Homework 3

Arturo Ortiz

2023-11-30

# Problem 1

r <- matrix(c(1.00,0.44,0.41,0.29,0.33,0.25,  
 0.44,1.00,0.35,0.35,0.32,0.33,  
 0.41,0.35,1.00,0.16,0.19,0.18,  
 0.29,0.35,0.16,1.00,0.59,0.47,  
 0.33,0.32,0.19,0.59,1.00,0.46,  
 0.25,0.33,0.18,0.47,0.46,1.00), nrow = 6, byrow = TRUE)  
colnames(r) <- c("French", "English", "History", "Arithmetic", "Algebra", "Geometry")  
rownames(r) <- c("French", "English", "History", "Arithmetic", "Algebra", "Geometry")  
r

## French English History Arithmetic Algebra Geometry  
## French 1.00 0.44 0.41 0.29 0.33 0.25  
## English 0.44 1.00 0.35 0.35 0.32 0.33  
## History 0.41 0.35 1.00 0.16 0.19 0.18  
## Arithmetic 0.29 0.35 0.16 1.00 0.59 0.47  
## Algebra 0.33 0.32 0.19 0.59 1.00 0.46  
## Geometry 0.25 0.33 0.18 0.47 0.46 1.00

## A

Students who performed well in the areas of geometry, algebra, and arithmetic can be used to understand factor 1. Students who performed well in the areas of French, English, and history can be used to interpret factor 2.

efa <- factanal(covmat = r, factors = 2)  
print(efa, cut = 0.5)

##   
## Call:  
## factanal(factors = 2, covmat = r)  
##   
## Uniquenesses:  
## French English History Arithmetic Algebra Geometry   
## 0.508 0.595 0.644 0.377 0.440 0.628   
##   
## Loadings:  
## Factor1 Factor2  
## French 0.661   
## English 0.551   
## History 0.591   
## Arithmetic 0.770   
## Algebra 0.715   
## Geometry 0.570   
##   
## Factor1 Factor2  
## SS loadings 1.593 1.215  
## Proportion Var 0.265 0.202  
## Cumulative Var 0.265 0.468  
##   
## The degrees of freedom for the model is 4 and the fit was 0.0101

## B

The class categories with the largest loadings can be used to name these factors. Math might be the first factor. Humanities can be recognized as Factor 2. Someone might suggest that factor 3 is trigonometry.

efa2 <- factanal(covmat = r, factors = 3)  
efa2

##   
## Call:  
## factanal(factors = 3, covmat = r)  
##   
## Uniquenesses:  
## French English History Arithmetic Algebra Geometry   
## 0.448 0.497 0.679 0.411 0.376 0.611   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## French 0.229 0.706   
## English 0.295 0.538 0.355   
## History 0.554   
## Arithmetic 0.741 0.169 0.109   
## Algebra 0.756 0.222   
## Geometry 0.567 0.185 0.183   
##   
## Factor1 Factor2 Factor3  
## SS loadings 1.588 1.207 0.181  
## Proportion Var 0.265 0.201 0.030  
## Cumulative Var 0.265 0.466 0.496  
##   
## The degrees of freedom for the model is 0 and the fit was 0.001

## C

The orthogonal rotation known as Varimax rotation is used. This rotation reveals which variables have a variance that is closest to one or zero among the resulting components. This shows which correlations between the components are high or low.

## D

Students who performed well in every subject except history might be used to understand factor 1. Students who performed well in French, English, and history but poorly in arithmetic, algebra, and geometry might be attributed to factor 2.

efa.nr <- factanal(covmat = r, factors = 2, rotation = "none")  
efa.nr

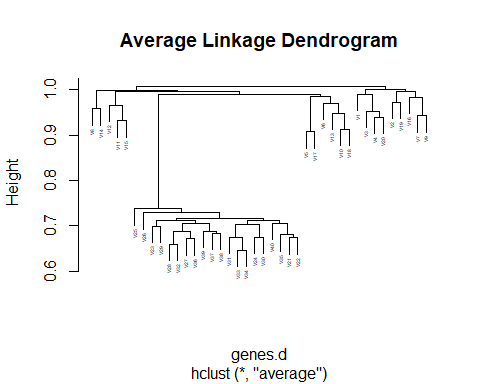
##   
## Call:  
## factanal(factors = 2, covmat = r, rotation = "none")  
##   
## Uniquenesses:  
## French English History Arithmetic Algebra Geometry   
## 0.508 0.595 0.644 0.377 0.440 0.628   
##   
## Loadings:  
## Factor1 Factor2  
## French 0.558 0.425   
## English 0.569 0.286   
## History 0.392 0.450   
## Arithmetic 0.738 -0.279   
## Algebra 0.718 -0.209   
## Geometry 0.595 -0.133   
##   
## Factor1 Factor2  
## SS loadings 2.204 0.603  
## Proportion Var 0.367 0.101  
## Cumulative Var 0.367 0.468  
##   
## The degrees of freedom for the model is 4 and the fit was 0.0101

# Problem 2

genes <- read.csv("https://raw.githubusercontent.com/EricBrownTTU/ISQS6350/main/geneexpression.csv", header = FALSE)

## A

genes.d <- as.dist(1 - cor(genes))  
genes.hca <- hclust(genes.d, "average")  
plot(genes.hca, main = "Average Linkage Dendrogram", cex = 0.3)



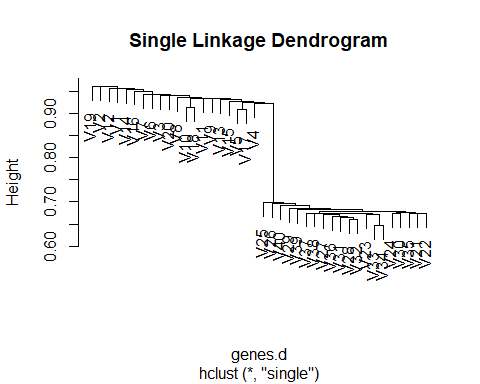
## B

The samples are divided into two groups based on genes. The patients in good health and the sick group are the two groups.

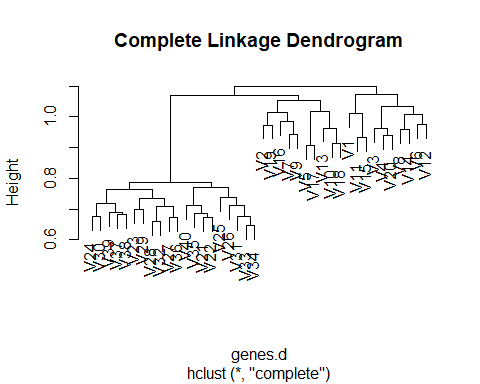
## C

Since both the Single Linkage and Complete Linkage indicate that the genes first split into two groups, this has no bearing on the conclusion drawn in section B.

genes.hcs <- hclust(genes.d, "single")  
plot(genes.hcs, main = "Single Linkage Dendrogram")

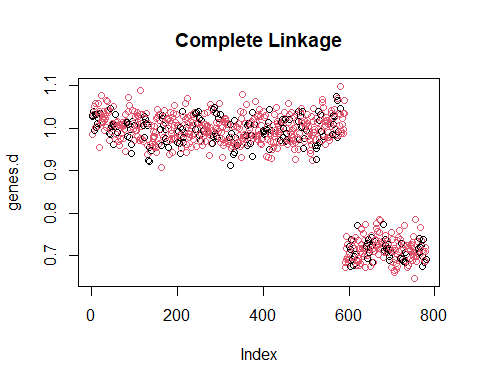


genes.hcc <- hclust(genes.d, "complete")  
plot(genes.hcc, main = "Complete Linkage Dendrogram")



## D

genes.ct <- cutree(genes.hcc, 2)  
plot(genes.d, col = genes.ct, main = "Complete Linkage")



# Problem 3

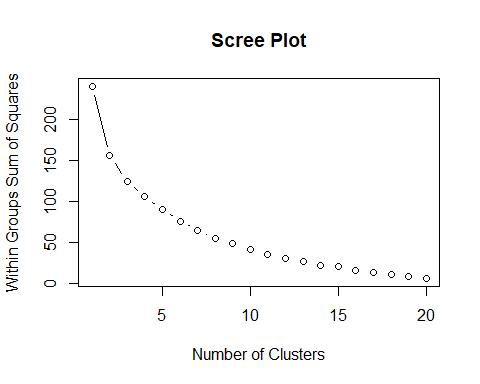
urlRemote <- "https://raw.githubusercontent.com/"  
pathGithub <- "EricBrownTTU/ISQS6350/main/"  
filename <- "protein.csv"  
protein1 <- read.csv(paste0(urlRemote, pathGithub, filename))  
protein <- protein1[,-1]

## A

protein.s <- scale(protein)  
pro.km <- kmeans(protein.s, centers = 4, nstart = 10)  
pro.km$tot.withinss

## [1] 105.5389

pro.wgss <- function(mydata, maxc){  
 wss <- numeric(maxc)  
 for (i in 1:maxc){  
 wss[i] <- kmeans(mydata, iter.max = 100,  
 centers = i, nstart = 10)$tot.withinss  
 }  
 plot(1:maxc, wss, type = "b",  
 xlab = "Number of Clusters",  
 ylab = "Within Groups Sum of Squares",  
 main = "Scree Plot")  
}  
pro.wgss(protein.s, maxc = 20)



## B

Every nation with an identical number underneath it belongs to the same cluster. Belgium, Denmark, Finland, France, Ireland, Norway, Sweden, and the United Kingdom make up Cluster 1. Spain and Portugal are in Cluster 2. Austria, Czechoslovakia, East Germany, The Netherlands, Poland, Switzerland, and West Germany are all included in Cluster 3. Albania, Bulgaria, Greece, Hungary, Italy, Romania, USSR, and Yugoslavia are included in Cluster 4.

protein2 <- read.csv(paste0("https://raw.githubusercontent.com/", "EricBrownTTU/ISQS6350/main/protein.csv"), row.names = "Country")  
protein2.s <- scale(protein2)  
protein2.km <- kmeans(protein2.s, centers = 4, nstart = 10)  
protein2.km$cluster

## Albania Austria Belgium Bulgaria Czechoslovakia   
## 4 1 1 4 1   
## Denmark East\_Germany Finland France Greece   
## 2 1 2 1 4   
## Hungary Ireland Italy The\_Netherlands Norway   
## 4 1 4 1 2   
## Poland Portugal Romania Spain Sweden   
## 1 3 4 3 2   
## Switzerland United\_Kingdom USSR West\_Germany Yugoslavia   
## 1 1 4 1 4

## C

Protein averages for red and white meat, milk, fish, cereals, and starchy foods are higher in Cluster 1, indicating that these nations consume a lot of animal products. With a small bit of red meat, nuts/seeds, and starchy foods, Cluster 2 has high average protein content in cereals, fish, eggs, and fruits/veggies. This indicates that the nations in this cluster have certain places that are more accessible to land or water, which affects how much protein they eat. The abundance of grains, milk, and red and white meat in Cluster 3 suggests these nations will be slightly more landlocked than those in Cluster 2. Cluster 4’s main diet consists of milk and grains, with traces of red and white meat as well as nuts and seeds, indicating that these nations get their protein from animals.

aggregate(protein2, by = list(cluster = protein2.km$cluster), mean)

## cluster Red\_Meat White\_Meat Eggs Milk Fish Cereals  
## 1 1 11.90909 10.81818 3.818182 19.45455 3.363636 26.63636  
## 2 2 10.00000 7.25000 3.500000 26.75000 8.500000 22.75000  
## 3 3 6.50000 3.50000 2.000000 7.00000 10.500000 28.00000  
## 4 4 7.62500 5.37500 2.125000 12.12500 1.875000 46.00000  
## Starchy\_Foods Pulses\_nuts\_oilseeds Fruits\_Vegetables Total  
## 1 5.090909 1.818182 4.454545 87.36364  
## 2 4.750000 1.250000 2.000000 86.75000  
## 3 6.000000 5.500000 7.500000 76.50000  
## 4 2.750000 5.125000 4.125000 87.12500

## D

Two clusters were selected for the data by the model-based clustering technique. Albania, Bulgaria, Greece, Hungary, Italy, Portugal, Romania, Spain, USSR, and Yugoslavia are all included in Cluster 1. Cluster 2 includes the following countries: Denmark, East Germany, Norway, Poland, Sweden, Switzerland, The Netherlands, United Kingdom, West Germany, Austria, Belgium, Czechoslovakia, and Finland.

library(mclust)

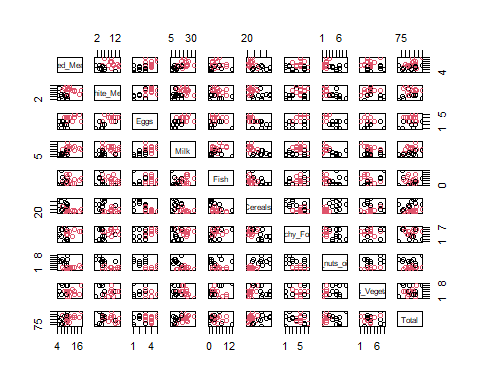
## Warning: package 'mclust' was built under R version 4.3.2

## Package 'mclust' version 6.0.1  
## Type 'citation("mclust")' for citing this R package in publications.

pro.mc <- Mclust(protein2[,1:9])  
table(pro.mc$classification, protein1$Country)

##   
## Albania Austria Belgium Bulgaria Czechoslovakia Denmark East\_Germany  
## 1 1 0 0 1 0 0 0  
## 2 0 1 1 0 1 1 1  
##   
## Finland France Greece Hungary Ireland Italy Norway Poland Portugal Romania  
## 1 0 0 1 1 0 1 0 0 1 1  
## 2 1 1 0 0 1 0 1 1 0 0  
##   
## Spain Sweden Switzerland The\_Netherlands United\_Kingdom USSR West\_Germany  
## 1 1 0 0 0 0 1 0  
## 2 0 1 1 1 1 0 1  
##   
## Yugoslavia  
## 1 1  
## 2 0

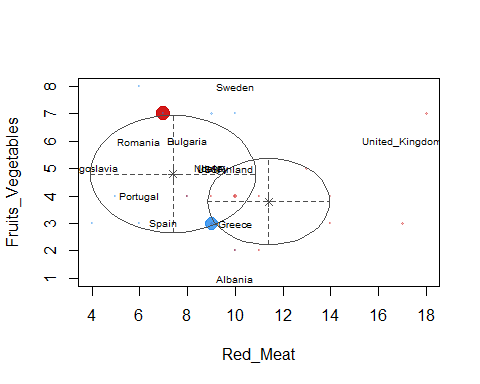
plot(protein2, col = pro.mc$classification)



## E

The most uncertain countries are USSR from cluster 1 and Poland from cluster 2. Poland had an uncertainty of 0.29%, whereas the USSR had 0.23%.

pro.mc2 <- Mclust(protein2[,1:9])  
plot(pro.mc2, what = "uncertainty", dimens = c(1,9))  
text(protein2, labels = rownames(protein2), cex = 0.6)



protein2.clust.data <- cbind(rownames(protein2), pro.mc2$classification,  
 pro.mc2$uncertainty)  
protein2.clust.data[order(pro.mc2$uncertainty),]

## [,1] [,2] [,3]   
## Albania "Albania" "1" "0"   
## Greece "Greece" "1" "0"   
## Yugoslavia "Yugoslavia" "1" "0"   
## Bulgaria "Bulgaria" "1" "1.61870516990348e-13"  
## Romania "Romania" "1" "4.03876931898139e-12"  
## Portugal "Portugal" "1" "4.29301039162056e-12"  
## Denmark "Denmark" "2" "7.63342722365223e-10"  
## Spain "Spain" "1" "1.57576796055992e-09"  
## Ireland "Ireland" "2" "2.77050693497927e-09"  
## Sweden "Sweden" "2" "4.47517911617012e-09"  
## West\_Germany "West\_Germany" "2" "2.52977037096258e-08"  
## The\_Netherlands "The\_Netherlands" "2" "4.35199024595434e-08"  
## Austria "Austria" "2" "4.6187358493377e-08"   
## Belgium "Belgium" "2" "2.26013116444435e-07"  
## East\_Germany "East\_Germany" "2" "2.96609156125172e-07"  
## Finland "Finland" "2" "3.0318905142046e-07"   
## Hungary "Hungary" "1" "3.15078838353244e-07"  
## United\_Kingdom "United\_Kingdom" "2" "3.68110132986654e-06"  
## Switzerland "Switzerland" "2" "5.96245735695611e-06"  
## France "France" "2" "7.60263417864948e-06"  
## Norway "Norway" "2" "2.1398613976209e-05"   
## Italy "Italy" "1" "2.2088801571396e-05"   
## Czechoslovakia "Czechoslovakia" "2" "3.47228371009312e-05"  
## USSR "USSR" "1" "0.00233632645322546"   
## Poland "Poland" "2" "0.00289318250546344"

# Problem 4

courses <- read.csv("https://raw.githubusercontent.com/EricBrownTTU/ISQS6350/main/Coursetopics.csv", header = TRUE)  
courses <- as.matrix(courses)  
library(arules)

## Warning: package 'arules' was built under R version 4.3.2

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

## A

There were 179 rules found based on the support, confidence, and minlen criteria applied to this collection of association rules.

myrules <- apriori(courses, parameter = list(support = 0.01, confidence = 0.25, minlen = 1))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.25 0.1 1 none FALSE TRUE 5 0.01 1  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 3   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[8 item(s), 365 transaction(s)] done [0.00s].  
## sorting and recoding items ... [8 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 done [0.00s].  
## writing ... [179 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

## B

The lift indicates that students are almost four times more likely to take the Data Mining course if they take the intro, regression, and prediction courses. Lift = 4.01, support = 0.014, and confidence = 0.71.

inspect(sort(myrules, by = "lift")[1])

## lhs rhs support confidence  
## [1] {Intro, Regression, Forecast} => {DataMining} 0.01369863 0.7142857   
## coverage lift count  
## [1] 0.01917808 4.010989 5

## C

The five rows represent the five possible classes a student might enroll in: Cat.Data, Forecast, DOE, Data Mining, and SW. Given that students are 2.4 times more likely to take SW following Intro and Regression, it is advised that they enroll in it. Additionally, SW has the most confidence (0.54).

course.rule <- subset(myrules, subset = lhs %ain% c("Intro", "Regression"))  
inspect(course.rule[1:5])

## lhs rhs support confidence coverage   
## [1] {Intro, Regression} => {Forecast} 0.01917808 0.2692308 0.07123288  
## [2] {Intro, Regression} => {DOE} 0.02465753 0.3461538 0.07123288  
## [3] {Intro, Regression} => {DataMining} 0.02739726 0.3846154 0.07123288  
## [4] {Intro, Regression} => {Cat.Data} 0.03287671 0.4615385 0.07123288  
## [5] {Intro, Regression} => {SW} 0.03835616 0.5384615 0.07123288  
## lift count  
## [1] 1.926848 7   
## [2] 2.005495 9   
## [3] 2.159763 10   
## [4] 2.216599 12   
## [5] 2.426401 14