**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6**

**实验成绩：**

**《应用多元统计分析实验》实验报告 6**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验名称 |  | | 实验成绩 |  | |
| 学号 |  | | 姓名 | 枫叶 | |
| 实验时间 | 2024年5月14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5-9  数据做距离判别 习题5-10 （1） 判别分析作业9   1. **实验软件**   R语言   1. **实验结果**  加载包 library(MASS) library(readxl) library(dplyr) library(purrr) #加载相关函数 source("正态总体假设检验函数汇集.R") source("距离判别函数汇集.R") 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5-9数据.xlsx") %>%  select(-"序号") discriminiant.distance(data[1:7,2:4],data[8:14,2:4])  ## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 ## blong 1 1 1 1 1 1 1 2 2 2 2 2 2 2 第二题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题5-10数据.xlsx") %>%  select(-"样品号") %>%  mutate(across(5,as.factor)) %>%  arrange(类别号) d <- group\_split(data,类别号) %>%  map(select,1:4) data\_new <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题5-10数据.xlsx",sheet = 2) %>%  select(-"新样本") #实验手册中的函数先计算似然比统计量再取对数，容易数值溢出，这里参考了李高荣多元统计教材中的函数，直接计算对数形式的统计量 multi.cov.test(data[,1:4],data$类别号,3)  ## $p.value ## [1] 0.04509344 ##  ## $correct.M ## [1] 21.37938  检验结果显示在0.05的显著性水平下应该拒绝原假设，即认为各总体协方差阵不完全相等，但经尝试，发现第二类的样本协方差阵接近奇异，如下所示，且尝试加上一个单位阵也无法改善奇异情况。  filter(data,类别号==2) %>%  select(-5) %>%  #as.matrix() %>%  #+diag(1,4) %>%  var() %>%  det()  ## [1] -4.099574e-10  由于样本协方差阵奇异，无法使用马氏距离进行判别，下面对函数作了修改，当协方差阵接近奇异时，使用shrink估计代替正常的估计量，由于对该估计方法的理论没有了解，无法判断会对判别效率产生什么影响  table(distinguish.distance.correct(data[,1:4],data$类别号) == data$类别号)  ## Estimating optimal shrinkage intensity lambda.var (variance vector): 0.3921  ##  ## Estimating optimal shrinkage intensity lambda (correlation matrix): 0.4978  ##  ## FALSE TRUE  ## 6 11  distinguish.distance.correct(data[,1:4],data$类别号,data\_new)  ## Estimating optimal shrinkage intensity lambda.var (variance vector): 0.3921  ##  ## Estimating optimal shrinkage intensity lambda (correlation matrix): 0.4978  ## 1 2 3 ## blong 1 2 3  在训练集上判别正确11个，错误6个，正确率为64.7%，对新样本的判别结果为1,2,3。可以看到判别结果较差。  cor(data[1:7,1:4])  ## X1 X2 X3 X4 ## X1 1.0000000 0.4287099 0.5090015 0.4877708 ## X2 0.4287099 1.0000000 0.9091831 0.9634985 ## X3 0.5090015 0.9091831 1.0000000 0.9299444 ## X4 0.4877708 0.9634985 0.9299444 1.0000000  cor(data[8:11,1:4])  ## X1 X2 X3 X4 ## X1 1.0000000 0.6206161 0.4333895 0.5961613 ## X2 0.6206161 1.0000000 0.9677675 0.9995225 ## X3 0.4333895 0.9677675 1.0000000 0.9735623 ## X4 0.5961613 0.9995225 0.9735623 1.0000000  cor(data[12:17,1:4])  ## X1 X2 X3 X4 ## X1 1.0000000 0.1179923 -0.58430030 0.52162750 ## X2 0.1179923 1.0000000 0.47079812 0.86956238 ## X3 -0.5843003 0.4707981 1.00000000 0.01890974 ## X4 0.5216275 0.8695624 0.01890974 1.00000000  考虑指标间的共线性问题，发现总体1中X2与X3和X4的相关性极强，总体2中X2、X3和X4之间的相关性极强，总体3中X2和X4的相关性极强，综合三个总体的情况，考虑删去X2  table(distinguish.distance(data[,c(1,3,4)],data$类别号) == data$类别号)  ##  ## FALSE TRUE  ## 2 15  删去X2后正确率达到了88.2%，进一步用该方案判别新样本  distinguish.distance(data[,c(1,3,4)],data$类别号,data\_new[,c(1,3,4)])  ## 1 2 3 ## blong 1 2 3  判别结果依旧为1,2,3 第三题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计判别分析作业9数据.xlsx") %>%  dplyr::select(-c(1,2)) %>%  mutate(across(9,as.factor)) data\_new <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计判别分析作业9数据.xlsx",sheet = 2) %>%  dplyr::select(-c(1,2)) model <- lda(Group~.,data,prior=c(1,1,1)/3) predict(model,data\_new)$class  ## [1] 2 3 ## Levels: 1 2 3  天津和陕西分别被判为第二类和第三类 第四题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计判别分析作业9数据.xlsx",sheet = 3) %>%  dplyr::select(-c(1,2)) %>%  mutate(across(9,as.factor)) data\_new <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计判别分析作业9数据.xlsx",sheet = 4) %>%  dplyr::select(-c(1,2)) model <- lda(Group~.,data,prior=c(1,1)/2) predict(model,data\_new)$class  ## [1] 2 2 ## Levels: 1 2  广东和西藏均被判为第二类 | | | | | |