# PCA

2023-03-16

## Set the working directory

### Read Data

#“‘{r, echo=FALSE} # Go to the MD project folder before running the script #dir <- “C:/Users/polcg/repos/MD” # for some reason the it can’t access current directory # Import preprocessed data # db <- read.csv(paste0(dir,“/data/Preprocessed_data.csv”),header=T, sep=“,”, strip.white = TRUE) # take out trailing whitespaces

## read dataframe back in from RDS file

#db <- readRDS(paste0(dir,“/data/Preprocessed_data_rds.rds”)) #“‘

## Try this to read the data

### We check for missing data

```
## [1] 0
```

```
##  [1] "age"
##  [2] "class.of.worker"
##  [3] "detailed.industry.recode"
##  [4] "detailed.occupation.recode"
##  [5] "education"
##  [6] "wage.per.hour"
##  [7] "marital.stat"
##  [8] "major.industry.code"
##  [9] "major.occupation.code"
## [10] "race"
## [11] "hispanic.origin"
## [12] "sex"
## [13] "full.or.part.time.employment.stat"
## [14] "capital.gains"
## [15] "capital.losses"
## [16] "dividends.from.stocks"
## [17] "tax.filer.stat"
## [18] "detailed.household.and.family.stat"
## [19] "detailed.household.summary.in.household"
## [20] "instance.weight"
## [21] "num.persons.worked.for.employer"
## [22] "country.of.birth.father"
## [23] "country.of.birth.mother"
## [24] "country.of.birth.self"
## [25] "citizenship"
```

```
## [26] "veterans.benefits"
## [27] "weeks.worked.in.year"
## [28] "income"
```

```
##                                   age                          class.of.worker
##                             "integer"                                 "factor"
##                detailed.industry.recode               detailed.occupation.recode
##                               "factor"                                 "factor"
##                             education                            wage.per.hour
##                               "factor"                               "integer"
##                           marital.stat                     major.industry.code
##                               "factor"                                 "factor"
##                  major.occupation.code                                     race
##                               "factor"                                 "factor"
##                         hispanic.origin                                      sex
##                               "factor"                                 "factor"
##       full.or.part.time.employment.stat                           capital.gains
##                               "factor"                               "integer"
##                          capital.losses                    dividends.from.stocks
##                             "integer"                               "integer"
##                          tax.filer.stat      detailed.household.and.family.stat
##                               "factor"                                 "factor"
## detailed.household.summary.in.household                          instance.weight
##                               "factor"                               "numeric"
##         num.persons.worked.for.employer                 country.of.birth.father
##                             "integer"                                 "factor"
##                 country.of.birth.mother                    country.of.birth.self
##                               "factor"                                 "factor"
##                            citizenship                        veterans.benefits
##                               "factor"                                 "factor"
##                     weeks.worked.in.year                                   income
##                             "integer"                                 "factor"
```

## We separate numerical and categorical features

```
##                             age                   wage.per.hour
##                               1                               6
##                  capital.gains                  capital.losses
##                              14                              15
##           dividends.from.stocks                 instance.weight
##                              16                              20
## num.persons.worked.for.employer            weeks.worked.in.year
##                              21                              27
```

```
##                         class.of.worker                 detailed.industry.recode
##                                       2                                        3
##               detailed.occupation.recode                                education
##                                       4                                        5
##                           marital.stat                     major.industry.code
##                                       7                                        8
##                   major.occupation.code                                     race
##                                       9                                       10
##                         hispanic.origin                                      sex
##                                      11                                       12
##       full.or.part.time.employment.stat                          tax.filer.stat
```

```
##                                    13                                    17
##      detailed.household.and.family.stat detailed.household.summary.in.household
##                                    18                                    19
##                 country.of.birth.father                 country.of.birth.mother
##                                    22                                    23
##                   country.of.birth.self                            citizenship
##                                    24                                    25
##                       veterans.benefits                                income
##                                    26                                    28
```

## Perform PCA

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6     PC7
## Standard deviation     1.3985 1.0942 1.0070 0.9990 0.9531 0.9393 0.89216
## Proportion of Variance 0.2445 0.1497 0.1268 0.1247 0.1135 0.1103 0.09949
## Cumulative Proportion  0.2445 0.3941 0.5209 0.6456 0.7592 0.8695 0.96895
##                            PC8
## Standard deviation     0.49838
## Proportion of Variance 0.03105
## Cumulative Proportion  1.00000
```

## Correlation matrix

```
## corrplot 0.92 loaded
```

## Plot the correlation matrix

```
corrplot(cor_mat, method="circle")
```

```r
# pca$sdev
variance<- pca$sdev^2
# variance
totalVariance<- sum(variance)
# totalVariance
percentageVar<- 100*variance/totalVariance
# percentageVar
barplot(percentageVar)
```

```r
# Create a screeplot
barplot(percentageVar, main = "Scree plot", xlab = "Dimensions", ylab = "Percentage of explained varianc
lines(percentageVar)
points(percentageVar)
text(seq_along(percentageVar), percentageVar, labels = paste0(round(percentageVar,1), "%"), pos=3)
axis(2, at=seq(0,max(percentageVar),5), labels=paste0(seq(0,max(percentageVar),5),'%'))
```

Scree plot
Percentage of explained variances
Dimensions
15%
12.7%
12.5%
11.4%
11%
9.9%
3.1%

### We use cumulative proportion of variance to easily spot how many components of pca to choose

```
SumPercentageVar <- cumsum(percentageVar)
barplot(SumPercentageVar, main = "Cumulative Scree plot", xlab = "Number of components", ylab = "Cumula
lines(SumPercentageVar)
points(SumPercentageVar)
text(seq_along(SumPercentageVar), SumPercentageVar, labels = paste0(round(SumPercentageVar,1), "%"), po
axis(2, at=seq(0,max(SumPercentageVar),5), labels=paste0(seq(0,max(SumPercentageVar),5),'%'))
```

In this case we would choose 5

```
# SELECTION OF THE SINGIFICNT DIMENSIONS (keep 80% of total inertia)

nd = 5
pca$rotation # Loadings of the pca (how much of that variable is represented by the principal component
```

```
##                                       PC1         PC2         PC3         PC4
## age                             0.27165455  0.26282697  0.37222236 -0.02847641
## wage.per.hour                   0.28115151 -0.19136156 -0.26687867 -0.17411384
## capital.gains                   0.12359532  0.62360052 -0.30312730 -0.05326405
## capital.losses                  0.15134175  0.02876875  0.73256422  0.40133193
## dividends.from.stocks           0.07584342  0.68800368 -0.02141987  0.04181086
## instance.weight                 0.06233485 -0.04821283 -0.39156372  0.89397569
## num.persons.worked.for.employer 0.62468737 -0.14303526 -0.08176079 -0.03879062
## weeks.worked.in.year            0.63961342 -0.09200473 -0.03498587 -0.05023641
##                                        PC5        PC6         PC7          PC8
## age                             0.51611330 -0.6143016  0.26472613 -0.062817628
## wage.per.hour                  -0.64575328 -0.5809296  0.16825794 -0.002747413
## capital.gains                  -0.11330970  0.3160451  0.62267807 -0.025919591
## capital.losses                 -0.47197125  0.1767846  0.15581006 -0.016246414
```

```
## dividends.from.stocks           -0.21046712 -0.1531580 -0.67151882 -0.000679995
## instance.weight                  0.11645127 -0.1637966  0.02664231  0.013235285
## num.persons.worked.for.employer  0.09547653  0.2489304 -0.16179480 -0.695598603
## weeks.worked.in.year             0.11857799  0.2043626 -0.10853429  0.714896898
```

At a quick glance it would seem that the most important variables in terms of the variance be num.persons.worked.for.employer and weeks.worked.in.year since they have the higher absolute value in PC1 which amounts for the most amount of variance.

## We project our data on only 5 dimensions of pca

```
# STORAGE OF THE EIGENVALUES, EIGENVECTORS AND PROJECTIONS IN THE nd DIMENSIONS
# View(pca$x)
dim(pca$x)
```

```
## [1] 20000     8
```

```
dim(numerical_data)
```

```
## [1] 20000     8
```

```
numerical_data[2000,]
```

```
##      age wage.per.hour capital.gains capital.losses dividends.from.stocks
## 2000  58             0             0              0                     0
##      instance.weight num.persons.worked.for.employer weeks.worked.in.year
## 2000         1706.64                                6                  40
```

```
pca$x[2000,]
```

```
##          PC1          PC2          PC3          PC4          PC5          PC6
##  1.680889370 -0.109800547  0.237097498 -0.186678084  1.020025087 -0.004955472
##          PC7          PC8
## -0.121030712 -0.762899412
```

```
Psi = pca$x[,1:nd]
dim(Psi)
```

```
## [1] 20000     5
```

```
Psi[2000,]
```

```
##        PC1        PC2        PC3        PC4        PC5
##  1.6808894 -0.1098005  0.2370975 -0.1866781  1.0200251
```

```
# Observations on significant PCAs.
SignificantCoord = pca$x[,1:nd]
CorrelationSignificant = cor(numerical_data,SignificantCoord)
```

## We store some data to plot later

```
iden = row.names(numerical_data)
etiq = names(numerical_data)
ze = rep(0,length(etiq)) # WE WILL NEED THIS VECTOR AFTERWARDS FOR THE GRAPHICS
```

## Plot of quality of variables using the square cosine

```
SquareCos = CorrelationSignificant^2
color_palette <- c("#FC8D62", "#FDCC8A", "#FFFFB3", "#A6D854", "#1A9850")
corrplot(SquareCos, is.corr=FALSE, title="Quality of the variables with the PCAs",col =color_palette, bg
```

Quality of the variables with the PCAs
PC1
PC2
PC3
PC4
PC5
age
wage.per.hour
capital.gains
capital.losses
dividends.from.stocks
instance.weight
num.persons.worked.for.employer
weeks.worked.in.year
0.8
0.64
0.48
0.32
0.16
0

## Individuals Plot

```
color_palette1 <- c("#4B5D67", "#322F3D", "#59405C", "#87556F")
colorIteration <- 1

for (i in seq(nd-1)) {
  for (j in seq(i+1, nd)) {
    plot(SignificantCoord[, i], SignificantCoord[, j],
         col = color_palette1[(colorIteration %% length(color_palette1)) + 1],
         main = "Individuals plot",
         xlab = paste0("PCA", i, " (", percentageVar[i], ' %)'),
         ylab = paste0("PCA", j, " (", percentageVar[j], ' %)'),
         pch = 19,
         cex = 2.5, asp = 0.5)
    abline(h = 0, v = 0, lty = "dashed")


    colorIteration <- colorIteration + 1
  }
}
```

Individuals plot
PCA2 (14.9647296009382 %)
PCA1 (24.4487886458602 %)
40
30
20
10
0
−5
0
5
10

Individuals plot
PCA3 (12.674884607360 5 %)
PCA1 (24.4487886458602 %)
15
10
5
0
−5
−10
−2
0
2
4
6
8

Individuals plot
PCA4 (12.4743501010451 %)
PCA1 (24.4487886458602 %)
15
10
5
0
−5
−10
−2
0
2
4
6
8

Individuals plot
PCA5 (11.3550912443 88 %)
PCA1 (24.4487886458602 %)
5
0
−5
−10
−15
−20
−2
0
2
4
6
8

Individuals plot
PCA3 (12.674884607360 5 %)
PCA2 (14.9647296009382 %)
60
40
20
0
−20
−40
−60
0
10
20
30
40

Individuals plot
PCA4 (12.4743501010451 %)
PCA2 (14.9647296009382 %)
60
40
20
0
−20
−40
−60
0
10
20
30
40

Individuals plot
PCA5 (11.3550912443 88 %)
PCA2 (14.9647296009382 %)
60
40
20
0
−20
−40
−60
−80
0
10
20
30
40

Individuals plot
PCA4 (12.4743501010451 %)
PCA3 (12.6748846073605 %)
30
20
10
0
−10
−20
−5
0
5
10

Individuals plot
PCA5 (11.3550912443 88 %)
PCA3 (12.6748846073605 %)
20
10
0
−10
−20
−30
−5
0
5
10

Individuals plot
PCA5 (11.3550912444388 %)
PCA4 (12.4743501010451 %)
10
5
0
−5
−10
−15
−20
−25
−4
−2
0
2
4
6
8

## Correlation Circle

```
eje1<-1
#eje2<-3
eje2<-2

plot(Psi[,eje1],Psi[,eje2])
text(Psi[,eje1],Psi[,eje2],labels=iden, cex=0.5)
axis(side=1, pos= 0, labels = F, col="cyan")
axis(side=3, pos= 0, labels = F, col="cyan")
axis(side=2, pos= 0, labels = F, col="cyan")
axis(side=4, pos= 0, labels = F, col="cyan")
```

For example let's take a look at 3402 to make sure it makes sense: as we can see PC1 corresponds roughly to 4 and PC2 is in the middle of -40 and -30

```
Psi[3402,]
```

```
##        PC1        PC2        PC3        PC4        PC5
##   4.141940  36.733831  -2.578435   1.528280 -10.646693
```

```
#Projection of variables

Phi = cor(numerical_data,Psi)
View(Phi)

#select your axis

X<-Phi[,eje1]
Y<-Phi[,eje2]
```

View

```
for (i in seq(nd-1)) {
  for (j in seq(i+1, nd)) {

    X<-Phi[,i]
    Y<-Phi[,j]
    plot(Psi[,i],Psi[,j],type="n", ylim = c(-1, 1), xlim = c(0, 1), xlab = paste("Dim", i), ylab = past
    axis(side=1, pos= 0, labels = F)
    axis(side=3, pos= 0, labels = F)
    axis(side=2, pos= 0, labels = F)
    axis(side=4, pos= 0, labels = F)
    arrows(ze, ze, X, Y, length = 0.07,col="blue")
    text(X,Y,labels=etiq,col="darkblue", cex=0.7)


  }
}
```

Projection of numeric variables in X: 1, Y: 3
capital.losses
age
dividends.from.stocks
weeks.worked.in.year
num.persons.worked.for.employer
wage.per.hour
capital.gains
instance.weight
Dim 3
Dim 1
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

Projection of numeric variables in X: 1, Y: 4
instance.weight
capital.losses
dividends.from.stocks
capital.gains
age
num.persons.worked.for.employer
weeks.worked.in.year
wage.per.hour
Dim 4
Dim 1
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

Projection of numeric variables in X: 1, Y: 5
Dim 5
Dim 1
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
age
instance.weight
weeks.worked.in.year
num.persons.worked.for.employer
capital.gains
dividends.from.stocks
capital.losses
wage.per.hour

Projection of numeric variables in X: 2, Y: 3
Dim 3
Dim 2
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
capital.losses
age
year
dividends.from.stocks
capital.gains
weight

Projection of numeric variables in X: 2, Y: 4
weight
capital.losses
dividends.from.stocks
capital.gains
age
Dim 4
Dim 2
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

Projection of numeric variables in X: 2, Y: 5
age
weight
capital.gains
dividends.from.stocks
capital.losses
Dim 5
Dim 2
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

Projection of numeric variables in X: 3, Y: 4
Dim 4
Dim 3
capital.losses
age
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

Projection of numeric variables in X: 3, Y: 5
Dim 5
Dim 3
age
capital.losses
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

```
# PROJECTION OF ILLUSTRATIVE qualitative variables on individuals' map
# PROJECCI? OF INDIVIDUALS DIFFERENTIATING THE Dictamen
# (we need a numeric Dictamen to color)


varcat=factor(db[,5])
plot(Psi[,1],Psi[,2],col=varcat)
axis(side=1, pos= 0, labels = F, col="darkgray")
axis(side=3, pos= 0, labels = F, col="darkgray")
axis(side=2, pos= 0, labels = F, col="darkgray")
axis(side=4, pos= 0, labels = F, col="darkgray")
legend("bottomleft",levels(factor(varcat)),pch=1,col=c(1,2), cex=0.6)
```

10th grade
11th grade
12th grade no diploma
1st 2nd 3rd or 4th grade
5th or 6th grade
7th and 8th grade
9th grade
Associates degree−academic program
Associates degree−occup /vocational
Bachelors degree(BA AB BS)
Children
Doctorate degree(PhD EdD)
High school graduate
Less than 1st grade
Masters degree(MA MS MEng MEd MSW MBA)
Prof school degree (MD DDS DVM LLB JD)
Some college but no degree
Psi[, 2]
Psi[, 1]
−2
0
2
4
6
8
0
10
20
30
40