Tarea 6

Hairo Ulises Miranda Belmonte

16 de abril de 2019

EJERCICIO 1

1. Muestra que la matriz de covarianza

$$\rho = \begin{pmatrix} 1 & .63 & .45 \\ .63 & 1 & .35 \\ .45 & .35 & 1 \end{pmatrix}$$

para $p = 3$, con las variables aleatorias estandarizadas Z_1, Z_2, Z_3, , puede ser generada por el modelo de factores con $m = 1$

$$Z_1 = .9F_1 + \epsilon_1$$

$$Z_2 = .7F_2 + \epsilon_2$$

$$Z_3 = .5F_3 + \epsilon_3$$

donde $Var(F_1) = 1, Cov(\epsilon, F_1) = 0$, y

$$\Psi = \begin{pmatrix} .19 & 0 & 0 \\ 0 & .51 & 0 \\ 0 & 0 & .75 \end{pmatrix}$$

esto se esribe como $\rho = LL' + \Psi$

Respuesta

Sea $L = [.9, .7, .5]'$, entonces:

$$\begin{pmatrix} .9 & .7 & .5 \end{pmatrix} \begin{pmatrix} .9 \\ .7 \\ .5 \end{pmatrix} = \begin{pmatrix} .81 & .63 & .45 \\ .63 & .49 & .35 \end{pmatrix}$$

Por lo tanto, si se intenta aproximar $\rho = LL' + \Psi$, se tiene:

$$\rho = \begin{pmatrix} .81 & .63 & .45 \\ .63 & .49 & .35 \\ .45 & .35 & .25 \end{pmatrix} + \begin{pmatrix} .19 & 0 & 0 \\ 0 & .51 & 0 \\ 0 & 0 & .75 \end{pmatrix} = \begin{pmatrix} 1 & .63 & .45 \\ .63 & 1 & .35 \\ .45 & .35 & 1 \end{pmatrix}$$

concluyendo que si se aproxima buen y si se genera el modelo que se indica con un factor.

EJERCICIO 2

Usa la información del ejercicio anterior

a) Calcula las comunalidades h_i^2 con $i = 1, 2, 3$

Sabemos que la diagonal de $LL^{'}$ son las comunalidades h_i^2, entonces:

$$h_1^2 = 0.81$$

$$h_2^2 = 0.49$$

$$h_3^2 = 0.25$$

b) Calcula $Cov(Z_i, F_1)$ ¿Cual variable podría llevar el mayor peso en la interpretación del factor común ? Porqué?

$$Cov(Z_i, F1) = L$$

$$L = \begin{pmatrix} .97 \\ .7 \\ .5 \end{pmatrix}$$

$$Cov(Z_1, F1) = l_{11} = .97$$

$$Cov(Z_2, F1) = l_{21} = .7$$

$$Cov(Z_3, F1) = l_{31} = .5$$

En conclusión, la variable Z_1 tiene más correlación respecto al factor F_1; de esta manera, puede tener mayor peso sobre ese factor.

EJERCICIO 3

Los valores y vectores propios de la matriz de correlaciones ???? en el ejercicio 1 son

$$\lambda_1 = 1.96; \;\; e_1^{'} = [.625, .593, .507]^{'}$$

$$\lambda_2 = 0.68; \;\; e_2^{'} = [-.219, -.491, .843]^{'}$$

$$\lambda_3 = 0.36; \;\; e_3^{'} = [.749, -.638, -.177]^{'}$$

a) Asumiendo un modelo de factores con m=1, calcula la matriz de cargas L y la matriz de varianzas específicas ???? usando el método por componentes principales. Compara los resultados con los del ejercicio 1.

$$L = \sqrt{(}\lambda_1)e_1$$

$$\sqrt{(1.96)} \begin{pmatrix} .625 \\ .593 \\ .507 \end{pmatrix} = \begin{pmatrix} .8750 \\ .8302 \\ .7098 \end{pmatrix}$$

Entonces:

$$\begin{pmatrix} .8750 \\ .8302 \\ .7098 \end{pmatrix} \approx \begin{pmatrix} .9 \\ .7 \\ .5 \end{pmatrix}$$

Como se observa al estimar el factor con PCA y contrastarlo respecto a las cargas del ejercicio anterior, se observa que la estimación de las cargas son ligeramenta distintas.

b) Qué proporción de la varianza poblacional total es explicada por el primer factor común?

Como se trabaja con la matriz de correlación, la proporción de la varianza total explicada se calcula de la siguiente forma:

$$\lambda_1/p$$

con $p = 3$, y $\lambda_1 = 1.96$; por lo tanto el total de la varianza explicada es:

$$\frac{\lambda_1}{p} = \frac{1.96}{3} = .65$$

EJERCICIO 4

4. (Solución única pero impropia: caso Heywood). Considere un modelo factorial con m=1 para la población con matriz de covarianza

$$\Sigma = \begin{pmatrix} 1 & 0.4 & 0.9 \\ 0.4 & 1 & 0.7 \\ 0.9 & 0.7 & 1 \end{pmatrix}$$

Muestra que existe una única elección de L y Ψ con $\Sigma = LL' + \Psi$, pero que $\psi_3 < 0$ por lo que la elección no es admisible.

Se realiza la siguiente representación matricial

$$\Sigma = LL' + \Psi = \begin{pmatrix} 1 = l_{11}^2 + \psi_1 & 0.4 = l_{11}l_{21} & 0.9 = l_{11}l_{31} \\ & 1 = l_{21}^2 + \psi_2 & 0.7 = l_{21}l_{31} \\ & & 1 = l_{31}^2 + \psi_3 \end{pmatrix}$$

1) Igualando l_{31} se tiene

$$\frac{.9}{l_{11}} = l_{31}$$

$$\frac{.7}{l_{21}} = l_{31}$$

$$\frac{.9}{l_{11}} = \frac{.7}{l_{21}}$$

$$\frac{l_{11}}{l_{21}} = \frac{.9}{.7}$$

2) sustituyendo con $l_{11}l_{21} = .4$

$$l_{11}^2 = \frac{l_{11}}{l_{21}}(l_{11}l_{21}) = \frac{.9}{.7}(.4) = .514$$

entonces

$$l_{11} = \pm .717$$

3) sustituyendo para encontrar l_{21}

$$l_{21} = \frac{.4}{\pm .717} = \pm .558$$

4) Finalmente

$$l_{11}l_{31} = .9$$

$$l_{31} = \frac{.9}{l_{11}}$$

$$l_{31} = \frac{.9}{\pm .717} = \pm 1.2155$$

entonces, las cargas son:

$$l_{11} = \pm .717$$

$$l_{21} = \pm .558$$

$$l_{31} = \pm 1.255$$

Calculando LL'

$$LL' = \begin{pmatrix} .717 \\ .558 \\ 1.255 \end{pmatrix} \begin{pmatrix} .717 & .558 & 1.255 \end{pmatrix} = \begin{pmatrix} .514 & 0.4 & 0.9 \\ 0.4 & .3111 & 0.7 \\ 0.9 & 0.7 & 1.575 \end{pmatrix}$$

sabemos que $1 = h_i^2 + \psi_i$; entonces, despejando ψ_i se tiene que $\psi_i = 1 - h_i^2$, para ψ_3

$$\psi_3 = 1 - 1.575 = -.575$$

Por lo tanto, queda demostrado que $\psi_3 < 0$.

EJERCICIO 5

1.El Proyecto de Evaluación de la Apertura Sintética de la Personalidad (SAPA) es una colección de datos psicológicos basada en la web.2 Un subconjunto de los datos está disponible en R como bfi en la biblioteca “psych”.

(a) Utilice el conjunto de datos /states.rds/.

```
# Bibliotecas

library("magrittr")
library("knitr")
library("kableExtra")
library("corrplot")
library("tidyverse")
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
# Direccion
#getwd()
#setwd("C:/Users/h_air/Desktop/CIMATMCE/
     # Semestre_2/Multivariado/Tarea/Dr. Rodrigo/Tarea 5")
```

```
library("psych")
data(bfi)
data("bfi.dictionary")
```

a) Utilice el comando complete.cases () para eliminar individuos en bfi con cualquier valor faltantes

Se retiran los valores faltantes y se calcula la matriz de correlación de los datos; se opta por la matriz de correlación por las diferentes unidades de medidas en las variables demograficas.

```
datos <- bfi[complete.cases(bfi),]
# se estandarizan por tene unidades de medidas distintas
R <- cor(datos)
```

b) Utilice el análisis de factores para agrupar elementos de naturaleza similar. Trata de interpretar la naturaleza de los ítems que se agrupan. Este es un ejercicio útil en psicología. El test de ji cuadrado para el número de factores puede no ser apropiado con una muestra tan grande.

```
datos <- bfi[complete.cases(bfi),]
# se estandarizan por tene unidades de medidas distintas
R <- cor(datos)
```

Primero observamos la existencia de estructura de correlación en las variables. Se realiza la prueba esferica de Barlett para probar la hipótesis nula de n correlación entre las variables.

```
cortest.bartlett(R, n=nrow(datos))
```

```
## $chisq
## [1] 17331.21
##
## $p.value
## [1] 0
##
## $df
## [1] 378
```

Se rechaza la H_o de no correlación; por lo tanto se procede al análsiis de factores.

Ahora se utiliza el pindice KMO, se observa un valor MSA de $.84$, el cual bajó la clasificación del KMO, indica que los datos son buenos para realizar el análisis de factores.

```
# Indice KMO
# - KMO > 0.90    Muy bueno
# - 0.80<KMO<0.90 Bueno
# - 0.70<KMO<0.80 Aceptable
# - 0.60<KMO<0.70 Regular
# - 0.50<KMO<0.60 Malo
# - KMO < 0.50    Inaceptable
KMO(datos)
```

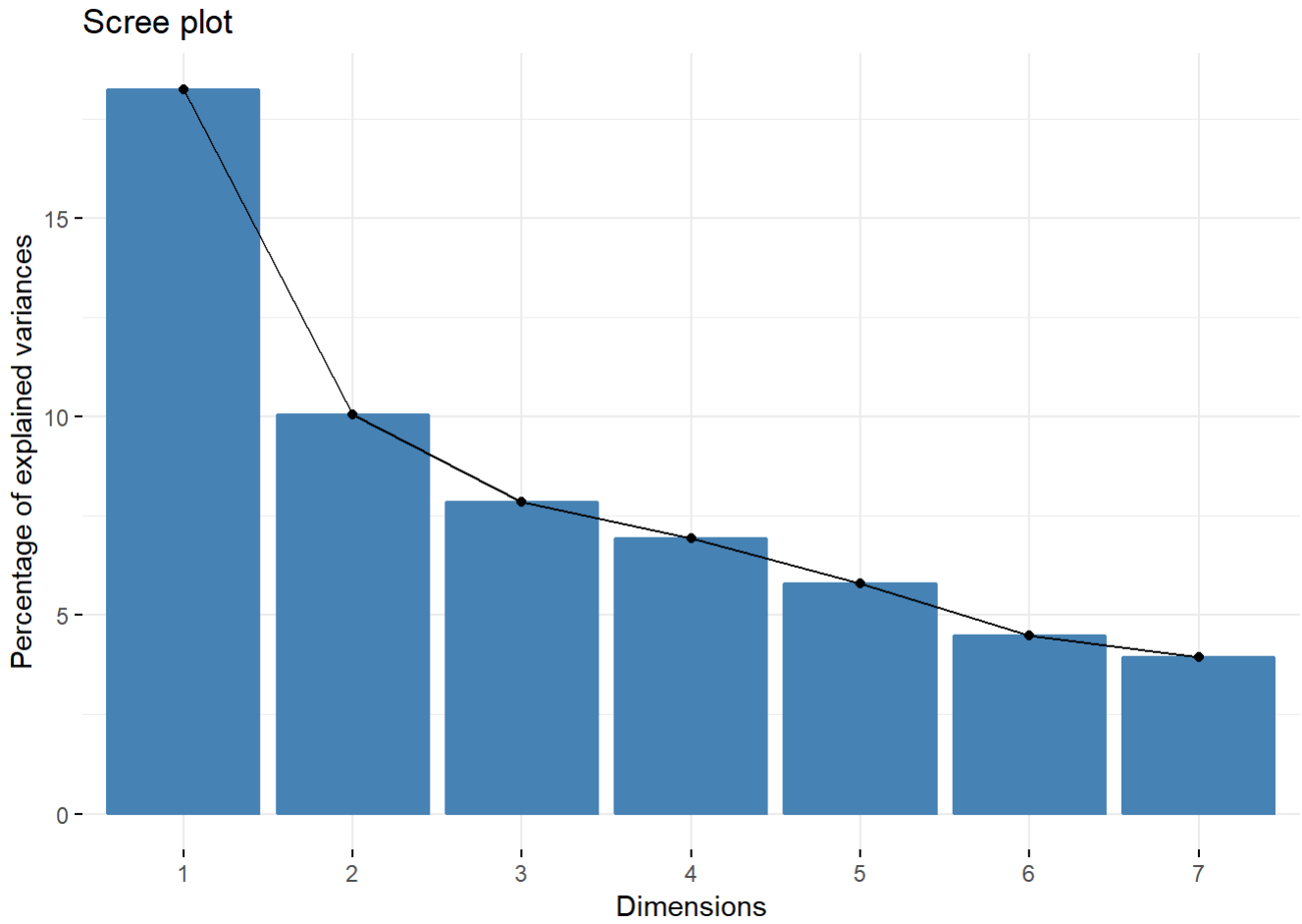
```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = datos)
## Overall MSA =  0.84
## MSA for each item =
##        A1        A2        A3        A4        A5        C1        C2
##      0.75      0.84      0.87      0.86      0.90      0.83      0.78
##        C3        C4        C5        E1        E2        E3        E4
##      0.84      0.83      0.86      0.85      0.88      0.89      0.87
##        E5        N1        N2        N3        N4        N5        O1
##      0.89      0.77      0.78      0.86      0.88      0.85      0.85
##        O2        O3        O4        O5    gender education       age
##      0.78      0.83      0.79      0.76      0.71      0.58      0.61
```

Se estiman factores por PCA. Se gráfica el screeplot para dar idea de cuantos factores utilizar; se sugieren 2 o 3.

```
library("factoextra")
library("ade4")
```

```
## Warning: package 'ade4' was built under R version 3.5.3
```

```
ven.pca <- dudi.pca(datos, scannf = FALSE, nf = 2,scale = T)
eig.val <- get_eigenvalue(ven.pca)
fviz_screeplot(ven.pca, ncp=7)
```



Se decide estimar 2 factores con PCA.

Nota: a manera de ejercicio, se programa la función para que arroje los factores estimados con PCA, contrastando respecto a los de la función dudi.pca. Se observa que básicamente son los mismos; no obstantes, en esta tarea se sigue utilizando los resultados de la función dudi.pca.

```
# factores con componentes principales
FactoresPCA <- function(Sn, m, p)
{
  Factores <- matrix(0L, p, m)
  propios <- Sn %>% eigen
  vectores <- propios$vectors
  valores <- propios$values
  for(i in 1:m) Factores[,i] <- sqrt(valores[i])*vectores[,i]
  phi <- diag(p)- Factores%*%t(Factores)
  phi <- phi %>% diag
  comunalidaes <- Factores%*%t(Factores)%>% diag
  LL <- Factores%*%t(Factores)
  Residual  = Sn - LL - diag(phi)
  return(list(Factor = Factores, Phi = phi, Comunalidades = comunalidaes, LL = LL, Residual = Re
sidual))
}
```

Cargas de los factores función propia:

```
# factores con mi función
RESULTADO <- FactoresPCA(R,2,dim(R)[2])
RESULTADO$Factor %>% head(5)
```

```
##            [,1]       [,2]
## [1,]  0.2482925  0.05542447
## [2,] -0.4889084 -0.34471675
## [3,] -0.5518702 -0.34350571
## [4,] -0.4442872 -0.15754724
## [5,] -0.6022609 -0.20568389
```

Utilizando dudi.pca la carga de los factores son:

```
# factores con paqueteria
ven.pca$co %>% head(5)
```

```
##         Comp1       Comp2
## A1  0.2482925  0.05542447
## A2 -0.4889084 -0.34471675
## A3 -0.5518702 -0.34350571
## A4 -0.4442872 -0.15754724
## A5 -0.6022609 -0.20568389
```

Se realiza el test ji cuadrado para el número de factores que ajusten a los datos; primero se ve si en este ejemplo se puede utilizar la prueba, haciendo uso del siguiente críterio:

$$m < \frac{1}{2}(2 * p + 1 - \sqrt{(}8 * p + 1)$$

```
m <- 2
p <- dim(datos)[2]
m < .5*(2*p + 1 - sqrt(8*p + 1))
```

```
## [1] TRUE
```

El test se puede realizar, ya que $m = 2$ y $\frac{1}{2}(2 * p + 1 - \sqrt{(}8 * p + 1) = 21$

Se realiza la prueba:

```
testFactores <- function(m, LL, phi, Sn, alpha, n, p)
{
  EstPrueba <- (n-1-((2*p+4*m+5)/6))*log(det(LL+phi)/det(Sn))
  gradosLibertad <- (((p-m)^2)-p-m)/2
  ValCrit <- qchisq(alpha,gradosLibertad, lower.tail = F)
  Resultado <- EstPrueba > ValCrit
  return(list(estadistico = EstPrueba, gl = gradosLibertad, critico = ValCrit, Rehaza = Resultad
o))
}

testFactores(m, RESULTADO$LL, diag(RESULTADO$Phi), R, .05,
             length(datos[,1]), dim(datos)[2])
```

```
## $estadistico
## [1] 3525.422
##
## $gl
## [1] 323
##
## $critico
## [1] 365.9123
##
## $Rehaza
## [1] TRUE
```

$$(n-1-(2p+4m+5)/6)ln(\frac{|\hat{L}\hat{L}^{'}+\hat{\Psi}|}{|S_n|}) = 3525.442$$

$$\chi^2_{\frac{(p-m)^2-p-m}{2}} = 365.9123$$

Se tiene evidencia suficiente para rechazar la hipótesis nula; de esta manera, dos factores no son suficiente. Sin embargo, dado a que los factores se estiman por PCA, se decide utilizar el críterio del codo, en el screeplot, y tomar 2 factores.

Utilizando las funciones en R, se puede calcular la calidad de la representación y las contribuciones de las variables sobre los factores; asimismo, la de los individuos.

La contribuciones relativas de las variables como coordenadas al cuadrado son:

```
# Tambien las podemos calcular como las coordenadas al cuadrado
head(ven.pca$co^2, 28)
```

```
##                Comp1        Comp2
## A1        0.061649176 0.0030718716
## A2        0.239031458 0.1188296375
## A3        0.304560673 0.1179961762
## A4        0.197391089 0.0248211336
## A5        0.362718211 0.0423058619
## C1        0.129999697 0.0128402773
## C2        0.120501995 0.0353048975
## C3        0.126846725 0.0018877523
## C4        0.241738961 0.0239965419
## C5        0.268164639 0.0332296779
## E1        0.202937052 0.0533502305
## E2        0.416841866 0.0017920767
## E3        0.325902494 0.1190768813
## E4        0.394183915 0.0388055235
## E5        0.316436746 0.0869119268
## N1        0.177868637 0.4310530102
## N2        0.166462271 0.4286418301
## N3        0.162534664 0.4641303472
## N4        0.299959236 0.2103717493
## N5        0.134854807 0.2877942286
## O1        0.125867676 0.0245083146
## O2        0.046621207 0.0210937227
## O3        0.176063722 0.0690812530
## O4        0.008071515 0.0807937837
## O5        0.049960368 0.0003349460
## gender    0.014861711 0.0743757148
## education 0.005156327 0.0004883784
## age       0.030172383 0.0051199474
```

Vriables que hablan sobre la amistad (A), extroversión (E) y conciencia (C), contribuyen más al factor 1; las variables que representan al neurotismo (N), aportan más al factor 2; el resto de variables como las de sinceridad y aspectos demográficos, parecen no aportar de forma considerable a los primeros dos factores.

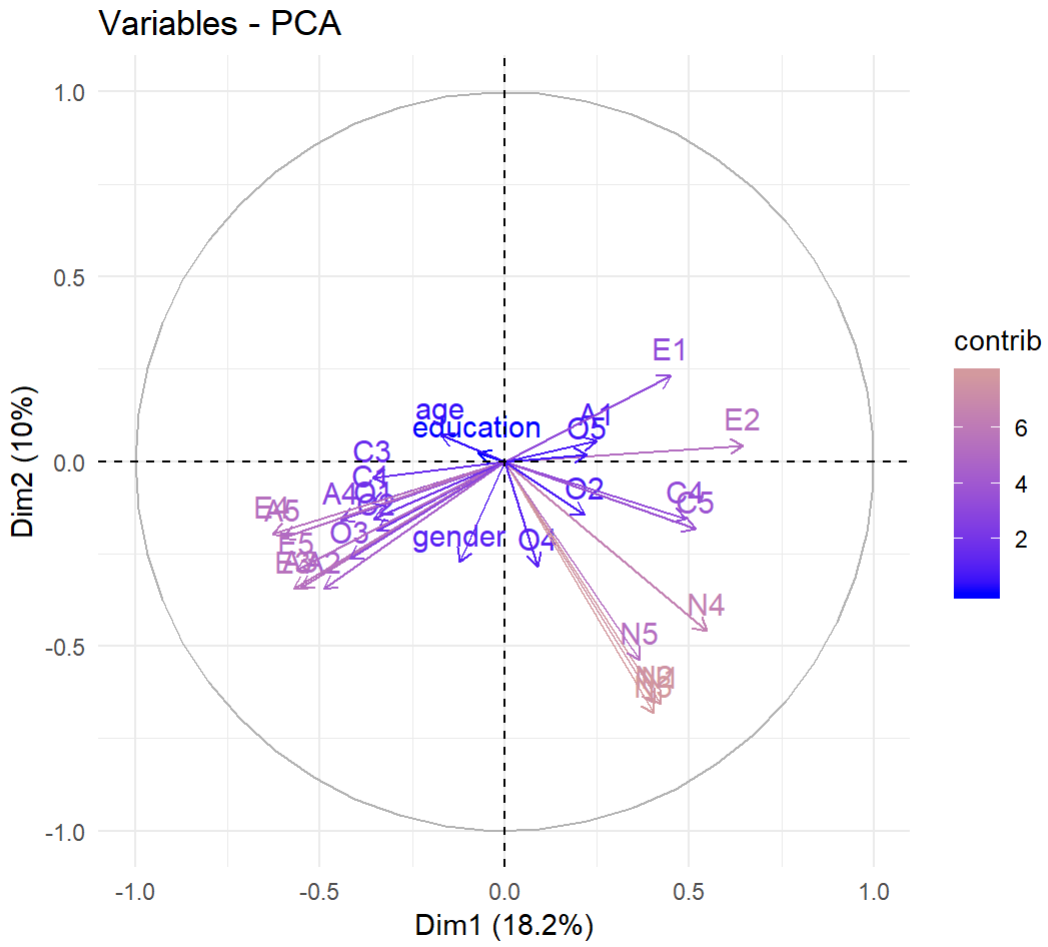
Bajo esta representación, al contar con 28 variables, los resultados no se aprecian lo suficiente; entonces, se decide realizar un análisis gráfico.

Se realiza el mapa de factores para $m = 2$, donde se observa que las variables (preguntas) de la misma clase se agrupan; no obstante, existen algunas carácteristicas que se mezclan; i.e., en la base se tienen 5 tipos de preguntas; sobre la amabilidad (A), conciencia (C), extraversión (E), neurotismo (N) y franquza (O).

En el mapa factorial se puede ver cosas interesantes, en la representción de los dos primeros factores, se observan grupos marcados, las variables que hablan sobre la amabilidad, lo extrovertido y lo conciente que es un individuo, en el lado izquierdo, contribuyendo más al primer factor; y las variables que hacen referencia al neurotismo contribuyen más al segundo factor.

Dentro de las que contribuyen más al primer factor, se encuentra un contraste, el cual hace referencia a lo extrovertido que son los individuos; i.e., las variables E1 y E2 respecto a la E3,E4 Y E5; lo cual tiene sentido, ya que E1, indica que si la persona habla mucho; E2, que si encuentra dificultad en hablar con otros; el resto de las variables en el grupo de extroversión indican cosas como: sabe cautivar a la gente o hace amigos de forma sencilla.

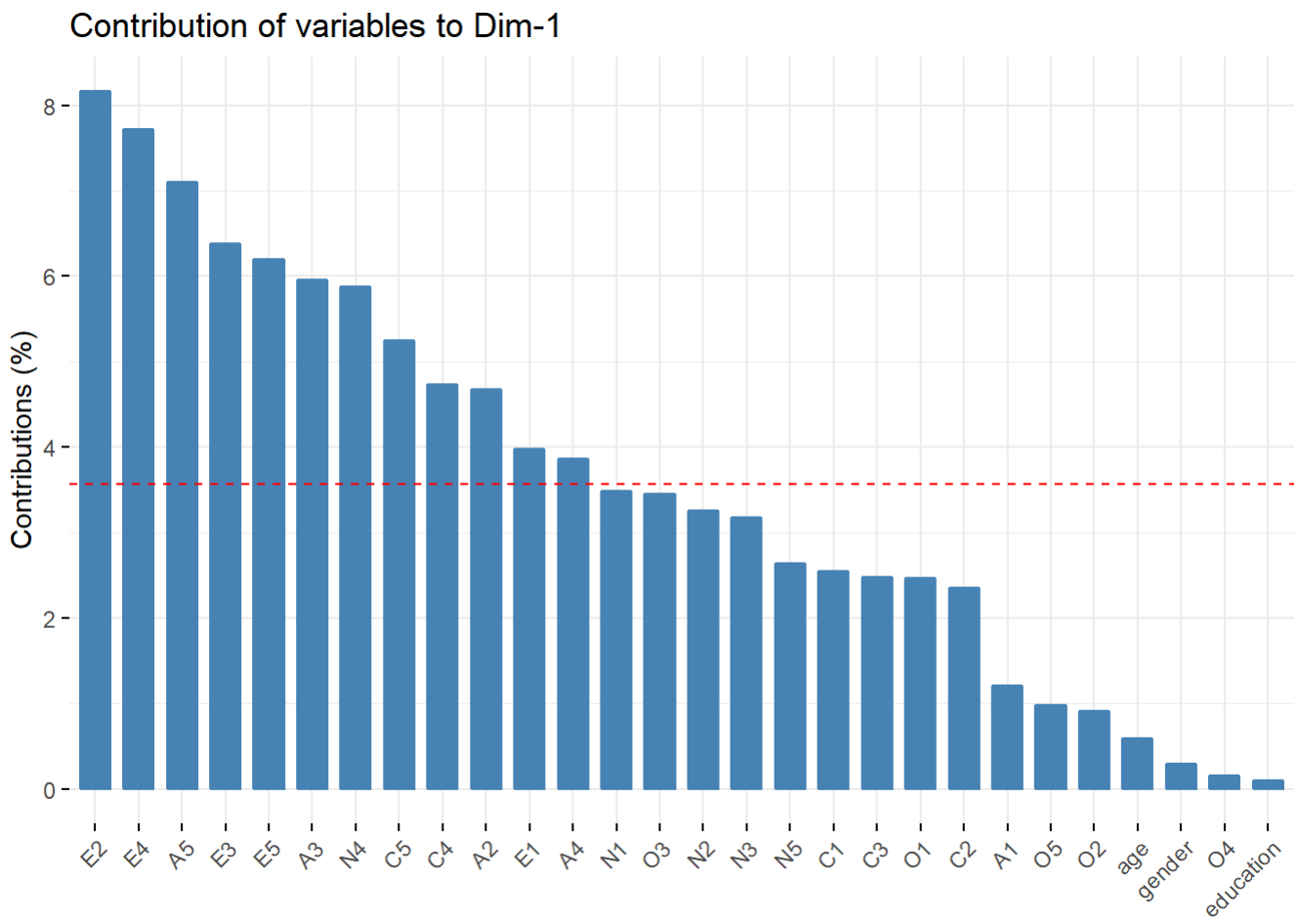
```
# Mapa factorial
fviz_pca_var(ven.pca, col.var="contrib")+
  scale_color_gradient2(low="blue", mid="yellow",high="red", midpoint=14)+
  theme_minimal()
```



Ahora, se realiza el análisis por factor; i.e, las contribuciones de las variables a las componentes principales. Aquí se espera que cada variable este sobre la linea roja, indicando el aporte en las misma cuantia para cada componente.

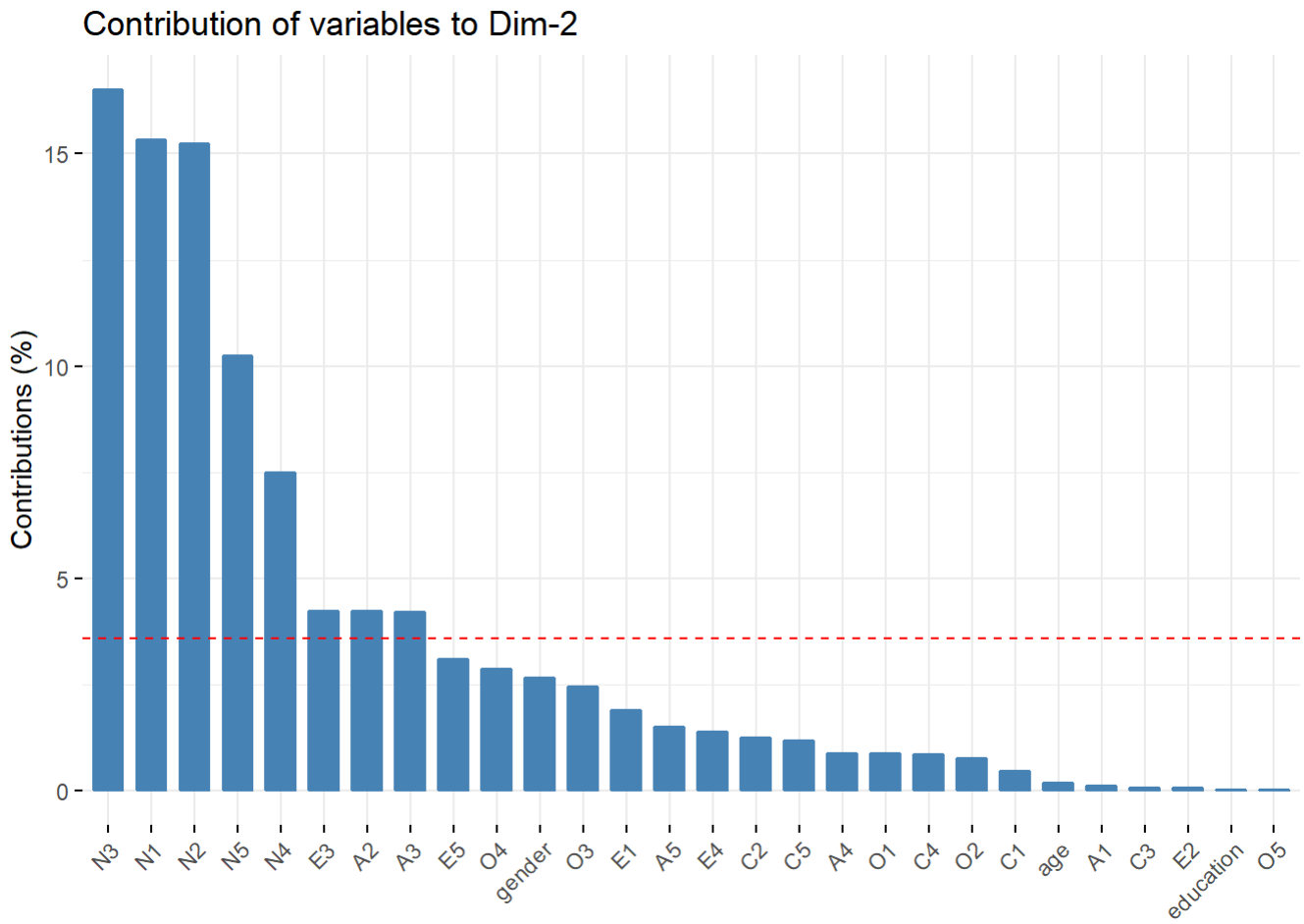
En el gráfico de abajo podemos observar lo ya mencionado sobre que variables contribuyen más al primer factor; aquellas preguntas que hacen referencia a la amabilidad, extroversión y concienia.

```
# Primera componente
fviz_contrib(ven.pca, choice = "var", axes = 1)
```



Ahora con la segunda componente; se detecta que la mayor contribución a este factor son las variables que hacen mención al neurotismo.

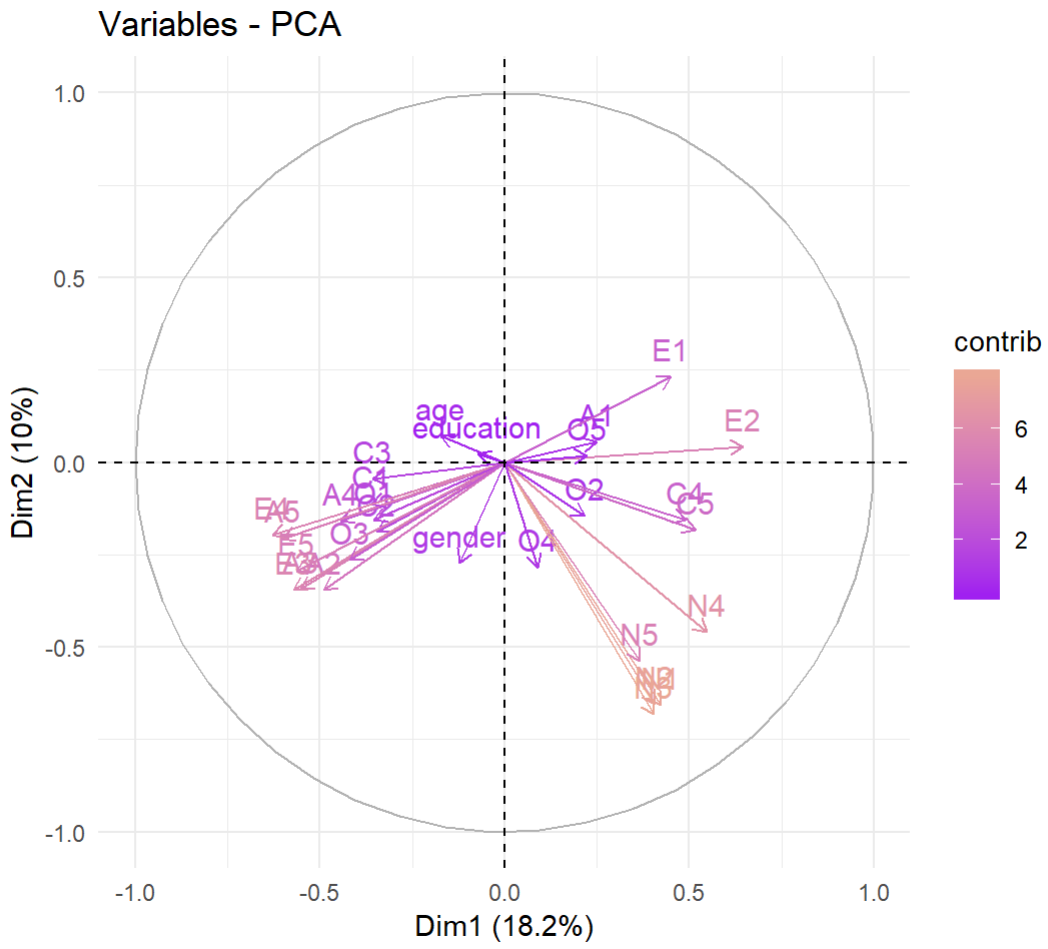
```
# Segunda componente
fviz_contrib(ven.pca, choice = "var", axes = 2)
```



c) Identifica las preguntas que tienen una preponderancia de acuerdo extremo y / o en desacuerdo las respuestas. Del mismo modo, identifica casos atípicos tales como personas que parecen responder de manera extrema. Es decir, las personas que tienden a estar totalmente de acuerdo o en desacuerdo con la mayoría de las preguntas.

En el inciso c) se indica que se encuentren las respuestas que tienen una preponderancia de acuerdo extremo o desacuerdo, y bajo la representación del gráfico de factores se epueden encontrar algunas, como las que se presentan a continuación:

```
# Mapa factorial
fviz_pca_var(ven.pca, col.var="contrib")+
  scale_color_gradient2(low="purple", mid="yellow",high="red", midpoint=14)+
  theme_minimal()
```



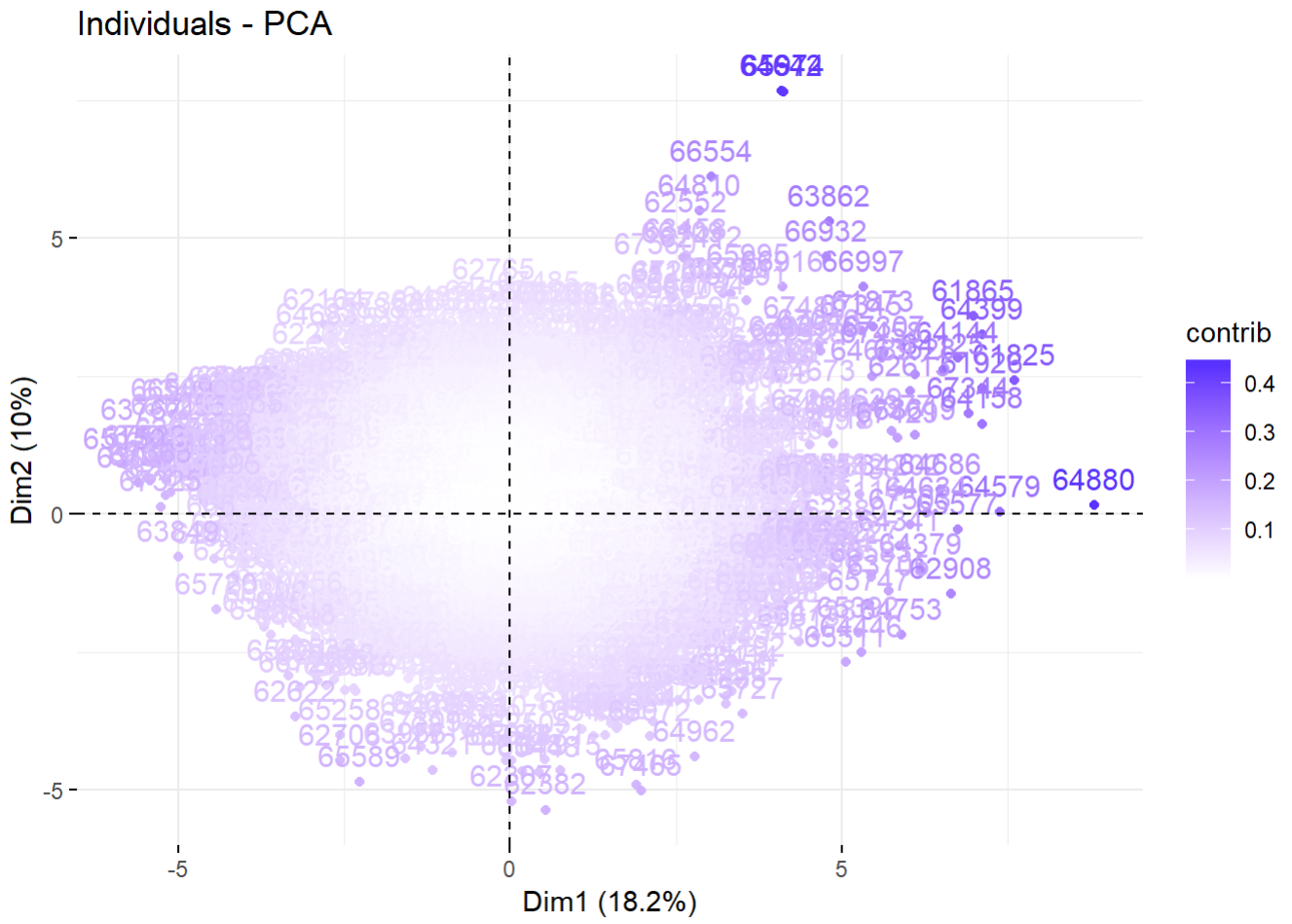
Entre estas se tienen:

E1: No habla mucho E2: Encunentra difícil hablar con los demas

Todas las de la categoria de neurotismo (N1,N2,N3,N4)

Se identifican casos atípicos tales como personas que parecen responder de manera extrema; es decir, las personas que tienden a estar totalmente de acuerdo o en desacuerdo con la mayoría de las preguntas. Se gráfica el mapa de factores por inidviduo para localizar las respuestas extremas; bajo esta representación se encuentran a individuos con respuestas extremas, cabe a clarar que más adelante se tiene en que respuestas los individuos contestan de forma extrema.

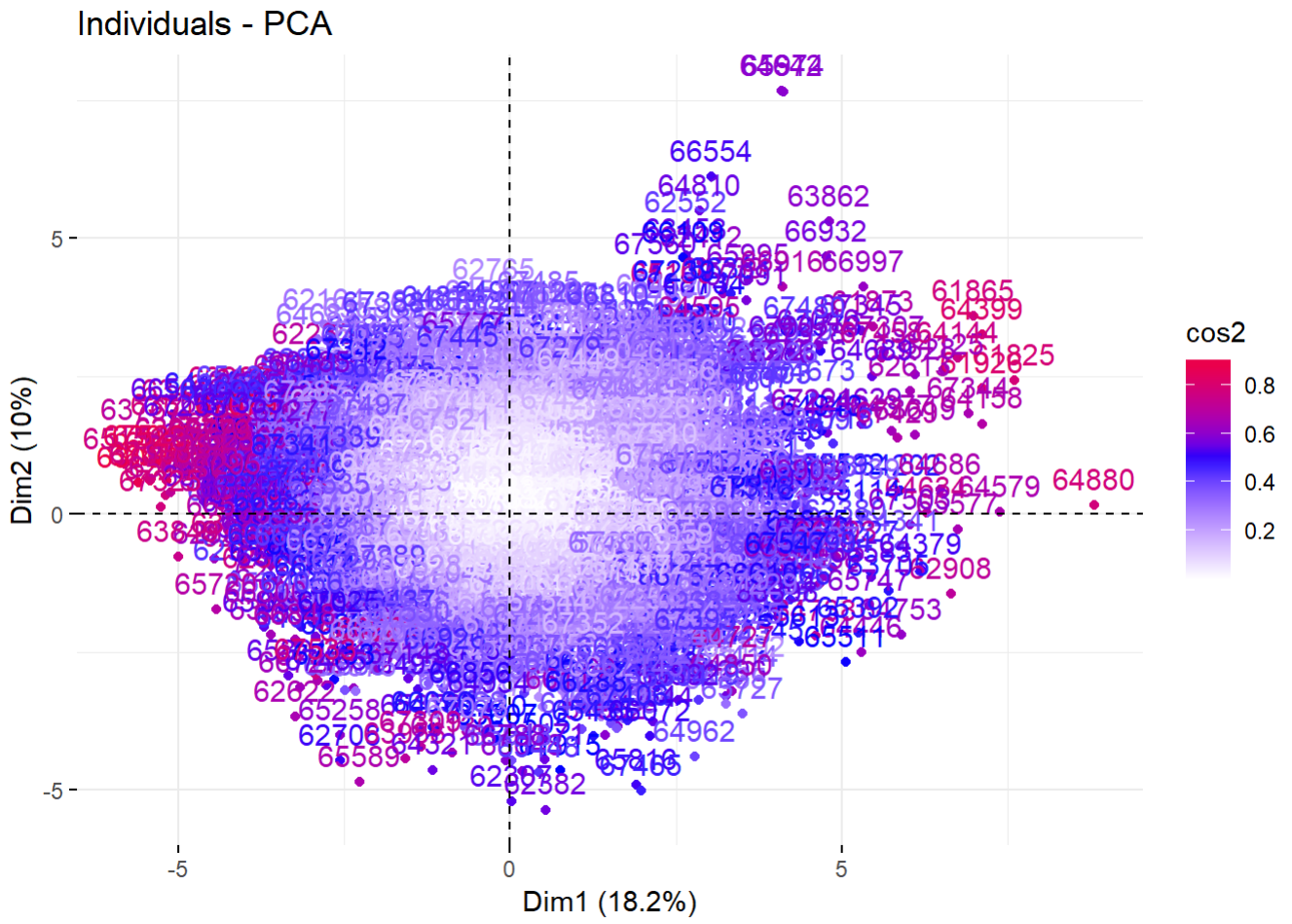
```
# Calidad de los individuos en el mapa factorial
fviz_pca_ind(ven.pca, col.ind="contrib") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.50)
```



Si se colorea por su contribución no se aprecia muy bien; entonces, se colorea por la calidad de la representación.

Entre los individuos con respuestas extremas se encuentran los dos en la esquina superior derecha, cuya etiqueta se traslapan; aismismo, el individuo 66554; por otro lado, en ese cuadrante pero cerca al eje de las abscisas, el individuo 64880.

```
# Calidad de los individuos en el mapa factorial
fviz_pca_ind(ven.pca, col.ind="cos2") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.50)
```

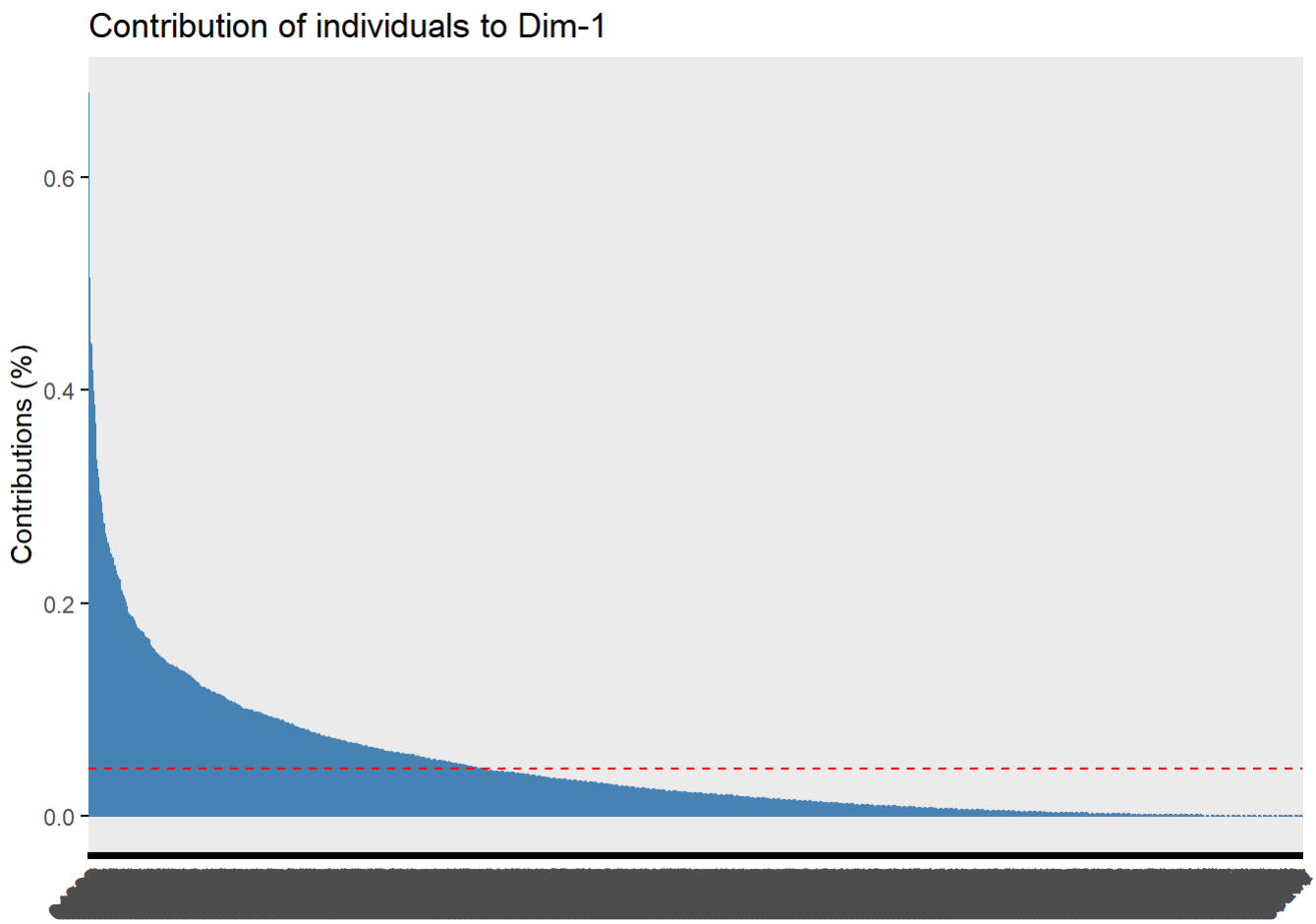



```
# NOTA: cos2 = la calidad de los individuos en el mapa de calor
```

Si se desea observar lo anterior para cada factor, dado la cantidad de individuos que se tienen, será difícil localizarlos; para sostener lo dicho, a continuación, se presentan las contribuciones de los individuos en cada factor.

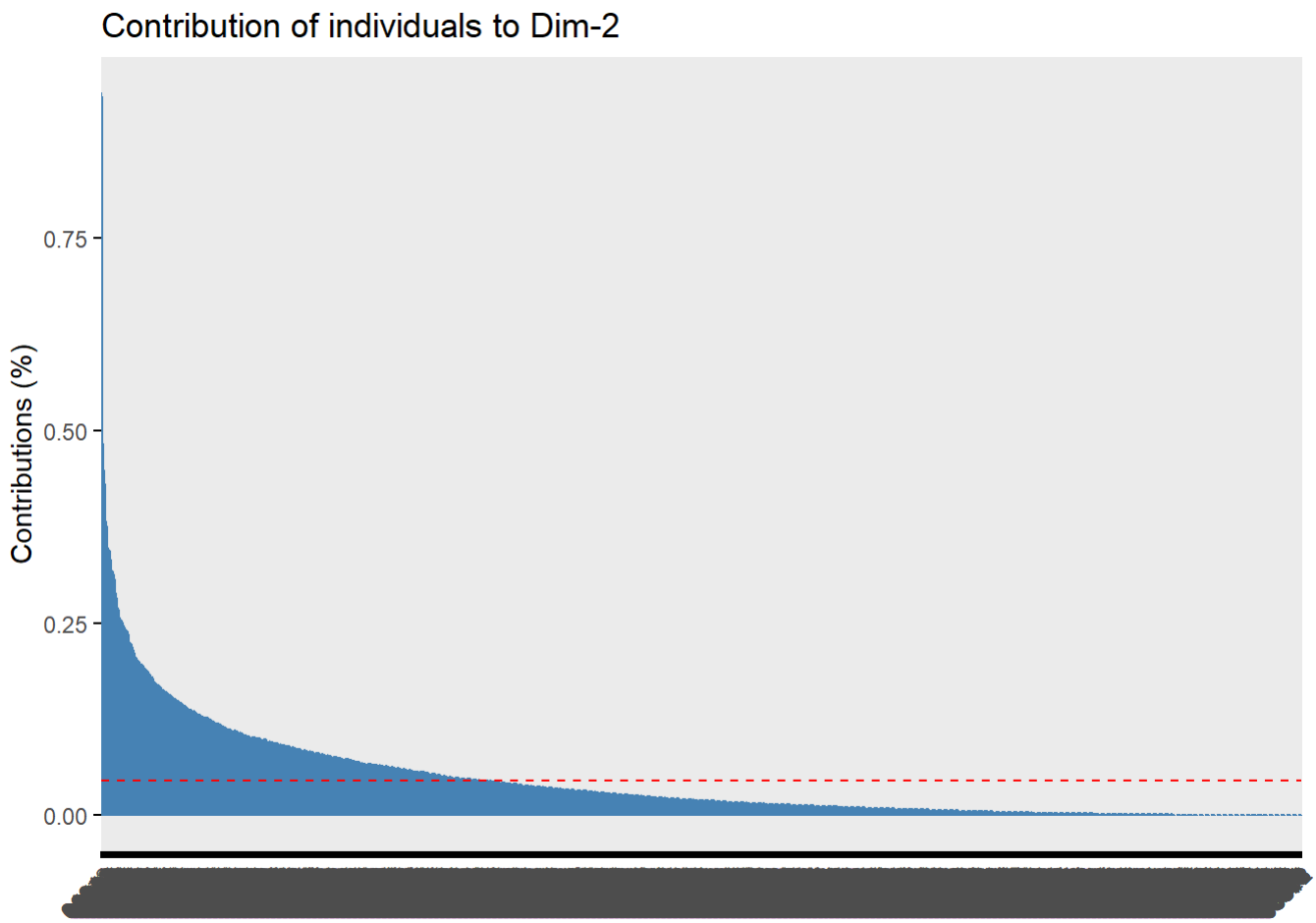
Primera componente:

```
# Primera componente
fviz_contrib(ven.pca, choice = "ind", axes = 1)
```



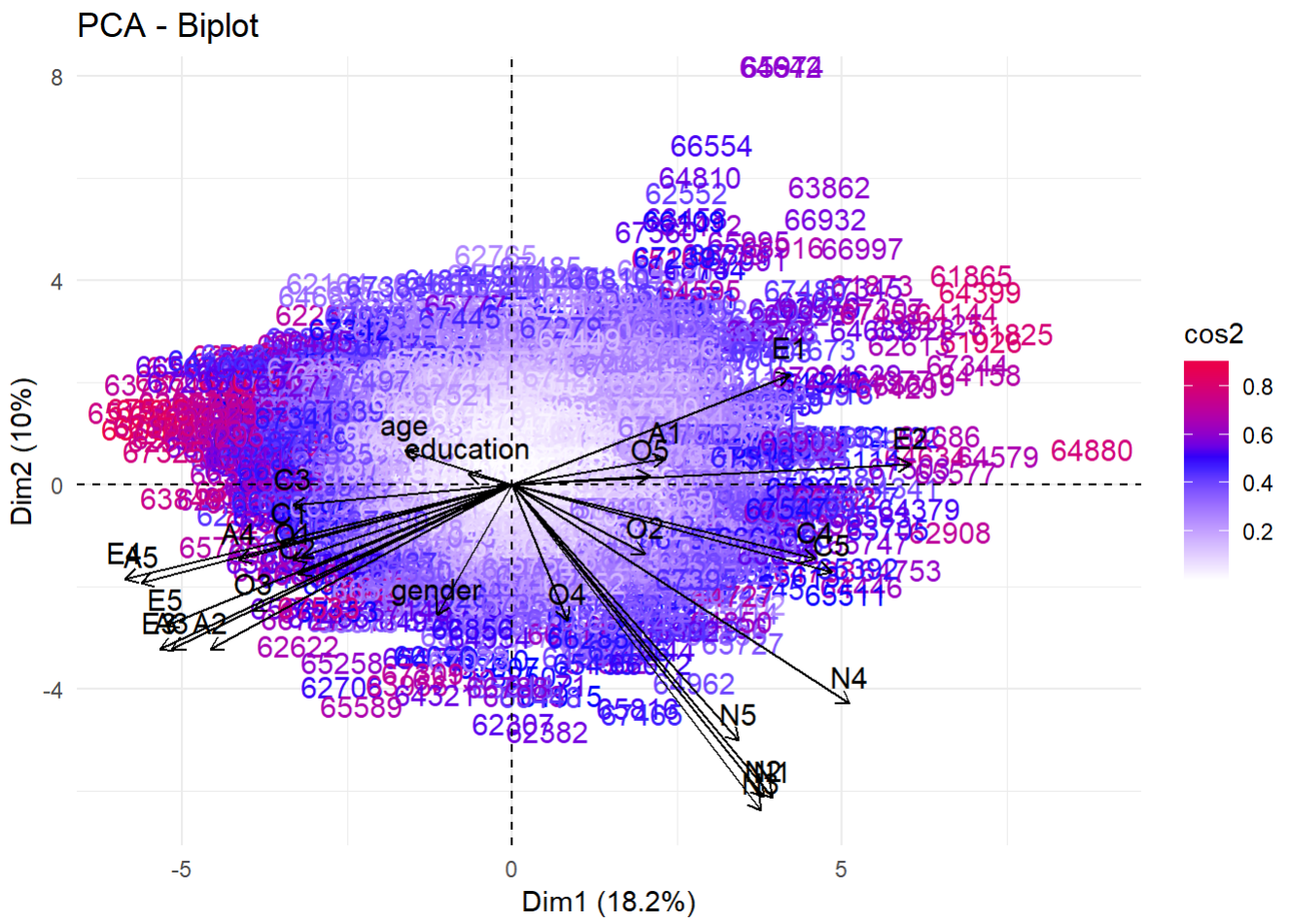
Segunda componente:

```
# Primera componente
fviz_contrib(ven.pca, choice = "ind", axes = 2)
```



Se localizan los individuos con respuestas extremas y cuales son esas respuestas.

```
# Biplot de individuos y variables
fviz_pca_biplot(ven.pca, geom = "text", col.ind="cos2", col.var = "black") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.50) +
  theme_minimal()
```



De forma de ejemplo, se realiza un filtrado en el individuo 66554, el cual tiene como pregunta extrema la E1, indicando que no habla mucho; la A1, que es indiferente a los demas; y el E2, que le cuesta hablar con otras personas. Este individuo cuenta con un nivel de educación universitaria (cursandola) y con edad de 23 años, lo cual carácteriza bien su respuestas. Esta persona cae dentro del contraste del primer factor, que representa el contrario de los individuos sociales.

```
library("tidyverse")
obs <- rownames_to_column(datos, var="label")
obs %>% filter(label==66554)
```

```
##   label A1 A2 A3 A4 A5 C1 C2 C3 C4 C5 E1 E2 E3 E4 E5 N1 N2 N3 N4 N5 O1 O2
## 1 66554  6  1  1  6  1  6  5  6  1  1  6  6  1  1  1  1  1  1  1  6  1
##   O3 O4 O5 gender education age
## 1  1  6  1      1         3  23
```

Otro individuo con respuestas extrema es el 64880, que también cae dentro de los contrastes del primer factor, al responder lo contrario de ser un individuo social y amable.

Por último, lo aterior se realizó estimando los factores con PCA; a continuación solo se estiman los factores con maxima verosimilitud y se contrastan los residuales respecto a los de la estimación por PCA.

Nota: no se realiza de nuevo el análisis, solo se contrastan residuales para ver que método aproxima mejor la matriz de correlación de los datos.

Por ML, dos factores:

```
var_vend.fa2<- factanal(covmat=R,factors=2)
CARGAS2<-var_vend.fa2$loadings # Cargas factoriales
LL <- CARGAS2%*%t(CARGAS2)
VAR_ESP2<-var_vend.fa2$uniquenesses # Singularidades
residuales <- R - LL - diag(VAR_ESP2)
residuales %>%  norm
```

```
## [1] 2.322079
```

la norma de los residuales es 2.322079, utilizando MV

```
residuales2 <- R - RESULTADO$LL - RESULTADO$Phi
residuales %>%  norm
```

```
## [1] 2.322079
```

la norma de los residuales con PCA es de 20.56397; por lo tanto, con MV los residuales son más pequeños, y por lo tanto es mejor la aproximación con MV de la matriz de correlación.

Este ejercicio se reliza con PCA para ejemplificar que MV aproxima mejor la matriz de covarianza o correlación, en el siguiente ejercicio se realiza con MV todo el análisis.

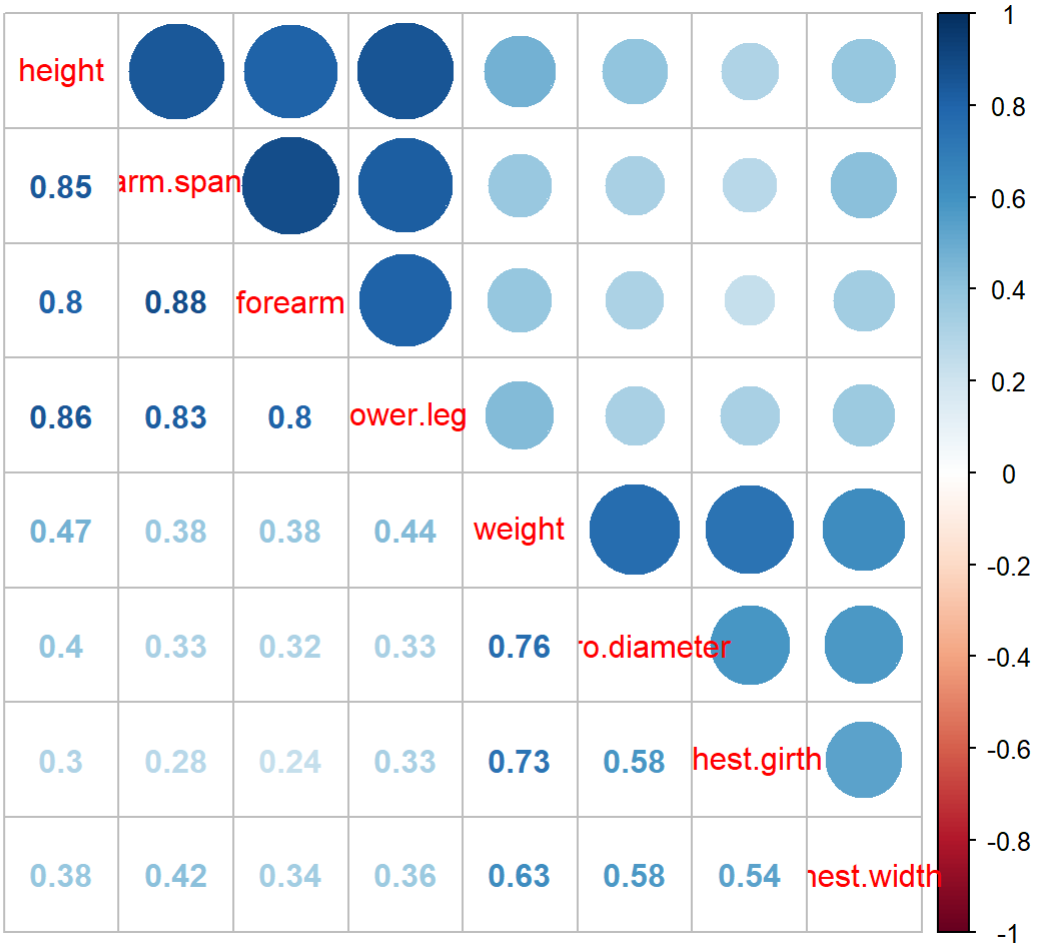
EJERCICIO 6

6. El conjunto de datos Harmon23.cor en el paquete “datasets” es una matriz de correlación de ocho mediciones físicas realizadas en 305 niñas entre las edades de 7 y 17 años.

Antes de realizar el análisis de factores se determina si existe correlación entre grupps de variables.

Se observa el gráfico de correlación y se ve que a lo menos existen variables con alto nivel de correlación; ejemplo, altura con la variabale arm.spam yforerman, entre otros.

```
library("corrplot")
R <- datasets::Harman23.cor
R <- R$cov
corrplot.mixed(R)
```



Se realiza la prueba de Bartlett para probar la hipótesis nula de que las variables no están correlacionadas.

Se encuetra evidencia suficiente para rechazar la hipótesis nula; por lo tanto, si existe correlación entre las variables.

```
n <- 305
cortest.bartlett(R,n)
```

```
## $chisq
## [1] 2085.74
##
## $p.value
## [1] 0
##
## $df
## [1] 28
```

Ahora se utiliza la prueba KMO para ver que tan bueno es realizar análisis de factores en los datos.

El valor MSA es de $.85$, por lo tanto, bajo el críterio de clasificación es bueno utilizar análisis de factores sobre este conjunto de datos.

```
# Indice KMO
# - KMO > 0.90    Muy bueno
# - 0.80<KMO<0.90 Bueno
# - 0.70<KMO<0.80 Aceptable
# - 0.60<KMO<0.70 Regular
# - 0.50<KMO<0.60 Malo
# - KMO < 0.50    Inaceptable
KMO(R)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R)
## Overall MSA =  0.85
## MSA for each item =
##         height       arm.span        forearm      lower.leg         weight
##           0.86           0.82           0.86           0.89           0.78
## bitro.diameter    chest.girth    chest.width
##           0.85           0.82           0.90
```

a) Realiza un análisis factorial de estos datos.

b) Varía el número de factores para encontrar un ajuste adecuado del modelo e interprete las cargas factoriales resultantes.

Para evaluar que número de factores utilizar, se realiza la prueba ji cuadradra. Antes de aplicar la prueba, se observa si es posible realizarla:

$$m < \frac{1}{2}(2 * p + 1 - \sqrt{(}8 * p + 1)$$

con 2 factores; si se puede.

```
m <- 2
p <- dim(R)[2]
m < .5*(2*p + 1 - sqrt(8*p + 1))
```

```
## [1] TRUE
```

con 3 factores, también.

```
m <- 3
p <- dim(R)[2]
m < .5*(2*p + 1 - sqrt(8*p + 1))
```

```
## [1] TRUE
```

con 5 factores, no se puede.

```
m <- 5
p <- dim(R)[2]
m < .5*(2*p + 1 - sqrt(8*p + 1))
```

```
## [1] FALSE
```

Se realiza la prueba con factores de 2 al 4, ya que con uno factor dan problemas conocidos como casos Heywood, y con 5 en adelate ya no es valido hacer la prueba ji cuadrada.

```
testFactores <- function(m, LL, phi, Sn, alpha, n, p)
{
  EstPrueba <- (n-1-((2*p+4*m+5)/6))*log(det(LL+phi)/det(Sn))
  gradosLibertad <- (((p-m)^2)-p-m)/2
  ValCrit <- qchisq(alpha,gradosLibertad, lower.tail = F)
  Resultado <- EstPrueba > ValCrit
  return(list(estadistico = EstPrueba, gl = gradosLibertad, critico = ValCrit, Rehaza = Resultad
o))
}
```

Se estiman factores con MV y se realiza la prueba para detérminar el número de factores, con $m = 2$

```
library("psych")
factors <- factanal(covmat = R, factors=2)
LL <- factors$loadings%*%t(factors$loadings)
phi <- factors$uniquenesses %>% diag
testFactores(2, LL, phi, R, .05, 305, dim(R)[2])
```

```
## $estadistico
## [1] 75.73361
##
## $gl
## [1] 13
##
## $critico
## [1] 22.36203
##
## $Rehaza
## [1] TRUE
```

Con dos factores si se tiene evidencia para rechazar la hipótesis nula de que el modelo con $m = 2$ factore se ajusta bien a los datos.

Con 3 factores, también se rechaza la hipótesis nula.

```
factors <- factanal(covmat = R, factors=3)
LL <- factors$loadings%*%t(factors$loadings)
phi <- factors$uniquenesses %>% diag
testFactores(3, LL, phi, R, .05, 305, dim(R)[2])
```

```
## $estadistico
## [1] 23.02947
##
## $gl
## [1] 7
##
## $critico
## [1] 14.06714
##
## $Rehaza
## [1] TRUE
```

Es con 4 factores cuando no se rechaza la hipótesis nula, y con 4 factores se ajustan bien a los datos.

```
factors <- factanal(covmat = R, factors=4)
LL <- factors$loadings%*%t(factors$loadings)
phi <- factors$uniquenesses %>% diag
testFactores(4, LL, phi, R, .05, 305, dim(R)[2])
```

```
## $estadistico
## [1] 4.961171
##
## $gl
## [1] 2
##
## $critico
## [1] 5.991465
##
## $Rehaza
## [1] FALSE
```

Se utilizan 4 factores para realizar el análisis.

```
factores <- factanal(covmat = R,factors=4)
cargas <- factores$loadings # Cargas factoriales
varEspecifica <- factores$uniquenesses # Singularidades
LL <- cargas%*%t(cargas)
cumunalidades <- LL %>% diag
```

Se calcula el residual de la aproximación a la matriz de correlación.

```
Rest <- LL + diag(varEspecifica)
round(R-Rest,digits=3)
```

```
##                height arm.span forearm lower.leg weight bitro.diameter
## height          0.000        0  -0.007     0.001  0.001          0.004
## arm.span        0.000        0   0.000     0.000  0.000          0.000
## forearm        -0.007        0   0.000     0.002  0.007          0.000
## lower.leg       0.001        0   0.002     0.000 -0.002         -0.001
## weight          0.001        0   0.007    -0.002  0.000         -0.002
## bitro.diameter  0.004        0   0.000    -0.001 -0.002          0.000
## chest.girth     0.000        0  -0.002     0.000  0.001          0.001
## chest.width    -0.011        0  -0.020     0.012  0.002          0.006
##                chest.girth chest.width
## height               0.000      -0.011
## arm.span             0.000       0.000
## forearm             -0.002      -0.020
## lower.leg            0.000       0.012
## weight               0.001       0.002
## bitro.diameter       0.001       0.006
## chest.girth          0.000      -0.006
## chest.width         -0.006       0.000
```

Se observa que la matriz de residuale tiene elementos casis cero y cero.

Se observan las cargas de los factores (coeficientes)

```
factors$loadings
```

```
##
## Loadings:
##                Factor1 Factor2 Factor3 Factor4
## height          0.879   0.277          -0.115
## arm.span        0.937   0.194           0.277
## forearm         0.875   0.191
## lower.leg       0.887   0.209   0.135  -0.188
## weight          0.246   0.882   0.111  -0.109
## bitro.diameter  0.187   0.822
## chest.girth     0.117   0.729   0.526
## chest.width     0.263   0.644           0.141
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      3.382   2.595   0.323   0.165
## Proportion Var   0.423   0.324   0.040   0.021
## Cumulative Var   0.423   0.747   0.787   0.808
```

Se observa que las variables como altura, el alcance del brazo (ar.spam), el antebrazo y la parte inferior de la pierna, son las variables que más contribuyen al primer factor; el segundo factor, las variables que más le contribuyen son: el peso y bitro.diameter; el tercer factor le contribuye más la variable que mide la circunferencia del pecho (chest.girth), y en el cuarto se puede decir que se centra en el ancho del pecho, que a la vez se relaciona con el alcance de los brazos.

Nombrando los factores se tiene:

Primer factor es el de extremidades.

Segundo factor el de masa corporar,

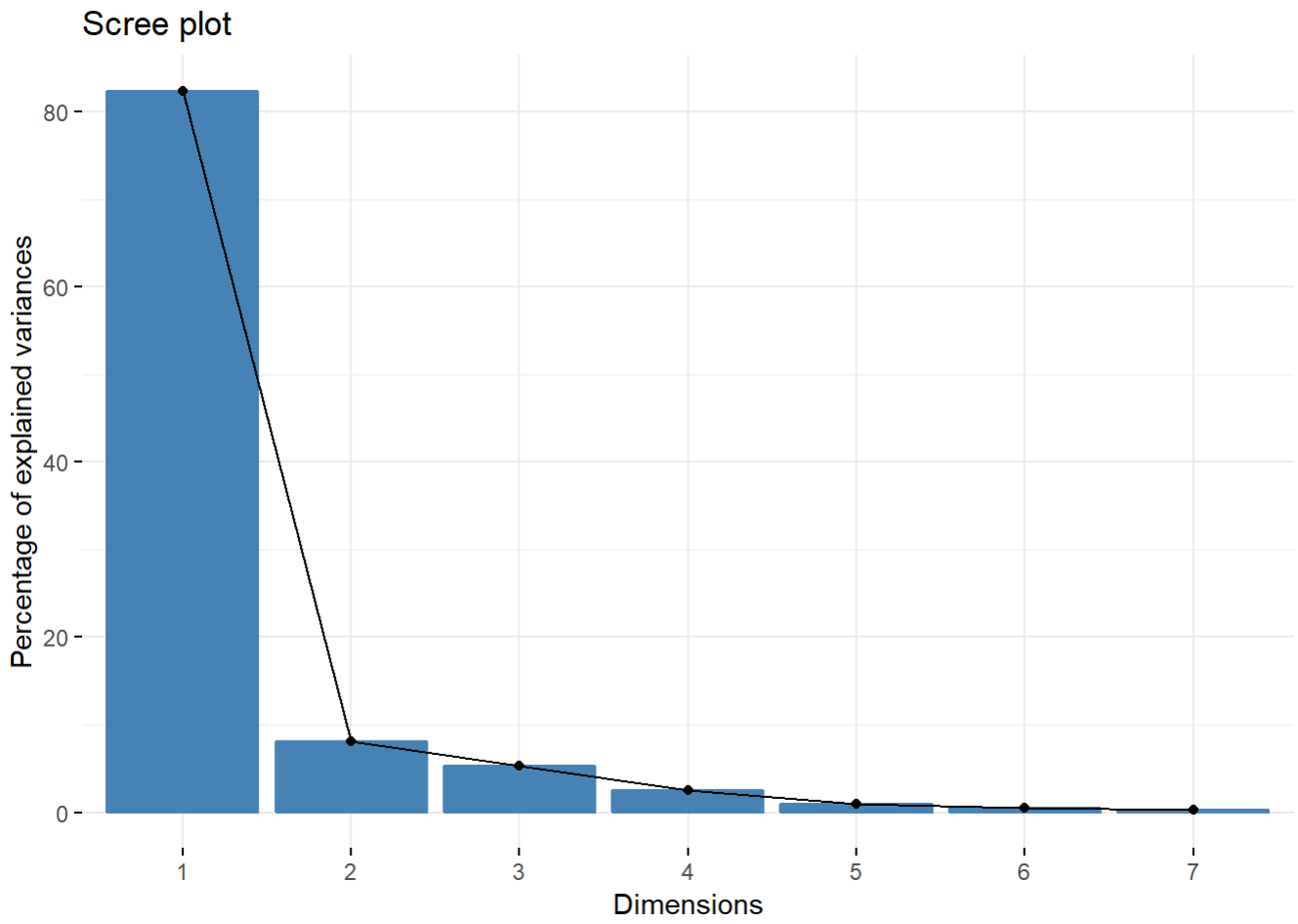
Tercer factor circunferencia del pecho

Cuarto factor el de ancho de pecho

No se pueden utilizar los factor scores dado que no tenemos los datos originales, entonces se tiene que utilizar la estimción con PCA.

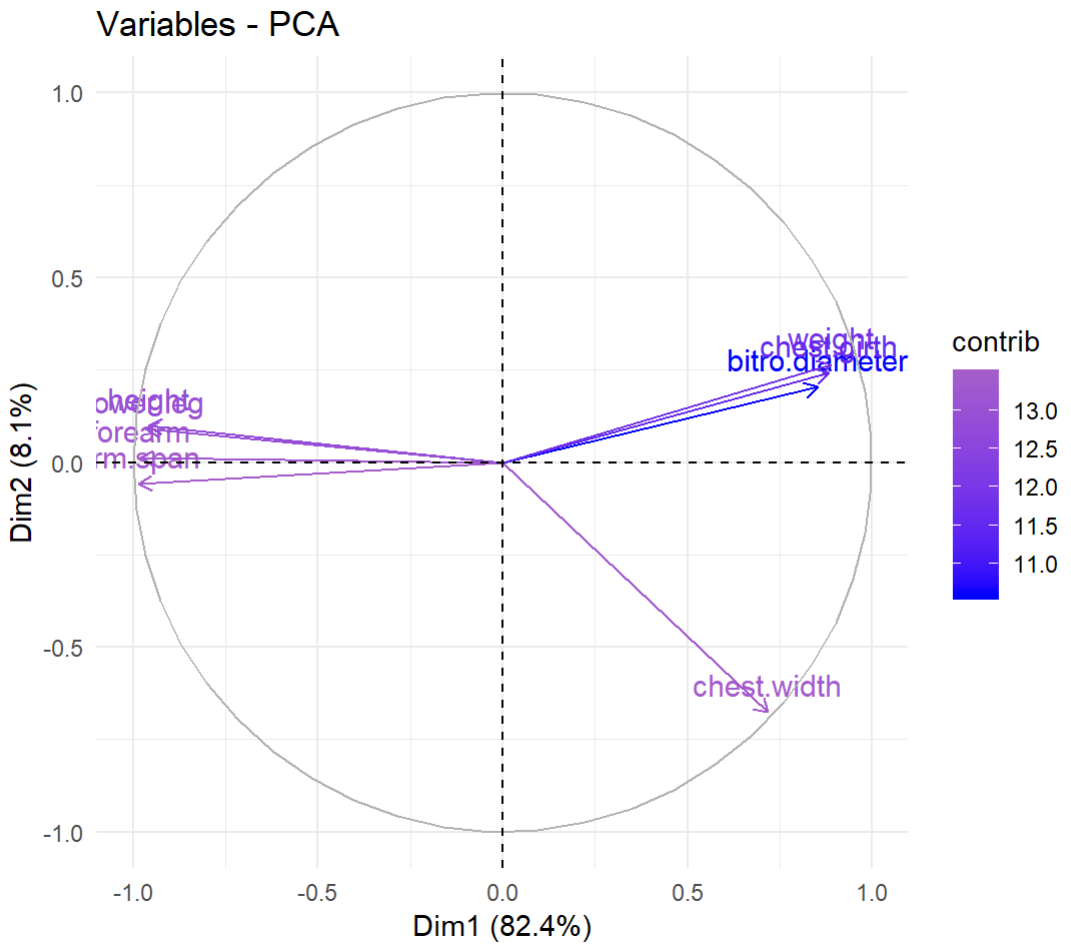
Para estimar los factores con PCA se utilizan criterios de PCA para seleccionar los factores.

```
library("factoextra")
library("ade4")
ven.pca <- dudi.pca(R, scannf = FALSE, nf = 2,scale = T)
eig.val <- get_eigenvalue(ven.pca)
fviz_screeplot(ven.pca)
```



Con el screeplot se observa que a lo mucho dos factores son los adecuados.

```
# Mapa factorial
fviz_pca_var(ven.pca, col.var="contrib")+
  scale_color_gradient2(low="blue", mid="yellow",high="red", midpoint=20)+
  theme_minimal()
```

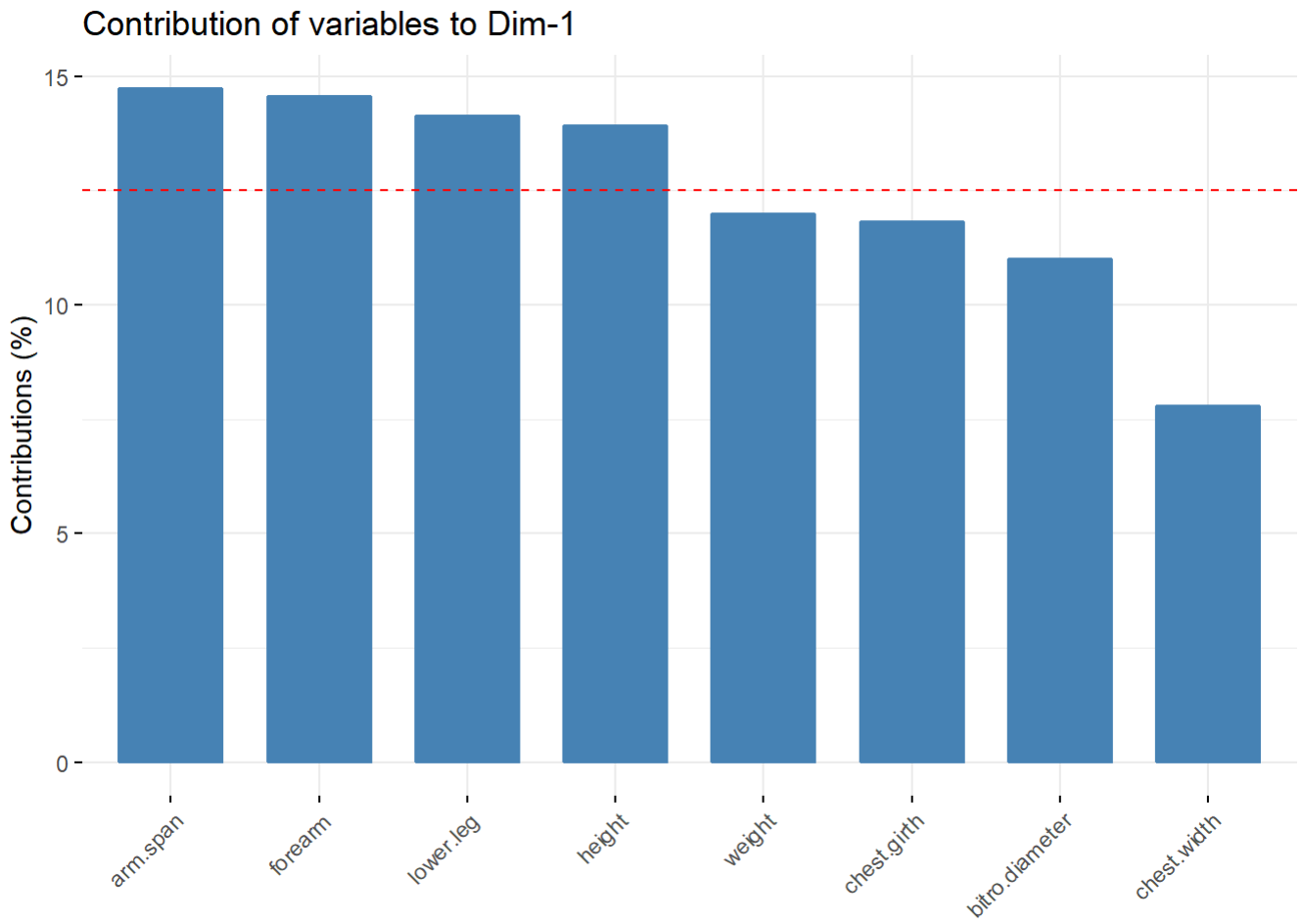


Con la representación del gráfico anterior se observar que el primer factor contiene el 82% de la varianza total explicada; con los dos factores se tiene que las variables como la altura, alcance del brazo, antebrazo, y el largo de la parte baja del pie se agrupan y son las variables que más aportan al primer factor; también en este factor se observa un contraste, que tiene relación con el medidas del pecho y altura, los cuales contribuyen en gran medida al primer factor; y el segundo factor la variable que más lo explica es la anchura del pecho (chest width).

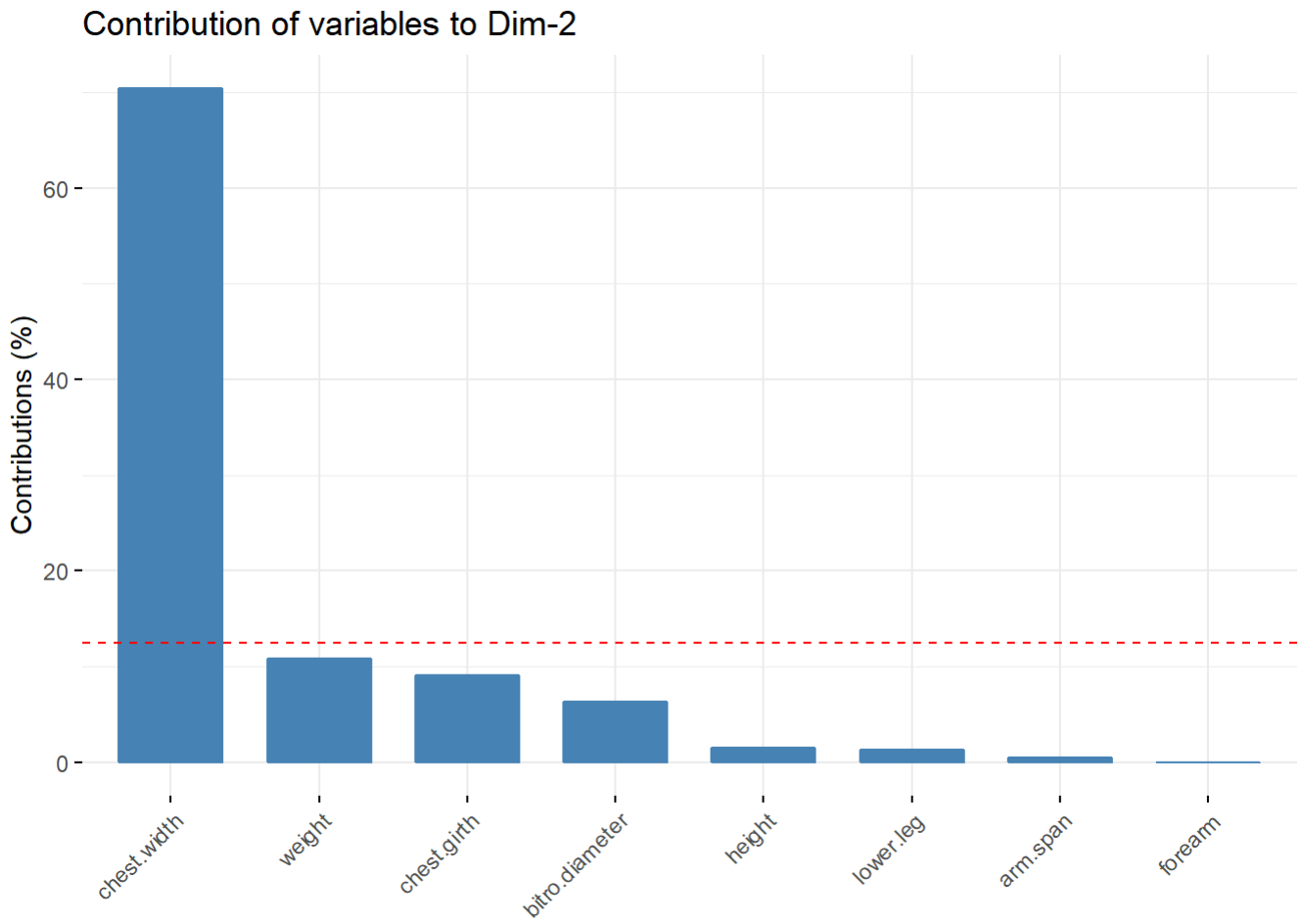
Ahora, se realiza el análisis por factor; i.e, las contribuciones de las variables a las componentes principales. Aquí se espera que cada variable este sobre la linea roja, indicando que aportan en las misma cuantia en cada componente.

Se ve que la distribución en la contribución tiende a ser homogenea para el primer factor; sin embargo, esto cambia con el segundo factor.

```
# Primera componente
fviz_contrib(ven.pca, choice = "var", axes = 1)
```




```
# Segunda componente
fviz_contrib(ven.pca, choice = "var", axes = 2)
```



En conclusión, nos quedamos con la interpretación de la estimación de MV, ya que con más factores se logró clasificar a los siguientes grupos.

Primer factor es el de extremidades.

Segundo factor el de masa corporar,

Tercer factor circunferencia del pecho

Cuarto factor el de ancho de pecho

Con dos factores y utilizando PCA, para estimar a los factores, se encuentra lo siguiente:

Primer factor es el individuos altos.

Segundo factor el de individuos bajos y robustos.

EJERCICIO 7

7) La matriz de correlación dada a continuación proviene de las puntuaciones de 220 chicos en seis asignaturas escolares: 1) Francés, 2) Inglés, 3) Historia, 4) Aritmética, 5) Álgebra y 6) Geometría

```
R <- matrix(c(1,.44,.41,.29,.33,.25,
              .44,1,.35,.35,.32,.33,
              .41,.35,1,.16,.19,.18,
              .29,.35,.16,1,.39,.47,
              .33,.32,.19,.59,1,.46,
              .25,.33,.18,.47,.46,1),6,6)
```

a) Encuentre la solución de dos factores de un análisis de factor de máxima verosimilitud.

En esta estimación no se rotan los factores.

```
factores <- factanal(covmat = R,factors=2,rotation = "none")
cargas <- factores$loadings # Cargas factoriales
varEspecifica <- factores$uniquenesses # Singularidades
LL <- cargas%*%t(cargas)
cumunalidades <- LL %>% diag
```

Se calcula el residual de la aproximación a la matriz de correlación.

```
Rest <- LL + diag(varEspecifica)
round(R-Rest,digits=3)
```

```
##        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
## [1,]  0.000 -0.007 -0.001  0.004  0.022 -0.012
## [2,] -0.007  0.000  0.013  0.023 -0.018 -0.003
## [3,] -0.001  0.013  0.000 -0.027 -0.018  0.022
## [4,]  0.004  0.023 -0.027  0.000  0.192 -0.002
## [5,]  0.022 -0.018 -0.018 -0.008  0.000  0.007
## [6,] -0.012 -0.003  0.022 -0.002  0.007  0.000
```

Se observa que los elementos de la matriz de residuales se aproximana a cero; por lo tanto, se realiza una representación adecuada de la matriz de correlación.

Se obtienen las cargas de los dos factores estimados por MV, las comunalidades y su varianza especifica.

Varianza Especifica:

```
varEspecifica %>% as.matrix
```

```
##           [,1]
## [1,] 0.4735747
## [2,] 0.5919930
## [3,] 0.6743551
## [4,] 0.5910608
## [5,] 0.6102397
## [6,] 0.4404532
```

Comunalidades:

```
cumunalidades %>% as.matrix
```

```
##           [,1]
## [1,] 0.5264255
## [2,] 0.4080069
## [3,] 0.3256450
## [4,] 0.4089387
## [5,] 0.3897596
## [6,] 0.5595469
```

```
rownames(cargas) <- c("Frances", "Ingles", "Historia", "Aritmética", "Algebra", "Geometría")
cargas
```

```
##
## Loadings:
##            Factor1 Factor2
## Frances     0.616   0.384
## Ingles      0.612   0.184
## Historia    0.443   0.360
## Aritmética  0.601  -0.219
## Algebra     0.602  -0.164
## Geometría   0.653  -0.364
##
##                Factor1 Factor2
## SS loadings       2.10   0.518
## Proportion Var    0.35   0.086
## Cumulative Var    0.35   0.436
```

Se observa que sin rotar los factores no se puede dar una interpretación con facilidad.

b) Mediante una inspección de las cargas, encuentre una rotación ortogonal que permite una interpretación más fácil de los resultados.

Se debe mencionar que para rotar los factores se utiliza el método varimax, el cual se encuentra por default en la función “factanal”.

```
factores <- factanal(covmat = R,factors=2,rotation = "varimax")
cargas <- factores$loadings # Cargas factoriales
varEspecifica <- factores$uniquenesses # Singularidades
LL <- cargas%*%t(cargas)
cumunalidades <- LL %>% diag
```

Se calcula el residual de la aproximación a la matriz de correlación.

```
Rest <- LL + diag(varEspecifica)
round(R-Rest,digits=3)
```

```
##        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
## [1,]  0.000 -0.007 -0.001  0.004  0.022 -0.012
## [2,] -0.007  0.000  0.013  0.023 -0.018 -0.003
## [3,] -0.001  0.013  0.000 -0.027 -0.018  0.022
## [4,]  0.004  0.023 -0.027  0.000  0.192 -0.002
## [5,]  0.022 -0.018 -0.018 -0.008  0.000  0.007
## [6,] -0.012 -0.003  0.022 -0.002  0.007  0.000
```

Se observa que los elementos de la matriz de residuales se aproximana cero; por lo tanto, se realiza una representación adecuada de la matriz de correlación.

Se estiman las cargas de los dos factores con MV, las comunalidades y su varianza especifica. Los cuales por propiedad no cambian ante una rotación de los factores.

Varianza Especifica:

```
varEspecifica %>% as.matrix
```

```
##           [,1]
## [1,] 0.4735747
## [2,] 0.5919930
## [3,] 0.6743551
## [4,] 0.5910608
## [5,] 0.6102397
## [6,] 0.4404532
```

Comunalidades:

```
cumunalidades %>% as.matrix
```

```
##           [,1]
## [1,] 0.5264255
## [2,] 0.4080069
## [3,] 0.3256450
## [4,] 0.4089387
## [5,] 0.3897596
## [6,] 0.5595469
```

Se observan a los factores rotados:

```
rownames(cargas) <- c("Frances", "Ingles", "Historia", "Aritmética", "Algebra", "Geometría")
cargas
```

```
##
## Loadings:
##            Factor1 Factor2
## Frances    0.224   0.690
## Ingles     0.350   0.534
## Historia   0.107   0.560
## Aritmética 0.601   0.219
## Algebra    0.567   0.262
## Geometría  0.735   0.141
##
##                Factor1 Factor2
## SS loadings      1.406   1.212
## Proportion Var   0.234   0.202
## Cumulative Var   0.234   0.436
```

```
rownames(cargas) <- c("Frances", "Ingles", "Historia", "Aritmética", "Algebra", "Geometría")
cargas
```

```
##
## Loadings:
##            Factor1 Factor2
## Frances    0.224   0.690
## Ingles     0.350   0.534
## Historia   0.107   0.560
## Aritmética 0.601   0.219
## Algebra    0.567   0.262
## Geometría  0.735   0.141
##
##                Factor1 Factor2
## SS loadings      1.406   1.212
## Proportion Var   0.234   0.202
## Cumulative Var   0.234   0.436
```

Bajo la rotación con el método de varimax los factores son más sencillos de interpretar.

Se observa que entre los dos primeros factores se explica un total de la varianza acumulada del 43.6%; las variables que más aportan al primer factor, en base al tamaño de las cargas de cada variable para cada factor son: Aritmética, Algebra y Geometría; en el segundo factor: Frances Ingles e Historia.

Entonces, el primer factor se relaciona con las habilidades matemáticas; el segundo factor, con las habilidades sociales y humanidades, donde se encuentran aspectos sociales y lingüísticos.