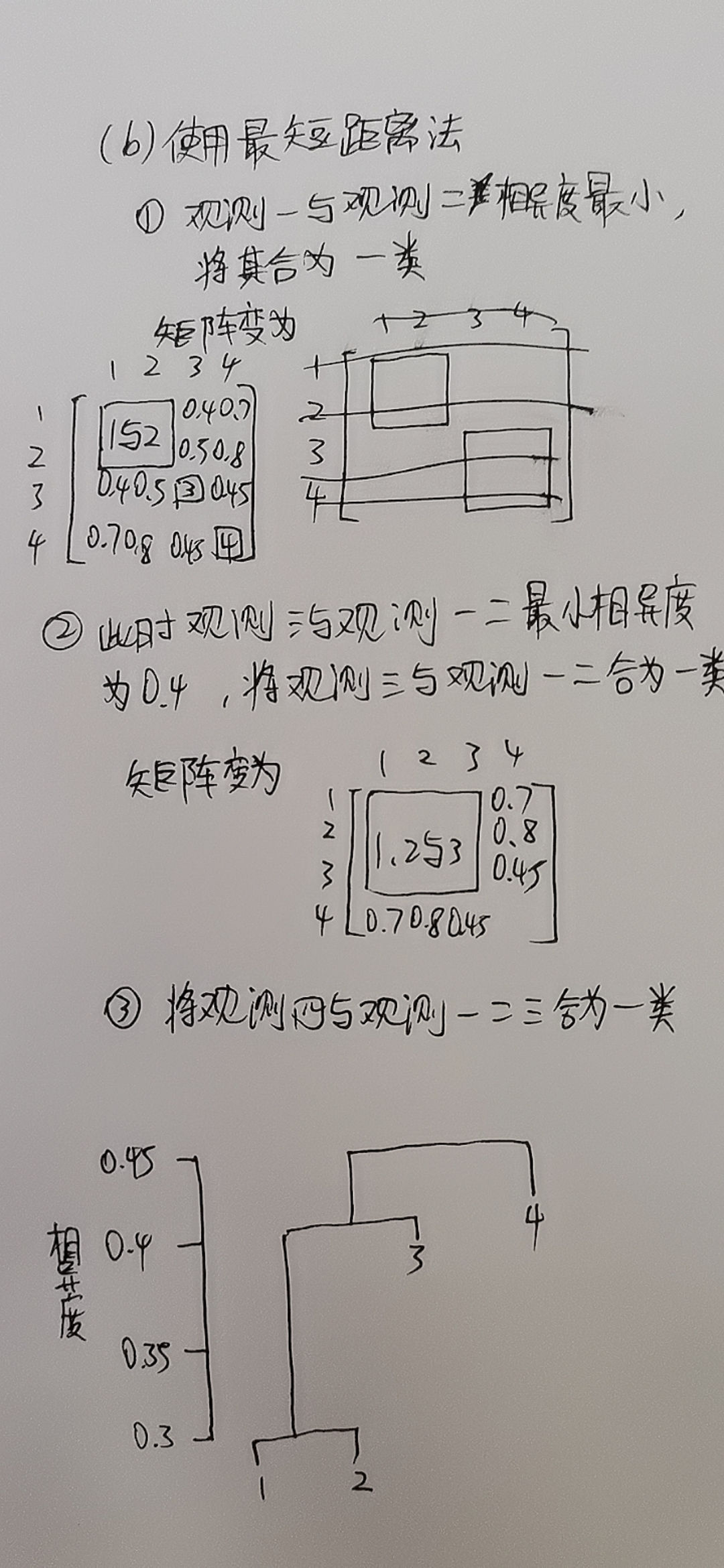
当我们每次迭代时，算法10.1中的2(b)步骤将会把观测分配到最近的类中心。此时很明显降低了(a)问中的右式，因此左式和右式相等，左式也就是我们的最小化目标会减小

2.问题（略）

（a）问题（略）



（b）问题（略）



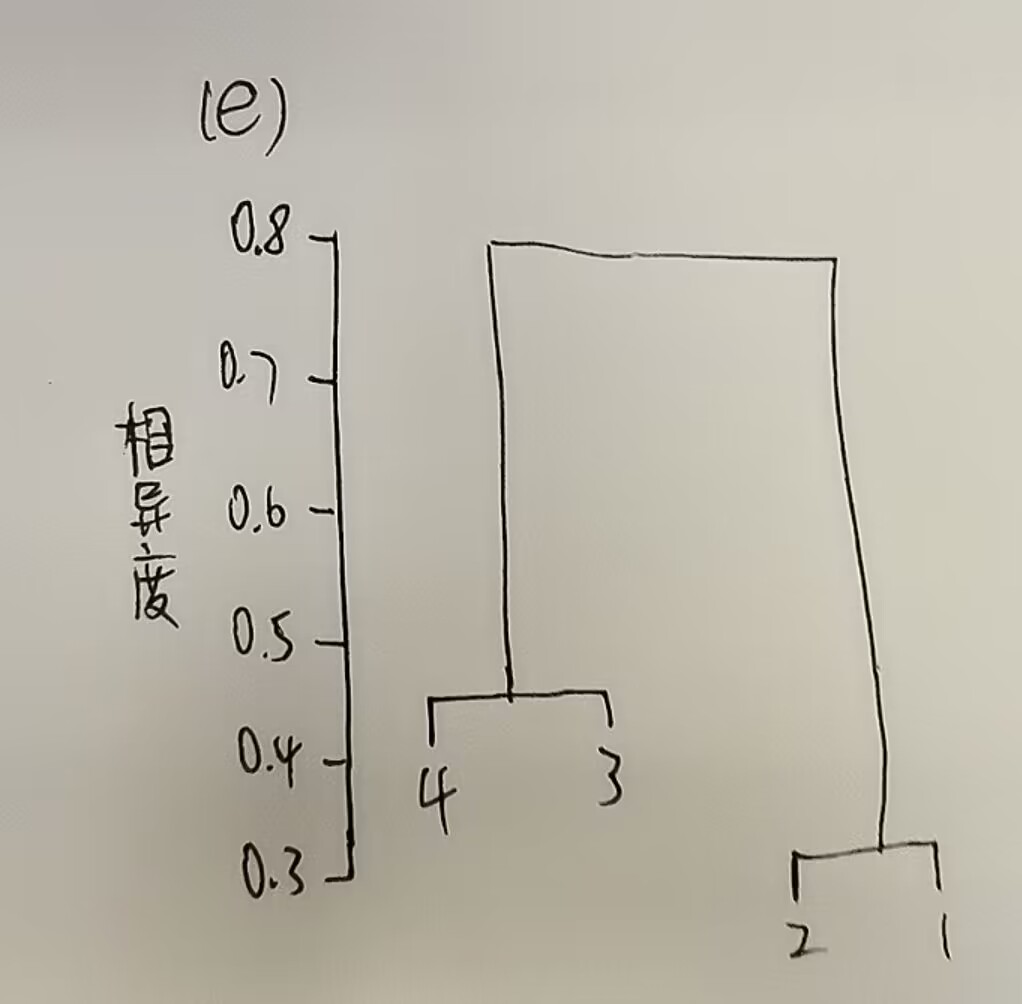
（c）问题（略）

切割K=2个类，第一个类包含观测一、观测二，第二个类包含观测三、观测四

（d）问题（略）

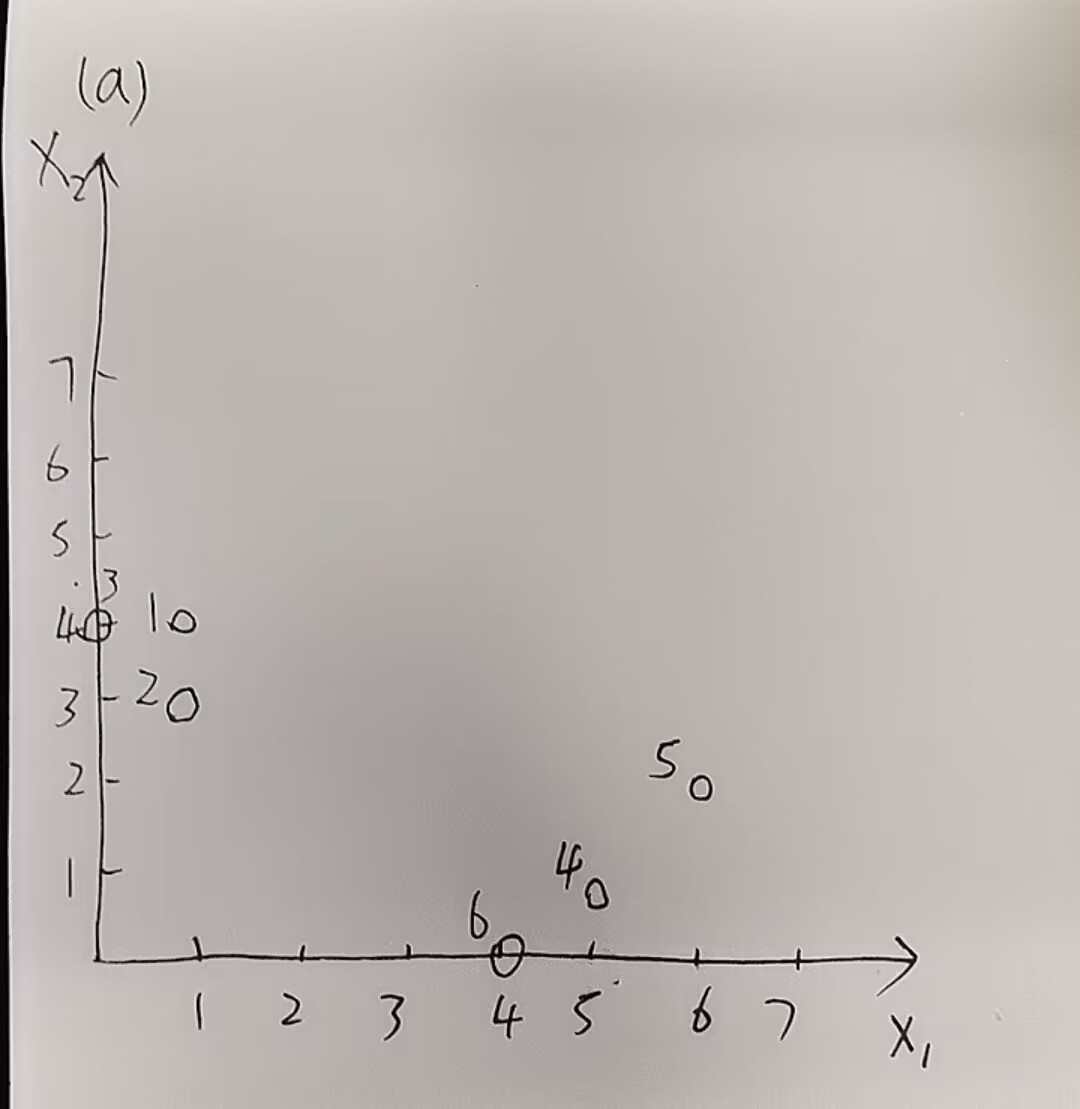
切割K=2个类，第一个类包含观测一、观测二、观测三，第二个类包含观测四

（e）问题（略）

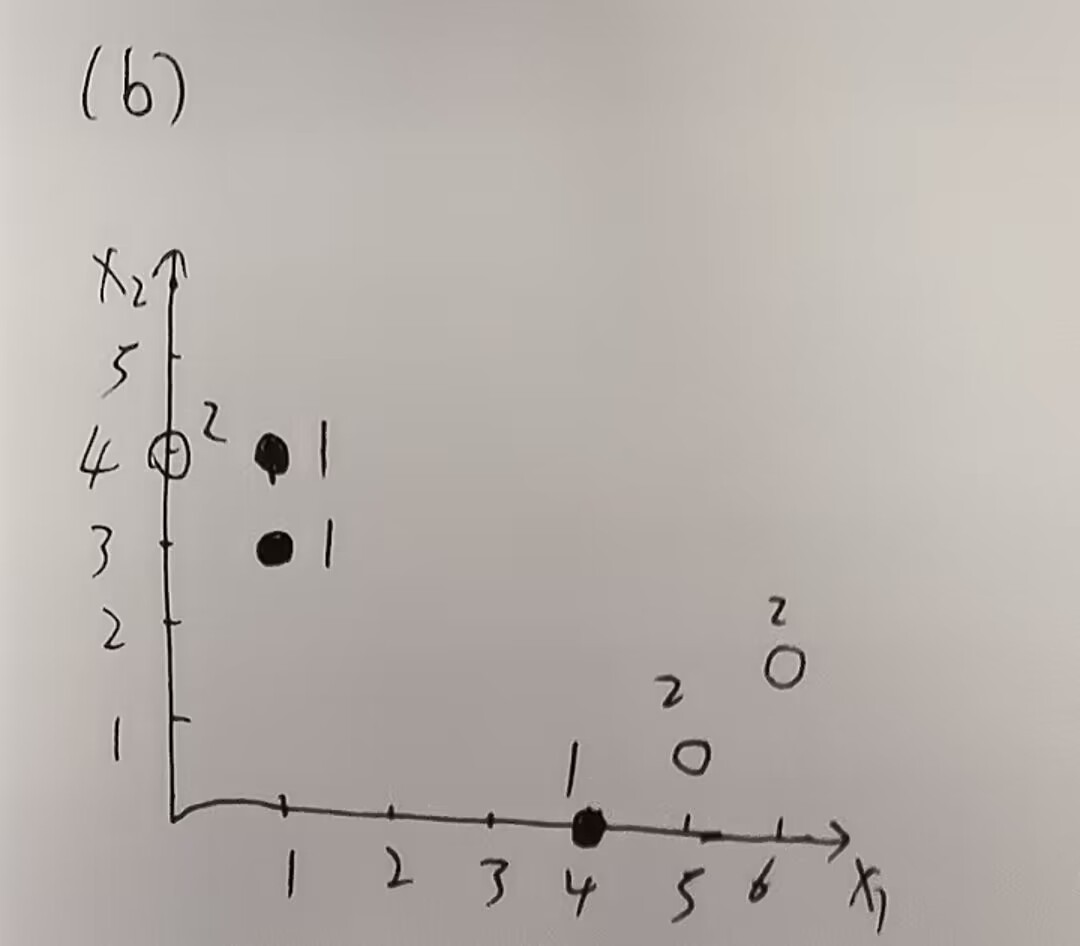


3.问题（略）

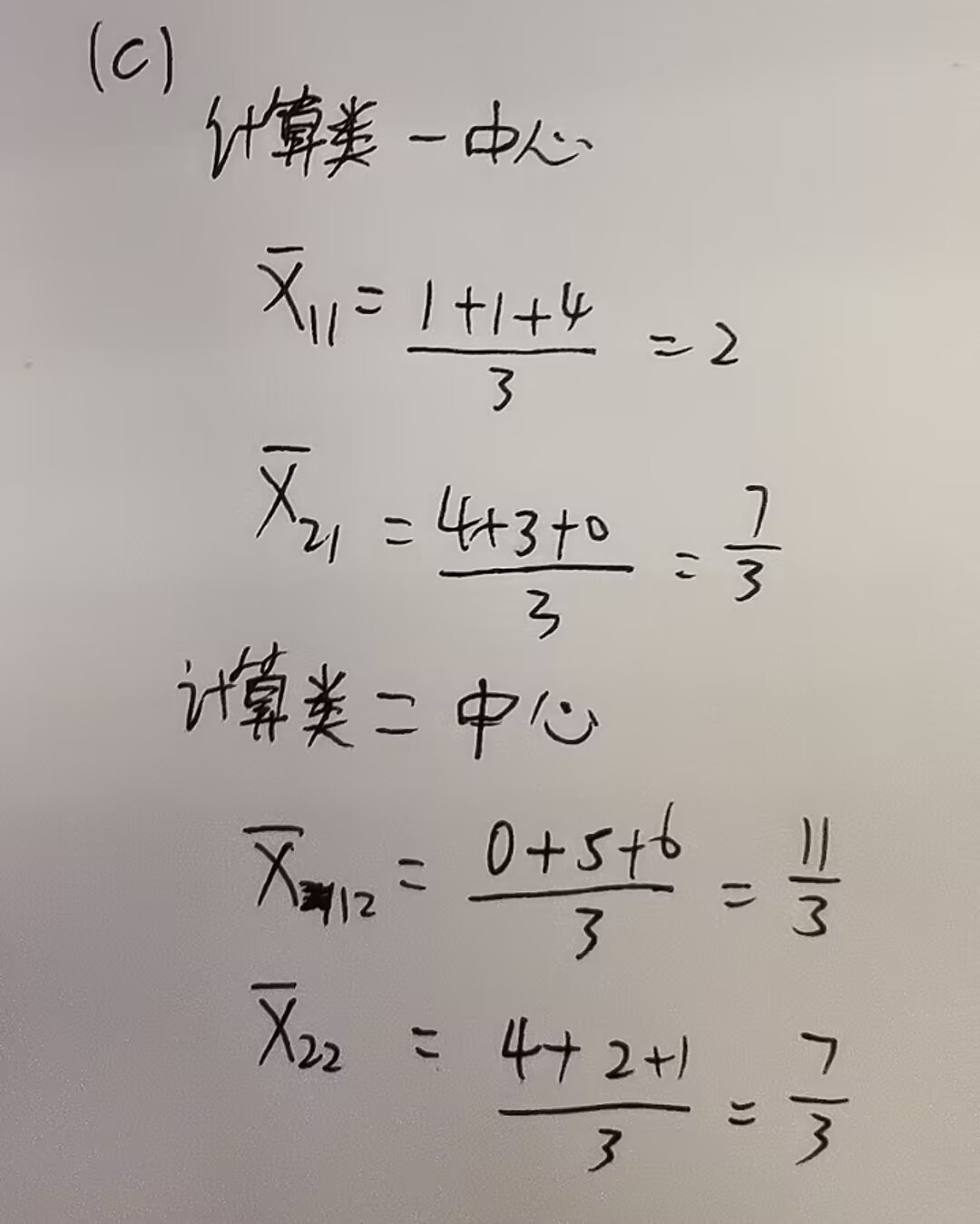
（a）问题（略）



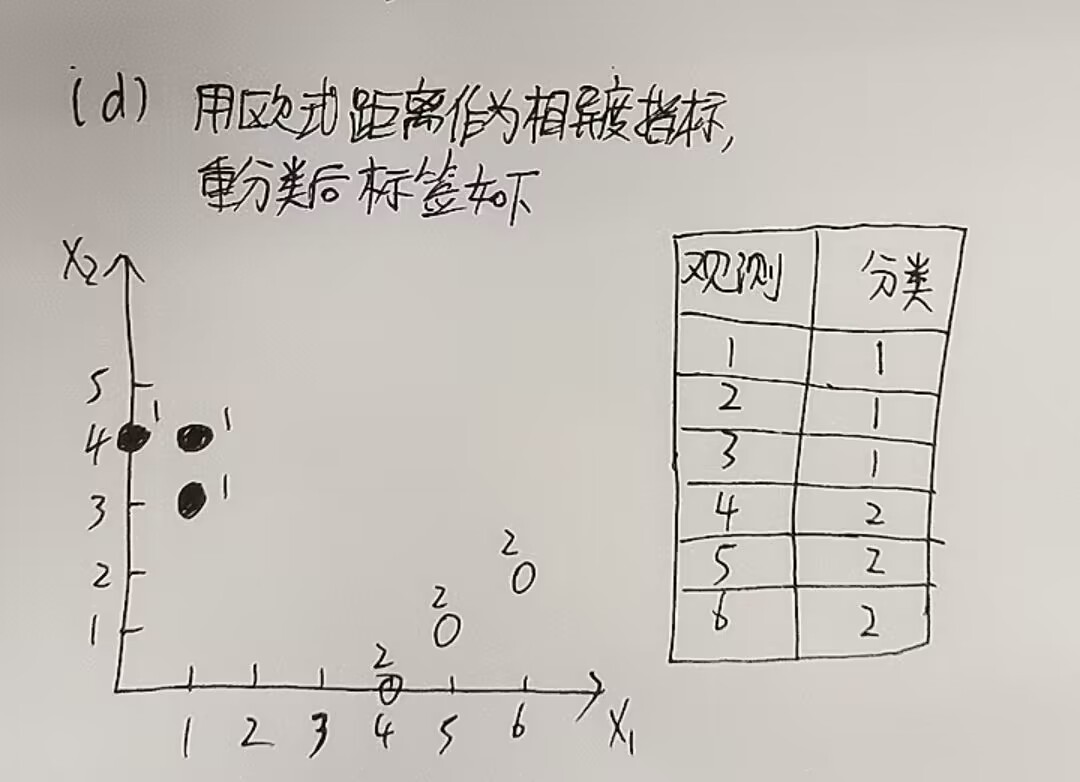
（b）问题（略）



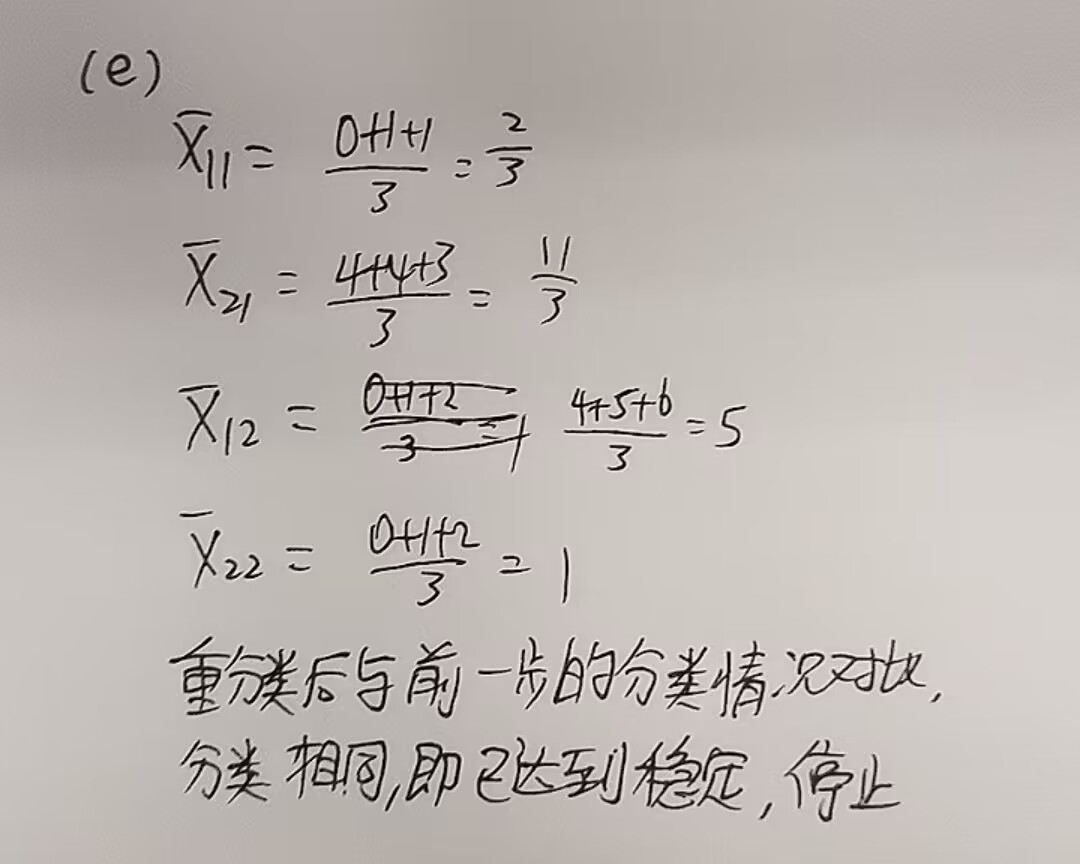
（c）问题（略）



（d）问题（略）



（e）问题（略）



（f）问题（略）

前问的图示已标明类别的颜色

4.问题（略）

（a）问题（略）

信息不充分无法回答。

由于将要汇合的两个集合不止一个观测，而是都包含多个观测，因而汇合位置会具体随集合中具体观测的位置变化而变化，故信息不充分无法回答

（b）问题（略）

将在相同位置汇合。

在相同位置汇合的原因是{5}和{6}是两个单独的观测，它们之间的最长距离法下的相异度和最短距离法下的相异度一定是相同的

5.问题（略）

根据书中图10-14的三种不同处理来进行

1. 不进行标准化

此时按照购买袜子和计算机的总数量来分类，分成第一类：{1，2，7，8}，对应最高的购买数8、11、8、9，第二类：{3，4，5，6}

1. 对各自标准差进行标准化

此时根据10-14的数据图，购买计算机的权重大幅增加，故据图能分成第一类：{5，6，7，8}，第二类：{1，2，3，4}

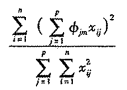
1. 以花费的美元作y轴

此时根据10-14的数据图，由于计算机的价值远大于袜子，购买计算机的权重几乎占满，故据图能分成第一类：{5，6，7，8}，第二类：{1，2，3，4}

6.问题（略）

（a）问题（略）

即第一主成分的累计方差占总方差的10%，见如下公式



意味着基因数据集中90%的信息通过将组织样本观察投影到第一个主成分超平面上而丢失，10%保留

（b）问题（略）

从题目论述的预分析中明显存在缺陷：即该第一主成分没有考虑机器A和机器B的因素（而预分析过程的论述中提到前期使用机器A比较多，后期使用机器B比较多），导致第一主成分存在很强的线性趋势，并且第一主成分只解释了10%的总方差。故很明显改进方案中需要考虑机器A和机器B的因素

（c）问题（略）

rm(list=ls())

set.seed(123)

Control <- matrix(rnorm(50 \* 1000), ncol = 50)

Treatment <- matrix(rnorm(50 \* 1000), ncol = 50)

X <- cbind(Control, Treatment)

X[1, ] <- seq(-18, 18 - .36, .36)

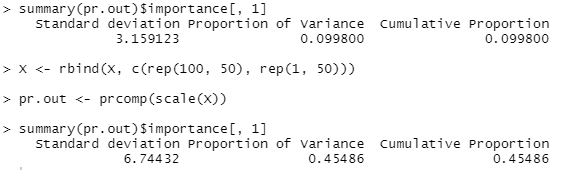
pr.out <- prcomp(scale(X))

summary(pr.out)$importance[, 1]

X <- rbind(X, c(rep(100, 50), rep(1, 50)))

pr.out <- prcomp(scale(X))

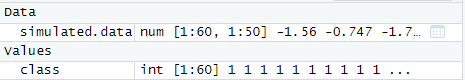
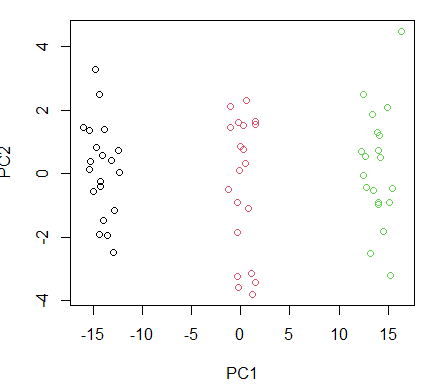
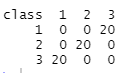
summary(pr.out)$importance[, 1]



该模拟实验前半段模拟第一主成分只占总方差10%，第二部分我们添加一个dataframe的绑定，将机器A和机器B的权值设为100：1，结果显示我们的第一主成分加入机器A和机器B的因素后能够解释方差的45%，大幅提升

**应用**

10.问题（略）

# 10.7.10
  
rm(list=ls())
  
library(ISLR)
  
  
（a）问题（略）  
set.seed(123)
  
# 使用不同均值来构造不同类的数据
  
# 60 \* 50, n=60, p=50
  
simulated.data = matrix(c(rnorm(20 \* 50, mean = -1),
  
 rnorm(20 \* 50, mean = 1),
  
 rnorm(20 \* 50, mean = 3)), ncol = 50, byrow = TRUE)
  
class = unlist(lapply(1:3,function(x){rep(x,20)}))
  
  
（b）问题（略）  
# 画图
  
pr.out2 = prcomp(simulated.data)
  
plot(pr.out2$x[,1:2],col=class)
  
  
（c）问题（略）  
# K=3的均值聚类
  
kmeans.out = kmeans(simulated.data, 3, nstart = 60)
  
table(class, kmeans.out$cluster)
  
  
# 结果发现与真实类相比所有类(每个类20个)都完全分类正确
  
  
（d）问题（略）  
# K=2的均值聚类
  
kmeans.out = kmeans(simulated.data, 2, nstart = 60)
  
table(kmeans.out$cluster)
  
  
# 结果发现确实分成了两类，第一类20个第二类40个
  
  
（e）问题（略）  
kmeans.out = kmeans(simulated.data, 4, nstart = 60)
  
table(kmeans.out$cluster)
  
  
# 结果发现分成了四类，第一类20个，第二类12个，第三类8个，第四类20个
  
  
（f）问题（略）  
kmeans.out = kmeans(pr.out2$x[,1:2], 3, nstart = 60)
  
table(kmeans.out$cluster)
  
table(class)
  
# 分类结果与真实相比完全正确
  
  
（g）问题（略）  
set.seed(123)
  
# 使用scale进行标准化
  
kmeans.out = kmeans(scale(simulated.data), 3, nstart = 60)
  
table(kmeans.out$cluster)
  
table(class)
  
  
plot(pr.out2$x[,1:2],col=kmeans.out$cluster)
  
  
# 分类结果与真实相比仍然完全正确



