Principal Component Analysis

By Saba Akram

12/22/2020

houses <- read.table("C:/Users/sabaa/Downloads/cadata.txt", quote="\"", comment.char="")  
head(houses)

## V1 V2 V3 V4 V5 V6 V7 V8 V9  
## 1 452600 8.3252 41 880 129 322 126 37.88 -122.23  
## 2 358500 8.3014 21 7099 1106 2401 1138 37.86 -122.22  
## 3 352100 7.2574 52 1467 190 496 177 37.85 -122.24  
## 4 341300 5.6431 52 1274 235 558 219 37.85 -122.25  
## 5 342200 3.8462 52 1627 280 565 259 37.85 -122.25  
## 6 269700 4.0368 52 919 213 413 193 37.85 -122.25

names(houses)<-c("MVAL","MINC","HAGE","ROOMS","BEDRMS","POPN","HHLDS","LAT","LONG")  
head(houses)

## MVAL MINC HAGE ROOMS BEDRMS POPN HHLDS LAT LONG  
## 1 452600 8.3252 41 880 129 322 126 37.88 -122.23  
## 2 358500 8.3014 21 7099 1106 2401 1138 37.86 -122.22  
## 3 352100 7.2574 52 1467 190 496 177 37.85 -122.24  
## 4 341300 5.6431 52 1274 235 558 219 37.85 -122.25  
## 5 342200 3.8462 52 1627 280 565 259 37.85 -122.25  
## 6 269700 4.0368 52 919 213 413 193 37.85 -122.25

houses$MINC\_Z<-(houses$MINC- mean(houses$MINC))/(sd(houses$MINC))  
houses$HAGE\_Z<-(houses$HAGE- mean(houses$HAGE))/(sd(houses$HAGE))  
houses$ROOMS\_Z<-(houses$ROOMS-mean(houses$ROOMS))/(sd(houses$ROOMS))  
houses$BEDRMS\_Z<-(houses$BEDRMS-mean(houses$BEDRMS))/(sd(houses$BEDRMS))  
houses$POPN\_Z<-(houses$POPN-mean(houses$POPN))/(sd(houses$POPN))  
houses$POPN\_Z<-(houses$POPN-mean(houses$POPN))/(sd(houses$POPN))  
houses$HHLDS\_Z<-(houses$HHLDS-mean(houses$HHLDS))/(sd(houses$HHLDS))  
houses$LAT\_Z<-(houses$LAT-mean(houses$LAT))/(sd(houses$LAT))  
houses$LONG\_Z<-(houses$LONG-mean(houses$LONG))/(sd(houses$LONG))  
  
# Randomly Selecting 90% of the training data sets.  
   
choose<-runif(dim(houses)[1],0,1)  
test.house<-houses[which(choose<.1),]  
train.house<-houses[which(choose>=.1),]  
  
# Principal Component Analysis.  
library(psych)  
pcal<-principal(train.house[,c(10:17)],nfactors = 8,rotate = "none",scores = TRUE)  
pcal

## Principal Components Analysis  
## Call: principal(r = train.house[, c(10:17)], nfactors = 8, rotate = "none",   
## scores = TRUE)  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 h2 u2 com  
## MINC\_Z 0.09 -0.06 0.92 0.37 -0.02 -0.02 0.04 -0.01 1 0.0e+00 1.4  
## HAGE\_Z -0.43 0.03 -0.41 0.80 0.01 0.03 0.01 0.00 1 2.2e-16 2.1  
## ROOMS\_Z 0.96 0.10 0.10 0.10 0.12 0.16 -0.12 0.02 1 5.6e-16 1.2  
## BEDRMS\_Z 0.97 0.09 -0.12 0.06 0.15 -0.07 0.05 -0.09 1 1.7e-15 1.1  
## POPN\_Z 0.93 0.04 -0.12 0.08 -0.33 0.04 0.00 -0.02 1 1.3e-15 1.3  
## HHLDS\_Z 0.97 0.09 -0.11 0.09 0.05 -0.12 0.07 0.08 1 1.7e-15 1.1  
## LAT\_Z -0.15 0.97 0.02 -0.09 0.02 0.13 0.11 0.00 1 1.2e-15 1.1  
## LONG\_Z 0.15 -0.97 -0.06 -0.06 0.04 0.14 0.11 0.01 1 8.9e-16 1.1  
##   
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8  
## SS loadings 3.91 1.91 1.07 0.82 0.15 0.08 0.05 0.01  
## Proportion Var 0.49 0.24 0.13 0.10 0.02 0.01 0.01 0.00  
## Cumulative Var 0.49 0.73 0.86 0.96 0.98 0.99 1.00 1.00  
## Proportion Explained 0.49 0.24 0.13 0.10 0.02 0.01 0.01 0.00  
## Cumulative Proportion 0.49 0.73 0.86 0.96 0.98 0.99 1.00 1.00  
##   
## Mean item complexity = 1.3  
## Test of the hypothesis that 8 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0   
## with the empirical chi square 0 with prob < NA   
##   
## Fit based upon off diagonal values = 1

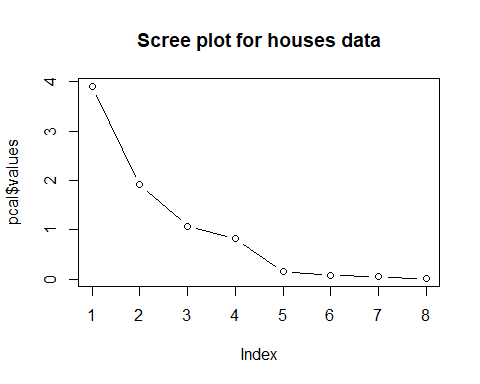
#PCA Results  
   
 #Eigen Values  
  
pcal$values

## [1] 3.90588026 1.90804016 1.07161349 0.82232940 0.14888011 0.08152021 0.04681217  
## [8] 0.01492420

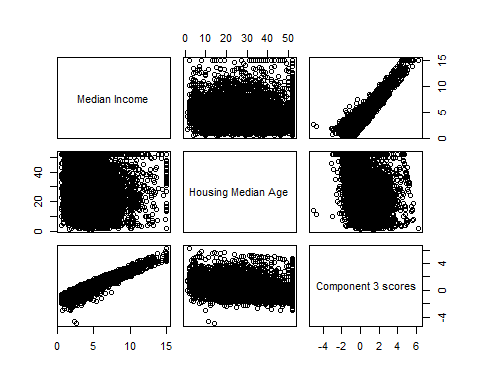
#Loading Matrix and Variance exlained  
  
pcal$loadings

##   
## Loadings:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8   
## MINC\_Z 0.921 0.371   
## HAGE\_Z -0.431 -0.409 0.803   
## ROOMS\_Z 0.956 0.104 0.103 0.123 0.159 -0.119   
## BEDRMS\_Z 0.969 -0.121 0.146   
## POPN\_Z 0.932 -0.121 -0.327   
## HHLDS\_Z 0.972 -0.113 -0.116   
## LAT\_Z -0.148 0.969 0.132 0.113   
## LONG\_Z 0.153 -0.968 0.136 0.109   
##   
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8  
## SS loadings 3.906 1.908 1.072 0.822 0.149 0.082 0.047 0.015  
## Proportion Var 0.488 0.239 0.134 0.103 0.019 0.010 0.006 0.002  
## Cumulative Var 0.488 0.727 0.861 0.963 0.982 0.992 0.998 1.000

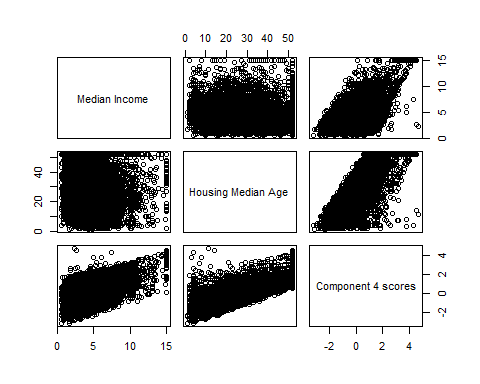
# Scree Plot  
  
plot(pcal$values,type="b",main = "Scree plot for houses data")



# plot factor scores  
pairs(~train.house$MINC+train.house$HAGE+pcal$scores[,3],labels=c("Median Income","Housing Median Age","Component 3 scores"))



pairs(~train.house$MINC+train.house$HAGE+pcal$scores[,4],labels=c("Median Income","Housing Median Age","Component 4 scores"))



# Calculate communalities  
  
comm3<-loadings(pcal)[2,1]^2+loadings(pcal)[2,2]^2+loadings(pcal)[2,3]^2  
comm4<-loadings(pcal)[2,1]^2+loadings(pcal)[2,2]^2+loadings(pcal)[2,3]^2+loadings(pcal)[2,4]^2  
comm3;comm4

## [1] 0.3537888

## [1] 0.9990335

# Validation of Principal components  
  
pca2<-principal(test.house[,c(10:17)], nfactors = 4,rotate = "none",scores = TRUE)  
pca2$loadings

##   
## Loadings:  
## PC1 PC2 PC3 PC4   
## MINC\_Z 0.932 0.348  
## HAGE\_Z -0.440 -0.384 0.809  
## ROOMS\_Z 0.958 0.114  
## BEDRMS\_Z 0.972 -0.124   
## POPN\_Z 0.938 -0.112   
## HHLDS\_Z 0.974 -0.115   
## LAT\_Z -0.110 0.974   
## LONG\_Z 0.121 -0.973   
##   
## PC1 PC2 PC3 PC4  
## SS loadings 3.917 1.915 1.068 0.817  
## Proportion Var 0.490 0.239 0.134 0.102  
## Cumulative Var 0.490 0.729 0.863 0.965

# Read in and prepare data for factor analysis.  
adult<- read.csv("D:/Data Mining and Predictive Analysis/Data sets/adult/Clem3Training")  
adult$"capnet"<-adult$capital.gain-adult$capital.loss  
adult.s<-adult[,c(1,3,5,13,16)]  
#Standardization of the data  
adult.s$age\_Z<-(adult.s$age-mean(adult.s$age))/(sd(adult$age))  
adult.s$dem\_Z<-(adult.s$demogweight-mean(adult.s$demogweight))/(sd(adult$demogweight))  
adult.s$educ\_Z<-(adult.s$education.num-mean(adult.s$education.num))/(sd(adult$education.num))  
adult.s$capnet\_Z<-(adult.s$capnet-mean(adult.s$capnet))/(sd(adult$capnet))  
adult.s$hours\_Z<-(adult.s$hours.per.week-mean(adult.s$hours.per.week))/(sd(adult$hours.per.week))  
 # Randomly select a training dataset  
dim(adult.s)

## [1] 25000 10

choose<-runif(dim(adult.s)[1],0,1)  
test.adult<-adult.s[which(choose<.1),c(6:10)]  
train.adult<-adult.s[which(choose>=.1),c(6:10)]  
  
# Bartlett's test for sphericity.  
  
library(psych)  
corrmat1<-cor(train.adult,method = "pearson")  
corrmat1

## age\_Z dem\_Z educ\_Z capnet\_Z hours\_Z  
## age\_Z 1.00000000 -0.074742938 0.03066279 0.06799947 0.071555771  
## dem\_Z -0.07474294 1.000000000 -0.04348495 0.00462801 -0.008977929  
## educ\_Z 0.03066279 -0.043484947 1.00000000 0.11544618 0.147195389  
## capnet\_Z 0.06799947 0.004628010 0.11544618 1.00000000 0.076253890  
## hours\_Z 0.07155577 -0.008977929 0.14719539 0.07625389 1.000000000

dim(train.adult)

## [1] 22477 5

cortest.bartlett(corrmat1,n=dim(train.adult)[1])

## $chisq  
## [1] 1253.601  
##   
## $p.value  
## [1] 3.936555e-263  
##   
## $df  
## [1] 10

# Factor Analysis with five components.  
library(GPArotation)  
fa1<-fa(train.adult,nfactors = 5,fm="pa",rotate = "none",SMC = FALSE)  
fa1$values

## [1] 1.2745153 1.0373121 0.9500960 0.9170091 0.8210675

fa1$loadings

##   
## Loadings:  
## PA1 PA2 PA3 PA4 PA5   
## age\_Z 0.421 -0.529 0.586 0.323 0.309  
## dem\_Z -0.225 0.777 0.431 0.306 0.257  
## educ\_Z 0.631 0.227 -0.422 -0.155 0.590  
## capnet\_Z 0.533 0.248 0.449 -0.599 -0.307  
## hours\_Z 0.603 0.202 -0.203 0.580 -0.466  
##   
## PA1 PA2 PA3 PA4 PA5  
## SS loadings 1.275 1.037 0.950 0.917 0.821  
## Proportion Var 0.255 0.207 0.190 0.183 0.164  
## Cumulative Var 0.255 0.462 0.652 0.836 1.000

# Proportion of variance and cumulative variance  
  
fa2<-fa(train.adult,nfactors = 2,fm="pa",max.iter = 200,rotate = "none")  
fa2$values

## [1] 0.537211520 0.350445640 0.035070434 -0.005519536 -0.030549032

fa2$loadings

##   
## Loadings:  
## PA1 PA2   
## age\_Z 0.594 -0.327  
## dem\_Z -0.111   
## educ\_Z 0.292 0.430  
## capnet\_Z 0.189 0.143  
## hours\_Z 0.225 0.193  
##   
## PA1 PA2  
## SS loadings 0.537 0.350  
## Proportion Var 0.107 0.070  
## Cumulative Var 0.107 0.178

# Varimax rotation  
fa2v<-fa(train.adult,nfactors = 2,fm="pa",max.iter = 200,rotate = "varimax")  
fa2v$loadings

##   
## Loadings:  
## PA1 PA2   
## age\_Z 0.678   
## dem\_Z -0.101   
## educ\_Z 0.520  
## capnet\_Z 0.222  
## hours\_Z 0.283  
##   
## PA1 PA2  
## SS loadings 0.484 0.404  
## Proportion Var 0.097 0.081  
## Cumulative Var 0.097 0.178

fa2v$communality

## age\_Z dem\_Z educ\_Z capnet\_Z hours\_Z   
## 0.46062215 0.01253402 0.27061486 0.05618058 0.08770555

# User-defined composites  
small.houses<-houses[,c(4:7)]  
a<-c(1/4,1/4,1/4,1/4)  
w<-t(a)\*small.houses