**云南大学数学与统计学院**

**实验报告**

**实验课名称： 应用多元统计分析实验**

**指导教师： 李会琼**

**专业（年级）： 统计学2021级**

**学生姓名： 枫叶 学号:**

**实验名称： 实验11**

**实验成绩：**

**《应用多元统计分析实验》实验报告 11**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验名称 | 多元正态分布均值向量的检验 | | 实验成绩 |  | |
| 学号 |  | | 姓名 | 枫叶 | |
| 实验时间 | 2024年6月7日 | 实验地点 | 格物楼3508 | 指导教师 | 李会琼 |
| 1. **实验目的**   学习使用R软件进行距离判别   1. **实验要求**   1. 对所使用的方法与所得到的结果进行适当的文字描述。  2. 在实验结果的相应部分附上完整的代码与适当的注释。  3. 采用一定的可视化方法体现出对应计算结果。   1. **实验内容**   习题7-11 习题7-12 例题7.3.3   1. **实验软件**   R语言   1. **实验结果**  加载包 library(readxl) library(dplyr) library(NbClust) 第一题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题7-11数据.xlsx") result <- select(data,-1) %>%  apply(2,scale) %>%  princomp() result$scores  ## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 ## [1,] 1.4752359 0.7586330 0.53804346 0.48981582 1.05856731 -0.002569752 ## [2,] 0.4982149 -2.5916439 0.22831171 0.85190555 0.16056855 -0.291077912 ## [3,] 1.0564432 -3.2255272 0.40941148 0.58246589 -0.92996315 0.059369772 ## [4,] 0.4598646 1.1836386 -0.99768111 1.59955537 0.01137638 0.074563880 ## [5,] 4.5284817 2.2624411 0.46764480 -0.75806395 -0.49630475 0.019121841 ## [6,] 0.3299732 -1.7736147 0.03113872 -0.93799882 0.36892036 0.206182306 ## [7,] -1.1025047 -0.3179334 0.28182618 -0.69170879 0.09142562 0.303286933 ## [8,] -2.1949756 2.2441391 1.09921449 0.55675111 -0.57189922 0.011282076 ## [9,] -0.8411700 0.8956587 0.35286803 0.12852355 0.52663782 -0.468728089 ## [10,] -2.0318632 0.8251577 0.23105405 -0.51413084 -0.64747223 -0.178638164 ## [11,] -0.7133301 -0.7556065 -0.12255137 -1.11097884 0.23427567 -0.382215050 ## [12,] -1.2014056 0.0303427 0.28702019 0.08174798 0.37043757 0.642349281 ## [13,] -0.2629643 0.4643149 -2.80630062 -0.27788402 -0.17656992 0.007072879 ## Comp.7 Comp.8 ## [1,] 0.394863770 0.0044105661 ## [2,] -0.127222867 0.0668743974 ## [3,] 0.082199156 -0.0240029426 ## [4,] -0.008633944 -0.0519754393 ## [5,] -0.121095417 0.0225801112 ## [6,] -0.027275780 -0.0668050892 ## [7,] -0.005088581 -0.0350216647 ## [8,] -0.039905790 -0.0524413761 ## [9,] -0.288181991 -0.0009076448 ## [10,] 0.279388740 0.0727275606 ## [11,] 0.017817773 -0.0295066733 ## [12,] -0.169328425 0.0786371100 ## [13,] 0.012463357 0.0154310848  样本主成分如上所示  summary(result)  ## Importance of components: ## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 ## Standard deviation 1.6929480 1.6354087 0.9266394 0.76988846 0.52980472 ## Proportion of Variance 0.3881141 0.3621802 0.1162769 0.08026528 0.03801052 ## Cumulative Proportion 0.3881141 0.7502943 0.8665712 0.94683649 0.98484701 ## Comp.6 Comp.7 Comp.8 ## Standard deviation 0.28273025 0.172362004 0.0474757436 ## Proportion of Variance 0.01082472 0.004023048 0.0003052219 ## Cumulative Proportion 0.99567173 0.999694778 1.0000000000  若要求损失信息不超过15%，则应选择前三个主成分，下面根据因子载荷考察其含义  result$loadings[1:8,1:3]^2 %>%  apply(1,sum)  ## X1 X2 X3 X4 X5 X6 X7 X8  ## 0.3256617 0.3273362 0.3469077 0.2491654 0.3635518 0.3074686 0.2083248 0.8715838  可以看到前三个主成分对大部分变量的解释力都不高，仅对变量8（能源利用率）有较强解释力  下面基于第一主成分对样本进行排序，理论上来说这是不稳妥的，因为第一主成分中有正有负，且考虑指标的正向性和负向性之后依旧不能统一方向，但目前暂无其他方法实现排序，故使用第一主成分的排序结果如下  order(result$scores[,1])  ## [1] 8 10 12 7 9 11 13 6 4 2 3 1 5  下面进一步进行分类，首先基于前两个主成分画出散点图来观察是否有明显聚类特征  plot(result$scores[,1:2])    由于散点图没有明显的类别特征，下面使用tracew统计量决定最佳类别数为3，并使用kmeans方法进行聚类  NbClust(result$scores[,1:2],distance = "euclidean",method = "average",min.nc=2,max.nc=10,index = "tracew")  ## $All.index ## 2 3 4 5 6 7 8 9 10  ## 44.2670 18.2875 10.6593 5.1927 4.1727 3.1725 1.6581 1.0523 0.7035  ##  ## $Best.nc ## Number\_clusters Value\_Index  ## 3.0000 18.3513  ##  ## $Best.partition ## [1] 1 2 2 1 3 2 1 1 1 1 1 1 1  model <- kmeans(result$scores[,1:2],3) data %>%  mutate(group=model$cluster) %>%  select(1,10)  ## # A tibble: 13 × 2 ## 行业 group ## <chr> <int> ## 1 1(冶金) 2 ## 2 2(电力) 1 ## 3 3(煤炭) 1 ## 4 4(化学) 2 ## 5 5(机械) 3 ## 6 6(建材) 1 ## 7 7(森工) 2 ## 8 8(食品) 2 ## 9 9(纺织) 2 ## 10 10(缝纫) 2 ## 11 11(皮革) 2 ## 12 12(造纸) 2 ## 13 13(文教艺术用品) 2  聚类结果如上所示，可以看到第一类为冶金、电力、煤炭、机械、建材几个行业，可以视之为重工业组，第二类为食品和缝纫，可以视之为生活用品组，第三类为化学、森工、纺织、皮革、造纸、文教艺术用品，可以视之为轻工业组 第二题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计表6.7数据.xlsx") result <- select(data,-1) %>%  apply(2,scale) %>%  princomp() result$scores[,1:2] %>%  plot()    与第一题一样地，观察前两个成分的散点图，有一定的类别轮廓，可以尝试分为四个类  model <- kmeans(result$scores[,1:2],4) data %>%  select(1) %>%  mutate(主成分聚类=model$cluster)  ## # A tibble: 16 × 2 ## 地区 主成分聚类 ## <chr> <int> ## 1 北京 2 ## 2 天津 4 ## 3 河北 3 ## 4 山西 3 ## 5 内蒙古 3 ## 6 辽宁 4 ## 7 吉林 1 ## 8 黑龙江 1 ## 9 上海 2 ## 10 江苏 4 ## 11 浙江 4 ## 12 安徽 1 ## 13 福建 1 ## 14 江西 1 ## 15 山东 4 ## 16 河南 3  其中第一类为北京、上海，第二类为天津、辽宁、江苏、浙江、山东，第三类为河北、山西、内蒙古、河南，第四类为吉林、黑龙江、安徽、福建、江西。其中第三类与基于average方法的系统聚类结果是完全一致的，而第二类与第四类在系统聚类中被分为了一类，北京和上海在系统中被分为了两类 第三题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计表7.6数据.xlsx") %>%  select(-1) result <- select(data,-4) %>%  apply(2,scale) %>%  princomp() summary(result)  ## Importance of components: ## Comp.1 Comp.2 Comp.3 ## Standard deviation 1.348115 0.9525822 0.0494597123 ## Proportion of Variance 0.666385 0.3327181 0.0008969632 ## Cumulative Proportion 0.666385 0.9991030 1.0000000000  和教材一致，前两个主成分的累积贡献率在99%以上  p\_data <- data.frame(y=scale(data$Y),result$scores[,1:2]) model <- lm(y~.-1,p\_data)  标准化变量下的回归系数如下所示  b\_scale <- c(result$loadings[1:3,1:2] %\*% model$coefficients) b\_scale  ## [1] 0.4805280 0.2211323 0.4825616  原始变量下的回归系数如下所示  b <- (b\_scale\*sd(data$Y)/apply(data[1:3],2,sd)) b  ## X1 X2 X3  ## 0.07277981 0.60922012 0.10625939  截距项为  mean(data$Y)-sum(b\*apply(data[1:3],2,mean))  ## [1] -9.130108 | | | | | |