Correspondance Analysis

Richa Singh

10/3/2018

## Correspondance Analysis

Correspondence Analysis is a generalized principal component analysis tailored for the analysis of qualitative data. Also commonly known as reciprocal averaging is a multivariate statistical technique proposed by Herman Otto. It is conceptually similar to principal component analysis, but applies to categorical rather than continuous data. In a similar manner to principal component analysis, it provides a means of displaying or summarising a set of data in two-dimensional graphical form.This is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information which is similar in nature to those produced by Factor Analysis techniques, and they allow you to explore the structure of categorical variables included in the table.

## Dataset

The dataset is about Drug-use-by age which consists of 17 observations and 28 variables.This dataset describes about the variation of drugs people consume across all age groups. The dataset consists of both the frequency and the percentage of the drug. As,in statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. So, I dropped the variables showing the percentage count of drug use and kept only the variables displaying the median of the frequency count of drug use across all age group.

rm(list = ls())  
graphics.off()

data <- read.csv("drug-use-by-age(1).csv")  
data[is.na(data)] <- 0  
data$meth.frequency <- sub("-", "0.0", data$meth.frequency)  
data$oxycontin.frequency <- sub("-", "0.0", data$oxycontin.frequency)  
data$inhalant.frequency <- sub("-", "0.0", data$inhalant.frequency)  
data$heroin.frequency <- sub("-", "0.0", data$heroin.frequency)  
data$crack.frequency <- sub("-", "0.0", data$crack.frequency)  
data$cocaine.frequency <- sub("-", "0.0", data$cocaine.frequency)  
datasupp <- data[c(3,5,7,9,11,13,15,17,19,21,23,25,27) ]  
data <- data[ -c(2,3,5,7,9,11,13,15,17,19,21,23,25,27) ]  
datae <- data[,2:14]

Converting the char column type to numeric

datae <- lapply(datae, function(x) as.numeric(as.character(x)))  
x<- as.data.frame(datae)

## Heat Map

A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors.

library(data4PCCAR)  
library(PTCA4CATA)  
library(corrplot)

## corrplot 0.84 loaded

library(ggplot2)  
library(ExPosition)

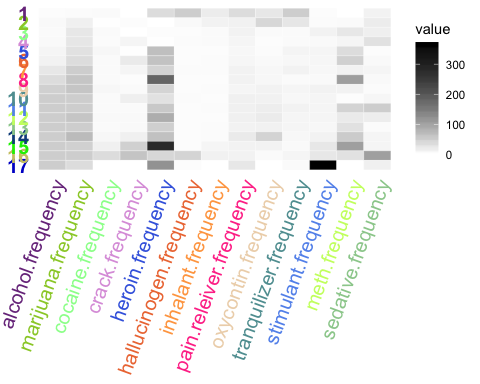
## Loading required package: prettyGraphs

library(InPosition)

##   
## Attaching package: 'InPosition'

## The following object is masked from 'package:PTCA4CATA':  
##   
## boot.ratio.test

col4J.drugs <- prettyGraphsColorSelection(NCOL(x))  
heatMapIJ <- makeggHeatMap4CT(x,  
 colorAttributes = col4J.drugs,  
 fontSize.x = 15)  
print(heatMapIJ)



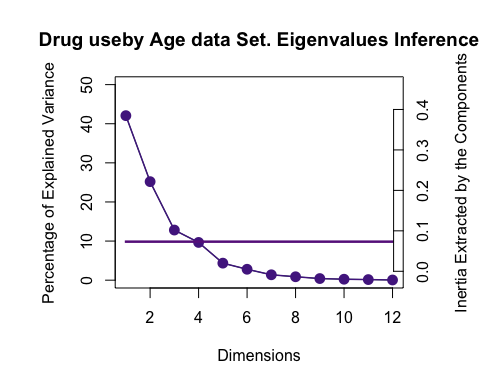
## Scree plot+ Inference Results

resCA.sym <- epCA(x, symmetric = TRUE,graphs = FALSE)  
# to run a plain CA but asymetric  
resCA.asym <- epCA(x, symmetric = FALSE,graphs = FALSE)  
  
drugs.sup <- supplementaryCols(SUP.DATA = datasupp, res = resCA.sym)  
  
rescainf <- epCA.inference.battery(x,graphs = FALSE)

## [1] "It is estimated that your iterations will take 0.03 minutes."  
## [1] "R is not in interactive() mode. Resample-based tests will be conducted. Please take note of the progress bar."  
## ===========================================================================

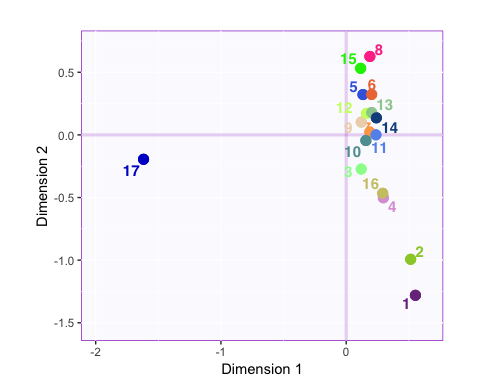
Fj.a <- resCA.asym$ExPosition.Data$fj  
Fi <- resCA.sym$ExPosition.Data$fi  
Fj <- resCA.sym$ExPosition.Data$fj  
  
color4drugs <-prettyGraphsColorSelection(NROW(data$age))  
  
constraints.sym <- minmaxHelper(mat1 = Fi, mat2 = Fj)  
constraints.asym <- minmaxHelper(mat1 = Fi, mat2 = Fj.a)  
constraints.sup <- minmaxHelper(mat1 = rbind(Fi, drugs.sup$fii),   
 mat2 = rbind(Fj, drugs.sup$fjj) )

library(PTCA4CATA)  
PlotScree(ev = resCA.sym$ExPosition.Data$eigs,   
 p.ev = rescainf$Inference.Data$components$p.vals,  
 title = 'Drug useby Age data Set. Eigenvalues Inference',  
 plotKaiser = TRUE  
)

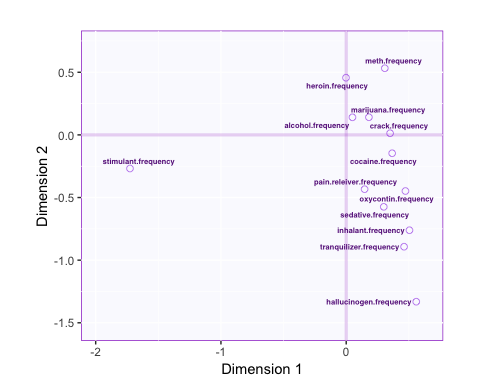


## Factor Maps for Symmetrical Graph

baseMap.i <- createFactorMap(Fi, constraints = constraints.sym,  
 col.points = color4drugs,  
 col.labels = color4drugs , cex = 5, text.cex = 4, pch = 20,  
 display.labels = TRUE , alpha.axes = 0.2,alpha.points = 1  
)  
print(baseMap.i$zeMap + baseMap.i$zeMap\_dots )



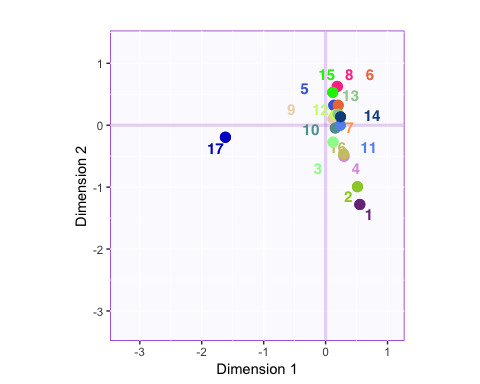
baseMap.j <- createFactorMap(Fj, constraints = constraints.sym,  
 color.points = "lightgreen",text.cex = 2,cex = 2,pch = 21)  
  
print(baseMap.j$zeMap)



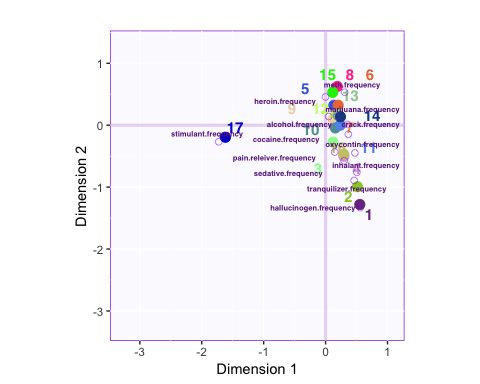
dev.new()  
print(baseMap.i$zeMap + baseMap.j$zeMap\_dots + baseMap.j$zeMap\_text)

## Factor Maps for Asymmetrical Graph

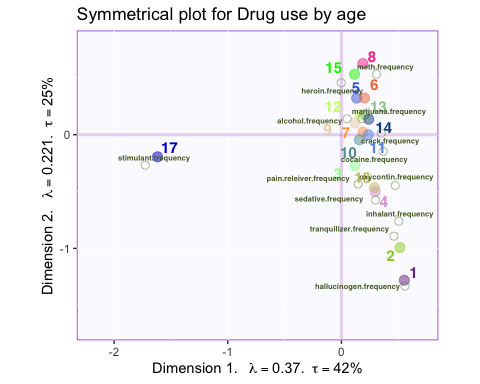
baseMap.i <- createFactorMap(Fi, constraints = constraints.asym,  
 col.points = color4drugs,  
 col.labels = color4drugs , cex = 5, text.cex = 4, pch = 20,  
 display.labels = TRUE , alpha.axes = 0.2,alpha.points = 1  
)  
print(baseMap.i$zeMap + baseMap.i$zeMap\_dots )



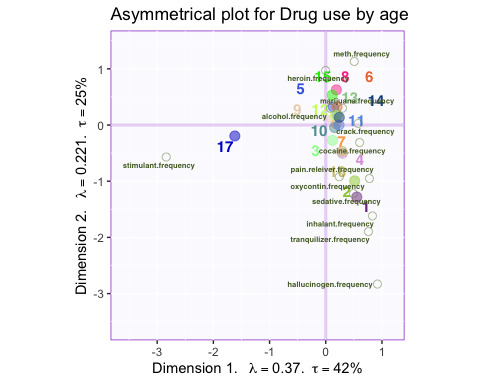
baseMap.j <- createFactorMap(Fj, constraints = constraints.asym,  
 color.points = "lightgreen",text.cex = 2,cex = 2,pch = 21)  
  
print(baseMap.i$zeMap + baseMap.j$zeMap\_dots + baseMap.j$zeMap\_text)



#only one can be asymmetric for createFactorMapIJ  
symMap <- createFactorMapIJ(Fi,Fj,  
 col.points.i = color4drugs,  
 col.labels.i = color4drugs, cex.i = 5, text.cex.i = 4, pch.i = 20,   
 pch.j = 21, text.cex.j =2, title = "Symmetrical plot for Drug use by age",  
 alpha.axes = 0.2,alpha.points.i = 1)  
  
asymMap <- createFactorMapIJ(Fi,Fj.a,  
 col.points.i = color4drugs,  
 col.labels.i = color4drugs, cex.i = 5, text.cex.i = 4, pch.i = 20,  
 pch.j = 21, text.cex.j = 2, title = "Asymmetrical plot for Drug use by age",  
 alpha.axes = 0.2,alpha.points.i = 1)  
labels4CA <- createxyLabels(resCA = resCA.sym)  
  
map.IJ.sym <- symMap$baseMap + symMap$I\_labels + symMap$I\_points +  
 symMap$J\_labels + symMap$J\_points + labels4CA  
print(map.IJ.sym)



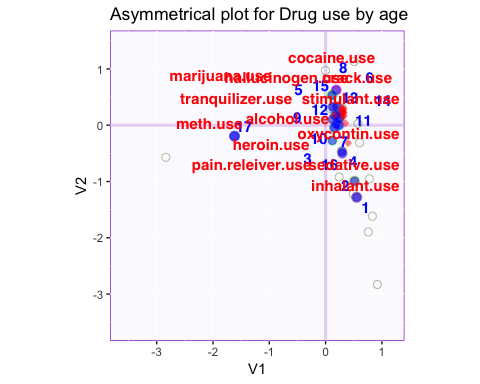
map.IJ.asym <- asymMap$baseMap + asymMap$I\_labels +   
 asymMap$I\_points + asymMap$J\_labels +   
 asymMap$J\_points + labels4CA  
print(map.IJ.asym)



mapSup <- createFactorMapIJ(Fi,   
 as.data.frame(drugs.sup$fjj),  
 col.points.i = 'blue',  
 col.labels.i = 'blue' ,  
 col.points.j = 'red',  
 col.labels.j = 'red',  
 constraints = constraints.sup  
)  
map.I.sup.asym <- asymMap$baseMap +   
 asymMap$I\_points + asymMap$J\_points +   
 mapSup$I\_labels + mapSup$I\_points + mapSup$J\_points + mapSup$J\_labels  
 labels4CA

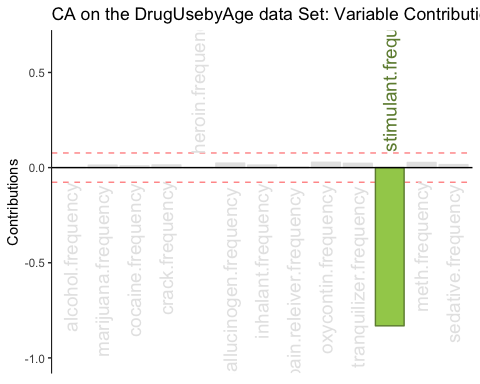
## $x  
## "Dimension " \* 1 \* ". " ~ ~lambda == 0.37 \* . ~ ~tau == 42 \*   
## "%"  
##   
## $y  
## "Dimension " \* 2 \* ". " ~ ~lambda == 0.221 \* . ~ ~tau == 25 \*   
## "%"  
##   
## attr(,"class")  
## [1] "labels"

print(map.I.sup.asym)

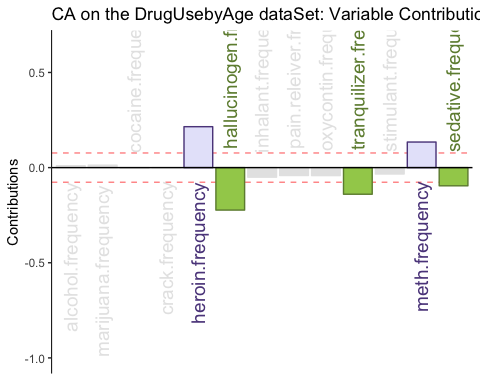


## Contribution Bars for variables

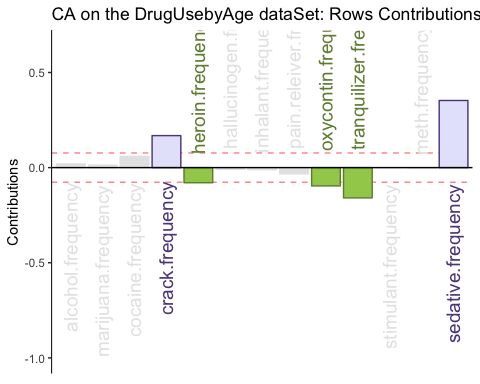
signed.ctrJ <- resCA.sym$ExPosition.Data$cj \* sign(resCA.sym$ExPosition.Data$fj)  
b003.ctrJ.s.1 <- PrettyBarPlot2(signed.ctrJ[,1],  
 threshold = 1 / NROW(signed.ctrJ),  
 font.size = 5,  
 #color4bar = gplots::col2hex(color4drugs), # we need hex code  
 main = 'CA on the DrugUsebyAge data Set: Variable Contributions (Signed)',  
 ylab = 'Contributions',  
 ylim = c(1.2\*min(signed.ctrJ), 1.2\*max(signed.ctrJ))  
)  
print(b003.ctrJ.s.1)



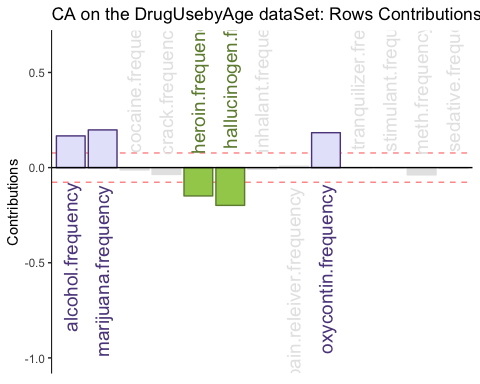
b004.ctrJ.s.2 <- PrettyBarPlot2(signed.ctrJ[,2],  
 threshold = 1 / NROW(signed.ctrJ),  
 font.size = 5,  
 #color4bar = gplots::col2hex(col4J.ibm), # we need hex code  
 main = 'CA on the DrugUsebyAge dataSet: Variable Contributions (Signed)',  
 ylab = 'Contributions',  
 ylim = c(1.2\*min(signed.ctrJ), 1.2\*max(signed.ctrJ))  
)  
print(b004.ctrJ.s.2)



b004.ctrJ.s.3 <- PrettyBarPlot2(signed.ctrJ[,3],  
 threshold = 1 / NROW(signed.ctrJ),  
 font.size = 5,  
 #color4bar = gplots::col2hex(col4J.ibm), # we need hex code  
 main = 'CA on the DrugUsebyAge dataSet: Rows Contributions (Signed)',  
 ylab = 'Contributions',  
 ylim = c(1.2\*min(signed.ctrJ), 1.2\*max(signed.ctrJ))  
)  
print(b004.ctrJ.s.3)

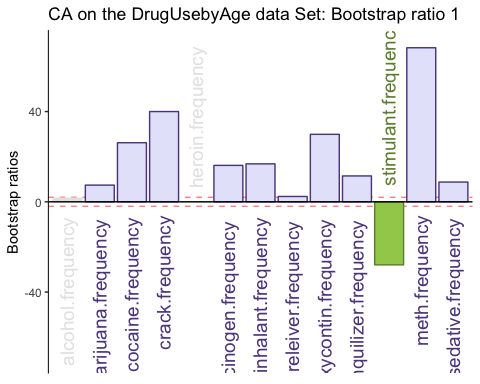


b004.ctrJ.s.4 <- PrettyBarPlot2(signed.ctrJ[,4],  
 threshold = 1 / NROW(signed.ctrJ),  
 font.size = 5,  
 #color4bar = gplots::col2hex(col4J.ibm), # we need hex code  
 main = 'CA on the DrugUsebyAge dataSet: Rows Contributions (Signed)',  
 ylab = 'Contributions',  
 ylim = c(1.2\*min(signed.ctrJ), 1.2\*max(signed.ctrJ))  
)  
print(b004.ctrJ.s.4)

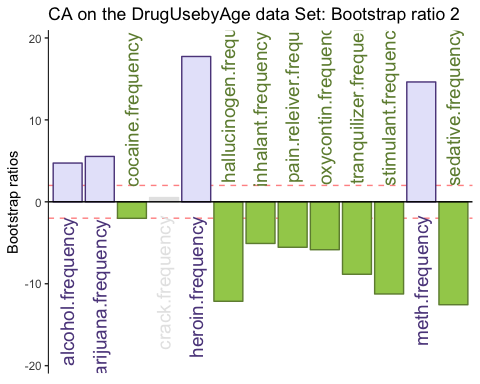


## Bootstrap Ratios for variables

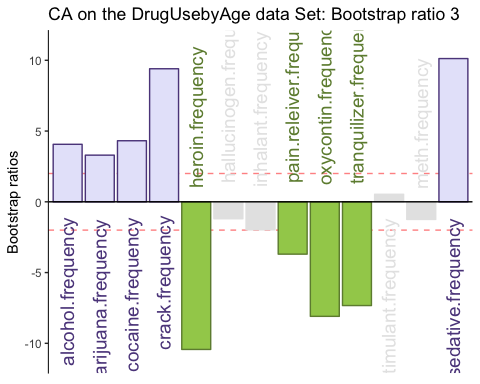
BR <- rescainf$Inference.Data$fj.boots$tests$boot.ratios  
laDim = 1  
ba001.BR1 <- PrettyBarPlot2(BR[,laDim],  
 threshold = 2,  
 font.size = 5,  
 #color4bar = gplots::col2hex(col4J.ibm),  
 main = paste0( 'CA on the DrugUsebyAge data Set: Bootstrap ratio ',laDim),  
 ylab = 'Bootstrap ratios'  
 #ylim = c(1.2\*min(BR[,laDim]), 1.2\*max(BR[,laDim]))  
)  
print(ba001.BR1)



#  
laDim = 2  
ba002.BR2 <- PrettyBarPlot2(BR[,laDim],  
 threshold = 2,  
 font.size = 5,  
 #color4bar = gplots::col2hex(col4J.ibm),  
 main = paste0(  
 'CA on the DrugUsebyAge data Set: Bootstrap ratio ',laDim),  
 ylab = 'Bootstrap ratios'  
)  
print(ba002.BR2)



laDim = 3  
ba002.BR3 <- PrettyBarPlot2(BR[,laDim],  
 threshold = 2,  
 font.size = 5,  
 main = paste0(  
 'CA on the DrugUsebyAge data Set: Bootstrap ratio ',laDim),  
 ylab = 'Bootstrap ratios'  
)  
print(ba002.BR3)



## Summary

From the Data Analysis we can infer that :

* Old aged people above the age of 65years are generally heavy on stimulants than the average use
* Kids around the age of 12 were consuming more hallucinogen drug than the average use
* Teenagers around the age of 13 were consuming more tranquilizer drug than the average use
* Teenagers around the age of 14 were consuming more sedative drug than the average use
* Teenagers around the age of 16 were consuming more heroin drug than the average use
* Teenagers around the age of 17 were consuming more heroin drug than the average use
* Youngsters around the age of 20 were consuming more alcholol and marijuana than the average use
* People around the age of 35-49 were consuming more heroin and meth drug than the average use
* People around the age of 22-23 were consuming more cocaine and crack drug than the average use
* Youngsters/People around the age of 24-34 were consuming more marijuana than the average use