Multivariate Data Analysis: Application to a Population Census

Homework II

Iker Caballero & Daniel González

Multivariate Data Analysis - Prof. Dante Conti

1 Introduction

In this second homework, we are asked to select a multivariate data set that allows us to carry on multivariate data analysis techniques such as principal component analysis (PCA) and multiple correspondence analysis (MCA).

Hence, we will first start by introducing the data set that we will use throughout this homework. Then, we will begin the multivariate analyses. We start the analysis with the PCA, and we go to the MCA afterward. Finally, we do a comparison of the different analyses in order, to summarize all the interpretations and results obtained and get more insights.

2 About the Dataset

The Adult data set was extracted by Barry Becker from the USA Census Bureau database, and it can be found in the UCI Machine Learning repository, as it is widely used for classification purposes in academic research. It contains data on different socioeconomic characteristics of individuals in 1994, containing categorical and numerical variables. We now present a definition of the variables:

- **V1:** If a person has a salary greater than or less than USD 50.000 (categorical)

- **Age:** Age of the person (continuous)

- **Fnlwgt:** Weight per individual in the census (continuous).

- **Education:** Level of education (categorical)

- **Education-num:** Years of education (continuous).

- **Occupation:** Occupation of the person (categorical). *Categories*: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

- **Race**: Race of the people (categorical). *Categories*: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **Sex:** Sex of the person (categorical). *Categories*: Female, Male.

- **Capital-gain:** Capital gained (continuous).

- **Capital-loss:** Capital lost (continuous).

- **Hours-per-week:** Hours per week worked (continuous).

- **Marital.status:** Marital status of the people (categorical). *Categories*: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

2.1 Data Quality

Regarding the quality of this database, we can see that it has no missing data, so we can continue working without eliminating any observations. Given that our purpose is to analyze individuals and variables through multivariate analysis methods, observations with high values in some variables are not eliminated, so that we can possibly characterize these individuals and obtain insights.

We will use the principal component analysis for analyzing this data, so we demean and scale the numerical variables in this procedure, as the magnitudes and scales of the variables are different.

3 Principal Component Analysis

The first method we apply to analyze the data is the principal component analysis, as mentioned previously. To perform this analysis, we use R software packages such as *FactoMinr*, and we obtain different results. The summary of these results is shown in Table 1.

Component	Eigenvalue	Percentage of variance	Cumulative percentage of variance
Component1	1.31	21.84	21.84
Component2	1.04	17.34	39.19
Component3	1.01	16.97	56.16
Component4	0.94	15.69	71.86
Component5	0.88	14.77	86.64
Component6	0.80	13.35	100

Table 1: PCA Summary

This table shows that, in order to explain 70-80% of the sample variance through the components, we would have to use the first four components, accounting for 71.86% of the sample variance explained. However, we decided to keep components 1 and 2 to visualize the individuals in a reduced dimensional space, so that the reader can simply interpret the results visually. Therefore, our interpretations and results will be restricted to these two for the sake of brevity.

In order to analyze the relations and the ”importance” of the different variables in the first two components, we can use Figure 1, as this biplot allows us to visualize the vectors for the variables in order to obtain insights about relations with components and among the variables themselves. The biplots that are shown in this analysis represent the first 2 components, which account for 39.1% of the sample variance.

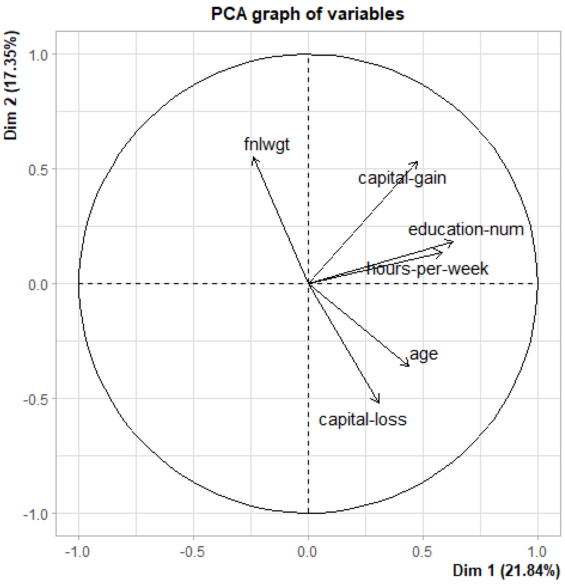

Figure 1: Biplot PCA

In the case of Component1, we see that hours-per-week and education years are the features with the closest angle with respect to the first dimension, which means that these are very related to the component. Moreover, we can see that the vectors' lengths are relatively high, indicating that these are very important factors in this component (referring to their weights in it). For Component2, we see that fnlwgt is the most correlated feature with relatively high relevance, but capital-gain and capital-loss do also matter in terms of correlation and weight inside the component.

Through the direction and angles between the vectors, we can also deduce the relations between variables. In this case, we can see that hours-per-week and years of education seem to be very positively correlated, which could be due to the logical relation between the years of education and professional career, which determines the hours worked per week. Additionally, we can see another strong negative correlation between fnlwgt and capital-loss, which is harder to explain given that we can also see that both fnlwgt and capital-loss are lowly correlated to capital-gain, so we cannot say that the weights are assigned according to a criterion opposite to capital losses (which would be capital gains. Using the same logic, we can see how the years of education and hours worked per week are merely uncorrelated with fnlwgt and capital-loss, which indicates that professions and education time could not have any (linear) relation with wealth loss and the weighting allocated to individuals in this census.

In Table 2, we also show the different correlation coefficients obtained for each variable and component in a numerical way, so that the reader can have a better idea of the relevance of each variable in each component:

Variable	Component1	Component2	Component3	Component4
Age	0.43	-0.35	-0.38	0.62
Fnlwgt	-0.24	0.55	0.52	0.55
Education-num	0.63	0.18	0.21	-0.35
Capital-gain	0.47	0.53	-0.32	0.21
Capital-loss	0.30	-0.51	0.62	0.21
Hours-per-week	0.58	0.13	0.21	-0.14

Table 2: Variables correlation

We can see that there is no variable that is highly correlated with one of the components (more than 0.75 in absolute value), but there are relatively high correlations for some variables. The analysis shows that education-number and hours-per-week are the variables more correlated to Component1, with values of 0.63 and 0.58 respectively. Regarding Component2, we can see that the most correlated variables are fnlwgt and capital-gain, with values of 0.55 and 0.53 respectively.

Now we can visualize the different results obtained with the PCA using biplots. Because of the correlation and the relevance of the different variables with these components, we can interpret the first component as representing professional career and education time, while the second component can be understood as a component related to census weighting and capital gains, but without a clear interpretation as the first one.

We first show some biplots in Figure 2 which allow us to characterize individuals depending on the continuous variables which are the most related to the components displayed in the biplot. The direction of the vectors and the position of the individuals with respect will make this possible, as the results of individual contributions to these variables are consistent with the direction and location of individuals.

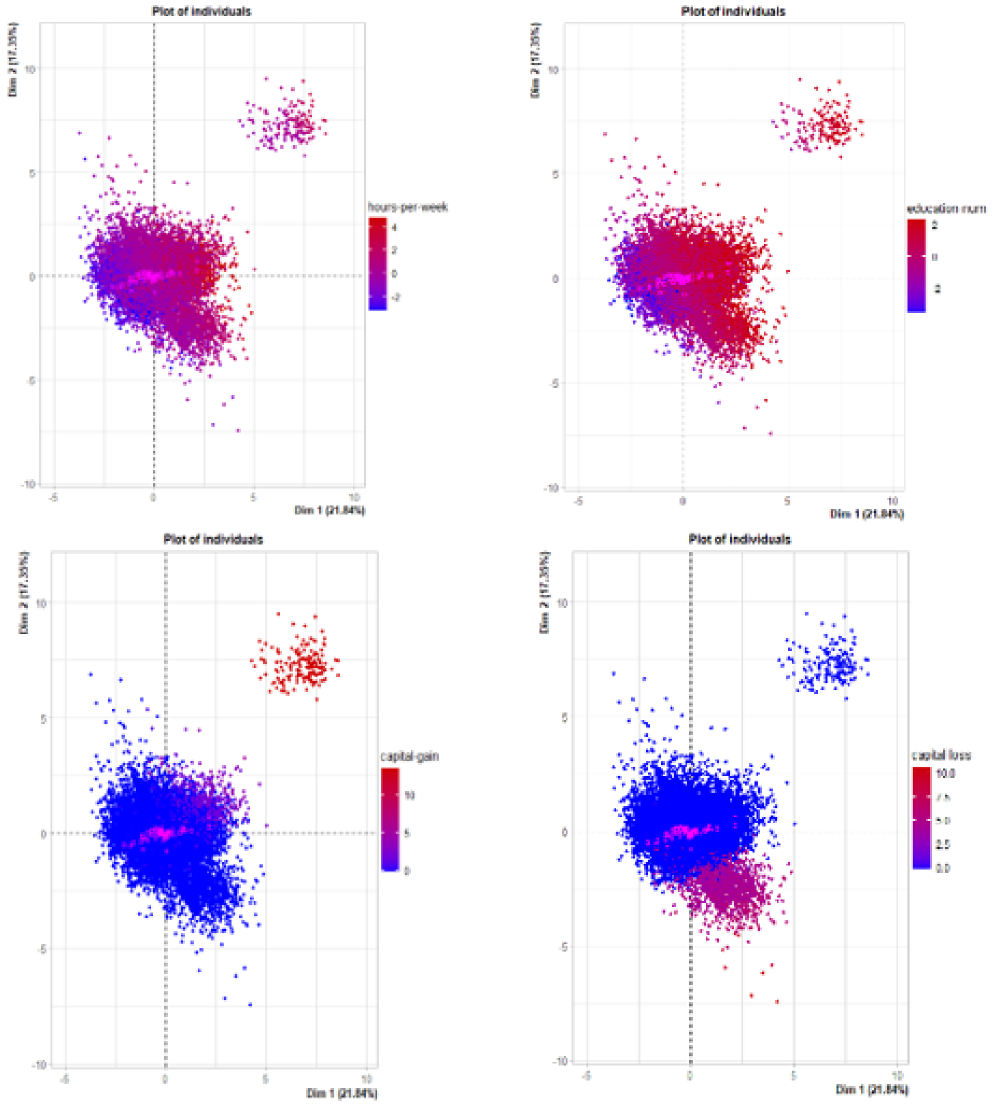

Figure 2: Biplot of individuals by education-number, capital-gain-loss, hours per week

The first two biplot shows the contribution of individuals with respect to hours-per-week and years of education respectively, showing that individuals who work more hours per week and have more years of education are located in the right direction from the origin.

When looking at the capital gains, one can see that the individuals that are in the top right part are the ones who earn the most capital gains. This indicates that the individuals that pertain to this cluster could be characterized as forming part of a socio-economic elite or upper class, so we would use this interpretation in the subsequent results.

At last, the individuals with more capital losses seem to be situated at the bottom part of the big individual cloud in the middle. These individuals have a more or less average years of education and a profession with average hours worked per week, while they also have a lower than average weighting in the sample and low capital gains (remember that the latter are not very correlated).

These are all expected results, given the interpretation of the dimensions and the direction of the vectors. Now, different plots for the individuals are displayed, colored by some categorical variables, so that we can also make an interpretation of the individuals regarding these variables.

In Figure 3, we plot the different individuals in the biplot which are divided by their annual salaries, depending on whether it is higher than USD 50.000 or not.

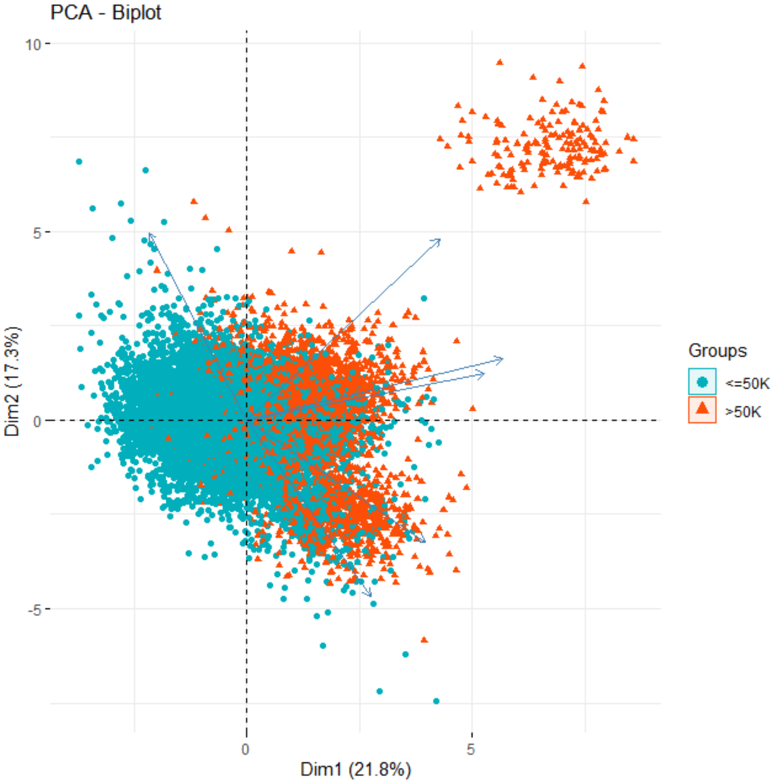

Figure 3: Biplot of individuals by salaries

We can observe that the right part of the biplot is related to people that have a salary greater than USD 50.000, as there is a greater concentration of individuals in that category. The people inside the other category are concentrated on the left-middle part of the biplot. We can see that, looking at the individuals in this other category, we can spot some outliers at the top-left and bottom-right parts, indicating individuals with high census weighting and capital gains but fewer years of education and hours worked per week and individuals with low census weighting and capital gains but more years of education and hours worked per week, respectively.

The top-right cluster of individuals all earn more than USD 50.000 per year, which seems to be logical as these individuals are the ones that pertain to the elite or upper socioeconomic class.

We now look at the occupations of the different individuals in the biplot shown in Figure 4:

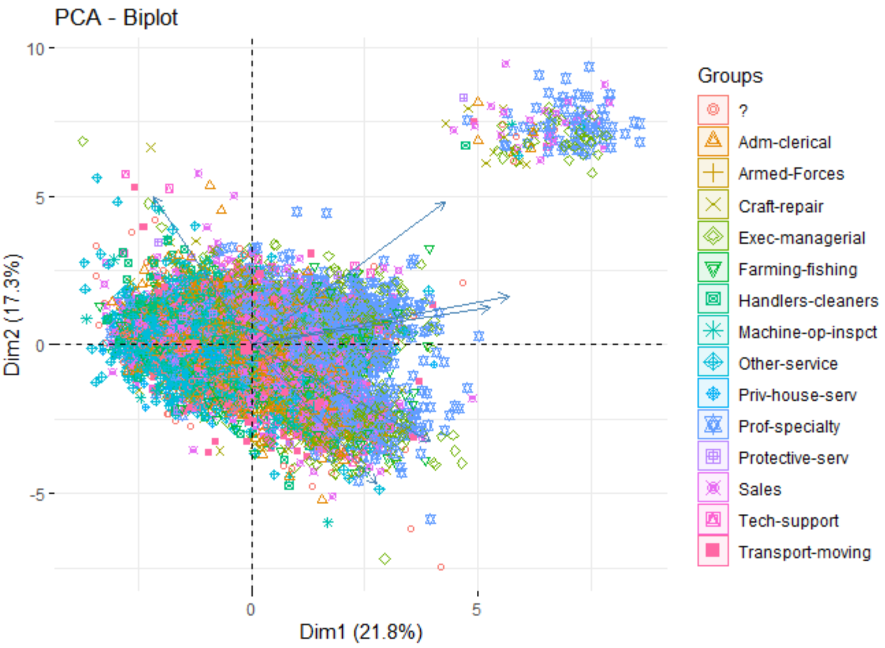

Figure 4: Biplot of individuals by occupation

In the case of the occupations, people that work in prof-specialty are mostly in the right part of the plot. Then there are some other kinds of jobs that are related to hard skills or high levels of education as exec-managerials. This makes sense because of the years of education and the dedication in hours that individuals with these professions need to offer.

In the top-right part cluster, we can identify a lot of individuals working in sales and as specialists. It makes sense because in sales people have the option to earn percentages per sale, while the specialized knowledge is also well paid for and has become very relevant in our nowadays societies. Nevertheless, there are some others also like clerical jobs or craft repair in that cluster too, which might not have a clear-cut explanation.

On the left side of the biplot, we can find more people that work in house services and other services, which are related to having fewer salaries, years of education, census weighting, and capital gains (as previously explained). However, one can see that this plot shows the heterogeneity present regarding professions, which might not be very useful for characterization (the salaries seem to be more useful for that purpose, for example).

Finally, in Figure 5, we focus on the sex of the individuals in the census:

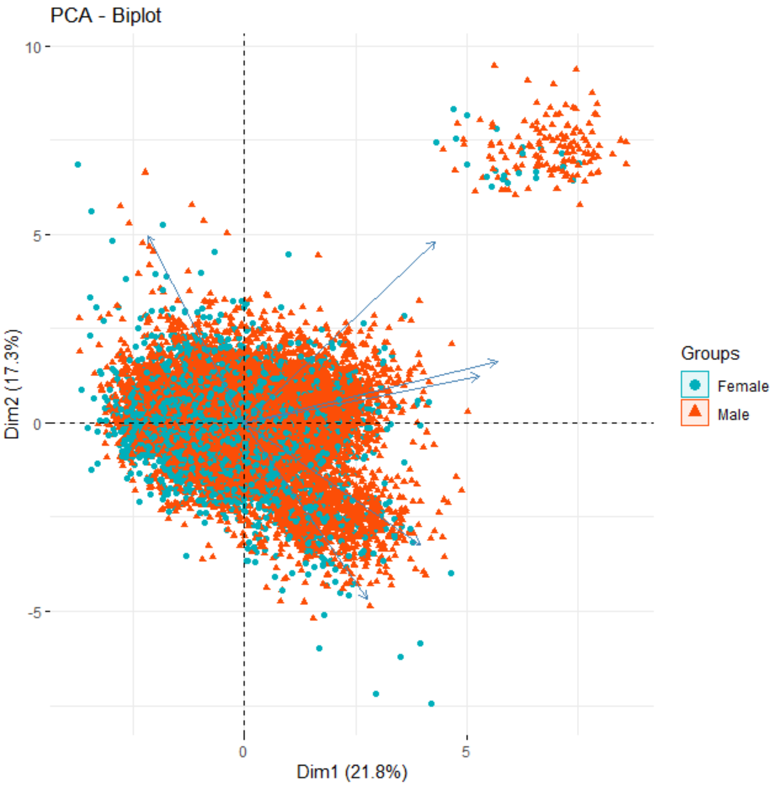

Figure 5: Biplot of individuals by sex

We can observe that a huge part of the males in the survey is on the right side of the biplot and females are a bit more on the left side. In the top-right cluster, we can see that most of the people there are males. This could have a controversial explanation: men might have more years of education and work more hours per week than women (as for this sample data) in 1994, as the concentration of men on the right side seems greater than that of women. The top-right cluster regarding the elite or a socioeconomic upper class is also principally composed of men, but we can clearly see women pertaining to that cluster too.

The analysis made possible the extraction of very interesting insights about the American population in this census with respect to continuous variables and categorical variables. Yet one can do a more specific analysis for categorical variables and individuals through the next analysis: the multiple correspondence analysis.

4 Multiple Correspondence Analysis

In this section, we show the results of the MCA performed with categorical variables and individuals. We also use numerical variables as supplementary variables, so that we can also get a sense of the information these yield in the analysis.

As before, the primary step is to define the number of dimensions to include. In order to choose it, we will compute the ratio $1/p$, where "p" is the number of categorical variables, which in this case would be $1/6$. The optimal number of dimensions for this problem will be the one that explains the 0.167 of the variance which is given for dimension 17, and where the cumulative variance is 49%. To represent the results visually, however, we will work with dimensions 1 and 2 because they are the most important in terms of the variance explained and allow us to plot.

We remark on an interesting fact that allows understanding the following results better: the different category points are located in such a way that the individuals will be represented in an opposite but symmetric way in these biplots. This happens because the biplot has different principal axes than the PCA analysis, as these represent the rows and columns and not principal components (which are linear combinations of the variables).

In Figure 6 we show a biplot that allows representing the different categories of all the categorical variables we are working with, while also highlighting the contribution of each one to the dimensions represented.

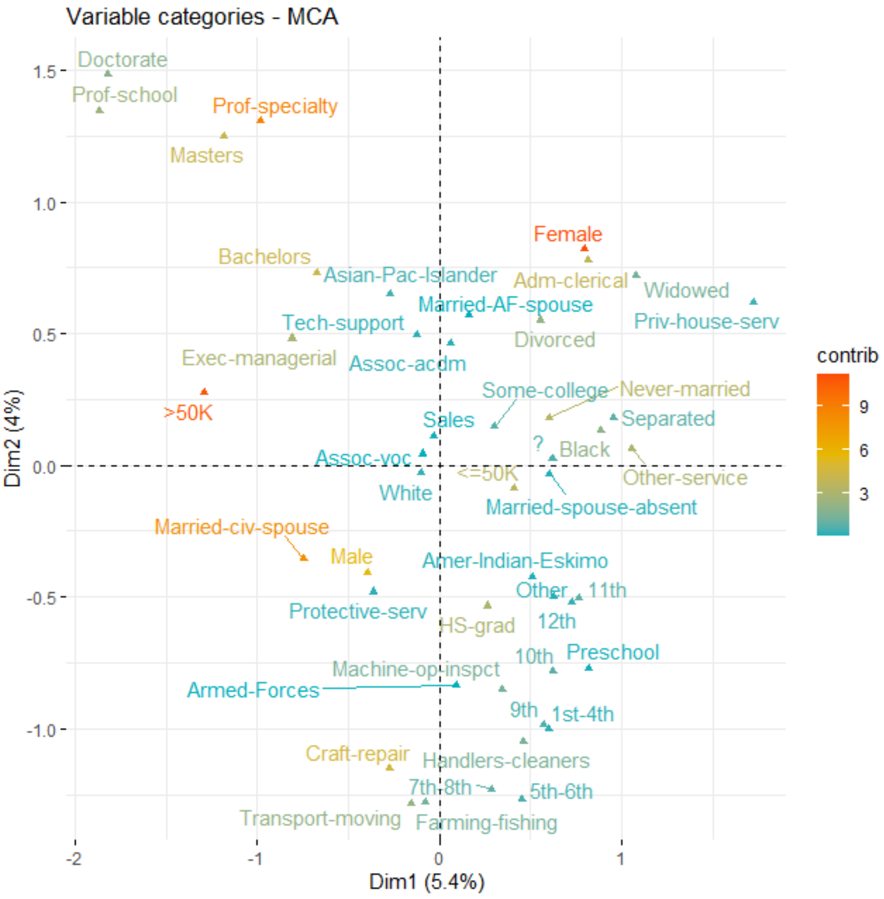

Figure 6: Biplot of variables by contribution

In this biplot, categories such as greater than USD 50.000, doctorate, male, and married are in the left part of the biplot. On the other hand, female, separated, salaries lower or equal to USD 50k, and low levels of education are located in the right part of the biplot. Categories in the middle of the biplot can be regarded as badly represented categories, so we can ignore them in our analysis.

In terms of contribution, we can see that for the dimension 1 categories that explain it more are salaries greater than USD 50.000, married, male, these on the left side, and low salaries and never married on the right side. We can get two important insights from these results. First, People that have higher salaries tend to be male and married. Second, people that have salaries less than USD 50.000 tend to be single people. Also, there are some other correlations between them like black, other-service (occupation), married-absent, and low salaries. This could be the characterization of a part of the population with these characteristics. So, black people tend to earn less than USD 50.000 per year, work in other services, and have marriage absent.

The most relevant categories for dimension 2 are female, prof-specialty, and craft-repair. Here we can find some categories related, such as prof-specialty with master, doctorate, and prof-school; craft-repair with mid-low levels of study; and, female with adm-clerical and divorced. In the first case, people that have master's and doctorates are closer because they are the highest level of education, and some of them are school professors. The second one is related to craft-repair which is an occupation that requires a few years of education. This is because it is an applied job and is done by hand, it is more related to the experience. Finally, the third relationship tells us that females are working in jobs that provide support to businesses, and some of them are also divorced.

Given that a biplot combining both individuals and the categories plotted before is not visually understandable, we prefer to plot the individuals in a separate plot in Figure 7.

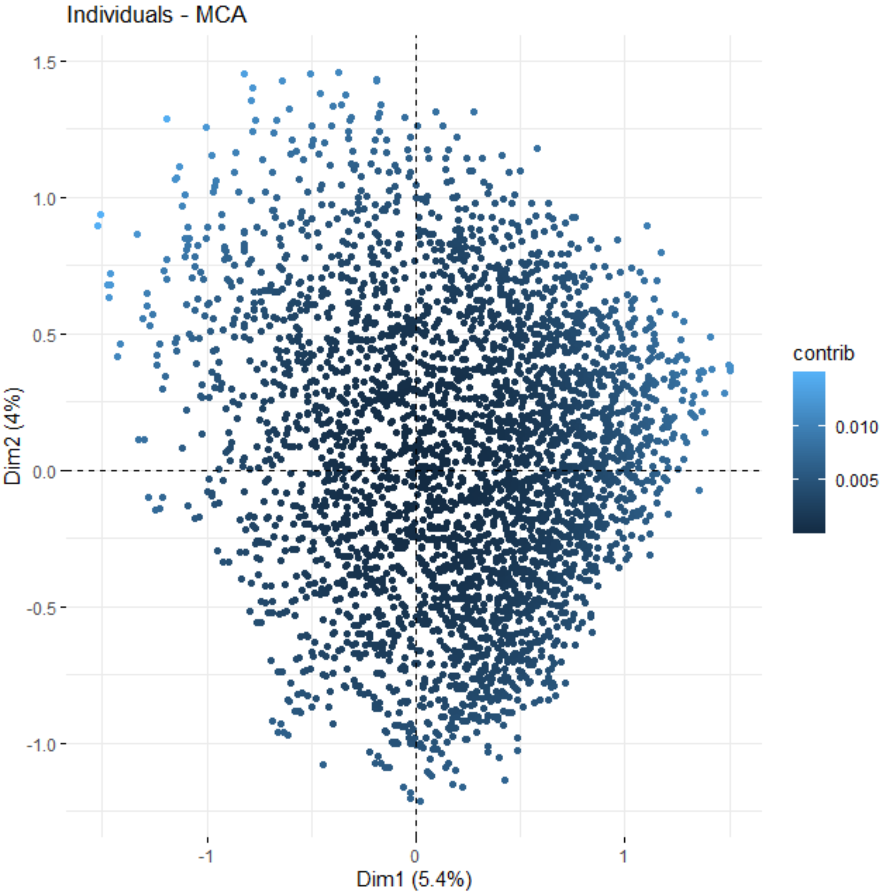

Figure 7: Biplot of individuals by contribution

Here we can appreciate the individuals in terms of contribution. We can see that individuals that contribute more to the variance of the dimensions are in the borders of the plot. This could be because they are located where the important categories are, and in that way, it produces the profiling of the people. Moreover, the individuals located in the middle (near the origin) contribute less than others.

To include the numerical variables used in the previous analysis, we represent them in vector form in Figure 8, where we can observe the relationship between the numerical variables in dimensions 1 and 2 through the angles, directions, and lengths of the vectors (like before).

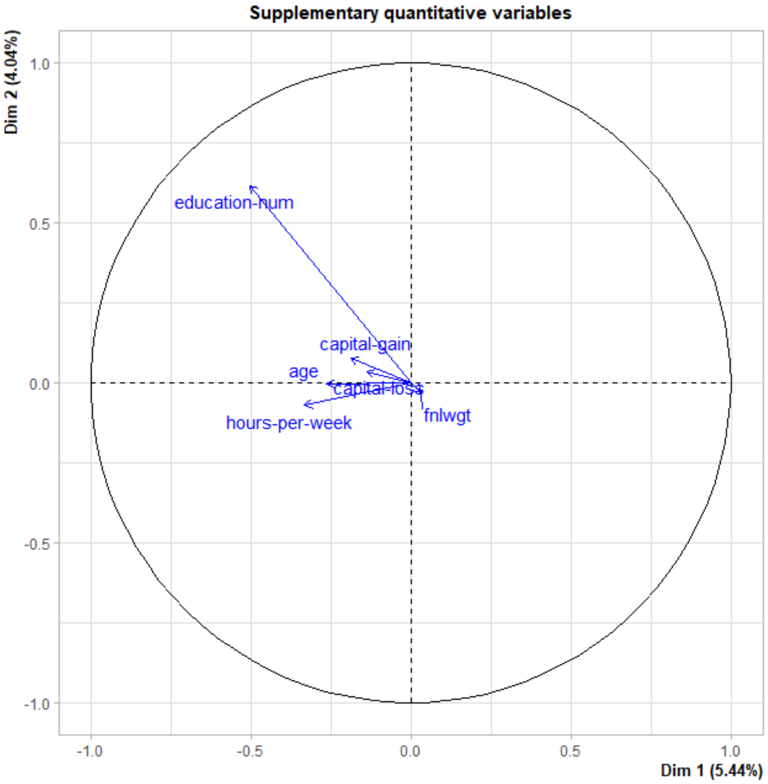

Figure 8: Biplot of variables by contribution

In this case, the years of education are related to both the first and the second dimension, given that the angle is merely the same with regard to both axes. Capital gains, age, hours worked per week and capital losses seem to be very correlated with the first dimension and among themselves. However, it seems that fnlwgt is not too relevant for the dimensions.

The following table represents the values of each numerical variable with dimensions:

Variable	Dimension1	Dimension2	Dimension3
Age	-0.264	-0.003	0.026
Fnlwgt	0.033	-0.029	0.020
Education-num	-0.503	0.617	-0.139
Capital-gain	-0.188	0.078	0.047
Capital-loss	-0.138	0.036	0.003
Hours-per-week	-0.332	-0.070	-0.068

Table 3: Numerical variables values

The variable education-number is correlated more to the first dimension in a negative way with a value of -0.5. The rest of the variables are related in a negative way too with low-mid values but fnlwgt, which has a correlation near zero. For dimension 2, we can see that the education-num is positively correlated with a value of 0.6, and the rest have low correlations with values near zero. Hence, the most important variable for the dimensions in the biplot will be the years of education. One can clearly see, then, why the categories of educational level that imply the most time of study are in the same direction as the vector education-num.

Finally, in Figure 9 we can observe that categorical variables that are more related to dimension 1 are v1 (whether the individual earns more or less than USD 50.000) and marital status, while variables that are more related to dimension 2 are occupation and education. Numerical variables are located near the origin so that we can tell that these variables are not well represented in these MCA biplots.

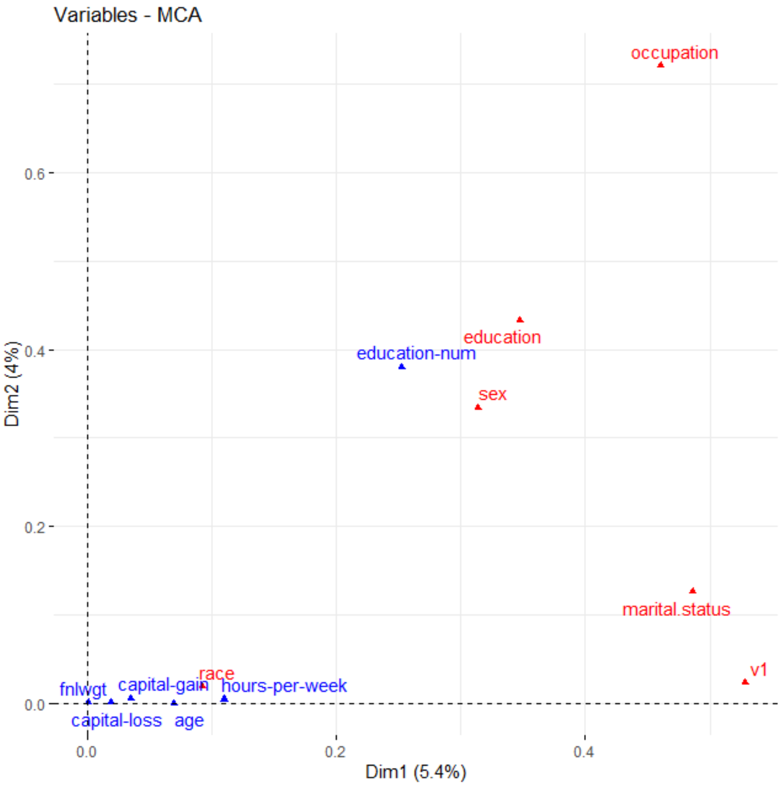

Figure 9: Biplot of variables by contribution

Once both analyses have been done, we can compare the results obtained from both so that the reader can make a global idea of the insights.

5 A Comparison between Results

The aim of this section is to compare the results obtained in both analysis so that one can understand the differences and the similarities that both yield.

In PCA we used 6 numerical variables: flnwght, capital-gain, capital-loss, education-number, hours-per-week, and age. As a result, we obtained that the first component could be explained principally through education-number and hours-per-week, which are related to the educational and professional field. The second one is more related to fnlwgtm which is a kind of factor of expansion of the survey (it gives similar values to similar individuals) and capital gains.

Also, when we plotted the individual features by categorical (as supplementary variables) and numerical variables we identify some patterns. First, people that have a salary greater than USD 50.000 are on the right side of the biplot and in the top right cluster, which are regarded as a socioeconomic elite or upper class, principally because of their high capital gains. Then, in the same area, we can identify some categories as males, people with high levels of education, and some specific jobs as managers, sales, and skilled ones. So, with this analysis, we could identify that people that work more hours, have more education, and have profits in their investments are the people located in the same right area of the biplot.

In the MCA analysis, we had a symmetrical rotation of the results, caused by the use of different principal axes. PCA works with Euclidean distances over individuals and it works with numerical variables, which form the principal components used as axes. In the case of MCA, we work with categories, so it uses a procedure to convert the factor into numbers (to be measurable) and it calculates the chi-square distance to get the results.

In this latter analysis, we observed that the most important categories in terms of contribution were females, salaries greater than USD 50.000, and prof-specialty. These categories are related to sex, v1 (salaries), and occupation variables. With this rotation, we obtained similar results as in PCA. In the end, it classifies individuals in the same areas. In the left part, we can find people with high salaries, males, married, high level of studies, and outstanding jobs. And, in the right part, we can find categories as female, lower levels of education, black, and separated.

Furthermore, the representation of the individuals in MCA is with dimensions 1 and 2 which explain less variance (9.4%) than PCA (39%. Consequently, we think that PCA makes a more understandable characterization of individuals and allows us to get more interpretable results for both variables and individuals (it is more insightful). For example, PCA identifies an important cluster in the top right side of the plot where are located people with salaries greater than USD 50.000, males with high education, and outstanding jobs. In the case of the MCA this cluster is not identified or is, at least, not as clear as with the PCA.

Appendix

Iker Caballero & Daniel González

2023-04-10

Load data

```
setwd("C:/Users/Ordenador/Desktop/Statistics Master/R and SAS/SAS/work")
library(readr)
library(dplyr)
adult <- read_csv("adult.data", col_names = FALSE)

colnames(adult) <- c('age','workclass','fnlwgt','education','education-num',
                     'marital-status','occupation','relationship','race',
                     'sex','capital-gain','capital-loss',
                     'hours-per-week','native-country','v1')
```

Data Preprocessing

```
dd<-adult

#set a list of numerical variables (with no missing values)
set.seed(123)

#separate numeric features
numeriques<-which(sapply(dd,is.numeric))
#numeriques

dcon<-dd[,numeriques]
sapply(dcon,class)
```

```
           age         fnlwgt  education-num   capital-gain   capital-loss
     "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
hours-per-week
     "numeric"
```

```
#separate categorical variables
df <- dd[,c(4,6,7,10,9,15)]
#convert as factor
df <- lapply( df, factor)
df<-as.data.frame(df)
```

```r
#scale numerical variables
dcon1<-as.data.frame(apply(dcon,2,scale))

#bind numerical and categorical into one data frame
df<-cbind(df,dcon1)

#head(df[,], 4)
summary(df)
```

```
        education                      marital.status             occupation
 HS-grad     :10501   Divorced             : 4443   Prof-specialty :4140
 Some-college: 7291   Married-AF-spouse    :   23   Craft-repair   :4099
 Bachelors   : 5355   Married-civ-spouse   :14976   Exec-managerial:4066
 Masters     : 1723   Married-spouse-absent:  418   Adm-clerical   :3770
 Assoc-voc   : 1382   Never-married        :10683   Sales          :3650
 11th        : 1175   Separated            : 1025   Other-service  :3295
 (Other)     : 5134   Widowed              :  993   (Other)        :9541
     sex                          race            v1            age
 Female:10771   Amer-Indian-Eskimo:  311   <=50K:24720   Min.   :-1.5822
 Male  :21790   Asian-Pac-Islander: 1039   >50K : 7841   1st Qu.:-0.7758
                Black             : 3124                 Median :-0.1160
                Other             :  271                 Mean   : 0.0000
                White             :27816                 3rd Qu.: 0.6905
                                                         Max.   : 3.7696

     fnlwgt        education-num       capital-gain      capital-loss
 Min.   :-1.6816   Min.   :-3.52960   Min.   :-0.1459   Min.   :-0.2167
 1st Qu.:-0.6817   1st Qu.:-0.42005   1st Qu.:-0.1459   1st Qu.:-0.2167
 Median :-0.1082   Median :-0.03136   Median :-0.1459   Median :-0.2167
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.4479   3rd Qu.: 0.74603   3rd Qu.:-0.1459   3rd Qu.:-0.2167
 Max.   :12.2684   Max.   : 2.30080   Max.   :13.3944   Max.   :10.5933

 hours-per-week
 Min.   :-3.19398
 1st Qu.:-0.03543
 Median :-0.03543
 Mean   : 0.00000
 3rd Qu.: 0.36951
 Max.   : 4.74289
```

```r
library(corrplot)

#calculate correlations
Mat_R<-cor(as.matrix(dcon1))
corrplot(Mat_R,upper = "pie",
         lower = "color",
         p.mat = Mat_R,
         tl.col="Black",
         tl.srt = 0,
         #order = "hclust",
         addrect = 3,
         pch.col = "black",
         insig = "p-value",
         sig.level = -1,)
```

PCA

Cumulative variance

```
# PCA of numeric variables
pc1 <- prcomp(dcon1,scale. = F)
class(pc1)
#attributes(pc1)

print(pc1)

#Total inertia
pc1$sdev
inerProj<- pc1$sdev^2
inerProj
totalIner<- sum(inerProj)
totalIner
pinerEix<- 100*inerProj/totalIner
pinerEix
barplot(pinerEix)

#Cummulated Inertia in subspaces, from first principal component to the 11th dimension subspace
barplot(100*cumsum(pc1$sdev[1:dim(dcon)[2]]^2)/dim(dcon)[2],xlab="Component")
percInerAccum<-100*cumsum(pc1$sdev[1:dim(dcon)[2]]^2)/dim(dcon)[2]
percInerAccum
```

PCA with Factor Miner

```
library(tidyverse)
library(FactoMineR)

#PCA
pca.car <- PCA(df,axes = c(1,2),quali.sup = c(1:6),scale.unit = F, graph = F)
```

```
#Significant components
pca.car$eig
barplot(pca.car$eig[,1],type="l",main="Screeplot")
```

```
#Significant components 70%
nd = 4

#Biplot of variables
plot(pca.car, axes = c(1, 2), choix = c("var"), title="Plot of variables",col.quali = c(1:6))
```

Biplots

```
#Biplot of individuals by most important numerical variables
plot(pca.car, axes = c(1, 2), choix = c("ind"), habillage=9, title="Plot of individuals", cex=0.7)
plot(pca.car, axes = c(1, 2), choix = c("ind"), habillage=10, title="Plot of individuals", cex=0.7)
plot(pca.car, axes = c(1, 2), choix = c("ind"), habillage=11, title="Plot of individuals", cex=0.7)
plot(pca.car, axes = c(1, 2), choix = c("ind"), habillage=12, title="Plot of individuals", cex=0.7)
```

Biplots by categorical Values

```
#Biplot of individuals by sex
library(ggplot2)
library(factoextra)
gp<-as.factor(dd$sex)
fviz_pca_biplot(pca.car,axes = c(1,2),
             col.ind = gp, # color by groups
             palette = c("#00AFBB", "#FC4E07"),
             addEllipses = TRUE, # Concentration ellipses
             ellipse.type = "confidence",
             legend.title = "Groups",
             repel = FALSE,label = T
)
```

```
gp<-as.factor(dd$v1)
fviz_pca_biplot(pca.car,axes = c(1,2),
             col.ind = gp, # color by groups
             palette = c("#00AFBB", "#FC4E07"),
             addEllipses = TRUE, # Concentration ellipses
             ellipse.type = "confidence",
             legend.title = "Groups",
             repel = FALSE,label = T
)
```

```
library(factoextra)
gp<-as.factor(dd$education)
fviz_pca_biplot(pca.car,axes = c(1,2),
             col.ind = gp, # color by groups
             #palette = c("#00AFBB", "#FC4E07"),
             addEllipses = TRUE, # Concentration ellipses
             ellipse.type = "confidence",
             legend.title = "Groups",
             repel = FALSE,label = T
)
```

```
library(factoextra)
gp<-as.factor(dd$occupation)
fviz_pca_biplot(pca.car,axes = c(1,2),
             col.ind = gp, # color by groups
             #palette = c("#00AFBB", "#FC4E07"),
             addEllipses = TRUE, # Concentration ellipses
             ellipse.type = "confidence",
             legend.title = "Groups",
             repel = FALSE,label = T
)
```

MCA

```
res.mca0<-MCA(df, graph=F,quanti.sup =c(7:12),ncp = Inf)

names(res.mca0)
print(res.mca0)
summary(res.mca0)

###EIGENVALUES
res.mca0$eig
head(res.mca0$eig)
length(res.mca0$eig[,1])
#barplot(res.mca0$eig[,1],main="EIGENVALUES",names.arg=1:nrow(res.mca0$eig))
#round(res.mca0$eig,2)
fviz_screeplot(res.mca0, addlabels = TRUE, ylim = c(0, 45))
###
totalInertia<- sum(res.mca0$eig[,1])
pinerEix<- 100*res.mca0$eig[,1]/totalInertia ### Equal to res.mca0$eig[,2]
(cumsum(pinerEix))

#The analysis will be performed with Dim1 and Dim2

#eigen gt 1/p=1/6=0.16
#1-17 dims eigen > 1/6

# Contributions
# CATEGORICAL VARIABLES - MODALITIES
#res.mca0$var$contrib
#(res.mca0$var$contrib[,1])
sum(res.mca0$var$contrib[,1]) ###Look that sum of total contributions for all variables in each Dim i i
# Individuals
#res.mca0$ind$contrib
sum(res.mca0$ind$contrib[,1])
```

```
### Biplots
#fviz_mca_biplot(res.mca0,repel = TRUE,
 #                ggtheme = theme_minimal())

##plot individuals
plot(res.mca0,invisible=c("var","quali.sup"),cex=0.7)
#res.mca0$ind$coord
ind <- get_mca_ind(res.mca0)
ind

#plot Variables
plot(res.mca0,invisible=c("ind","quali.sup"), cex=0.5)
```

```
### Step by step analysis for variables
var <- get_mca_var(res.mca0)
var

# coordinates of the variables
#res.mca0$var$coord
fviz_mca_var(res.mca0, choice = "mca.cor",
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())

#Coordinates of variable categories
head(round(var$coord, 2), 4)
fviz_mca_var(res.mca0,
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())
```

Biplots

```
#Contribution of variable categories to the dimensions
#head(round(var$contrib,2), 4)
# Total contribution to dimension 1 and 2
fviz_contrib(res.mca0, choice = "var", axes = 1, top = 15)
fviz_contrib(res.mca0, choice = "var", axes = 2, top = 15)

fviz_mca_var(res.mca0, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, # avoid text overlapping (slow)
             ggtheme = theme_minimal()
)

fviz_mca_var(res.mca0, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, # avoid text overlapping (slow)
             ggtheme = theme_minimal()
)

fviz_mca_ind(res.mca0,
             label = "none", # hide individual labels
             #habillage = "occupation",
             col.ind = "contrib",# color by groups
             #palette = c("#00AFBB", "#E7B800"),
             #addEllipses = TRUE, ellipse.type = "confidence",
             ggtheme = theme_minimal())


fviz_mca_ind(res.mca0,
             label = "none", # hide individual labels
             habillage = "occupation", # color by groups
             #palette = c("#00AFBB", "#E7B800"),
             addEllipses = TRUE, ellipse.type = "confidence",
             ggtheme = theme_minimal())
```

```
fviz_mca_ind(res.mca0,
             label = "none", # hide individual labels
             habillage = "education", # color by groups
             #palette = c("#00AFBB", "#E7B800","#00AF00"),
             addEllipses = TRUE, ellipse.type = "confidence",
             ggtheme = theme_minimal())

fviz_mca_ind(res.mca0,
             label = "none", # hide individual labels
             habillage = "marital.status", # color by groups
             #palette = c("#00AFBB", "#E7B800","#00AF00"),
             addEllipses = TRUE, ellipse.type = "confidence",
             ggtheme = theme_minimal())

fviz_mca_ind(res.mca0,
             label = "none", # hide individual labels
             habillage = "race", # color by groups
             #palette = c("#00AFBB", "#E7B800","#00AF00"),
             addEllipses = TRUE, ellipse.type = "confidence",
             ggtheme = theme_minimal())

fviz_ellipses(res.mca0, c("sex", "v1"),
              geom = "point")

###### Dimension description#######
####### correlated variables with a given dimension:
res.desc <- dimdesc(res.mca0, axes = c(1,2))
# Description of dimension 1
res.desc[[1]]
# Description of dimension 2
res.desc[[2]]
```