Clase 36: Técnicas de Reducción de Dimensionalidad en R

Herramientas Estadísticas y Forecast (HEF)

María José García

Análisis de Componentes Principales

Ejemplo en R

En medicina es muy importante poder detectar predisposición a desarrollar cualquier tipo de cáncer. La secuenciación genómica es una prueba diagnóstica de precisión que sirve para identificar las principales alteraciones en los genes de las células cancerosas. Las pruebas de secuenciación genómica pueden ayudar a identificar las alteraciones genómicas específicas en el cáncer de cada persona y, potencialmente, a seleccionar las opciones de tratamiento dirigido o los ensayos clínicos disponibles.

Para este ejemplo, se dispone de información sobre datos de microarray de líneas celulares de cáncer. Específicamente se cuenta con medidas sobre expresión de 6830 genes (variables) sobre 64 líneas celulares cancerígenas (observaciones). Mediante el análisis de componentes principales se busca encontrar patrones o agrupaciones mediante la representación de las dos primeras componentes principales.

Lectura de datos

```
load("datos_Clase36.RData")
dim(datos)
```

```
## [1]   64 6831
```

```
### Nombre de las columnas
head(names(datos))
```

```
## [1] "type" "G1"   "G2"   "G3"   "G4"   "G5"
```

```
### Tipos de cáncer distintos en el conjunto de datos
unique(datos$type)
```

```
##  [1] "CNS"          "RENAL"        "BREAST"       "NSCLC"        "UNKNOWN"
##  [6] "OVARIAN"      "MELANOMA"     "PROSTATE"     "LEUKEMIA"     "K562B-repro"
## [11] "K562A-repro" "COLON"        "MCF7A-repro" "MCF7D-repro"
```

```r
# Número de muestras por tipo de cáncer
table(datos$type)
```

```
##
##      BREAST         CNS       COLON K562A-repro K562B-repro    LEUKEMIA
##           7           5           7           1           1           6
## MCF7A-repro MCF7D-repro    MELANOMA       NSCLC     OVARIAN    PROSTATE
##           1           1           8           9           6           2
##       RENAL     UNKNOWN
##           9           1
```

```r
# Media de la expresión de cada gen (muestra de los 6 primeros).
# (MARGIN = 2 para que se aplique la función a las columnas)
apply(X = datos[,-1], MARGIN = 2, FUN = mean)[1:6]
```

```
##           G1           G2           G3           G4           G5           G6
## -0.019063414 -0.027813101 -0.019922789 -0.328672789  0.026092836  0.006717837
```

```r
# Varianza de la expresión de cada gen (muestra de los 6 primeros)
apply(X = datos[,-1], MARGIN = 2, FUN = var)[1:6]
```

```
##        G1        G2        G3        G4        G5        G6
## 0.1947740 0.5737041 0.1877537 1.1922566 0.2352962 0.1228028
```

Análisis de Componentes Principales

Recordar que el ACP sólo puede aplicarse a datos numéricos.

```
pca_datos <- prcomp(datos[,-1], scale = TRUE)

names(pca_datos)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
# Cantidad de componentes principales
dim(pca_datos$rotation)
```

```
## [1] 6830   64
```

Hay un total de 64 componentes principales distintas, ya que en general pueden haber $min(n-1, p)$ componentes en un set de datos $n \times p$.

```r
# Muestra de los primeros 6 elementos del vector de cargas de los 5 primero
head(pca_datos$rotation)[, 1:5]
```

```
##              PC1          PC2         PC3          PC4          PC5
## G1 -0.010682370  0.001324406 0.008503514 -0.003524094 -0.010126893
## G2 -0.002312078  0.001675266 0.010256593  0.002603645 -0.011400802
## G3 -0.005879750 -0.006289434 0.010055415 -0.010681458  0.010264980
## G4  0.003278071  0.002666138 0.008361513 -0.007475761  0.011248268
## G5 -0.007677535 -0.002508097 0.013820836  0.009509144  0.004094756
## G6  0.002266671 -0.009677933 0.010818283 -0.012751147 -0.007196820
```

```r
# Vectores de puntajes
head(pca_datos$x)[,1:5]
```

```
##          PC1       PC2          PC3          PC4        PC5
## P1 -19.68245  3.527748  -9.7354382   0.8177816 -12.511081
## P2 -22.90812  6.390938 -13.3725378  -5.5911088  -7.972471
## P3 -27.24077  2.445809  -3.5053437   1.3311502 -12.466296
## P4 -42.48098 -9.691742  -0.8830921  -3.4180227 -41.938370
## P5 -54.98387 -5.158121 -20.9291076 -15.7253986 -10.361364
## P6 -26.96488  6.727122 -21.6422924 -13.7323153   7.934827
```

```r
# Desviación estándar de cada componente principal (6 primeros)
pca_datos$sdev[1:6]
```

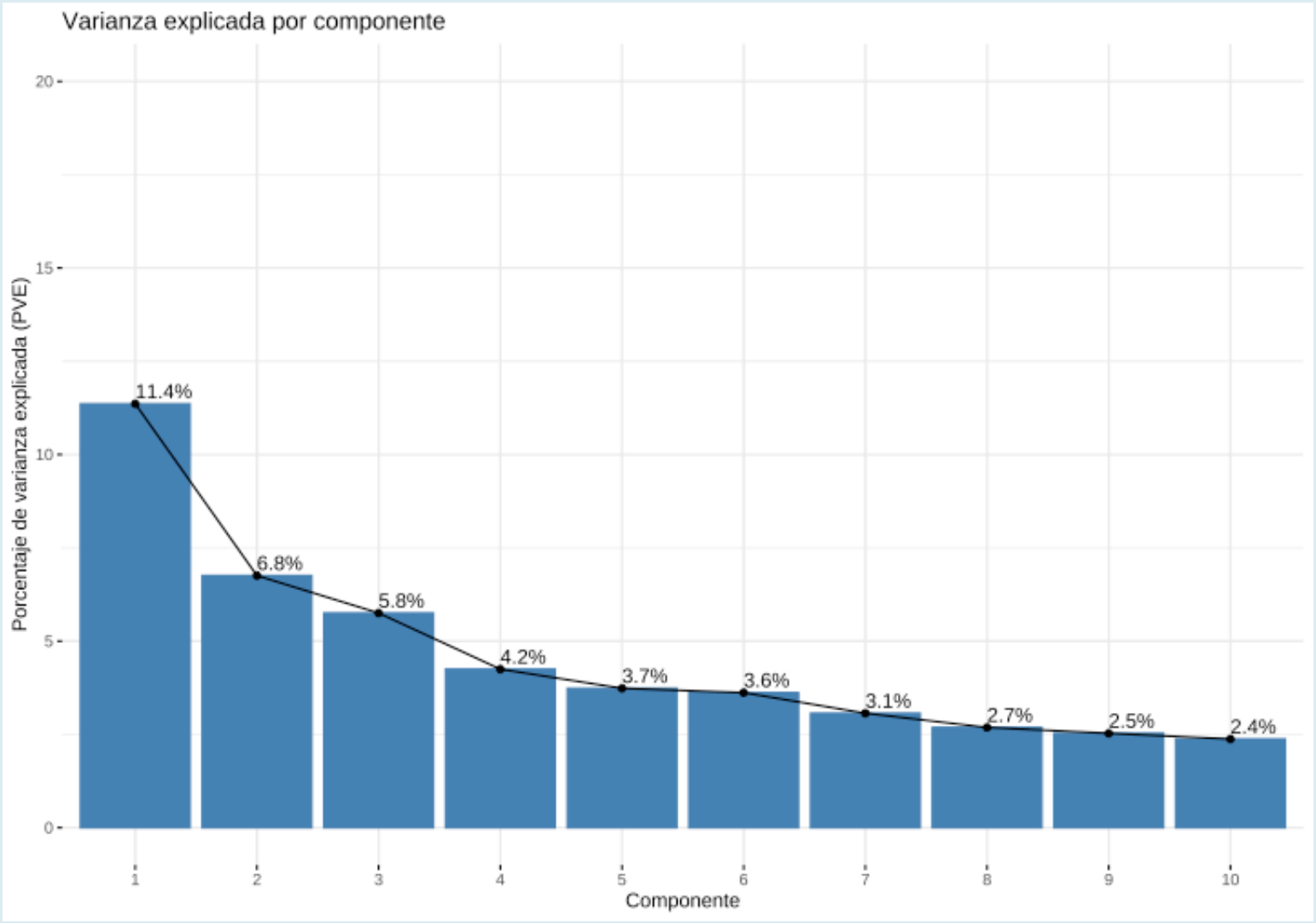
```
## [1] 27.85347 21.48136 19.82046 17.03256 15.97181 15.72108
```

```r
# Varianza explicada por cada componente principal (6 primeros)
(pca_datos$sdev^2)[1:6]
```

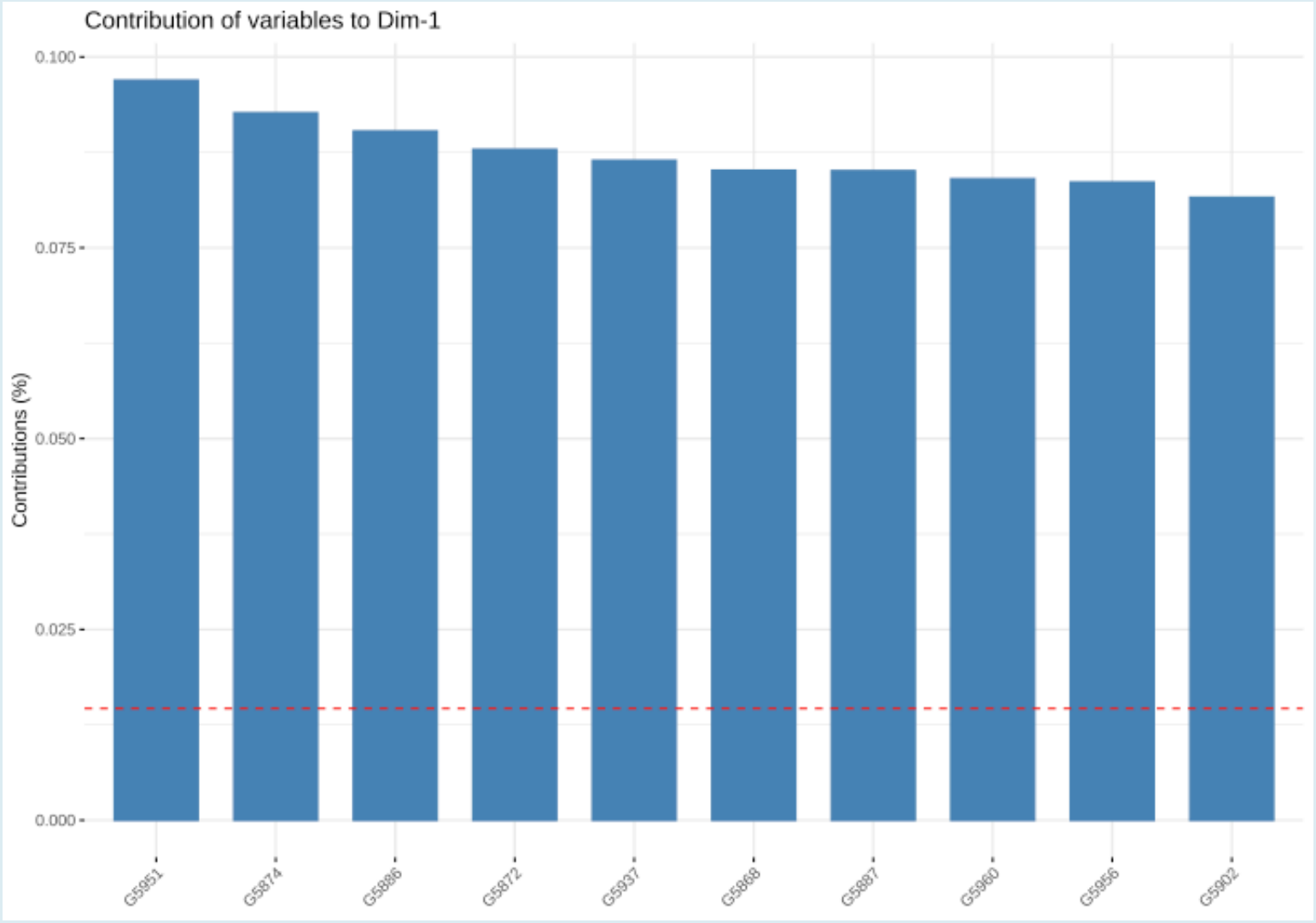
```
## [1] 775.8157 461.4486 392.8508 290.1080 255.0986 247.1524
```

```r
# Resumen del ACP
summary(pca_datos)
```

```r
library(factoextra)
fviz_screeplot(pca_datos, addlabels = TRUE, ylim = c(0, 20),
               main = "Varianza explicada por componente",
               xlab = "Componente", ylab = "Porcentaje de varianza explicad
```



```r
# Top 10 variables que más contribuyen a PC1
fviz_contrib(pca_datos, choice = "var", axes = 1, top = 10)
```

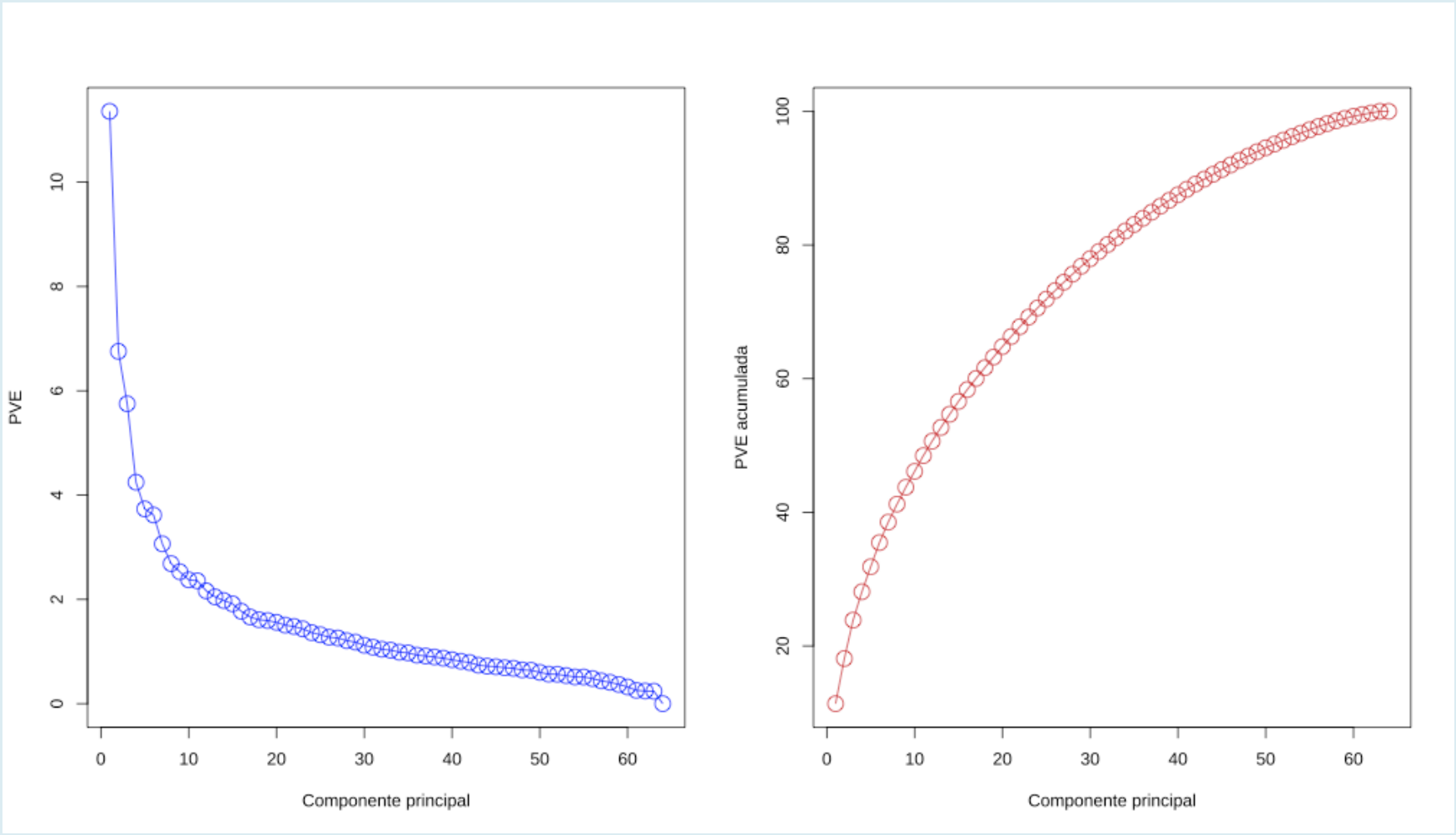


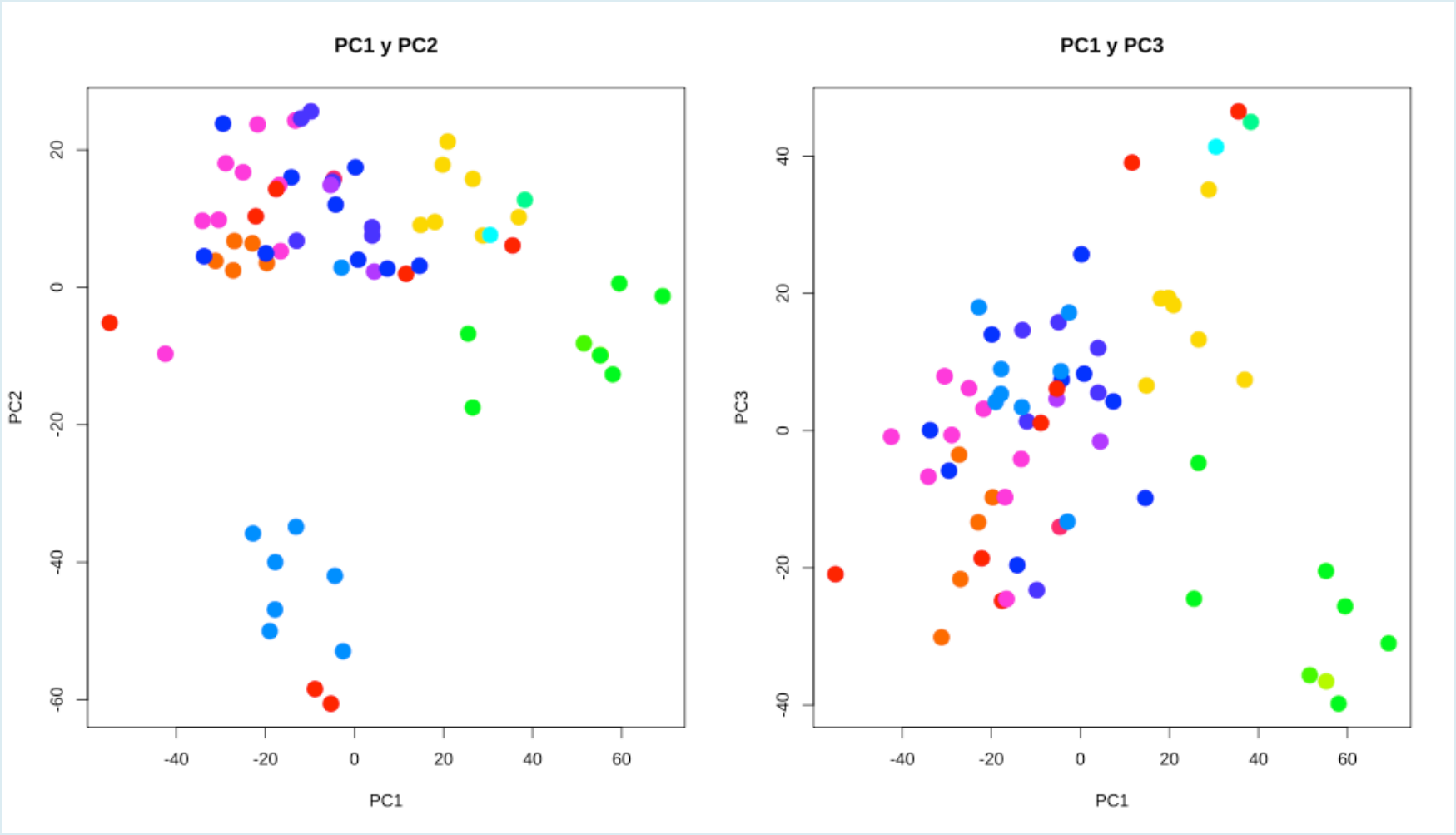
```r
# % varianza explicada
PVE <- 100*pca_datos$sdev^2/sum(pca_datos$sdev^2)
PVE[1:10]
```

```
##  [1] 11.358942  6.756203  5.751842  4.247554  3.734972  3.618630  3.066222
##  [8]  2.685903  2.529498  2.375869
```

```r
# % varianza acumulada
cumsum(PVE[1:10])
```

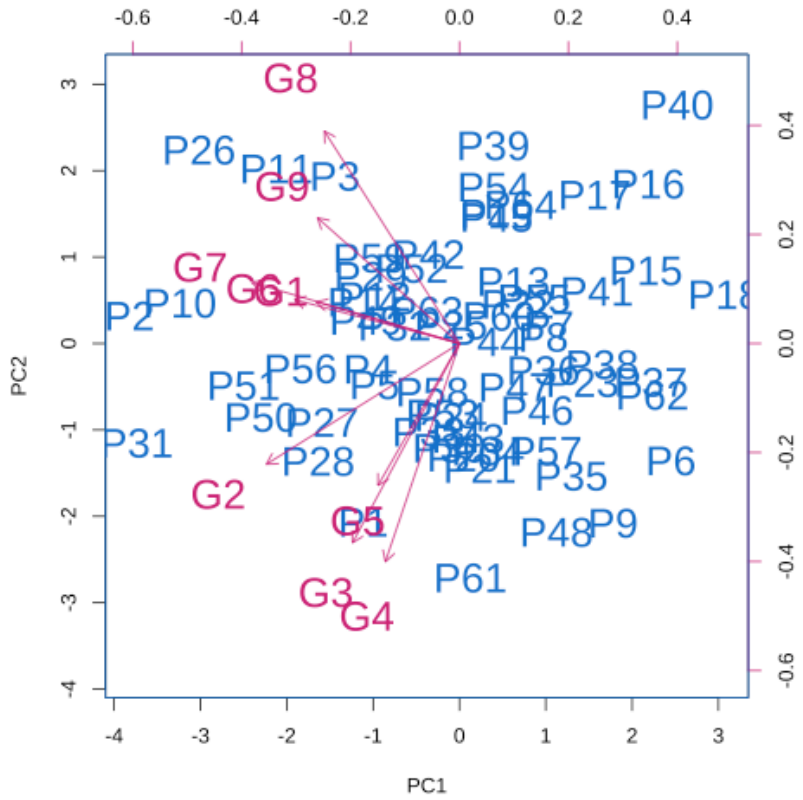
```
##  [1] 11.35894 18.11514 23.86699 28.11454 31.84951 35.46814 38.53437 41.22027
##  [9] 43.74977 46.12564
```


PVE
0
2
4
6
8
10
0
10
20
30
40
50
60
Componente principal
PVE acumulada
20
40
60
80
100
0
10
20
30
40
50
60
Componente principal

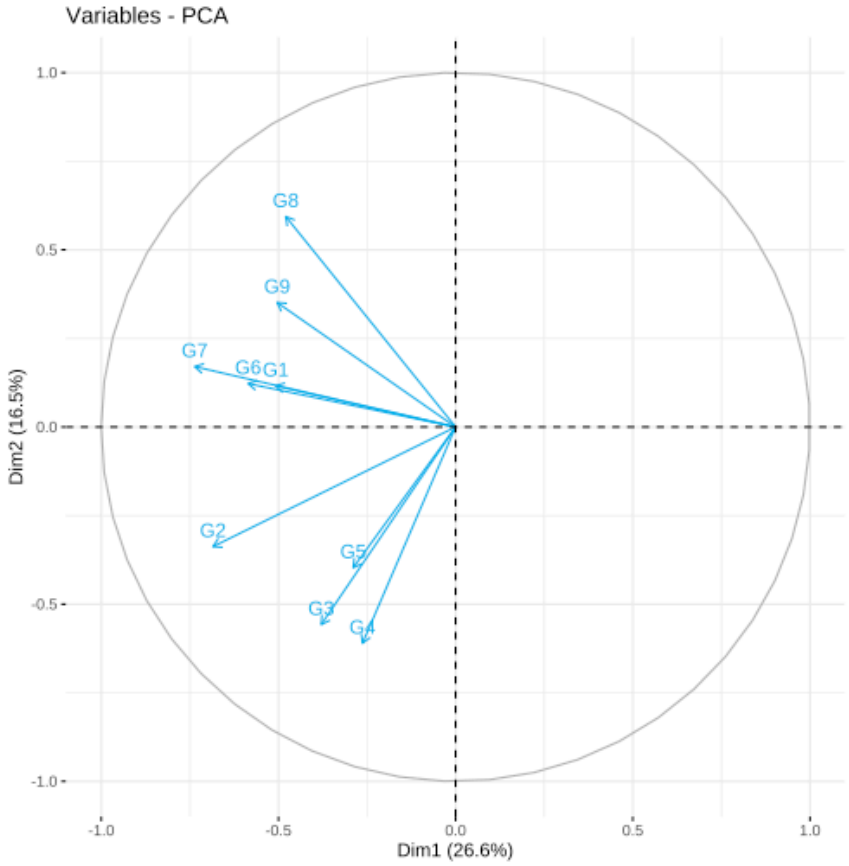

PC1 y PC2
PC2
PC1
20
0
-20
-40
-60
-40
-20
0
20
40
60
PC1 y PC3
PC3
PC1
40
20
0
-20
-40
-40
-20
0
20
40
60

Considere sólo los primeros 9 genes

```
PCA <- prcomp(datos[, 2:10], scale = TRUE)
biplot(PCA, scale = 0, cex = 2, col = c("dodgerblue3", "deeppink3"))
```

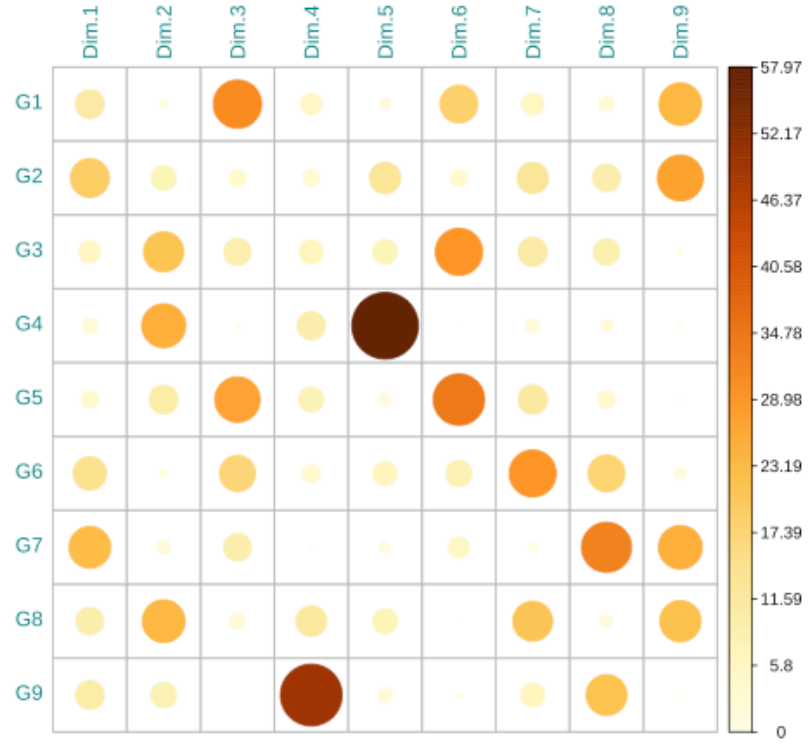


```
fviz_pca_var(PCA, col.var = "deepskyblue2", cex = 3)
```



La contribución de cada variable en cada componenete se obtiene de la siguiente manera:

```
Comp <- get_pca_var(PCA)
library(corrplot)
corrplot(Comp$contrib, is.corr = FALSE, tl.col = "darkcyan")
```



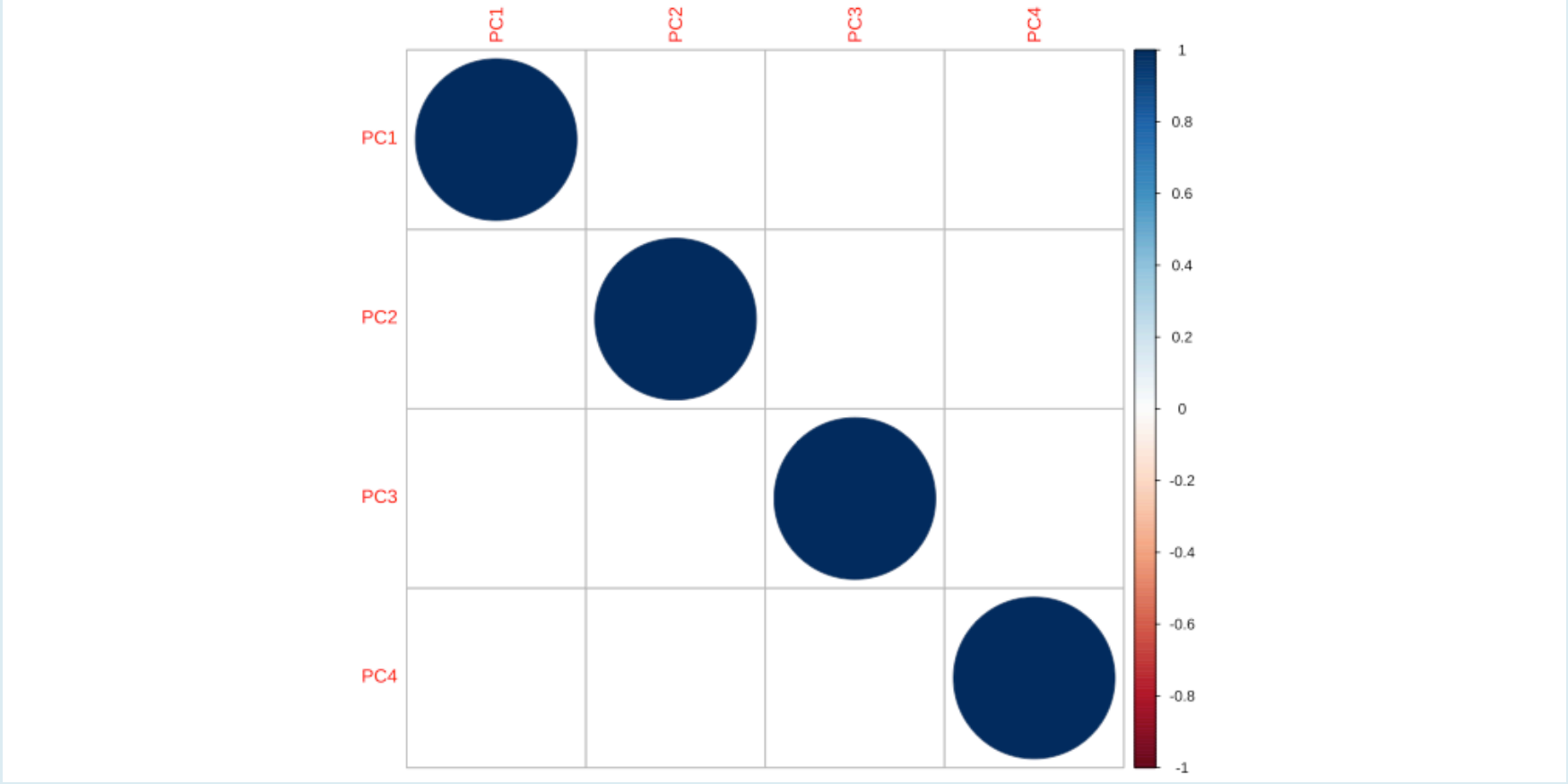
De vuelta a los datos completos, con 6.830 genes, los vectores proyectados ortogonales a los ejes originales se obtienen de la siguiente manera:

```
# Vectores proyectados ortogonales
head(scale(datos[,-1])%*%pca_datos$rotation[,1:4])
```

```
##           PC1       PC2         PC3         PC4
## P1 -19.68245  3.527748  -9.7354382   0.8177816
## P2 -22.90812  6.390938 -13.3725378  -5.5911088
## P3 -27.24077  2.445809  -3.5053437   1.3311502
## P4 -42.48098 -9.691742  -0.8830921  -3.4180227
## P5 -54.98387 -5.158121 -20.9291076 -15.7253986
## P6 -26.96488  6.727122 -21.6422924 -13.7323153
```

Es posible calcular la correlación entre ellos, lo que arroja que no tienen correlación.

```
corrplot(cor(scale(datos[,-1])%*%pca_datos$rotation[,1:4]))
```




```
# Son no correlacionados
```

Análisis Factorial

Ejemplo en R

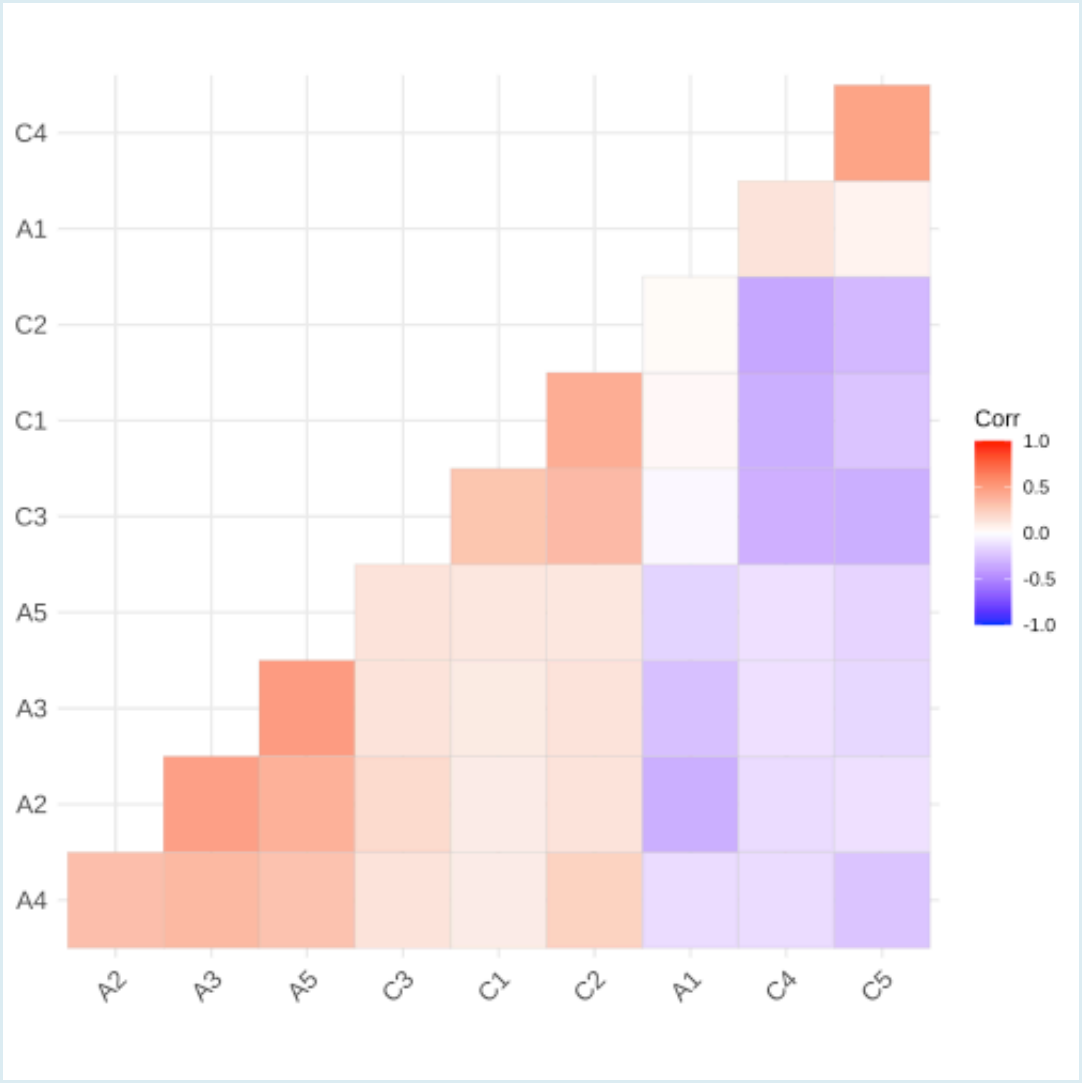
Se tiene información de sujetos que respondieron un informe de personalidad en base a ítems provenientes del Banco Internacional de Ítems de Personalidad (ipip.ori.org).

En particular, el conjunto de datos que se utilizará en este ejemplo contiene 200 observaciones con 10 respuestas a ítems de una prueba de personalidad.

```
library(polycor)
```

```
## Warning: package 'polycor' was built under R version 4.1.2
```

```
library(ggcorrplot)
# Matriz de correlación policórica
mat_cor <- hetcor(datos_af)$correlations
```


C4
A1
C2
C1
C3
A5
A3
A2
A4
A2
A3
A5
C3
C1
C2
A1
C4
C5
Corr
1.0
0.5
0.0
-0.5
-1.0

```
##
## Call:
## factanal(x = scale(datos_af), factors = 2)
##
## Uniquenesses:
##    A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
## 0.857 0.568 0.427 0.744 0.608 0.702 0.615 0.699 0.570 0.652
##
## Loadings:
##    Factor1 Factor2
## A1         -0.379
## A2  0.126   0.645
## A3          0.751
## A4  0.218   0.457
## A5  0.124   0.614
## C1  0.544
## C2  0.614
## C3  0.536   0.117
## C4 -0.649
## C5 -0.574  -0.137
##
##                Factor1 Factor2
## SS loadings       1.80   1.759
## Proportion Var    0.18   0.176
## Cumulative Var    0.18   0.356
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 455.22 on 26 degrees of freedom.
## The p-value is 6.01e-80
```

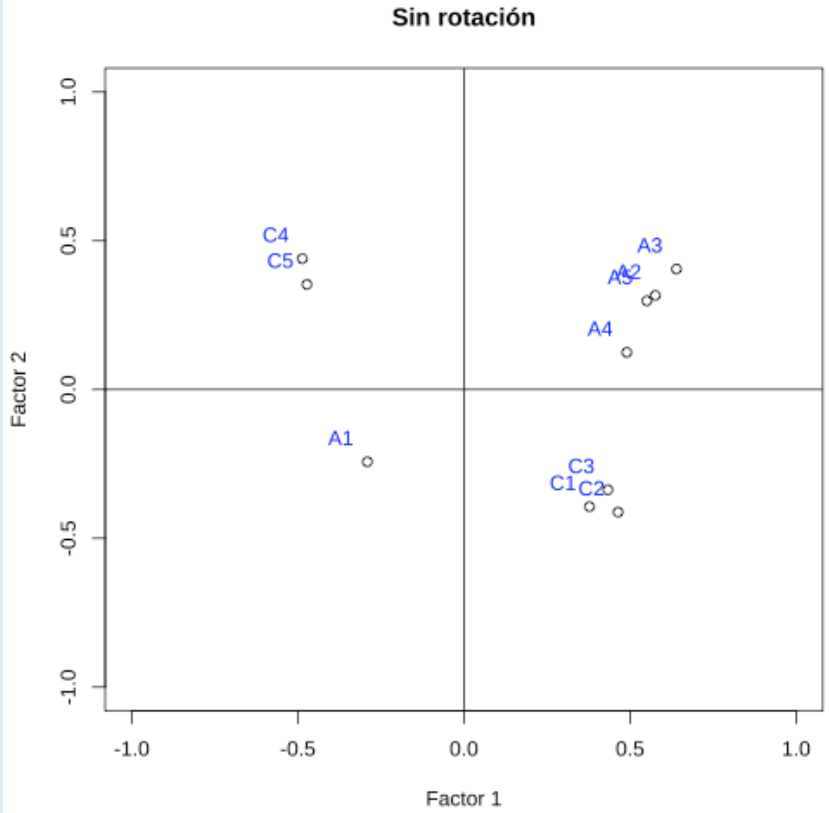
```r
# Comunalidad
apply(AF$loadings^2, 1 , sum)
```

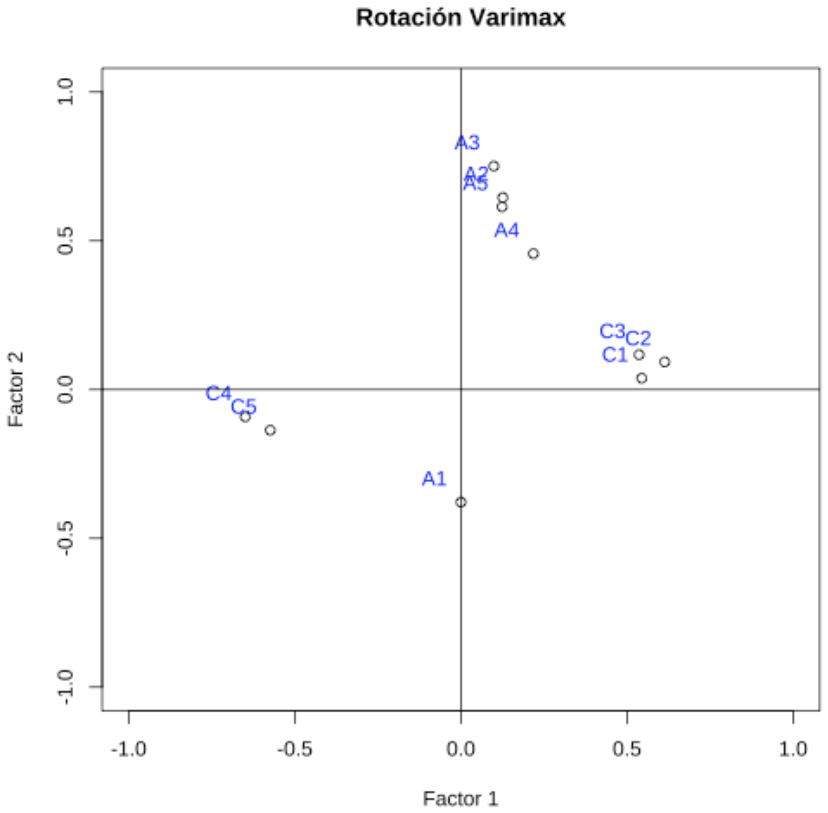
```
##        A1        A2        A3        A4        A5        C1        C2        C3
## 0.1433073 0.4321431 0.5734580 0.2560630 0.3922204 0.2977793 0.3851442 0.3014407
##        C4        C5
## 0.4300033 0.3480782
```

```r
# Unicidad
AF$uniquenesses
```

```
##        A1        A2        A3        A4        A5        C1        C2        C3
## 0.8566938 0.5678578 0.4265425 0.7439373 0.6077782 0.7022203 0.6148562 0.6985596
##        C4        C5
## 0.5699978 0.6519214
```

```r
# Vectores rotados
FA <- factanal(scale(datos_af), factors = 2, rotation = "none")
FA.varimax <- factanal(scale(datos_af), factors = 2, rotation = "varimax")
```





Trabajo en grupos

Los datos TIC2021 tienen información recabada el año 2021 en los 27 países de la Unión Europea sobre 7 variables: 4 relacionadas con el uso de las TIC por parte de las empresas y 3 relativas al uso de dichas tecnologías por parte de las personas y a la equipación TIC de los hogares. Las variables se describen a continuación:

- **ebroad:** Porcentaje de empresas con acceso a banda ancha (fija o móvil) considerando todas las empresas sin sector financiero (10 o más empleados y trabajadores por cuenta propia).
- **esales:** Porcentaje de empresas con ventas de comercio electrónico de al menos 1% de facturación considerando todas las empresas sin sector financiero (10 o más empleados y trabajadores por cuenta propia).
- **esocmedia:** Porcentaje de empresas que usan alguna red social a partir de 2014 considerando todas las empresas sin sector financiero (10 o más empleados y trabajadores por cuenta propia).
- **eweb:** Porcentaje de empresas que tienen un sitio web o página web propia considerando todas las empresas sin sector financiero (10 o más empleados y trabajadores por cuenta propia).
- **hbroad:** Porcentaje de hogares con conexión de banda ancha.
- **hiacc:** Porcentaje de hogares con acceso a Internet.

Ejercicio 1

Suponga que para modelar la información contenida en los datos puede considerar, a lo más, 3 variables. Realice un análisis de componentes principales de modo de representar la información de las 7 variables en sólo 3.

Ejercicio 2

Con los mismos datos, realice un análisis factorial para verificar la existencia de agrupaciones de variables. Comente sus hallazgos en relación a la consecuencia de los grupos encontrados.

¡Gracias!