Tarea4

Javier Montiel González, Eliza Zenteno Munuzuri, Andrés Cruz

24/10/2020

Ejercicio 1

Matriz de correlaciones

```
##      [,1] [,2] [,3]
## [1,] 1.00 0.63 0.45
## [2,] 0.63 1.00 0.35
## [3,] 0.45 0.35 1.00
```

Al aplicar el método de máxima verosimilitud del análisis de factores con m=1 obtenemos lo siguiente:

```
##
## Call:
## factanal(factors = 1, covmat = rho)
##
## Uniquenesses:
## [1] 0.19 0.51 0.75
##
## Loadings:
##      Factor1
## [1,] 0.9
## [2,] 0.7
## [3,] 0.5
##
##                Factor1
## SS loadings      1.550
## Proportion Var   0.517
##
## The degrees of freedom for the model is 0 and the fit was 0
```

Calculamos $LL' + \Psi$

```
L%*%t(L)+psi
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000003 0.6299992 0.4499997
## [2,] 0.6299992 1.0000022 0.3500010
## [3,] 0.4499997 0.3500010 0.9999999
```

Por lo tanto, $\rho = LL' + \Psi$, salvo por errores numéricos.

Ejercicio 2

Se tiene la siguiente matriz de factores no rotada, obtenida utilizando el método de componentes principales y considerando 4 factores.

```
library(stats)
factores <- matrix(c( 0.881, 0.828, 0.664, 0.792, 0.731, 0.476, -0.347, 0.508, -0.711, 0.564, -0.647, 0
row.names(factores) <- paste0("X",1:6)
```

```
colnames(factores) <- paste0("F",1:4)
factores
```

```
##       F1     F2     F3     F4
## X1 0.881 -0.347 -0.165  0.268
## X2 0.828  0.508 -0.070 -0.200
## X3 0.664 -0.711  0.154 -0.031
## X4 0.792  0.564 -0.179 -0.029
## X5 0.731 -0.647  0.117 -0.125
## X6 0.476  0.804  0.329  0.135
```

Obtener las comunalidades, el eigenvalor y el porcentaje de variación explicada

comunalidades:

```
diag(factores%*% t(factores))
```

```
##       X1       X2       X3       X4       X5       X6
## 0.995619 0.988548 0.971094 0.978242 0.982284 0.999458
```

Los altos valores de la comunalidades indican que los 4 factores considerados explican una buena proporcion dela varianza de cada variable.

Ejercicio 3

a)$(I + L^T\psi L)^{-1}(L^T\psi^{-1}L) = I - (I + L^T\psi^{-1}L)^{-1}$

$(I + L^T\psi L)^{-1}(L^T\psi^{-1}L) = (I + L^T\psi L)^{-1}(I + L^T\psi^{-1}L - I) = (I + L^T\psi L)^{-1}((I + L^T\psi^{-1}L) - I) = (I + L^T\psi L)^{-1}(I + L^T\psi^{-1}L) - (I + L^T\psi^{-1}L)^{-1}I = I - (I + L^T\psi^{-1}L)^{-1}$

b) $L^T(LL^T + \psi)^{-1} = (I + L^T\psi^{-1}L)^{-1}L\psi^{-1}$

$L^T(LL^T + \psi)^{-1} = L^T(LL^T)^{-1} + L^T\psi^{-1} = L^{-1} + L^T\psi^{-1} = (L^{-T}\psi L^{-1} + I)L^T\psi^{-1} = (I + L^T\psi^{-1}L)^{-1}L\psi^{-1}$

Ejercicio 4.

El siguiente ejemplo muestra un caso que se conoce como el caso de Heywood. Consideren un modelo factorial con m = 1 para la población con matriz de covarianza Σ

```
c1 <-c(1,0.4,0.9)
c2 <-c(0.4, 1, 0.7)
c3 <-c(0.9, 0.7, 1)

rbind(c1,c2,c3)
```

```
##    [,1] [,2] [,3]
## c1  1.0  0.4  0.9
## c2  0.4  1.0  0.7
## c3  0.9  0.7  1.0
```

Considerando $L = [l_{1,1}, l_{1,2}, l_{1,3}]'$

A partir de la matriz $\sum$ se definen las siguientes ecuaciones: (1) $1 = l_{1,1}^2 + \psi_1$ (2) $0.4 = l_{1,1}l_{1,2}$ (3) $1 = l_{1,2}^2 + \psi_2$ (4) $0.9 = l_{1,1}l_{1,3}$ (5) $0.7 = l_{1,2}l_{1,3}$ (6) $1 = l_{1,3}^2 + \psi_3$

Despejando $l_{1,1}$ de (1), sustituimos en (2) y (4) de donde se obtiene que: $l_{1,2} = \frac{0.4}{\sqrt{(1-\psi_1)}}$, $l_{1,3} = \frac{0.9}{\sqrt{(1-\psi_1)}}$.

Sustituyendo ambos valores en (5) se obtiene que $\psi = 0.48572$

Resulta que: $l_{1,1} = +/-0.7171$ $\psi_1 = 0.4857$ $l_{1,2} = +/-0.5577$ $\psi_2 = 0.6888$ $l_{1,3} = +/-1.2549$ $\psi_3 = -0.5750$

Ocurre que $\psi_3 \leq 0$ pero no es valido ya que se trata de una varianza.

Ejercicio 5

Datos de monitoreo atmosférico (REDMA)

Para la creación de la base de datos se tomaron los regitros de las estaciones que midieran los tres contaminantes (PM10, PST y PM25) considerados para el año 2019. Además, se descartaron las estaciones que tuvieran mediciones nulas en todos sus registros del año para alguno de los contaminantes. Así pues, las tres estaciones que se escogieron fueron: Tlanepantla (TLA), UAM Iztapalapa (UIZ) y Xalostoc (XAL). Se junto toda la información en una sola base con los siguientes campos: FECHA, ESTACIÓN, PM10, PST y PM25, con un toral de 151 observaciones.

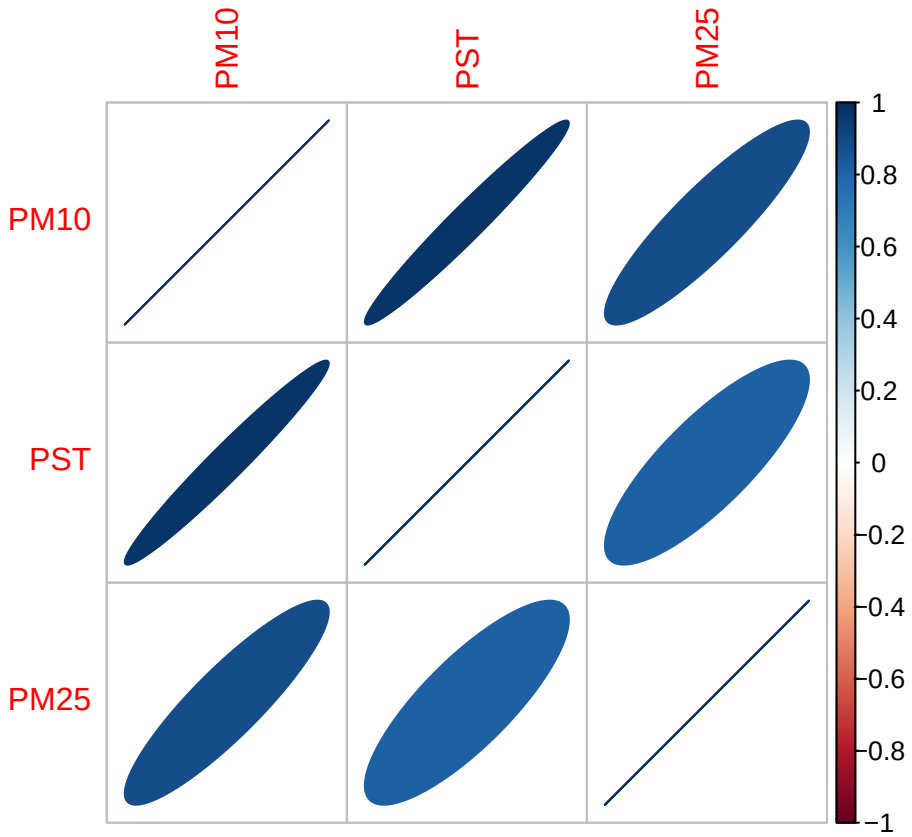
Calculamos la correlación de las variable consideradas (PM10, PST y PM25)

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## corrplot 0.84 loaded
```



Nótese que las variables PM10 y PST están altamente correlacionadas y en un menor grado PM10 y PM25, por lo que se podrían tener hasta dos factores.

Aplicamos el método de máxima verosimilitud a la matriz de correlaciones y notamos que el método solo soporta un factor.

```
##
## Call:
## factanal(factors = 1, covmat = R)
##
## Uniquenesses:
##  PM10   PST  PM25
## 0.005 0.051 0.214
##
## Loadings:
##      Factor1
## PM10 0.998
## PST  0.974
## PM25 0.887
##
##                Factor1
## SS loadings       2.73
## Proportion Var    0.91
##
## The degrees of freedom for the model is 0 and the fit was 0.315
```

Observamos que la proporción de la varianza explicada por el modelo es de 0.91. Así pues, con un factor parece ser suficiente para describir las variables. Las cargas asignadas nos sugieren que el factor asigna pesos similares a PM10 y PST con PM25 casi una decima menos a PST. Dicho factor podría considerarse como la contaminación del aire en la CDMX.

Por otro lado, al aplicar el método de componentes principales a la matriz de covarianzas obtenemos lo siguiente:

```
##   PM10    PST   PM25
## -24.27 -44.67  -9.02
```

La varianza explicada por el primer eigenvalor es de 0.9783, por lo que tomar un factor es razonable para describir a las variables. Resulta que en este caso las cargas tienen valores negativos y la variable PST tiene un mayor peso. En este caso, el factor podría considerarse como la calidad del aire, la cual es menor cuando hay una mayor concentración de particulas contaminantes en el aire.

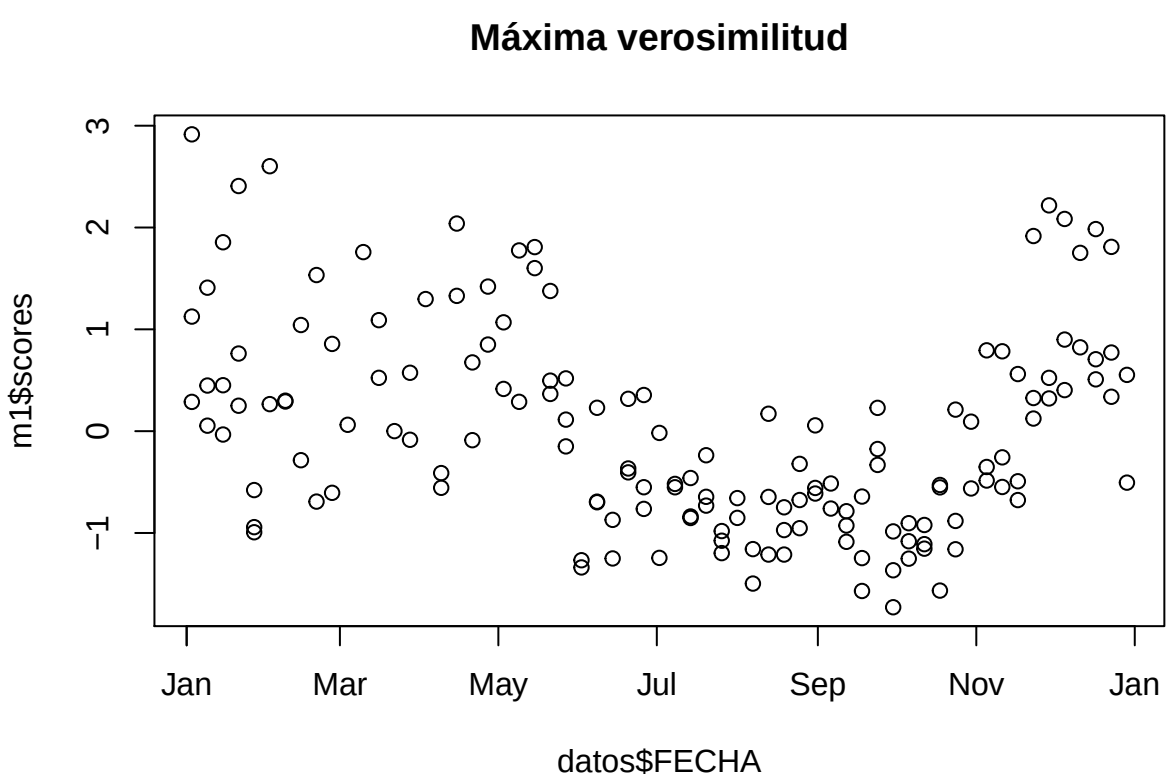
Calculamos los scores:

#Máxima Verosimilutud

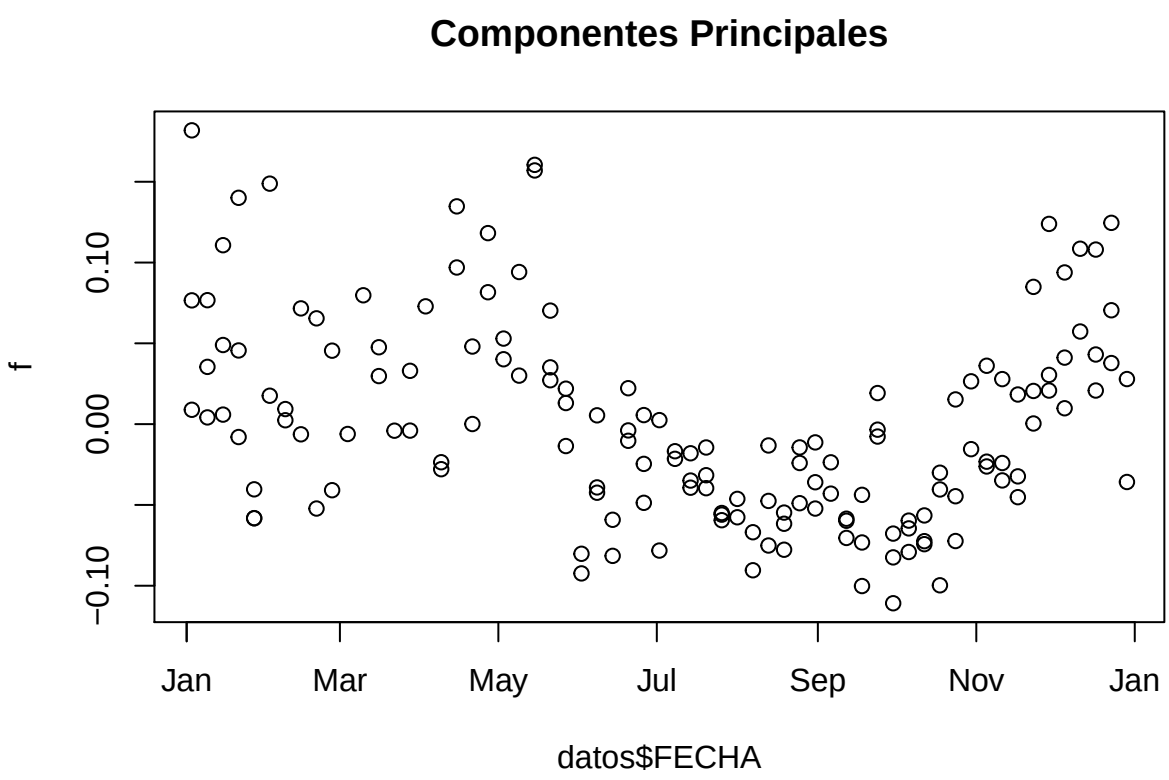
```
##          Factor1
## [1,]  1.12533091
## [2,]  0.05394559
## [3,]  0.45102629
## [4,]  0.76219484
## [5,] -0.99290075
## [6,]  0.26554906
```

#Componentes principales

```
## [1]  0.076584615  0.004169147  0.048964136  0.045595140 -0.058181560
## [6]  0.017573878
```



Observamos que de acuerdo al índice de contaminación que se creo, las estaciones monitorean una mayor contaminación en invierno con valores distintos en cada una y conforme se acerca la primavera la contaminación disminuye. Lo anterior es consistente con un fénomeno físico conocido como inversión térmica.

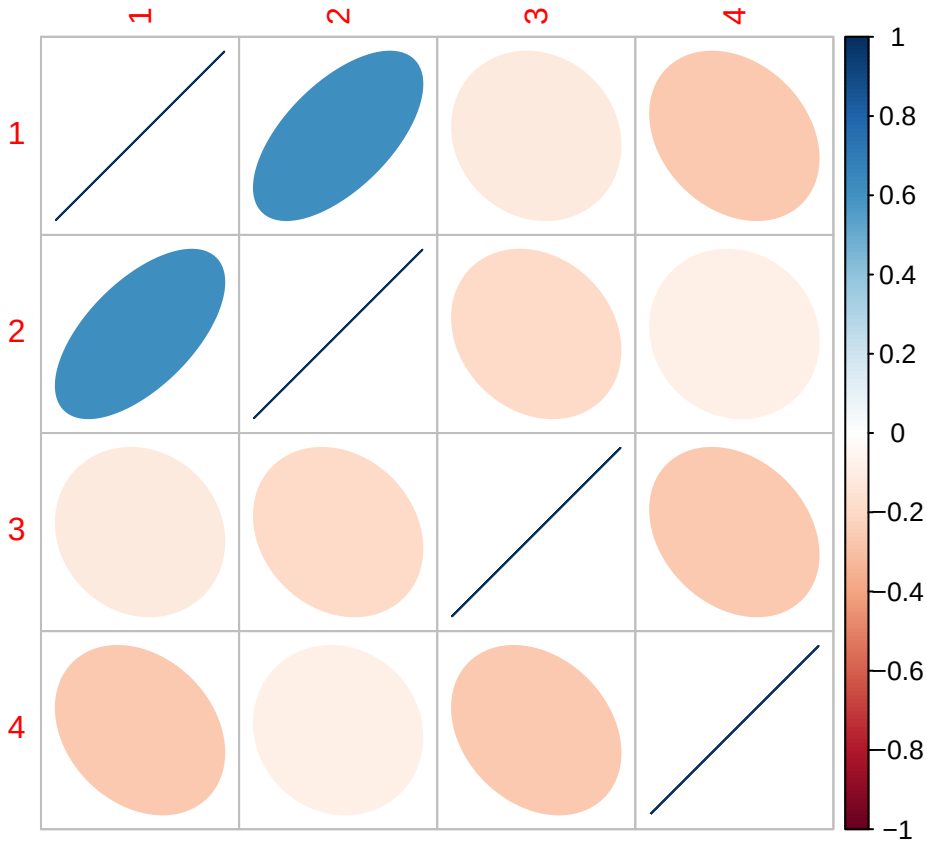


Por otro lado, para el índice obtenido con la matriz de covarianzas obtenemos un comportamiento inverso. Dado que se está considerando el factor como la calidad del aire, tiene sentido que tenga ese comportamiento.

Ejercicio 6

Contamos con la siguiente matriz de correlaciones

```
##        [,1]   [,2]   [,3]   [,4]
## [1,]  1.000  0.615 -0.111 -0.266
## [2,]  0.615  1.000 -0.195 -0.085
## [3,] -0.111 -0.195  1.000 -0.269
## [4,] -0.266 -0.085 -0.269  1.000
```



Ahora calcularemos las correlaciones canonicas construyendo la matriz A y calculando los valores propios

```
R11 <- R[1:2, 1:2]
R12 <- R[1:2, 3:4]
R21 <- R[3:4, 1:2]
R22 <- R[3:4, 3:4]

(A <- solve(R11) %*% R12 %*% solve(R22) %*% R21)
```

```
##                [,1]       [,2]
## [1,]  0.1068048090 0.04756232
## [2,] -0.0002003833 0.02914188
```

```
eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 0.10668190 0.02926479
##
## $vectors
##               [,1]       [,2]
## [1,]  0.999996661 -0.5228640
## [2,] -0.002584248  0.8524161
```

```
(B <- solve(R22) %*% R21 %*% solve(R11) %*% R12)
```

```
##            [,1]       [,2]
## [1,] 0.04504612 0.03795943
## [2,] 0.02562458 0.09090056
```

```r
eigen(B)
```

```
## eigen() decomposition
## $values
## [1] 0.10668190 0.02926479
##
## $vectors
##            [,1]       [,2]
## [1,] -0.5243952 -0.9233797
## [2,] -0.8514750  0.3838878
```

a. Así, encontramos las correlaciones canónicas muestrales que corresponden a la raíz cuadrada de los valores propios $\sqrt{0.1067} = 0.327$ y $\sqrt{0.0293} = 0.17$

b. EL primer par canónico es:

$$U_1 = 0.0450X_1^{(1)} + 0.0256X_2^{(1)}$$
$$V_1 = -0.524X_1^{(2)} - 0.851X_2^{(1)}$$

Notamos que la primer variable canónica le da más peso a los homicidios no primarios y la segunda variable le da mayor importancia a la probabilidad de castigo.Consideramos que las variables representan índices sobre la composición de los homicidios así como de las características de la acción punitiva. Además, la segunda variable canónica tiene signos negativos pues el segundo grupo de variables se correlacionan de forma negativa con el primer grupo. La correlación canónica es relativamente pequeña.

Ejercicio 7

Primero cargaremos los datos

```r
data <- read.csv("./bank-full.csv", header=T, sep=";")
```

Y obtenemos un primer resumen de los datos

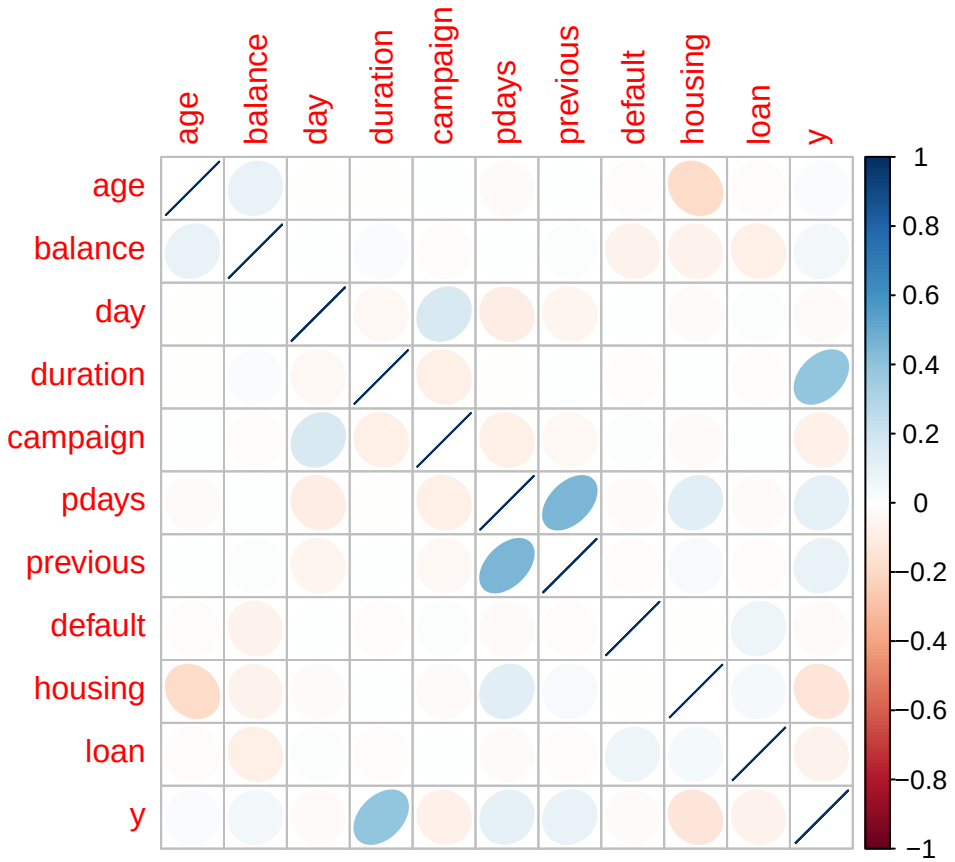
```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
##  $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Nuestra primera idea fue tranformar las variables de tipo factor a indicadores y codificar las variables binarias (con valores "yes" o "no") con ceros y unos. Despues de probar con cada una de las variables categóricas (job, marital, education, contact, poutcom) nos encontramos con el problema de que al intentar calcular

las correlaciones canónicas , teníamos matrices numericamente singulares. Inclusive invocando la funcion stats::concor, la ejecución nos regresaba una dirección canónica con coeficientes faltantes.

Para poder avanzar tuvimos que dejar de lado las variables categoricas y realizar el analisis con las variables restantes.Así, trabajamos con los siguientes datos transformados.

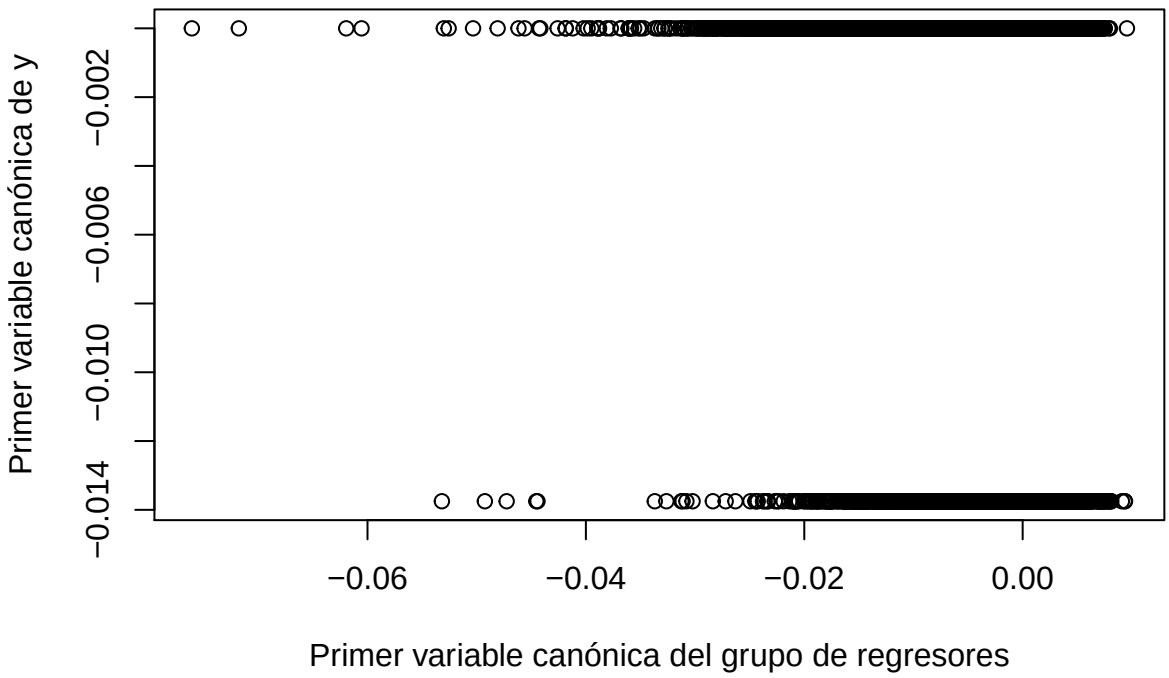
```
## 'data.frame':    45211 obs. of  11 variables:
##  $ age     : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ balance : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ day     : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ duration: int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays   : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ default : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ housing : num  1 1 1 1 0 1 1 1 1 1 ...
##  $ loan    : num  0 0 1 0 0 0 1 0 0 0 ...
##  $ y       : num  0 0 0 0 0 0 0 0 0 0 ...
```



Ahora que tenemos una representación más amigable de los datos, consideramos que para atacar el problema de clasificación procederemos a realizar el análisis de correlación canonica tomando a "y" como uno de los grupos y las demas variables en el otro grupo. Esto se realiza con la intención de utilizar las variables canónicas como sustitutas que permitan representar la relación entre ambos grupos de variables de forma simplificada.

```
## $cor
## [1] 0.4178867
##
```

```
## $xcoef
##                    [,1]          [,2]          [,3]          [,4]
## age       2.630110e-06 -9.095025e-05  3.088164e-05 -3.650389e-04
## balance  -9.592031e-08  1.541749e-06 -2.616722e-09 -1.148303e-07
## day       4.536952e-06  4.887947e-07  5.648640e-04  2.096993e-05
## duration -1.606658e-05 -1.730954e-06  9.709221e-07  4.839347e-06
## campaign  1.089705e-04  1.174008e-05 -6.585215e-06  9.113006e-05
## pdays    -9.950745e-06 -1.072057e-06  6.013353e-07 -8.321628e-06
## previous -2.439593e-04 -2.628328e-05  1.474275e-05 -2.040187e-04
## default   7.405003e-04  7.977880e-05 -4.474931e-05  6.192670e-04
## housing   3.239948e-03  3.490602e-04 -1.957939e-04  2.709510e-03
## loan      1.468681e-03  1.582303e-04 -8.875410e-05  1.228231e-03
##                    [,5]          [,6]          [,7]          [,8]
## age       3.174004e-05 -8.367042e-05 -5.223470e-05 -1.623397e-05
## balance   3.008394e-08 -2.893814e-08 -3.490523e-08 -1.031576e-07
## day      -9.243516e-05  4.723459e-05  7.367421e-06  2.788807e-06
## duration  2.751775e-06 -3.259034e-06 -1.758131e-06 -5.331369e-07
## campaign  1.535108e-03  1.398898e-04 -4.671154e-06  2.099044e-05
## pdays     7.911731e-07  4.518470e-05 -2.505959e-05 -1.492407e-06
## previous  1.939694e-05 -5.247210e-05  2.267336e-03 -1.527486e-05
## default  -5.887638e-05  1.592709e-04  7.900994e-05 -3.543259e-02
## housing  -2.576048e-04  6.968658e-04  3.456961e-04  7.962109e-05
## loan     -1.167732e-04  3.158919e-04  1.567054e-04  3.609254e-05
##                    [,9]         [,10]
## age      -2.215157e-04  4.741932e-05
## balance  -1.135426e-07  1.294338e-07
## day      -8.329595e-06 -5.590071e-06
## duration -5.056417e-06  1.907111e-06
## campaign  1.968683e-05 -1.811781e-05
## pdays     2.945944e-06  2.256756e-06
## previous -1.218679e-04  2.199496e-05
## default  -8.652628e-05 -2.621963e-03
## housing  -8.687498e-03 -7.292321e-04
## loan      4.693876e-04  1.275669e-02
##
## $ycoef
##             [,1]
## [1,] -0.01375033
##
## $xcenter
##          age      balance          day     duration     campaign
## 4.093621e+01 1.362272e+03 1.580642e+01 2.581631e+02 2.763841e+00
##        pdays     previous      default      housing         loan
## 4.019783e+01 5.803234e-01 1.802659e-02 5.558382e-01 1.602265e-01
##
## $ycenter
## [1] 0
```



Las variables canónicas pueden ser utilizadas en un modélo posterior que defina un nivel de corte para la clasificación dentro de las personas que compran o no el servicio

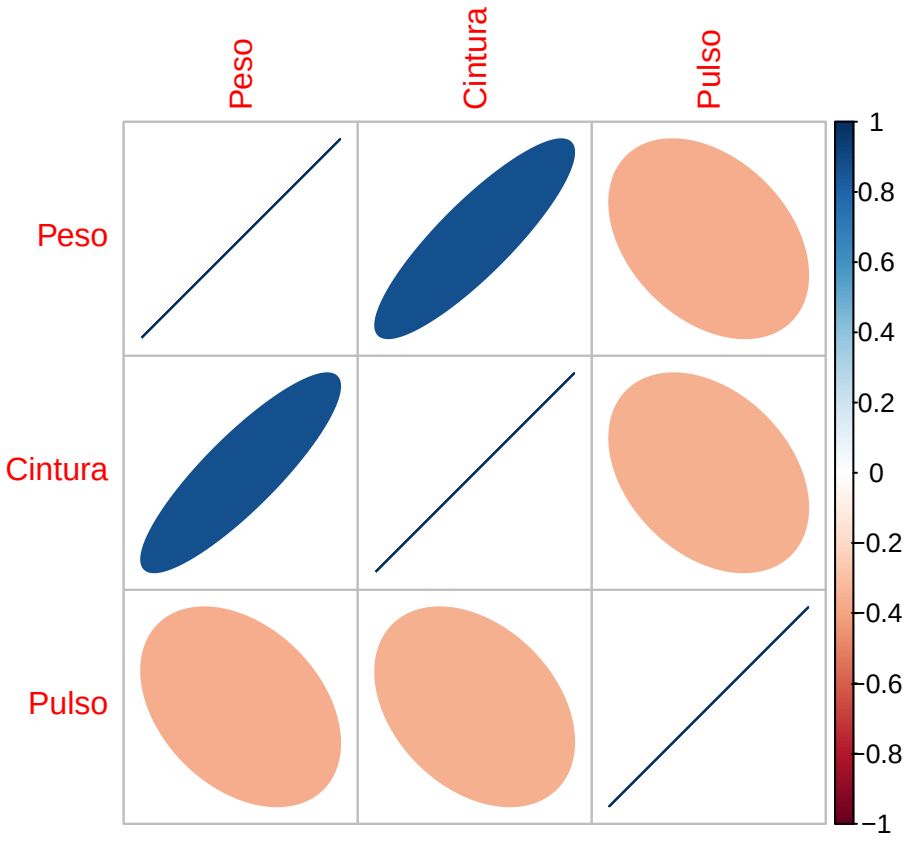
Ejercicio 8

Se tienen tres medidas fisiológicas y tres variables de ejercicios medidas en 20 hombres de 30-40 años en un gimnasio. Los datos están en el archivo FitnessClubdata.dat . Objetivo: determinar si las variables fisiológicas se relacionan de alguna forma con las variables de ejercicio.

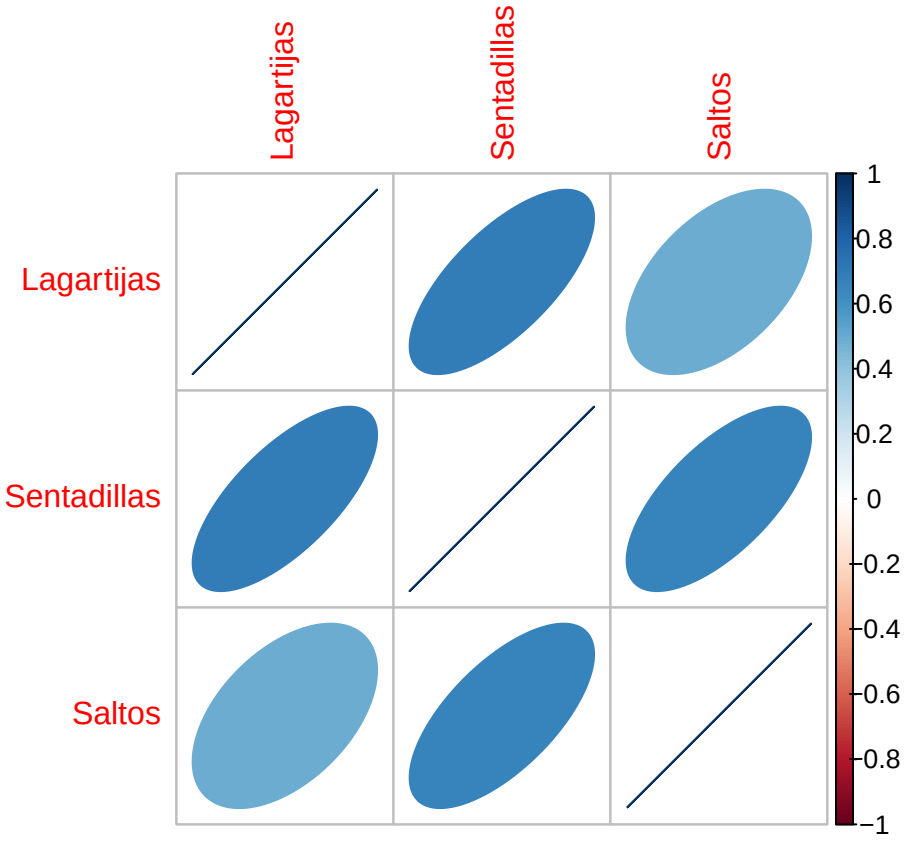
a. Analizar la matriz de correlaciones relevantes entre las variables de los dos grupos (dentro y entre grupos de variables).

```
library(corrplot)
datos <- read.table("FitnessClubData.dat", header = TRUE)

#Calculamos la matriz correlacion por grupo de variables
#para variables fisiologicas
R1<-cor(datos[,1:3])
#Grafica de correlacion
corrplot(R1, method = "ellipse")
```




```r
#para variables de ejercicio
R2<-cor(datos[,4:6])
#Grafica de correlacion
corrplot(R2, method = "ellipse")
```

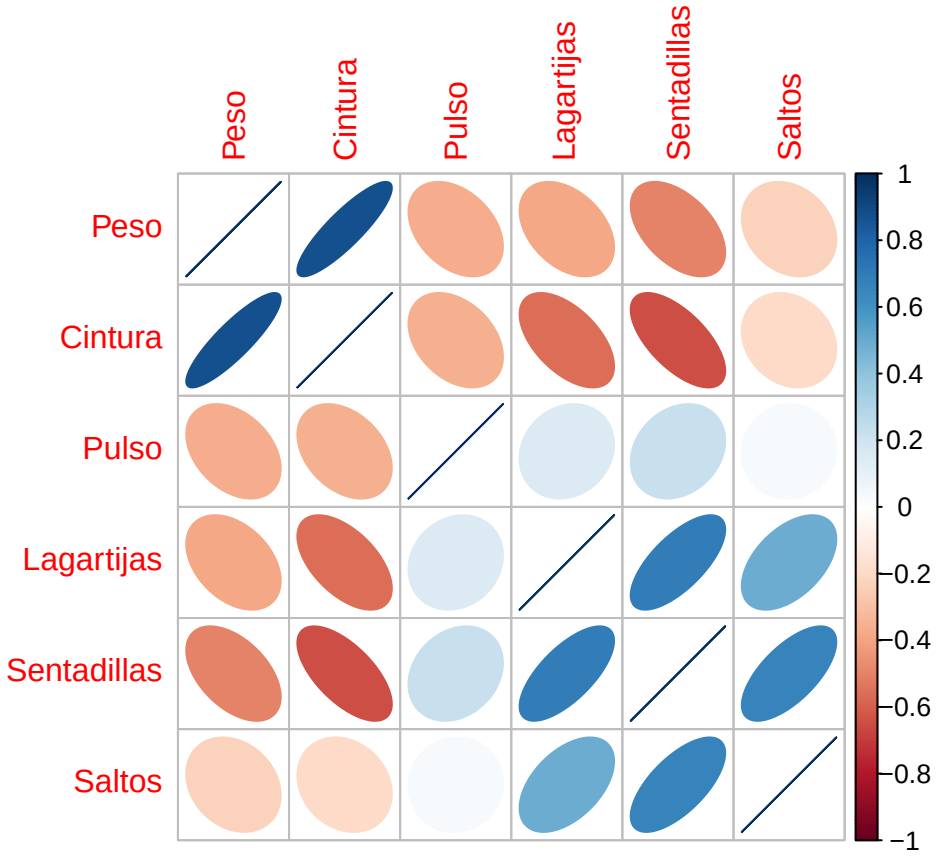


Para las variables fisiológicas, vemos que hay una alta correlación positiva entre las medidas de cintura y peso, sin embargo el pulso tiene una correlación negativa pero no muy marcada con respecto a las otras variables.

Se podría decir que una persona de mayor peso tendrá una mayor medida de cintura, y en general un pulso menor a comparación con una persona de menor peso, pero esta diferencia en el pulso probablemente no será muy marcada.

En las variables de ejercicio vemos que las correlciones son altas y positivas. En general una persona con buena capacidad de hacer lagartijas tiene igual una buena capacidad de llevar a cabo sentadillas o saltos. Se podrían tomar la relación entre las 3 variables como un un indicador de buena capacida física.

```
#Calculamos la matriz correlacion entre todas las variables
R<-cor(datos)
#Grafica de dispersion
#plot(cont)
#Grafica de correlacion
corrplot(R, method = "ellipse")
```



Al analizar la correlacion incluyendo variables tanto fisiologicas como de ejercicio vemos que las variables de ejercicicio tienen una correlación negativa con la varible de peso, se podría tomar como un indicador que de que a mayor peso se suele tener un menor rendimiento físico. La cintura vimos se correlaciona de forma positiva con el peso por lo que esta variables igual tiene correlaciones negativas con las variables de ejercicio. La variable pulso tiene una correlación negativa con el resto de las variables fisiológicas por lo que tiene una correlación positiva, con las variables de ejercicio aunque no es una correlación fuerte.