Universidad de Granada
Escuela Internacional de Posgrado
Máster en Estadística Aplicada
Materia: Técnicas Estadísticas Multivariantes.
Alumno: Francisco Javier Márquez Rosales

Análisis Factorial:

Ejercicios de Análisis Factorial:

Octubre, 2022

Ejercicios de Análisis Factorial

1. Realizar un resumen de la teoría de Análisis Factorial.
2. Una empresa dedicada al diseño de automóviles desea estudiar cuales son los deseos del comprador. Para ello realiza una encuesta a 20 individuos preguntándoles 10 características de sus productos que valoran de 1 a 5. Los datos se encuentran en el fichero factorial2. Intentar encontrar los factores de comportamiento latentes en los encuestados, identificar los factores.
3. Se estudian 100 individuos para comprobar la idea que los consumidores tienen sobre una empresa. Para ello se estudian siete variables sobre la empresa. Se desea realizar un análisis factorial para intentar reducir la dimensión de 7 a menos variables. Los datos están en el fichero Factorial 3.

Ejercicio 1: Resumen de la teoría de Análisis Factorial

Análisis factorial es un método multivariante de reducción de datos usado para explicar las correlaciones entre las variables observables (p) en términos de un número menor de variables no observadas, hipotéticas o latentes (m) llamadas factores. Las variables observadas se modelan como combinaciones lineales de factores, más expresiones de error.

Hay muchos tipos de modelo de Análisis Factorial, según las hipótesis adoptadas sobre los elementos con los que se define el Modelo u otras circunstancias, El modelo inicial que describiremos es un Modelo no-restringido (“unrestricted”) y aleatorio.

Elementos del Modelo:

- $X\ vector\ (px1)\ de\ variables\ Xi\ i=1,\ \ldots,\ p,\ de\ la\ población.$
- $F\ vector\ (mx1)\ constituido\ por\ las\ variables\ latentes\ o\ \text{“}factores\ comunes\text{”},$ $Fi\ i=1,\ \ldots,\ m,\ para\ m<p.\ F\ es\ el\ vector\ de\ factores\ comunes.$
- $\varepsilon\ vector\ (px1),\ con\ componentes\ ei\ i=1,\ \ldots,\ p,\ que\ designa\ los\ errores$ $asociados\ a\ cada\ variable\ Xi.\ Llamados\ vector\ de\ errores\ o\ factores\ específicos.$
- $L\ matriz\ (pxm),\ de\ rango\ m,\ llamada\ \text{“}matriz\ de\ factor\ loadings\text{”},\ sus\ elementos$ $son\ los\ \text{“}loadings\text{”}\ (cargas)\ de\ la\ variables\ Xi,\ respecto\ del\ factor\ común\ Fj.$

Características de los elementos del Modelo:

- Los vectores $X,\ F\ y\ \varepsilon$ son aleatorios.
- La matriz L es no aleatoria.
- Los elementos de L no tienen ninguna restricción
- X es observable y $F\ y\ \varepsilon$ no son observables.

Los elementos $X,\ F,\ \varepsilon\ y\ L$ se integran n un modelo lineal del tipo:

$$X=\mu+FL+\varepsilon$$

Método de estimación en el análisis factorial

Cuando no se puede determinar explícitamente la normalidad de los elementos aleatorios del modelo básico X, F y e, la metodología de “Componentes Principales” es la opción adecuada a implementar. En cambio, cuando se supone normalidad multivariante (modelo factorial de Lawley-Maxwell) el método de estimación por MáximaVerosimilitud, es el corresponde de manera natural a la situación y es el idóneo de uso. A continuación, mencionaremos los pasos usados por el método de los Componentes Principales.

El método de las componentes principales

Este procedimiento de estimación de Análisis Factorial, denominado de componentes principales, sigue los pasos siguientes (Tomado de guía de clases):

1. Dada la matriz de covarianzas muestrales S, se calculan los autovalores-autovectores muestrales, como soluciones, respectivamente, de $|S-\lambda I|=0$ y $(S-\lambda I)e=0$. Se obtienen así los pares $(\hat{\lambda}_i;\hat{e}_i)$, con $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$, $i=1,\ldots,p$.

2. Se toman, cuando $m<p$, los primeros $\hat{\lambda}_1,\ldots,\hat{\lambda}_m$ y se construye:

$$\hat{L}=\left[\sqrt{\hat{\lambda}_1}\hat{e}_1,\ldots,\sqrt{\hat{\lambda}_m}\hat{e}_m\right]_{p\times m}.$$

3. Las varianzas específicas muestrales, estimaciones de las ψ_i, se construyen mediante las relaciones

$$\hat{\psi}_i=s_{ii}-\sum_{j=1}^{m}\hat{l}_{ij}^2;\ i=1,\ldots,p;$$

   es decir, la diagonal principal de la matriz $S-\hat{L}\hat{L}'$.

   Por tanto $\hat{\Psi}=diag(\psi_1,\ldots,\psi_p)$, es la matriz estimada de unicidades.

4. Las comunalidades estimadas $\hat{h}_i^2$ se construyen entonces como:

$$\hat{h}_i^2=\hat{l}_{i1}^2+\ldots+\hat{l}_{im}^2.$$

Ejercicio 2:

Una empresa dedicada al diseño de automóviles desea estudiar cuales son los deseos del comprador. Para ello realiza una encuesta a 20 individuos preguntándoles 10 características de sus productos que valoran de 1 a 5. Los datos se encuentran en el fichero factorial2. Intentar encontrar los factores de comportamiento latentes en los encuestados, identificar los factores.

solución:

Empezamos haciendo la lectura del conjunto de datos y sobre el mismo entendiendo las características principales de su estructura.

```
datfact <- read.table("factorial2.txt", header=TRUE, sep="\t")
attach(datfact)
str(datfact)
## 'data.frame':    20 obs. of  10 variables:
##  $ precio      : int  4 5 2 1 1 5 4 3 4 5 ...
##  $ financiacion: int  1 5 1 1 1 5 5 2 4 5 ...
##  $ consumo     : int  4 4 3 1 2 5 4 3 4 5 ...
##  $ gasolina    : int  3 4 1 1 1 5 4 1 3 5 ...
##  $ seguridad   : int  3 3 4 4 5 3 2 4 4 2 ...
##  $ confor      : int  2 3 2 4 5 3 2 4 4 2 ...
##  $ capacidad   : int  4 4 1 2 4 4 5 2 3 3 ...
##  $ prestaciones: int  4 1 5 5 3 2 1 5 1 2 ...
##  $ juvenil     : int  4 1 4 5 3 2 1 5 1 2 ...
##  $ aerodinamica: int  4 3 5 4 2 1 1 5 1 2 ...
```

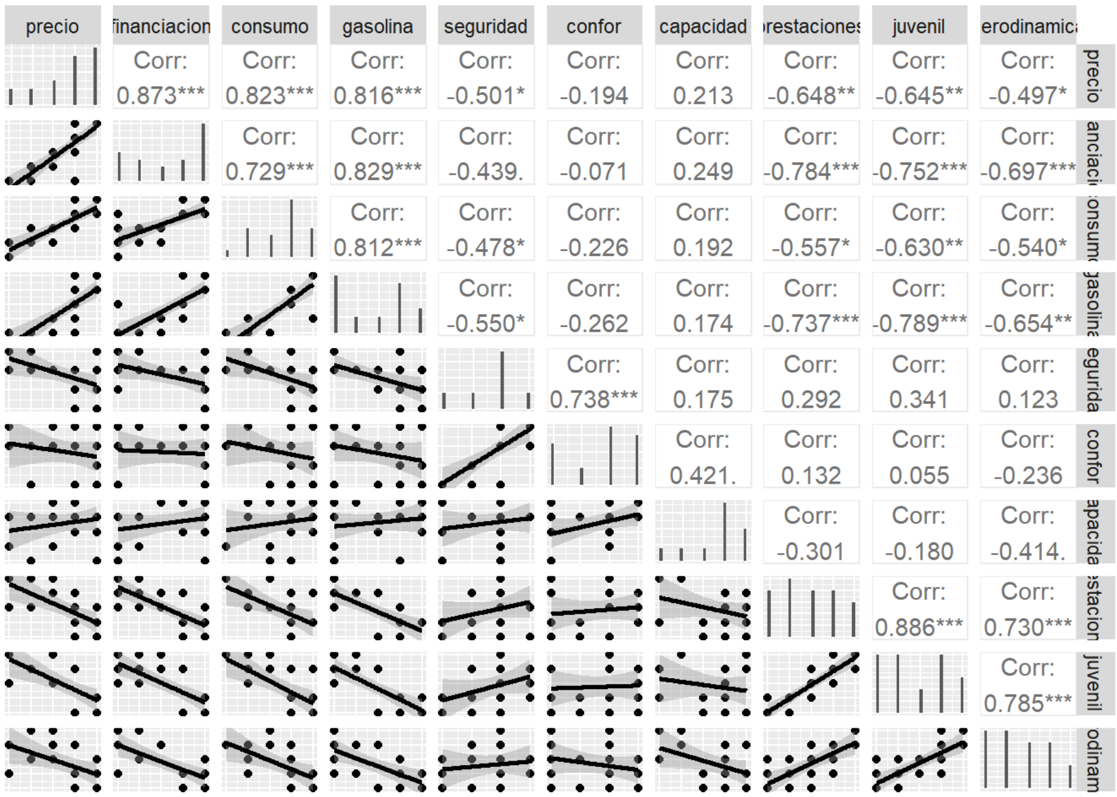
Efectivamente tenemos 20 observaciones y 10 características. Se comprueba que los datos son para cada individuo valores entre 1 y 5 y que en el conjunto no se presentan valores ausentes.

A continuación, para pasar al análisis factorial, debemos validar las hipótesis relacionadas con las variables: Linealidad, Normalidad y Correlación.

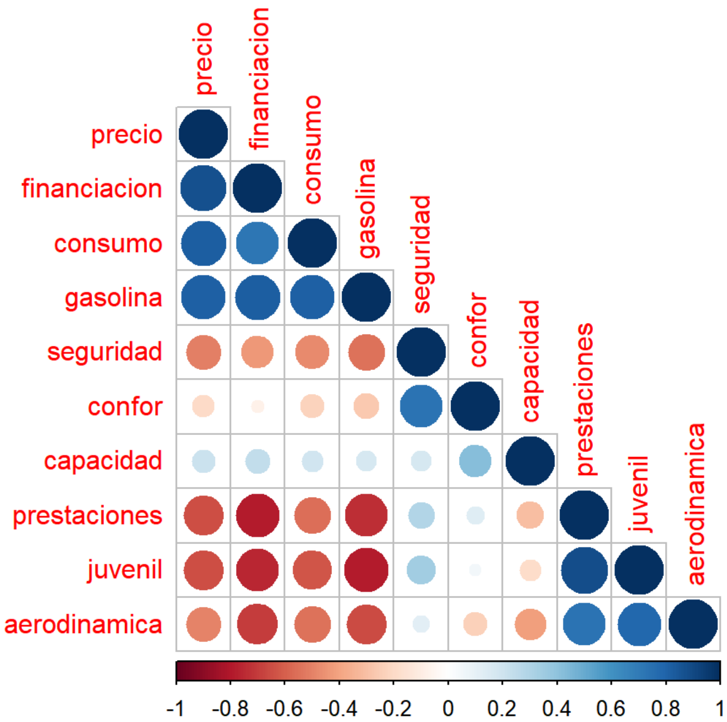
Iniciamos con un estudio de las correlaciones bivariantes:

```
require(GGally)
## Loading required package: GGally
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 4.1.3
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
require(corrplot)
## Loading required package: corrplot
## Warning: package 'corrplot' was built under R version 4.1.3
## corrplot 0.92 loaded
ggpairs(datfact, lower = list(continuous = "smooth"),
diag = list(continuous = "barDiag"), axisLabels = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
corrplot(cor(datfact),sig.level=0.05,typ="lower")
```



del resultado anterior podemos observar como, entre los pares de características, la mayoría de las combinaciones presenta una relación lineal. En algunos casos positiva y un para un grupo más numeroso negativa. La forma de la distribución de las variables, inicialmente, no sugiere que presenten un comportamiento normal. Para comprobar esta hipótesis vamos a realizar el test correspondiente.

Test de Normalidad:

Debemos comprobar si la distribución que siguen los datos se pudiera considerar como 'Distribución Normal'. Para ello generamos los respectivos gráficos la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es H0: los datos siguen una distribución normal. La prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

```
library(mvnormtest)

mshapiro.test(t(datfact))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.50098, p-value = 3.244e-07
```

Por el resultado de la prueba, debemos rechazar la hipótesis nula por lo que debemos asumir que las variables no siguen esta distribución. Como se sabe, para el análisis factorial, aun cuando tengamos este resultado el análisis es factible teniendo en cuenta que no se podrá usar los resultados de la metodología máximo verosímil.

Ejecutamos a continuación el estudio de la correlación. El cual para nuestro caso lo haremos realizando el estudio de la matriz de correlaciones.

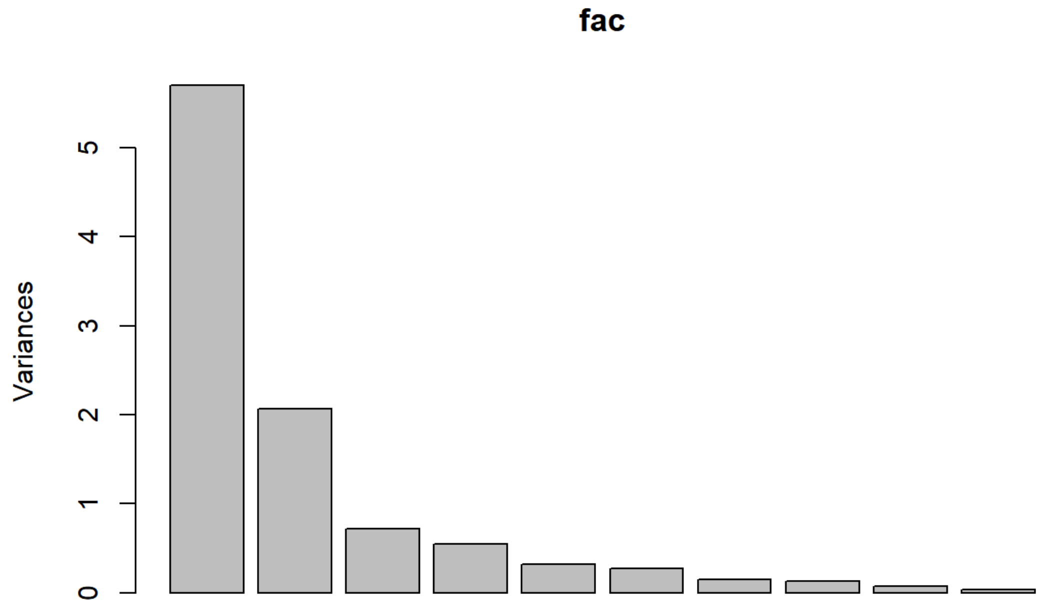
```
Redatfac<-cor(datfact,method = c("pearson"))
```

```
symnum(Redatfac)
##             prc f cns g s cnf cp prs j a
## precio      1
## financiacion +   1
## consumo     +   , 1
## gasolina    +   + +   1
## seguridad   .   . .   . 1
## confor                  , 1
## capacidad                 .   1
## prestaciones ,   , .   ,       .  1
## juvenil     ,   , ,   , .        +   1
## aerodinamica .   , .   ,       .  ,   , 1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

El resultado muestra como algunos pares de variables son significativos y que el determinante de la matriz es pequeño, por lo que se aceptará que la matriz es significativa. En otras palabras, se cumple la hipótesis.

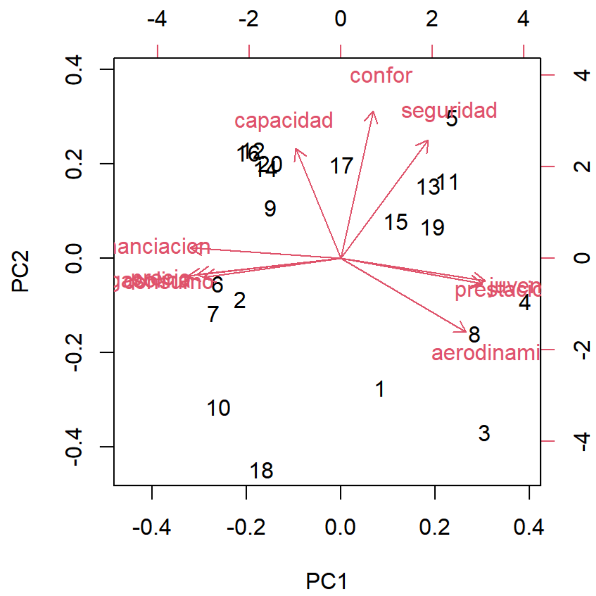
Ahora pasamos a obtener el análisis factorial, aplicando la técnica de Componentes Principales. Primero, con base en el resumen del análisis y la gráfica de los autovalores (plot) buscamos determinar el número de factores a considerar.

```
fac<-prcomp(datfact, retx=,center=TRUE,scale.=TRUE,tol=NULL)
summary(fac)
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.3877 1.4385 0.84881 0.74013 0.56192 0.52031 0.38267
## Proportion of Variance 0.5701 0.2069 0.07205 0.05478 0.03158 0.02707 0.01464
## Cumulative Proportion  0.5701 0.7770 0.84908 0.90386 0.93544 0.96251 0.97715
##                           PC8     PC9    PC10
## Standard deviation     0.3578 0.26145 0.17915
## Proportion of Variance 0.0128 0.00684 0.00321
## Cumulative Proportion  0.9899 0.99679 1.00000
plot(fac)
```



Obtenemos ahora el biplot de los componentes principales.

```
biplot(fac)
```



De los resultados anteriores vemos que la forma como se distribuyen los factores sugiere que podemos considerar tres factores. Para comprobar la validez de esta decisión, vamos a comprobar si la hipótesis del número de factores considerado es correcta. Para ello recurriremos al paquete factanal:

```
facmle<-vector("list",3)
for (i in 1:3) {
  facmle<-factanal(datfact,i)
  }
facmle
```

```
##
## Call:
## factanal(x = datfact, factors = i)
##
## Uniquenesses:
##       precio financiacion      consumo     gasolina    seguridad       confor
##        0.058        0.122        0.275        0.159        0.301        0.005
##    capacidad prestaciones      juvenil aerodinamica
##        0.715        0.144        0.100        0.220
##
## Loadings:
##              Factor1 Factor2 Factor3
## precio        0.908  -0.338
## financiacion  0.749  -0.561
## consumo       0.756  -0.377  -0.109
## gasolina      0.677  -0.599  -0.155
## seguridad    -0.459   0.178   0.676
## confor       -0.150           0.986
## capacidad     0.181  -0.209   0.457
## prestaciones -0.385   0.839
## juvenil      -0.385   0.867
## aerodinamica -0.304   0.775  -0.295
##
##                Factor1 Factor2 Factor3
## SS loadings      3.069   3.062   1.769
## Proportion Var   0.307   0.306   0.177
## Cumulative Var   0.307   0.613   0.790
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 16.72 on 18 degrees of freedom.
## The p-value is 0.542
```

El resultado de la prueba nos indica que no hay suficiente evidencia para descartar la hipótesis nula, en otras palabras, es correcto considerar los tres primeros factores principales. Es de destacar que estos tres primeros factores explican el 79% de la variabilidad total.

Considerando los tres factores principales, las cargas factoriales serán:

```
fac<-prcomp(datfact, retx=,center=TRUE,scale.=TRUE,tol=NULL,rank. = 3)
fac
```

```
## Standard deviations (1, .., p=10):
##  [1] 2.3877077 1.4384703 0.8488130 0.7401283 0.5619199 0.5203111 0.3826696
##  [8] 0.3577946 0.2614498 0.1791511
##
## Rotation (n x k) = (10 x 3):
##                     PC1         PC2         PC3
## precio       -0.3678778 -0.06850674  0.31336900
## financiacion -0.3866893  0.04415747  0.03957517
## consumo      -0.3521317 -0.08035169  0.31548847
## gasolina     -0.3908007 -0.07570314  0.02422468
## seguridad     0.2229614  0.50100370 -0.17648676
## confor        0.0826891  0.62572362  0.00103351
## capacidad    -0.1156815  0.46704502  0.63433016
## prestaciones  0.3609384 -0.10858780  0.33098781
## juvenil       0.3681529 -0.09531830  0.43029392
## aerodinamica  0.3204229 -0.31563730  0.26811319
```

Estas cargas factoriales corresponden a los autovectores de la matriz de correlaciones. Para obtener las cargas factoriales debemos multiplicar cada columna por la raíz cuadrada del autovalor correspondiente.

```
cargas11<-matrix(0,10,3)
for (i in 1:3) cargas11[,i]<-fac$rotation[,i]*fac$sdev[i]
cargas12<-varimax(cargas11,normalize=T)$loadings
print(cargas12,cutoff=0.3)
```

```
##
## Loadings:
##       [,1]   [,2]   [,3]
##  [1,] -0.627 -0.538  0.411
##  [2,] -0.805 -0.339  0.307
##  [3,] -0.590 -0.538  0.394
##  [4,] -0.776 -0.483
##  [5,]         0.878
##  [6,]         0.838  0.380
##  [7,]                0.847
##  [8,]  0.908
##  [9,]  0.954
## [10,]  0.876
```

```
##
##                  [,1]  [,2]  [,3]
## SS loadings    4.579 2.520 1.392
## Proportion Var 0.458 0.252 0.139
## Cumulative Var 0.458 0.710 0.849
```

Podemos obtener las comunalidades y las unicidades con base en el procedimiento indicado en las notas de la materia (pag.50)

```
comunalidad<-matrix(0,10,2)
for (i in 1:10)
  {for (j in 1:3)
        {comunalidad[i,1]=comunalidad[i,1]+cargas11[i,j]^2
        comunalidad[i,2]=1-comunalidad[i,1]}}
comunalidad
```

```
##             [,1]       [,2]
##  [1,] 0.8520221 0.14797789
##  [2,] 0.8576481 0.14235192
##  [3,] 0.7919950 0.20800495
##  [4,] 0.8829902 0.11700975
##  [5,] 0.8252336 0.17476639
##  [6,] 0.8491350 0.15086504
##  [7,] 0.8175543 0.18244572
##  [8,] 0.8460554 0.15394456
##  [9,] 0.9249136 0.07508643
## [10,] 0.8432811 0.15671894
```

Obtenemos también la contribución de cada factor en la explicación de cada variable, tanto para el total de la varianza de la variable (cargas13) como para el total explicado por el modelo factorial (cargas14):

```
cargas13<-matrix(0,10,3)
cargas14<-matrix(0,10,3)
  for (i in 1:10)
    {cargas13[i,]<-cargas12[i,]^2
    cargas14[i,]<-cargas13[i,]/comunalidad[i,1]}
print('cargas13')
```

```
## [1] "cargas13"
```

```
cargas13
```

```
##              [,1]       [,2]         [,3]
##  [1,] 0.39347785 0.28986793 1.686763e-01
##  [2,] 0.64865266 0.11493293 9.406248e-02
##  [3,] 0.34754744 0.28928274 1.551649e-01
```

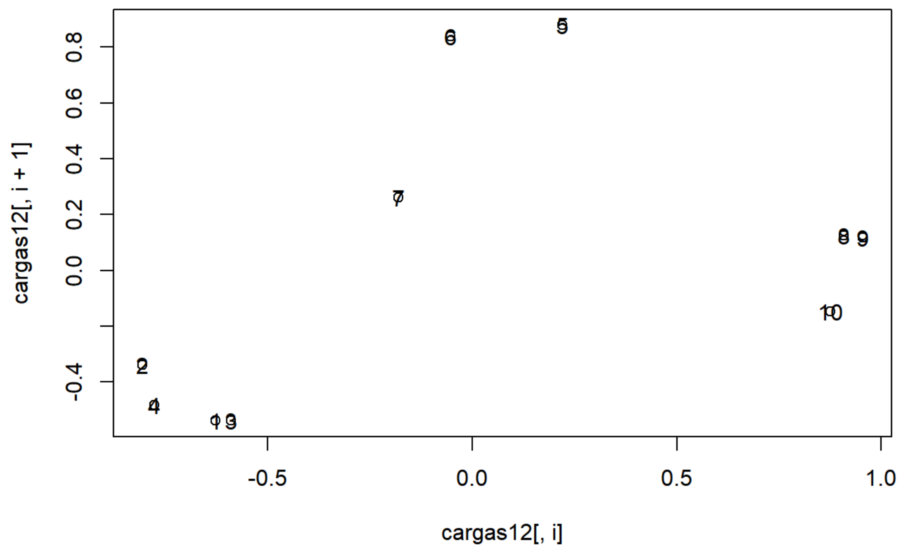
```
##   [4,] 0.60282385 0.23369272 4.647368e-02
##   [5,] 0.04840316 0.77069532 6.135133e-03
##   [6,] 0.00275162 0.70171912 1.446642e-01
##   [7,] 0.03234835 0.06847615 7.167298e-01
##   [8,] 0.82515995 0.01568513 5.210361e-03
##   [9,] 0.91080048 0.01410569 7.405625e-06
##  [10,] 0.76718387 0.02138705 5.471014e-02
print('cargas14')
## [1] "cargas14"
cargas14
##              [,1]       [,2]         [,3]
##   [1,] 0.461816483 0.34021174 1.979718e-01
##   [2,] 0.756315648 0.13400943 1.096749e-01
##   [3,] 0.438825274 0.36525827 1.959165e-01
##   [4,] 0.682707255 0.26466059 5.263216e-02
##   [5,] 0.058653890 0.93391169 7.434420e-03
##   [6,] 0.003240498 0.82639292 1.703666e-01
##   [7,] 0.039567224 0.08375731 8.766755e-01
##   [8,] 0.975302459 0.01853913 6.158415e-03
##   [9,] 0.984741173 0.01525082 8.006830e-06
##  [10,] 0.909760585 0.02536171 6.487770e-02
```

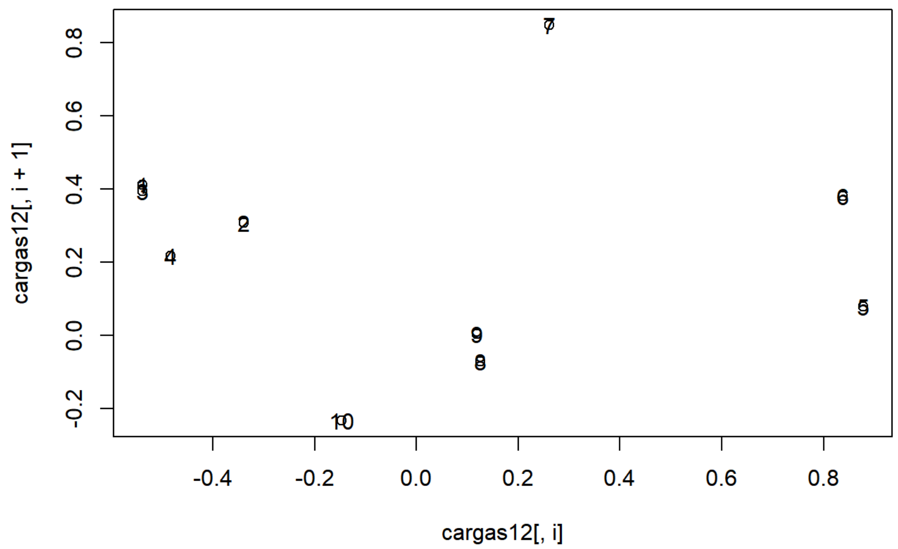
Vemos como el primer factor, explica de forma adecuada la variabilidad (casi total) de la variable 9 (Juvenil), con una proporción de Cargas13=0.91.

Completamos el análisis realizando otras gráficas.

Obtenemos las gráficas bidimensionales de las cargas factoriales.

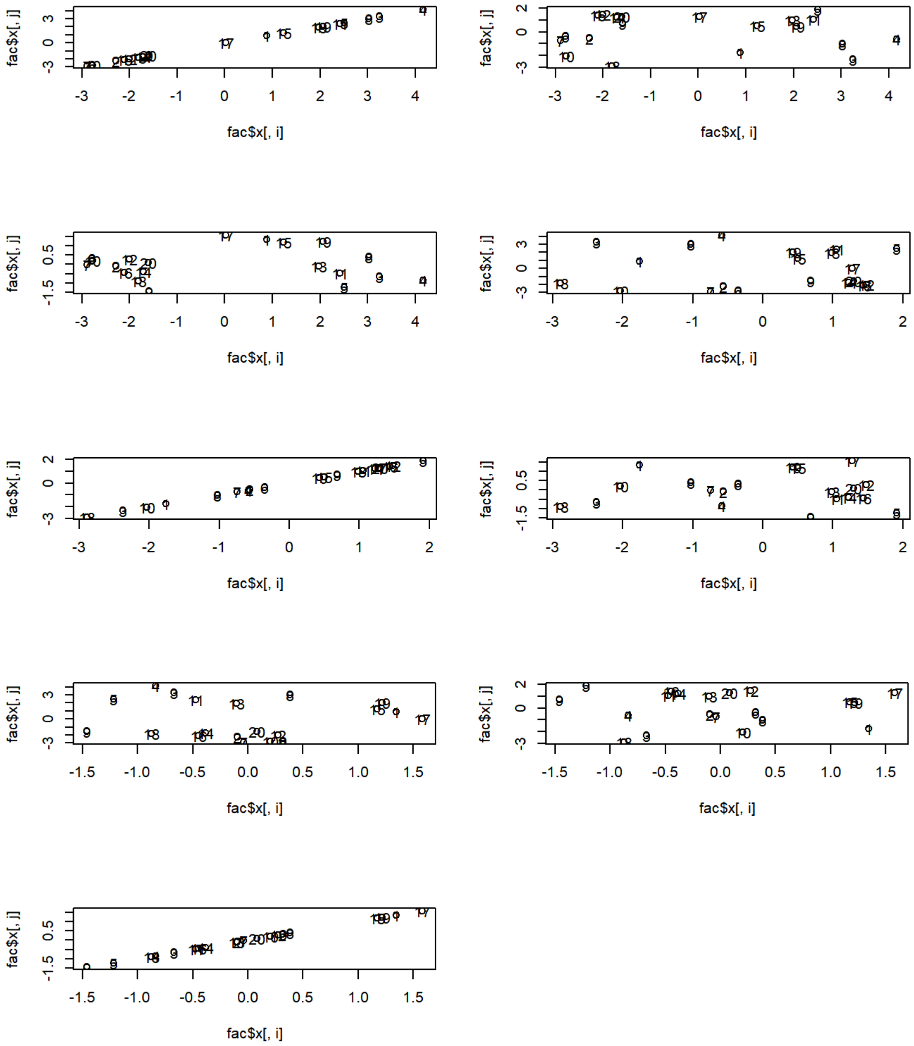
```
for (i in 1:2){
  plot(cargas12[,i],cargas12[,i+1])
  text(cargas12[,i],cargas12[,i+1])}
```





Ahora, graficamos las puntuaciones factoriales de los individuos:

```
# puntuaciones factoriales
par(mfrow=c(3,2))
for (i in 1:3){
  for(j in 1:3){
    {plot(fac$x[,i],fac$x[,j])}
    text(fac$x[,i],fac$x[,j])}}
```



En conclusión: identificamos tres factores de comportamiento latente los cuales, con base en los resultados anteriores, pueden ser identificados como:

Factor 1: Aspecto. Que relaciona a las características como Juvenil, Aerodinámico, etc.

Factor 2: Utilidad. Que relaciona a aspectos como Capacidad, confort, seguridad, etc

Factor 3: Economía. Que relaciona a aspectos como precio, financiación, etc.

Ejercicio 3:

Se estudian 100 individuos para comprobar la idea que los consumidores tienen sobre una empresa. Para ello se estudian siete variables sobre la empresa. Se desea realizar un análisis factorial para intentar reducir la dimensión de 7 a menos factores. Los datos están en el fichero Factorial 3.

solución:

Empezamos haciendo la lectura del conjunto de datos y sobre el mismo entendiendo las características principales de los datos a analizar

```
require(foreign) #Utilizamos este paquete para leer los datos desdel el archivo .SAV ya que .TXT tiene arreglos
```

```
## Loading required package: foreign

## Warning: package 'foreign' was built under R version 4.1.3
```

```
datfact2 <- read.spss("factorial3.sav", to.data.frame = TRUE)

attach(datfact2)

str(datfact2)
```

```
## 'data.frame':    100 obs. of  7 variables:

##  $ V1: num  4.1 1.8 3.4 2.7 6 1.9 4.6 1.3 5.5 4 ...

##  $ V2: num  0.6 3 5.2 1 0.9 3.3 2.4 4.2 1.6 3.5 ...

##  $ V3: num  6.9 6.3 5.7 7.1 9.6 7.9 9.5 6.2 9.4 6.5 ...

##  $ V4: num  4.7 6.6 6 5.9 7.8 4.8 6.6 5.1 4.7 6 ...

##  $ V5: num  2.4 2.5 4.3 1.8 3.4 2.6 3.5 2.8 3.5 3.7 ...

##  $ V6: num  2.3 4 2.7 2.3 4.6 1.9 4.5 2.2 3 3.2 ...

##  $ V7: num  5.2 8.4 8.2 7.8 4.5 9.7 7.6 6.9 7.6 8.7 ...
```

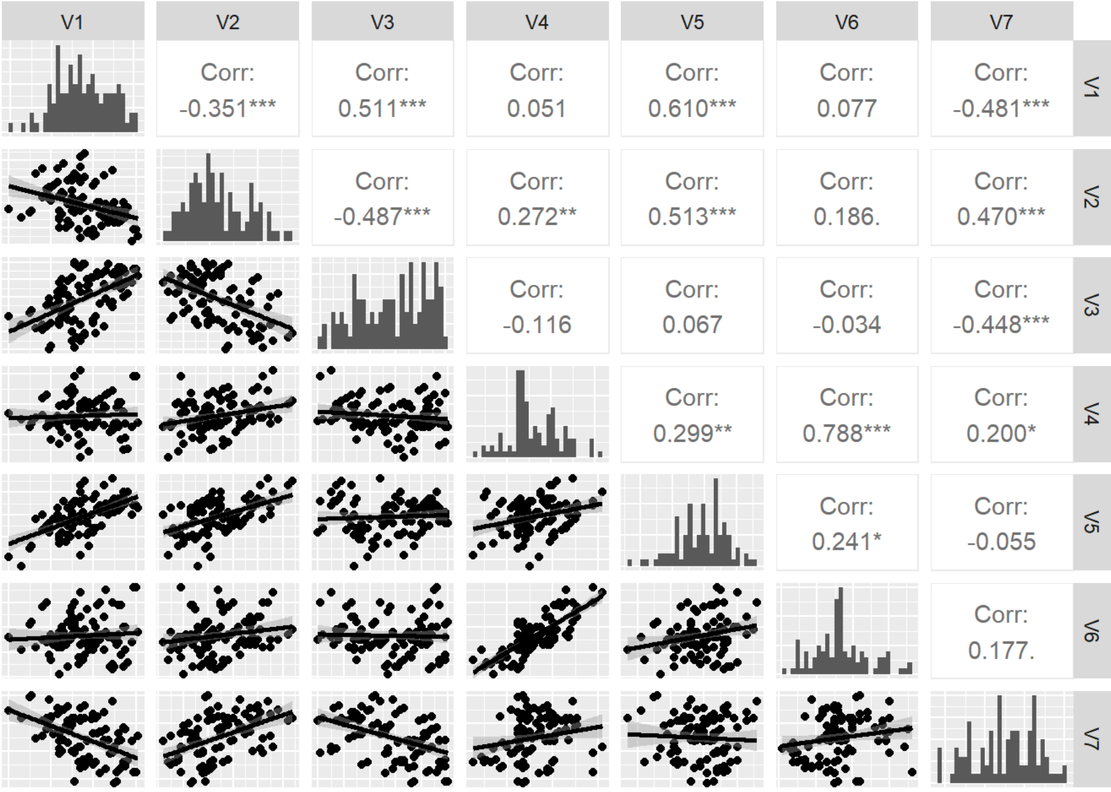
Efectivamente tenemos 100 observaciones y 7 características. Se comprueba que los datos son para cada individuo valores numéricos y que presentan un valor decimal. En el conjunto no se presentan valores ausentes. Leimos el conjunto desde el archivo spss (.sav) ya que el archivo .txt requiere trabajo de ordenamiento para su lectura.

A continuación, para pasar al análisis factorial, debemos validar las hipótesis relacionadas con las variables: Linealidad, Normalidad y Correlación.

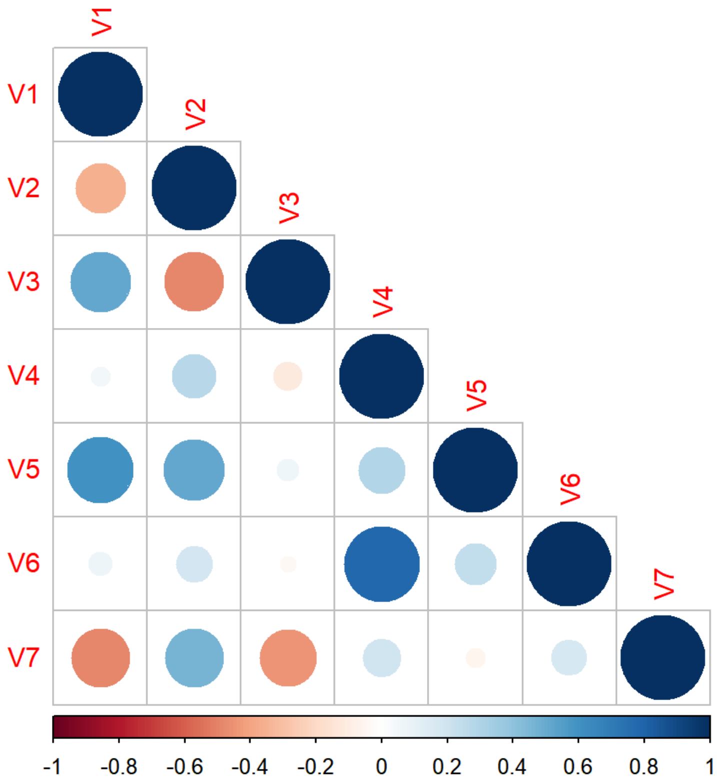
Iniciamos con un estudio de las correlaciones bivariantes:

```
require(GGally)

require(corrplot)

ggpairs(datfact2, lower = list(continuous = "smooth"),

diag = list(continuous = "barDiag"), axisLabels = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
corrplot(cor(datfact2),sig.level=0.05,typ="lower")
```



del resultado anterior podemos observar como, entre los pares de características, para algunas de las combinaciones se presenta una relación lineal. La forma de la distribución de las variables, inicialmente, no sugiere que presenten un comportamiento normal en todos los casos. Para comprobar esta hipótesis vamos a realizar el test correspondiente.

Test de Normalidad:

Debemos comprobar si la distribución que siguen los datos se pudiera considerar como ‘Distribución Normal’. Para ello generamos los respectivos gráficos la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es H0: los datos siguen una distribución normal. La prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

Normalidad:

Debemos comprobar si la distribución que siguen los datos se pudiera considerar como ‘Distribución Normal’. Para ello generamos los respectivos gráficos la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es H0: los datos siguen una distribución normal.

```r
library(mvnormtest)

mshapiro.test(t(datfact2))
```

```
##

##  Shapiro-Wilk normality test

##

## data:  Z

## W = 0.59595, p-value = 3.609e-15
```

Por el resultado del test, p-value = 3.609e-15, debemos rechazar la hipótesis nula por lo que debemos asumir que las variables no siguen una distribución normal. Como se sabe, para el análisis factorial, aun cuando tengamos este resultado el análisis es factible teniendo en cuenta que no se podrá usar los resultados de la metodología máximo verosimil.

Ejecutamos a continuación el estudio de la correlación. El cual para nuestro caso lo haremos realizando el estudio de la matriz de correlaciones.

```r
Redatfac2<-cor(datfact2,method = c("pearson"))

symnum(Redatfac2)
```

```
##    V1 V2 V3 V4 V5 V6 V7

## V1 1

## V2 .  1

## V3 .  .  1

## V4          1

## V5 ,  .        1

## V6          ,     1

## V7 .  .  .           1

## attr(,"legend")

## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

El resultado muestra como el determinante de la matriz es pequeño, pero no es concluyente para aceptar la hipótesis de correlación. En este caso aplicaremos adicionalmente el contraste de esfericidad de Barlet.

```r
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.3
##
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
cortest.bartlett(datfact2,n=NULL)
```

```
## R was not square, finding R from data
## $chisq
## [1] 564.5078
##
## $p.value
## [1] 4.573487e-106
##
## $df
## [1] 21
```

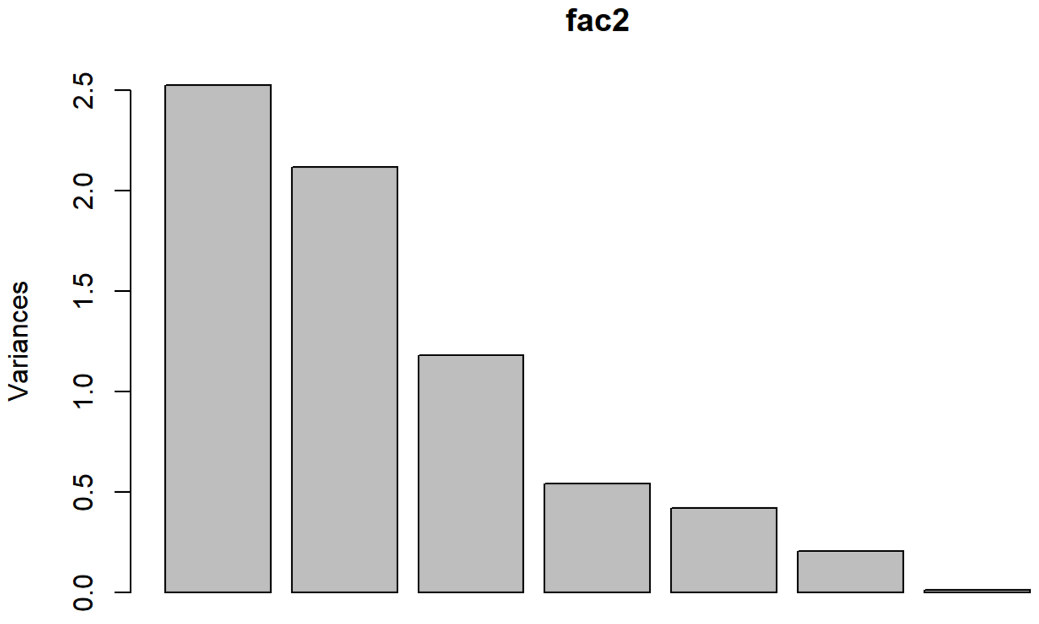
Este test contrasta si la matriz de correlaciones es igual a la identidad (en cuyo caso no existiría correlación) frente a que la matriz de correlaciones sea significativa. Dado el valor p-value menor a 0.05, se rechaza la hipótesis nula y se aceptará que la matriz es significativa. En otras palabras, se cumple la hipótesis.

Ahora pasamos a obtener el análisis factorial, aplicando la técnica de Componentes Principales. Primero, con base en el resumen del análisis y la gráfica de los autovalores (plot) buscamos determinar el número de factores a considerar.

```
fac2<-prcomp(datfact2, retx=,center=TRUE,scale.=TRUE,tol=NULL)
summary(fac2)
```

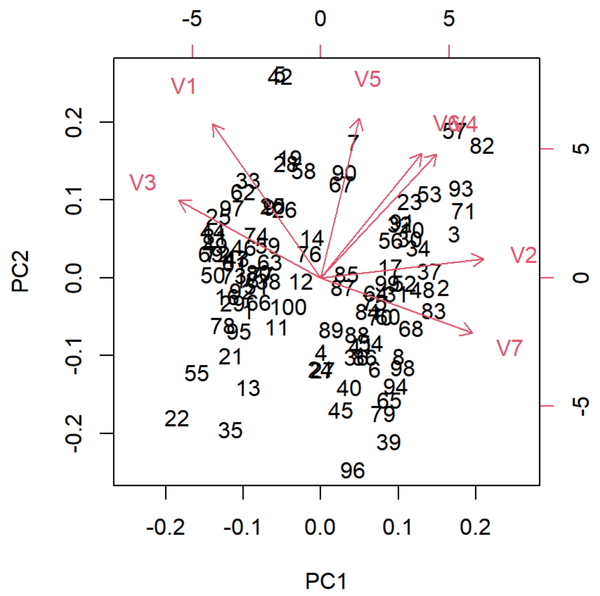
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.5895 1.4559 1.0864 0.73575 0.64697 0.4521 0.09721
## Proportion of Variance 0.3609 0.3028 0.1686 0.07733 0.05979 0.0292 0.00135
## Cumulative Proportion  0.3609 0.6637 0.8323 0.90966 0.96945 0.9987 1.00000
```

```
plot(fac2)
```



Obtenemos ahora el biplot de los componentes principales.

```
biplot(fac2)
```



De los resultados anteriores vemos que la forma como se distribuyen los factores sugiere que podemos considerar tres factores. Para comprobar la validez de esta decisión, vamos a comprobar si la hipótesis del número de factores considerado es correcta. Para ello recurriremos al paquete factanal:

```
facmle<-vector("list",3)
for (i in 1:3) {
  facmle<-factanal(datfact2,i)
  }
facmle
```

```
##
## Call:
## factanal(x = datfact2, factors = i)
##
## Uniquenesses:
##    V1    V2    V3    V4    V5    V6    V7
## 0.019 0.005 0.622 0.027 0.005 0.360 0.638
##
## Loadings:
##    Factor1 Factor2 Factor3
## V1 -0.832           0.530
## V2  0.803   0.103   0.582
## V3 -0.613
## V4  0.121   0.971   0.128
## V5          0.191   0.975
## V6          0.792
## V7  0.585   0.139
##
##                Factor1 Factor2 Factor3
## SS loadings      2.082   1.645   1.597
## Proportion Var   0.297   0.235   0.228
## Cumulative Var   0.297   0.532   0.761
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 3.85 on 3 degrees of freedom.
## The p-value is 0.278
```

El resultado del test nos indica que no hay suficiente evidencia para descartar la hipótesis nula (>0.05), en otras palabras, es correcto considerar los tres primeros factores principales. Es de destacar que estos tres primeros factores explican el 76% de la variabilidad total.

Como el objetivo principal es reducir la cantidad de 7 a menos factores, se probaron también los siguientes escenarios en cuanto a número de factores:

4 factores: el test produjo el siguiente resultado: Error in factanal(datfact2, i) : 4 factors are too many for 7 variables

2 factores: el test produjo el siguiente resultado: The p-value is 4.13e-17. Lo que rechazaría la hipótesis nula indicando que ese número es inadecuado.

Dado estos resultados, se valida que el número mínimos adecuado de factores en este caso es 3.

Teniendo una estructura factorial con 3 factores. Las cargas factoriales serán:

```
fac2<-prcomp(datfact2, retx=,center=TRUE,scale.=TRUE,tol=NULL,rank. = 3)
fac2
```

```
## Standard deviations (1, .., p=7):
## [1] 1.58952391 1.45586032 1.08635017 0.73575470 0.64696562 0.45208286 0.09720795
##
## Rotation (n x k) = (7 x 3):
##           PC1         PC2          PC3
## V1 -0.3329558  0.51540285 -0.185758972
## V2  0.4987557  0.06357912 -0.467551719
## V3 -0.4357607  0.25747969  0.157209950
## V4  0.3543065  0.41442357  0.415692916
## V5  0.1168717  0.53427953 -0.548858562
## V6  0.3091411  0.41567963  0.498089410
## V7  0.4644733 -0.18447422  0.005885988
```

Estas cargas factoriales corresponden a los autovectores de la matriz de correlaciones. Para obtener las cargas factoriales debemos multiplicar cada columna por la raíz cuadrada del autovalor correspondiente.

```
cargas21<-matrix(0,7,3)
for (i in 1:3) cargas21[,i]<-fac2$rotation[,i]*fac2$sdev[i]
cargas22<-varimax(cargas21,normalize=T)$loadings
print(cargas22,cutoff=0.3)
```

```
##
## Loadings:
##      [,1]   [,2]   [,3]
## [1,] -0.755  0.556
## [2,]  0.753  0.563
## [3,] -0.806
## [4,]                0.921
## [5,]         0.980
## [6,]                0.945
## [7,]  0.759
##
##                 [,1]  [,2]  [,3]
## SS loadings     2.38 1.619 1.827
## Proportion Var 0.34 0.231 0.261
```

```
## Cumulative Var 0.34 0.571 0.832
```

Podemos obtener las comunalidades y las unicidades con base en el procedimiento indicado en las notas de la materia (pag.50)

```
comunalidad<-matrix(0,7,2)
for (i in 1:7)
  {for (j in 1:3)
        {comunalidad[i,1]=comunalidad[i,1]+cargas21[i,j]^2
        comunalidad[i,2]=1-comunalidad[i,1]}}
comunalidad
##             [,1]        [,2]
## [1,] 0.8838512 0.116148796
## [2,] 0.8950622 0.104937819
## [3,] 0.6494502 0.350549813
## [4,] 0.8851246 0.114875422
## [5,] 0.9950572 0.004942813
## [6,] 0.9004826 0.099517396
## [7,] 0.6172443 0.382755732
```

Obtenemos también la contribución de cada factor en la explicación de cada variable, tanto para el total de la varianza de la variable (cargas23) como para el total explicado por el modelo factorial (cargas24):

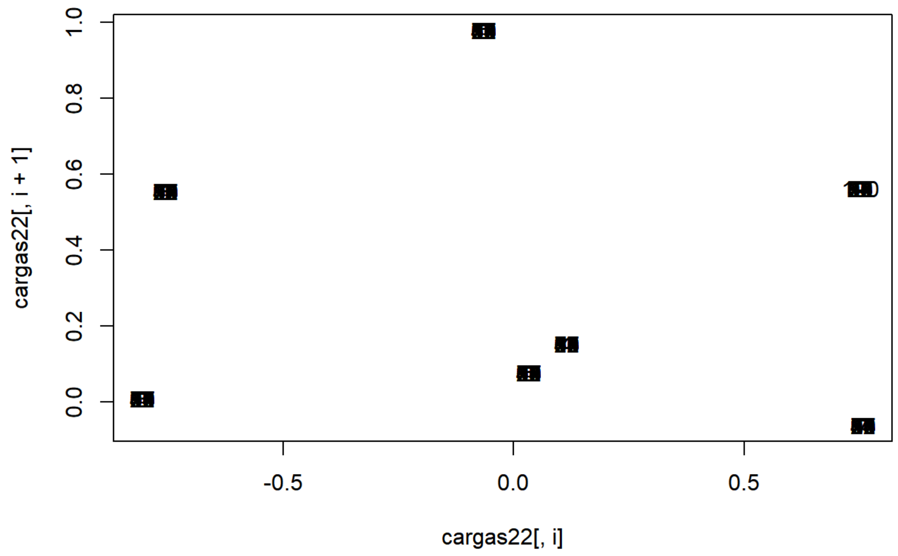
```
cargas23<-matrix(0,7,3)
cargas24<-matrix(0,7,3)
  for (i in 1:7)
    {cargas23[i,]<-cargas22[i,]^2
    cargas24[i,]<-cargas23[i,]/comunalidad[i,1]}
print('cargas23')
## [1] "cargas23"
cargas23
##              [,1]         [,2]         [,3]
## [1,] 0.569876901 3.088782e-01 5.096092e-03
## [2,] 0.566488047 3.167951e-01 1.177903e-02
## [3,] 0.649349774 7.109162e-05 2.932115e-05
## [4,] 0.013340485 2.337548e-02 8.484086e-01
## [5,] 0.004141033 9.598664e-01 3.104978e-02
## [6,] 0.001111131 5.853743e-03 8.935177e-01
## [7,] 0.575831150 3.860993e-03 3.755212e-02
print('cargas24')
## [1] "cargas24"
cargas24
```

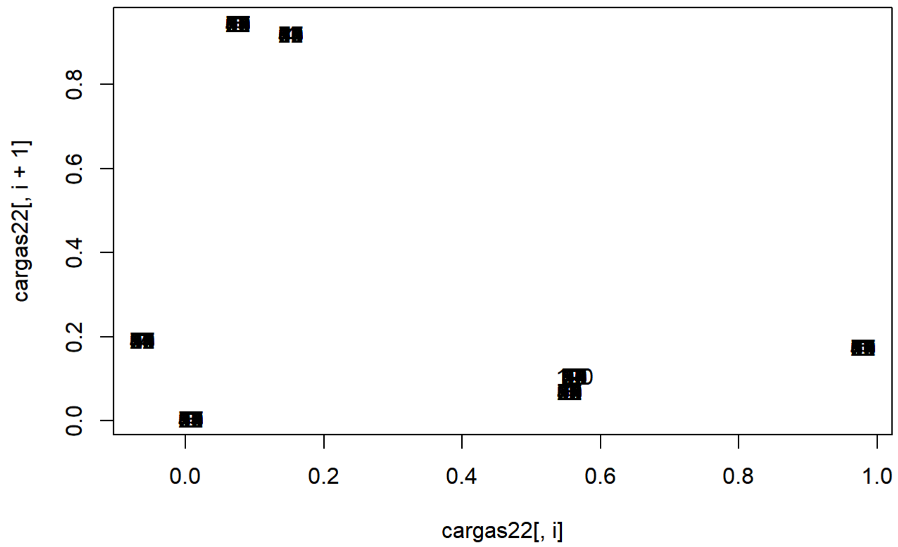
```
##            [,1]         [,2]         [,3]
## [1,] 0.644765655 0.3494685650 5.765780e-03
## [2,] 0.632903567 0.3539364244 1.316001e-02
## [3,] 0.999845388 0.0001094643 4.514766e-05
## [4,] 0.015071873 0.0264092489 9.585189e-01
## [5,] 0.004161603 0.9646343767 3.120402e-02
## [6,] 0.001233928 0.0065006729 9.922654e-01
## [7,] 0.932906436 0.0062552105 6.083835e-02
```

Vemos como el primer factor, explica de forma adecuada gran parte de la variabilidad de la variable V3, con una proporción de Cargas23=0.65.

A continuación, Obtenemos las gráficas bidimensionales de las cargas factoriales.

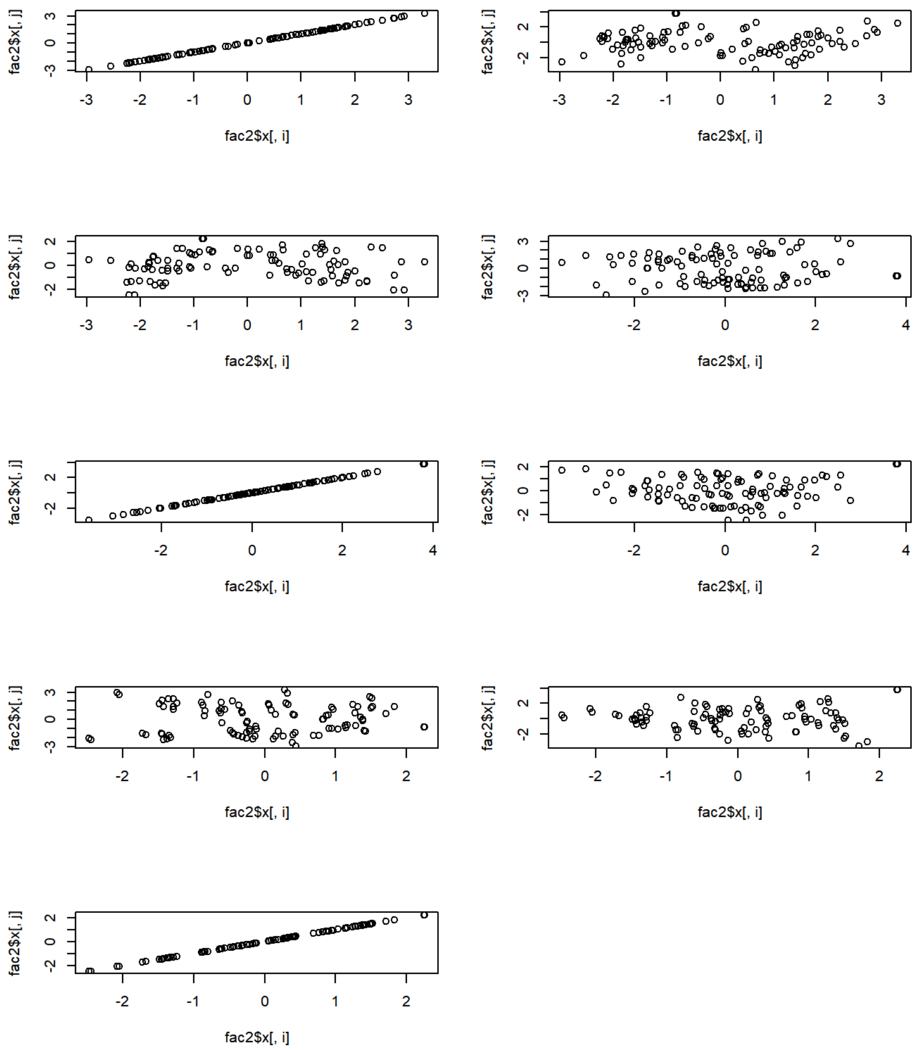
```
for (i in 1:2){
  plot(cargas22[,i],cargas22[,i+1])
  text(cargas22[,i],cargas22[,i+1],labels=row.names(datfact2))}
```





Ahora, graficamos las puntuaciones factoriales de los individuos:

```
# puntuaciones factoriales
par(mfrow=c(3,2))
for (i in 1:3){
  for(j in 1:3){
    {plot(fac2$x[,i],fac2$x[,j])}
    text(fac2$x[,i],fac2$x[,j],labels=row.names(fac2$x))}}
```



En conclusión:

identificamos tres factores de comportamiento latente los cuales, se comprobó que es número mínimo adecuado para este caso con la aplicación del contraste de hipótesis del paquete factanal. De esta forma pueden ser identificados los factores como:

Factor 1: Que relaciona a las características V4,V5,V6

Factor 2: Que relaciona a aspectos V1 y V3

Factor 3: Que relaciona a las variables V2 y V7