Tema: Reducción de dimensionalidad

Ejercicios de repaso

MSc. Alvaro Chirino

Octubre, 2022

Análisis de Componentes principales

Ejercicio 1

Usando la sección de discriminación de las EH defina un indicador basado en el primer componente principal.

- ¿Qué porcentaje de varianza se explica?
- Existe alguna relación con el nivel de pobreza
- Existe alguna relación con el sexo de la persona

Ejercicio 2

Usando la base de datos de https://www.kaggle.com/datasets/mohansacharya/graduate-admissions selecciones todas las variables excepto el *serial* y el *chance of admit*, realice el ACP y describa:

- Componentes a retener con el criterio de los eigen valores (eigen valores mayores a 1)
- ¿Qué explican los componentes retenidos?

Ejercicio 3

Indagar que es el biplot y cual su interpretación en los componentes principales. Graficar el biplot para los dos ejercicios previos. (Sugerencia: explorar la librería FactoMineR)

Análisis de correspondencia

Función de referencia para el AC y ACM

```
md_ac<-function(N,bdacm=NULL,acm=F,dm=2){
  require(Matrix)#rango de la matriz
  require(ggplot2)# visual
  require(ggrepel)# mejor manejo de las etiquetas
  library(rgl)# 3D
  require(stringi) # limpieza para texto
  library(fastDummies) # ACM
```

```r
if(acm){
  bdacm<-bdacm %>% to_factor()
  N<-dummy_cols(bdacm,remove_selected_columns = T)
  N<-as.matrix(N)
}
I<-nrow(N)
J<-ncol(N)
P<-N/sum(N)
r<-apply(P,1,sum)
c<-apply(P,2,sum)
Dr12<-diag(r^(-0.5))
Dc12<-diag(c^(-0.5))
S<-Dr12%*%(P-r%*%t(c))%*%Dc12
aux<-svd(S)
U<-aux$u
V<-aux$v
D<-aux$d
#Filas y columnas comparables
FF<-Dr12%*%U%*%diag(D)
GG<-Dc12%*%V%*%diag(D)
#Filas y columnas independientes (ACM)
X<-Dr12%*%U
Y<-Dc12%*%V # principalmente
rlab<-rownames(N)
clab<-colnames(N)
#inercia
inercia<-round(prop.table(aux$d^2)*100,3)
if(!acm){
  #FF GG to data frame
  rr<-rankMatrix(S)[1]
  df1<-data.frame(FF[,1:rr])
  df2<-data.frame(GG[,1:rr])
  colnames(df1)<-paste0("d",1:rr)
  colnames(df2)<-paste0("d",1:rr)
  df1<-df1 %>% mutate(vv="v1",cc="darkblue")
  df2<-df2 %>% mutate(vv="v2",cc="darkred")
  bd<-bind_rows(data.frame(df1,vlab=rlab,masa=r),
  data.frame(df2,vlab=clab,masa=c))
  bd<-bd %>% mutate(aux=1)
}
if(acm){
  #Y to data frame
  rr<-rankMatrix(S)[1]
  bd<-data.frame(Y[,1:rr])
  colnames(bd)<-paste0("d",1:rr)
  auxnn<-NULL
  auxll<-NULL
  k<-1
  for(i in names(bdacm)){
    auxnn[k]<-length(unique(bdacm[[i]]))
    auxll<-c(auxll,levels(bdacm[[i]]))
    k<-k+1
  }
```

```
    bd<-bd %>% mutate(vv=rep(names(bdacm),auxnn),
                      cc=rep(1:length(names(bdacm)),auxnn),
                      vlab=auxll,masa=c,aux=1)
  }
  #acentos
    bd$vlab3d<-stri_trans_general(bd$vlab,id = "Latin-ASCII")
  if(dm==1){
    f1<-ggplot(bd,aes(d1,aux,col=vv,size=masa,label=vlab))+geom_label_repel(max.overlaps=Inf)+scale_siz
  }
  if(dm==2){
    f1<-ggplot(bd,aes(d1,d2,size=masa,col=vv,label=vlab))+geom_label_repel(max.overlaps=Inf)+scale_size
  }
  if(dm==3){
    plot3d(bd$d1,bd$d2,bd$d3,type="n",
           xlab=substitute(paste(lambda[1],"=",ine1,"%"),list(ine1=inercia[1])),
           ylab=substitute(paste(lambda[2],"=",ine1,"%"),list(ine1=inercia[2])),
    zlab=substitute(paste(lambda[3],"=",ine1,"%"),list(ine1=inercia[3])))
    text3d(bd$d1,bd$d2,bd$d3,bd$vlab3d,col=bd$cc,cex=bd$masa*5)
  }
  if(dm<3){print(f1)}
}
```

Ejercicio 1

Usando la encuesta a hogares plantear una relación entre 2 y 3 variables respectivamente y realizar el AC Y ACM. Elija una relación *interesante* y comentar los resultados.

Ejercicio 2

Realizar los ajustes necesarios en la función md_ac para conocer el rango de la matriz S y obtener las coordenadas F, G, X e Y.

Ejercicio 3

Realizar los ajustes necesarios en la función md_ac para trabajar con variables discretas.

Ejercicio 4

Usando la EH realice el ACM usando las variables área, quintiles para el ingreso laboral (ylab) y años de educación, tomar en cuenta a las personas de 25 años o más