南京信息工程大学

数据分析课程设计报告

学生姓名 赖莹

学　　号 201983430049

院　　系 数学与统计学院

年级专业 1 班

一、数据的描述性分析

实验目的：掌握和理解相关系数和及数据的数字特征等。

实验题目 1.1

1， 通过模拟方法生成二元正态分布向量（样本容量为 500），其均值设定为(0，0)’，而协方差矩阵分别为如下情形：

1),$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$; 2),$\begin{pmatrix}0.2 & 0\\ 0 & 0.2\end{pmatrix}$; 3),$\begin{pmatrix}4 & 0\\ 0 & 4\end{pmatrix}$;

4),$\begin{pmatrix}0.2 & 0\\ 0 & 4\end{pmatrix}$; 5),$\begin{pmatrix}4 & 0\\ 0 & 0.2\end{pmatrix}$;

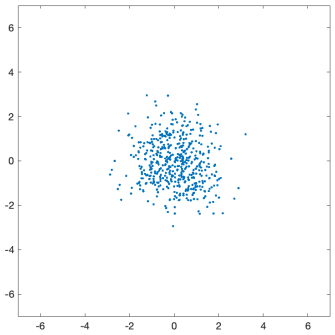
6),$\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$; 7),$\begin{pmatrix}0.3 & 0.5\\ 0.5 & 4\end{pmatrix}$; 8), $\begin{pmatrix}\mathbf{4} & \mathbf{0.5}\\ \mathbf{0.5} & \mathbf{0.3}\end{pmatrix}$,

请画出图像。

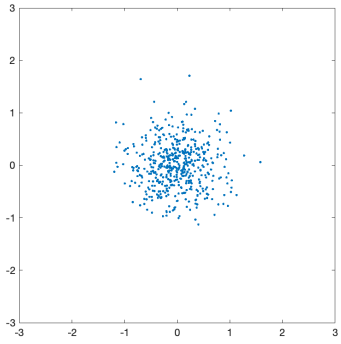
实验过程描述

参考 exmapl_1.m 的代码

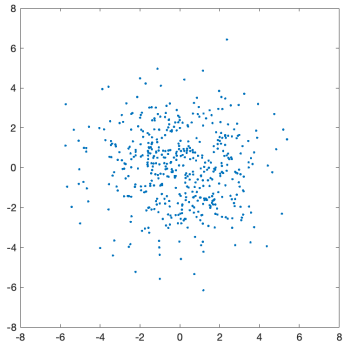
结果分析



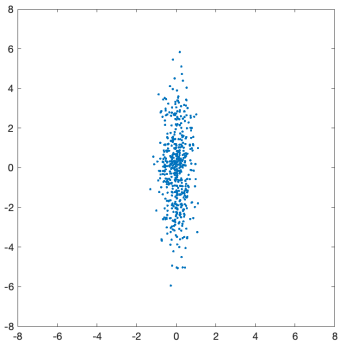
1



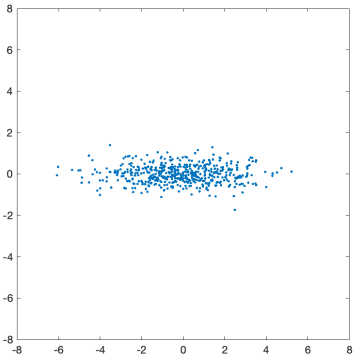
2



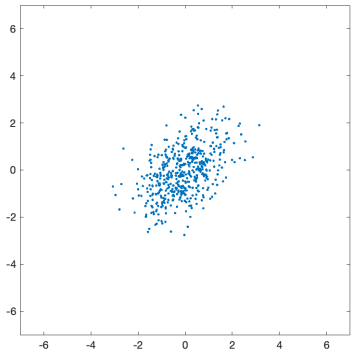
3



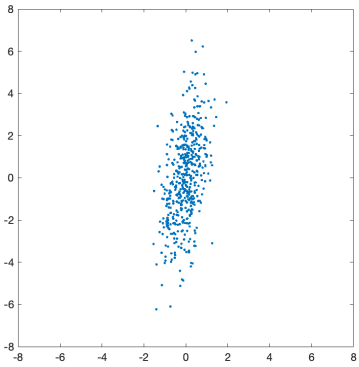
4



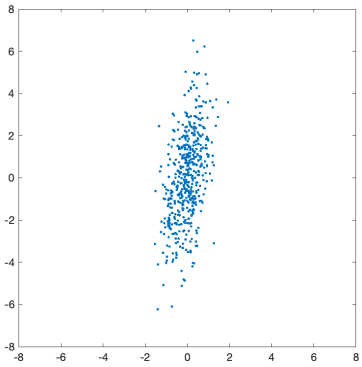
5



6



7



8

实验题目 1.2

2， 对 data_1 的数据剔除 ID 为 841 的观测值，

1） 对四个变量 EXPE、QUAL、LOYA 和 SATI 画出散点图矩阵；

2） 画出各变量的箱线图；

3） 计算观测数据的 pearson 相关系数矩阵，并做相关性的显著性检验。

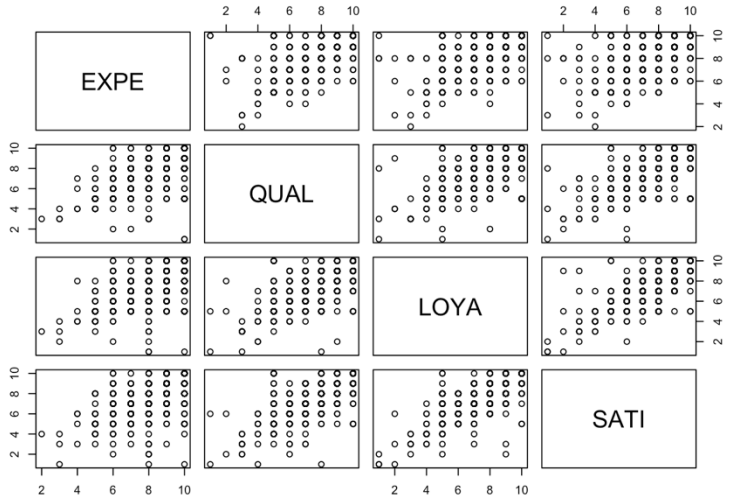
实验过程描述

首先用 Excel 去除 ID 为 841 的观测值，然后利用 R 语言中 pairs 函数画出散点图矩阵。再利用 ggplot2 库中的 boxplot 函数画出箱型图。

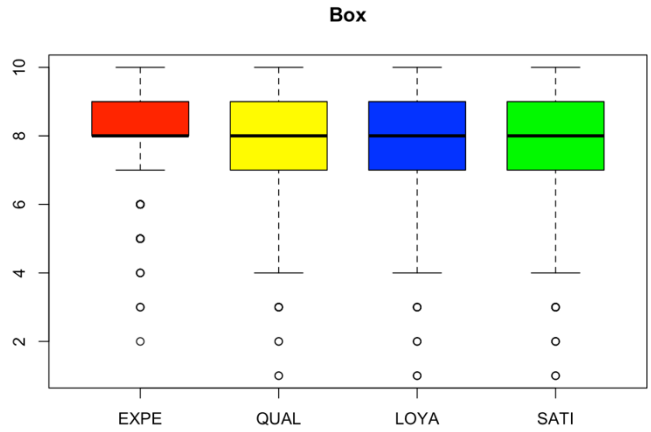
Pearson 矩阵：用于度量两组数据的变量 X 和 Y 之间的线性相关的程度。它是两个变量的协方差与其标准差的乘积之比； 因此，它本质上是协方差的归一化度量，因此结果始终具有介于-1 和 1 之间的值。与协方差本身一样，该度量只能反映变量的线性相关性，而忽略了许多其他类型的关系或相关性。最后使用 cor 函数算出 Pearson 矩阵。

结果分析

散点图矩阵如下图所示：



箱型图如下所示：



根据该箱型图可以看出每个变量都有异常点。

Pearson 相关系数矩阵：

```
> cor_pearson <- cor(data1, method = 'pearson')
> cor_pearson
                ID       EXPE       QUAL       LOYA          SATI
ID    1.000000000 0.01661423 0.01084361 0.04465975 -0.004057933
EXPE  0.016614233 1.00000000 0.59134837 0.53623883  0.546957739
QUAL  0.010843613 0.59134837 1.00000000 0.65150392  0.710389048
LOYA  0.044659747 0.53623883 0.65150392 1.00000000  0.755214067
SATI -0.004057933 0.54695774 0.71038905 0.75521407  1.000000000
```

SATI 与影响变量 EXPE、QUAL、VALU 之间的相关系数分别为： 0.5469、0.7014、0.7552。这说明 SATI 与各影响变量均存在着较显著的正相关关系。

实验题目 1.3

1.3 已知 8 个乳房肿瘤病灶组织的样本，其中前 3 个为良性肿瘤，后 5 个为恶性肿瘤。数据为细胞核显微图像的5 个量化特征：细胞核直径，质地，周长，面积，光滑度。已知样本的数据如下：

13.54, 14.36, 87.46, 566.3, 0.09779
13.08, 15.71, 85.63, 520, 0.1075
9.504, 12.44, 60.34, 273.9, 0.1024
17.99, 10.38, 122.8, 1001, 0.1184
20.57, 17.77, 132.9, 1326, 0.08474
19.69, 21.25, 130, 1203, 0.1096
11.42, 20.38, 77.58, 386.1, 0.1425
20.29, 14.34, 135.1, 1297, 0.1003

试根据已知样本利用距离判别（分别用协方差矩阵相等、协方差矩阵不等）对下面未知种类的三个样本进行分类：

16.6, 28.08, 108.3, 858.1, 0.08455
20.6, 29.33, 140.1, 1265, 0.1178
7.76, 24.54, 47.92, 181, 0.05263

结果：

首先，需要将数据集中的所有样本加载到 R 中

```
# 加载数据
data <- read.csv("data.csv")
# 将数据分为训练集和测试集
train <- data[1:8, ]
test <- data[9:11, ]
# 将训练集分为良性和恶性两类
benign <- train[1:3, ]
malignant <- train[4:8, ]
```

接下来，可以使用以下代码来计算每个类的均值向量和协方差矩阵：

```
# 计算良性类的均值向量和协方差矩阵
mean_benign <- colMeans(benign)
cov_benign <- cov(benign)
# 计算恶性类的均值向量和协方差矩阵
mean_malignant <- colMeans(malignant)
cov_malignant <- cov(malignant)
```

然后，可以使用 lda() 函数来构建距离判别模型。我们假设我们希望使用所有样本的 5 个特征来构建模型，并且希望使用协方差矩阵相等的方法。

```
library(MASS)
model <- lda(Type ~ ., data = samples, method = "equal")
```

最后，可以将未知种类的三个样本储存在数据框中，并使用 predict() 函数来对未知种类的样本进行分类。

二、主成分分析部分

实验目的：利用主成分分析进行数据降维。

实验题目 2.1

考虑在中国 31 省/市 8 个不同生活支出中的应用，这些指标包括食品，衣着，居住，家庭设备及服务，交通和通讯，文教娱乐用品及服务，医疗保健和其他商品及服务。

实验过程描述

首先在 R 语言载入这些指标数据。用如下命令画出的箱线图

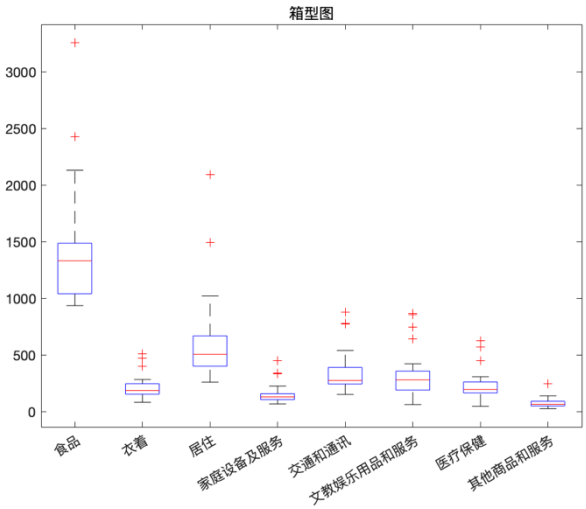
boxplot(ratings,'orientation','horizontal','labels',categories)

接下来做主成分分析。它利用正交变换来对一系列可能相关的变量的观测值进行线性变换，从而投影为一系列线性不相关变量的值，这些不相关变量称为主成分（Principal Components）。具体地，主成分可以看做一个线性方程，其包含一系列线性系数来指示投影方向。

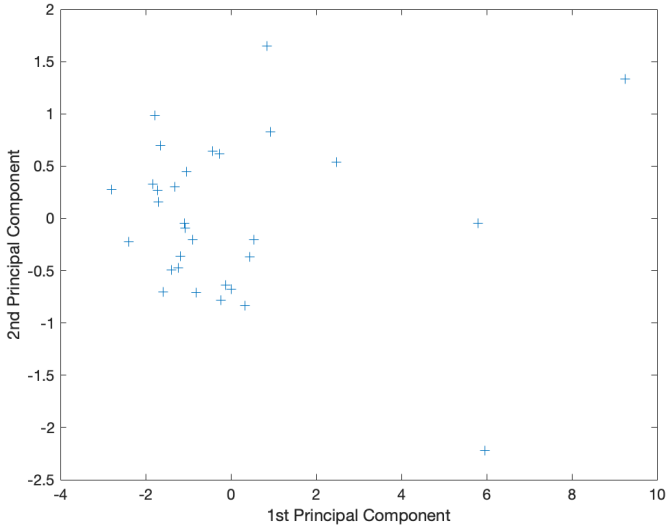
基本思想：

1. 将坐标轴中心移到数据的中心，然后旋转坐标轴，使得数据在C_1 轴上的方差最大，即全部 n 个数据个体在该方向上的投影最为分散。意味着更多的信息被保留下来。C_1成为第一主成分。
2. C_2 第二主成分：找一个C_2，使得C_2与C_1的协方差（相关系数）为 0，以免与C_1信息重叠，并且使数据在该方向的方差尽量最大。
3. 以此类推，找到第三主成分，第四主成分……第p个主成分。p个随机变量可以有p个主成分。

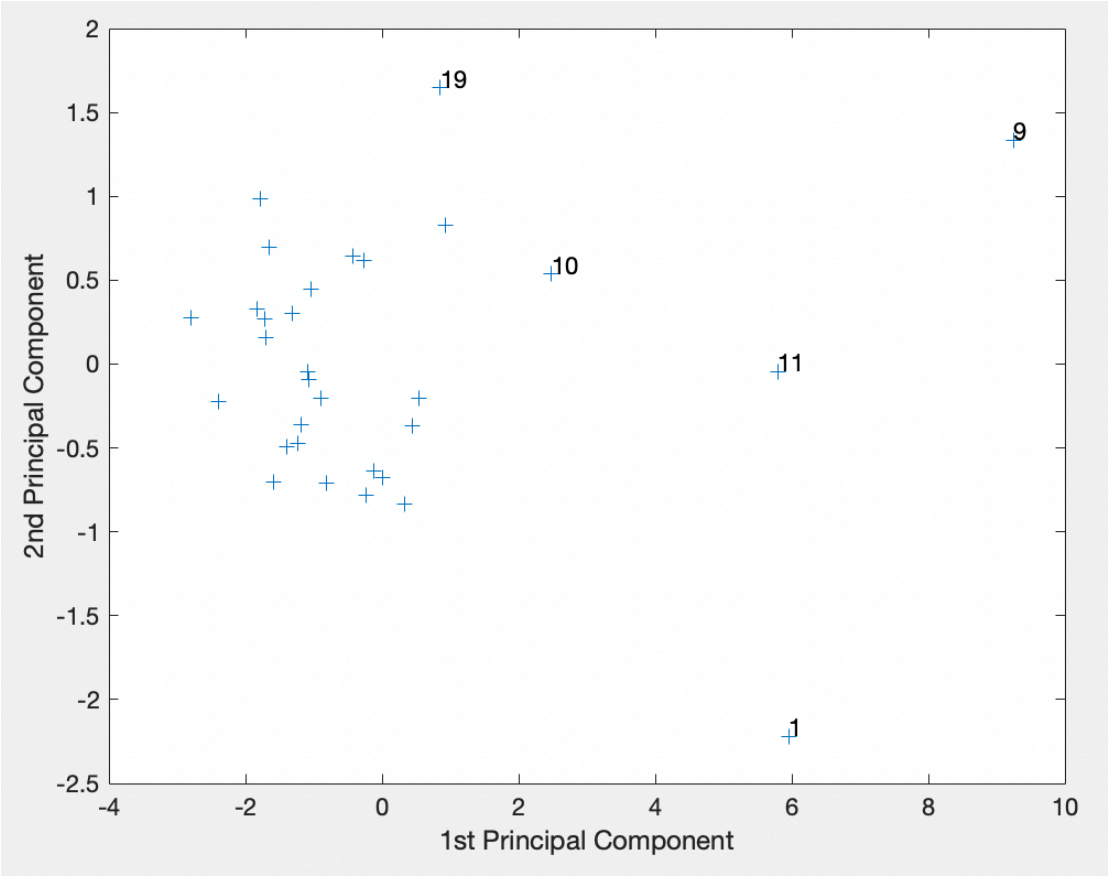
结果分析



根据以上箱型图可以看出食品开销是最大的，其次是居住开销，剩余的几类开销基本上相似。



上图为前两个主成分的图。下面使用 gname 命令确定的点。



因为在载入数据时，没有载入具体城市名，图像中只能显示城市标号，对应城市分别是：

1. 北京；
9. 上海；
10. 江苏；
11. 浙江；
19. 广东

这五个省份/市正是我国经济最发达的省/市，人口也较多。

去掉上面的五个省份/市，计算他们的方差和累积贡献率，分别如下：

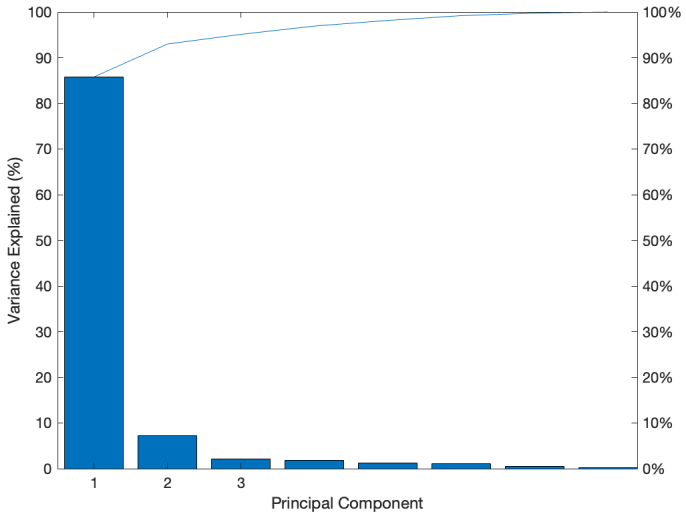
```
variances =

    6.8645
    0.5751
    0.1689
    0.1450
    0.0989
    0.0838
    0.0429
    0.0209
```

```
percent_explained =

   85.8068
    7.1889
    2.1115
    1.8121
    1.2359
    1.0477
    0.5362
    0.2609
```

最后利用 pareto 命令画出方差贡献率图。



实验题目 2.2

2.2 （该题建议采用 matlab） 参考资料中 TrainDatabase 图像库作为训练图像进行训练，该图像库中有 10 个人，每人有两幅图像，每幅图像大小为 $\mathbf{200\times180}$ 的。在 TestDatabase 中有 10 幅测试图像，其大小也为 $\mathbf{200\times180}$ 的。

1） 请利用 TrainDatabase 中的图像进行主成分分析，并将 TestDatabase 中的测试图像进行分类，选前三个最大的特征值。请显示特征脸，并给出分类依据；

2） 若选前两个和前四个最大的特征值，结果如何，请列出相应的结果。

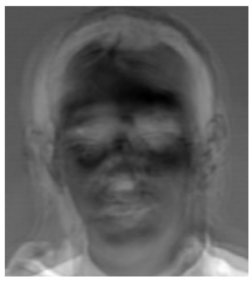
实验过程描述

得到特征值结果如下：

lambda = 19×1
10^7 ×

	1
1	3.0591
2	1.5884
3	0.6876
4	0.3895
5	0.3218
6	0.2863
7	0.2460
8	0.2266

得到特征脸结果如下：



分类依据：

考虑对上述数据的更高维逼近。此时考虑利用 d' 维数据（一般 $d'<d$ ）

$$x = m + \sum_{i=1}^{d'} a_i e_i \qquad (3)$$

来逼近原数据 $x_1, x_2, \cdots x_n$ 。也就是要确定 $e_1, e_2, \cdots e_{d'}$ 使下面准则极小

$$J_{d'} = \sum_{k=1}^{n} \left\| (m + \sum_{i=1}^{d'} a_{ki} e_i) - x_k \right\|^2$$

类似地可以证明，在 $e_1, e_2, \cdots e_{d'}$ 分别为散布矩阵 S 的 d' 个最大的特征值所对应的特征向量时，上式取得最小值。此时 a_{ki} 的值取为：

$$a_{ki} =< e_i, (x_k - m) >= e_i^T (x_k - m), \qquad i = 1,2,\cdots d' 。$$

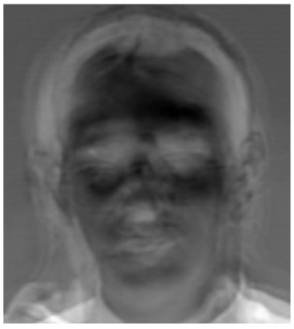
因为散布矩阵 S 为实对称矩阵，上面所求的 $e_1, e_2, \cdots e_{d'}$ 是相互正交的，也就是说它们是原数据的 d' 维逼近空间 $V = \{x \mid x = m + \sum_{i=1}^{d'} a_i e_i\}$ 的基向量。公式（3）中的 a_i 是向量 x 对应于 e_i 的系数，称该系数为主成分。也把原始数据在 d' 维空间 $V = \{x \mid x = m + \sum_{i=1}^{d'} a_i e_i\}$ 的逼近表示称为主成分分析。

选前两个特征值结果：

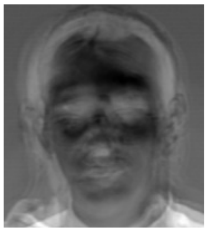
lambda = 19×1
10^7 ×

	1
1	3.0591
2	1.5884
3	0.6876
4	0.3895
5	0.3218
6	0.2863
7	0.2460
8	0.2266

根据数据知特征值不变
得到特征脸：



选前四个特征值的情况：
注意到特征值依旧不变，最后得到特征脸如下：



三、C 均值聚类部分

实验目的：掌握利用 C 均值聚类分析的方法。

实验题目 3.1

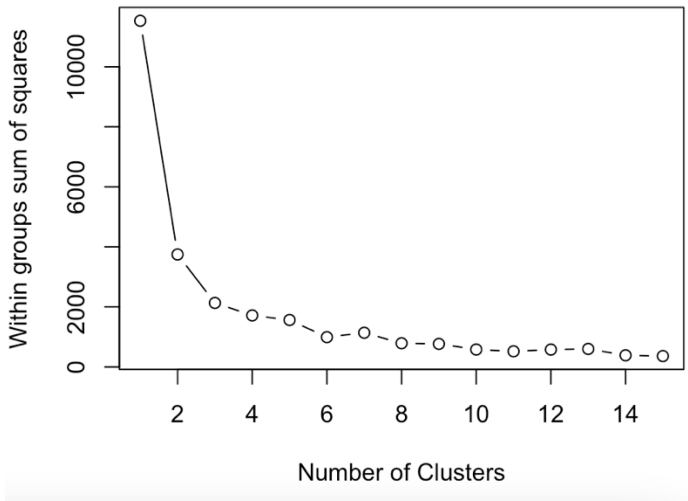
课本 220 页习题 6.5.

实验过程描述

在聚类分析中，K-means 聚类算法是最常用的，它需要分析者先确定要将这组数据分成多少类，也即聚类的个数，这个通常可以用因子分析的方法来确定。比如可以用“nFactors”包的函数来确定最佳的因子个数，将因子数作为聚类数，不过关于聚类个数的确定还要考虑数据的实际情况与自身需求，这样分析才会更具有现实意义。 另外，也可以通过绘制碎石图来确定聚类个数，这和主成分的思想相似。

结果分析

（1） 用$X=(X_2,X_4)^T$二个变量聚类

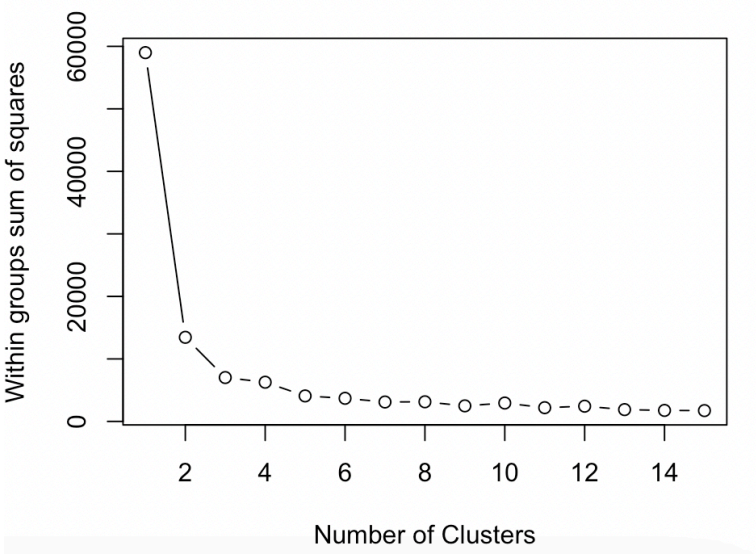


代码：

```
data31 <- data.frame(data3[,3],data3[,5])# 取后 X2,X4 列
wss <- (nrow(data31)-1)*sum(apply(data31,2,var)) # 计算离均差平方和
for (i in 2:15) wss[i] <- sum(kmeans(data31,
                                     centers=i)$withinss) #计算不同聚类个数的组内平方和
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares") # 绘图
# K-Means 聚类分析
fit1 <- kmeans(data31, 3) # 根据上面聚类图，聚类个数选为 3
# 获取聚类均值
aggregate(data31,by=list(fit1$cluster),FUN=mean) # aggregate()是一个分类汇总函数
res <- data.frame(data31, fit1$cluster)
```

```
> aggregate(data31,by=list(fit1$cluster),FUN=mean) # aggregate()是一个分类汇总函数
  Group.1 data3...3. data3...5.
1       1   30.41667  20.520833
2       2   34.56250   2.458333
3       3   26.88679  13.094340
```

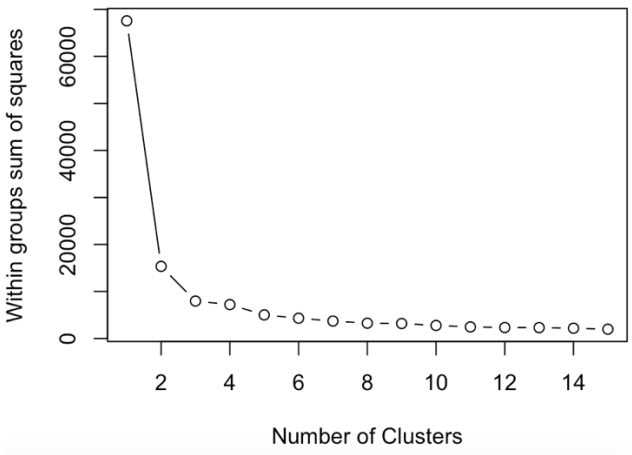
（2） 用$X=(X_1,X_2,X_3)^T$ 三个变量聚类



代码同上一致，一般需要控制组内平方和的值要小，同时聚类的个数也不能太多，所以从图中可以看出聚类个数定在 3 比较好。

```
> aggregate(data32,by=list(fit1$cluster),FUN=mean) # aggregate()是一个分类汇总函数
  Group.1      X50      X33      X14
1       1 58.46552 27.15517 43.63793
2       2 68.35714 30.64286 56.54762
3       3 50.06122 34.32653 14.63265
```

（3） 用$X=(X_1,X_2,X_3,X_4)^T$四个变量聚类



代码同上一致，一般需要控制组内平方和的值要小，同时聚类的个数也不能太多，所以从图中可以看出聚类个数定在 3 比较好。

```
> aggregate(data33,by=list(fit1$cluster),FUN=mean) # aggregate()是一个分类汇总函数
  Group.1      X50      X33      X14        X2
1       1 58.46552 27.15517 43.63793 14.327586
2       2 68.35714 30.64286 56.54762 20.119048
3       3 50.06122 34.32653 14.63265  2.469388
```

（4） 将以上各情况下的聚类结果与数据集中的实际分类情况比较，是否所用变量越多，聚类效果就越好？

答：从上述结果能看出并非使用变量越多效果越好。

实验题目 3.2

3.2 （该题建议采用matlab）自己拍一张照片，利用C均值算法进行图像的分割和向量量化（分类数自选）.

实验过程描述

C 均值算法是一种用于图像分割和向量量化的聚类算法。该算法的工作原理是将图像中的像素分成若干个聚类，每个聚类的像素具有相似的属性。

首先，将图像转换为灰度图像（如果需要的话），并将其转换为向量。接着，设定聚类数量（K）和最大迭代次数。然后，初始化隶属矩阵，并对其进行更新。在每次迭代中，首先更新聚类中心（也称为质心或均值），然后更新隶属矩阵。更新聚类中心时，需要计算每个像素到每个聚类中心的欧几里得距离，然后将像素分配给距离最小的聚类。更新隶属矩阵时，需要计算每个像素在每个聚类中的隶属度。

算法迭代过程中会不断更新聚类中心和隶属矩阵，直到满足收敛条件为止。收敛条件可以是达到最大迭代次数或者隶属矩阵的变化小于某个阈值。

最后，利用隶属矩阵对图像进行量化。将每个像素分配给隶属度最大的聚类，然后将像素赋值为该聚类的聚类中心。将量化后的图像转换回其原始大小，并显示原图像和量化后的图像。

注意，需要仔细选择聚类数量（K）和最大迭代次数，以确保算法性能良好。此外，还可以调整收敛阈值，以控制所需的精度水平。

原图：

参考代码（examp_quanity.m, examp_seq.m）首先通过将每个像素分配给具有最高成员值的集群来量化图像。然后它将量化图像重塑回其原始大小并显示原始图像和量化图像。

结果分析

图像量化结果：


BELLINGHAM
22


BELLINGHAM
22


BELLINGHAM
22


BELLINGHAM
22


BELLINGHAM
22


BELLINGHAM