ANÁLISIS DE COMPONENTES PRINCIPALES EN R

Uno de los problemas cuando se tienen una gran cantidad de datos multivariados es que generalmente hay demasiadas variables para hacer que la aplicación de técnicas tenga éxito en proporcionar una evaluación informativa inicial de los datos. Resulta evidente que los diagramas de dispersión, las matrices de diagrama de dispersión y otros gráficos son probablemente más útiles cuando el número de variables en los datos (*dimensionalidad de los datos)*, es relativamente pequeño. El tener demasiadas variables también puede causar problemas para otras técnicas multivariantes que el investigador puede querer aplicar a los datos.

Esta situación nos lleva al *Análisis de Componentes Principales* (ACP o *Principal Components Analysis*), una técnica multivariante que permite conocer la estructura que guardan un conjunto de datos multivariantes para posteriormente reducir la dimensionalidad del mismo, procurando retener la mayor cantidad de la variación original posible presente. Este objetivo se logra transformando el conjunto de datos originales a un nuevo conjunto más reducido. Los *componentes principales* son combinaciones lineales de las variables originales, que no están correlacionadas y están ordenadas de manera que las primeras representan la mayor parte de la variación en todas las variables originales. En el mejor de los casos posibles, el resultado de un ACP sería la creación de un pequeño número de nuevas variables que se pueden utilizar como sustituto para el gran número original de variables y, en consecuencia, proporcionar una base más simple para, por ejemplo, la representación gráfica o resumir los datos, y también quizá cuando se desea utilizar otras técnicas de análisis multivariados con los datos.

La expectativa general del ACP es que los primeros componentes representarán una proporción sustancial de la variación en las variables originales, y en consecuencia, se puede utilizar para generar un resumen con una baja de dimensionalidad que podría resultar útil más adelante en la investigación. Por ejemplo, se cuenta con un conjunto de datos consistente en puntajes de diferentes evaluaciones académicas de los estudiantes de una universidad. Una situación de interés podría ser construir un índice informativo del rendimiento global del estudiante. Una posibilidad sencilla sería la puntuación media de cada estudiante, aunque si el rango de calificaciones observadas varía de curso a curso, podría ser más razonable ponderar las calificaciones de alguna manera antes de calcular el promedio o, alternativamente, estandarizar los resultados de los exámenes antes de intentar combinarlos. El ACP podría aplicarse como parte un mecanismo para establecer, a partir de los resultados de las evaluaciones observadas y empleando el primer componente principal, una medida del rendimiento que discrimine al máximo a los estudiantes.

Otra aplicación puede ser en el campo de la economía, donde los datos complejos son a menudo resumidos por algún tipo de índice; por ejemplo, índices de precios, de salarios, de costo de vida, etc. Al evaluar los cambios en los precios a lo largo del tiempo, el economista deseará tener en cuenta el

hecho de que los precios de algunos productos básicos son más variables que otros, o que los precios de algunos de los productos se consideran más importantes que otros. En cada caso el índice tendrá que ser ponderado en consecuencia. En tales ejemplos, el primer componente principal a menudo puede satisfacer las necesidades del investigador.

No siempre es el primer componente principal el de mayor interés para un investigador. Un taxonomista, por ejemplo, al investigar la variación en las mediciones morfológicas en animales para los cuales todas las correlaciones entre parejas es probable que sean positivas, a menudo se ocupará más del segundo y subsiguientes componentes, ya que pueden proporcionar una descripción conveniente de aspectos de la forma de un animal. A menudo será de mayor interés para el investigador la forma que el tamaño del animal. Lo mismo puede ocurrir en el caso de la psiquiatría ya que, debido a las correlaciones positivas, las puntuaciones del primer componente sólo pueden proporcionar un índice de la gravedad de los síntomas y son los componentes restantes los que darán al psiquiatra información importante sobre el patrón de los síntomas.

Los componentes principales se usan más comúnmente (y de manera correcta) como un medio para construir una representación gráfica e informativa de los datos o como entrada a algún otro análisis como el Análisis de Regresión.

En algunas disciplinas, en particular la psicología y otras ciencias del comportamiento, los componentes principales pueden considerarse un fin en sí mismos y los investigadores pueden intentar interpretarlos de manera similar a los factores en un Análisis de Factores Exploratorio. Por lo tanto, el ACP es mayoritariamente una técnica multivariada exploratoria.

El siguiente documento contiene algunas notas y ejemplos resueltos sobre la forma de trabajar el Análisis de Componentes Principales en el programa estadístico R.

1. Medidas corporales de gorriones hembras

Recordando que mediante 5 variables medidas a 49 hembras de una especie de gorriones se busca encontrar alguna relación entre ellas. Dichas variables son: *Peso*, *Longitud del ala*, *Longitud de pico y cabeza*, *Longitud del humero* y *Longitud del esternón*. Por lo tanto, el primer paso será importar el archivo que contiene los datos de las variables medidas. Recuerde que es necesario indicar la ruta en donde se encuentra almacenado el archivo *Gorriones.csv*.

```
library(readr)
Gorriones <- read_csv("C:/Users/CIMAT/EME/Gorriones.csv")
View(Gorriones)
```

1.1 Verificar que los datos son adecuados

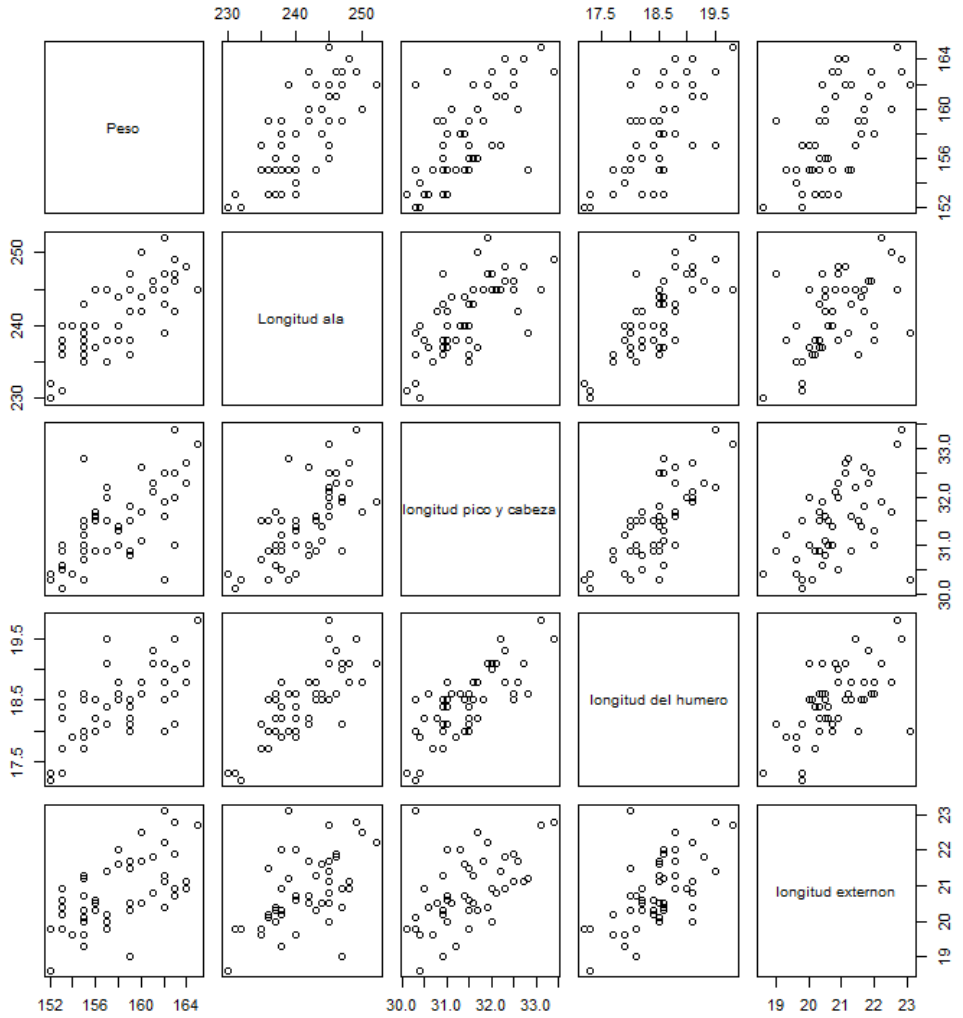
Antes de aplicar el ACP debe comprobarse si es pertinente realizar el análisis, es decir, si la relación entre las variables analizadas es lo suficientemente grande como para justificar el trabajo a realizar.

En un primer intento se puede recurrir a la matriz de correlación e identificar aquellas relaciones que son significativas y con grado alto de correlación. Esto nos permitirá evaluar, junto con alguna otra prueba, la conveniencia de eliminar aquellas variables con baja colinealidad. Este análisis previo que se realiza permitirá justificar la inclusión o exclusión de variables al comparar modelos y revisar los ajustes que se deban realizar.

```
install.packages("Hmisc")
library("Hmisc")

rcorr(as.matrix(Gorriones))
```

```
##                          Peso   Longitud ala   longitud pico y cabeza
## Peso                     1.00           0.73                     0.66
## Longitud ala             0.73           1.00                     0.67
## longitud pico y cabeza   0.66           0.67                     1.00
## longitud del humero      0.65           0.77                     0.76
## longitud externon        0.61           0.53                     0.53

##                         longitud del humero   longitud externon
## Peso                                   0.65                0.61
## Longitud ala                           0.77                0.53
## longitud pico y cabeza                 0.76                0.53
## longitud del humero                    1.00                0.61
## longitud externon                      0.61                1.00

## n= 49


## P
##                          Peso   Longitud ala   longitud pico y cabeza
## Peso                                  0e+00                     0e+00
## Longitud ala             0e+00                                  0e+00
## longitud pico y cabeza   0e+00          0e+00
## longitud del humero      0e+00          0e+00                     0e+00
## longitud externon        0e+00          0e+00                     1e-04

##                         longitud del humero longitud externon
## Peso                                  0e+00               0e+00
## Longitud ala                          0e+00               0e+00
## longitud pico y cabeza                0e+00               1e-04
## longitud del humero                                       0e+00
## longitud externon                     0e+00
```

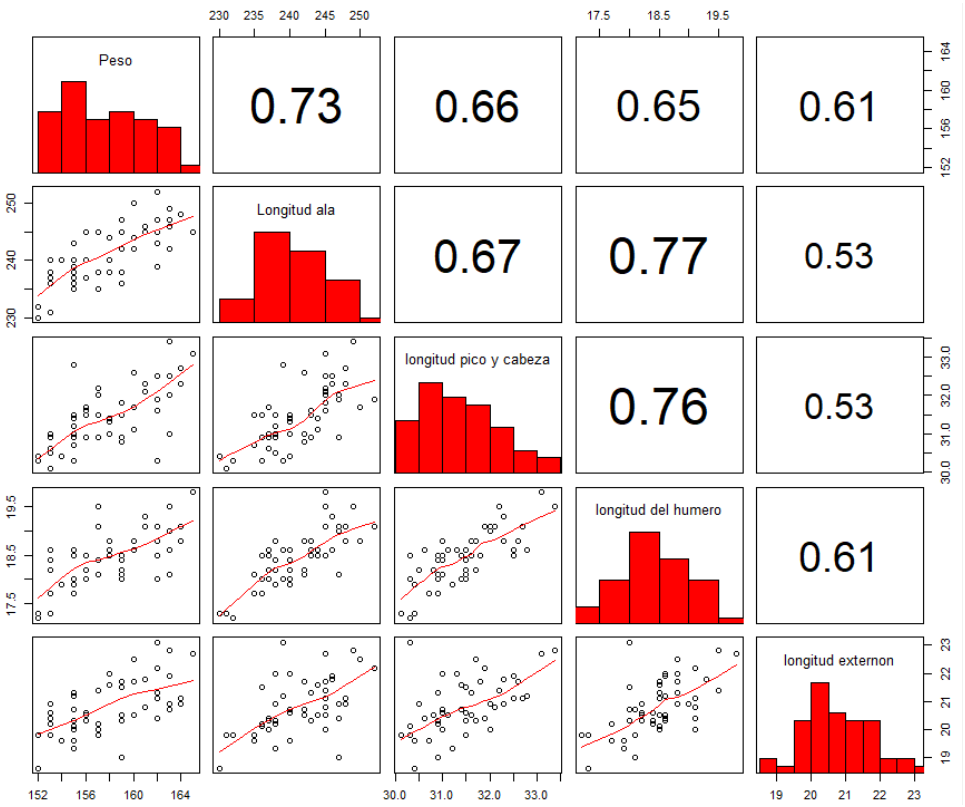
```
plot(Gorriones)        # Grafico de matriz de dispersión de las variables
```



Otra opción, aplicando lo estudiado en la unidad anterior, es obtener la matriz de diagramas de dispersión con histogramas y valores de correlación.

```
# Matriz de dispersión con histograma de la variable en la diagonal

panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col="red", ...)
}
```

```
# Matriz de dispersión con valores de correlación entre variables

panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}


# Matriz de diagramas de dispersión con histograma y valores de correlacion

par(fig=c(0,1,0,1))
pairs(Gorriones, lower.panel=panel.smooth, upper.panel=panel.cor,
      diag.panel=panel.hist)
```



Se aprecia tanto en la matriz de los *p–valores* como en las gráficas anteriores que los índices de correlación entre pares de variables son significativos estadísticamente, están por arriba de 0.50 y de forma positiva. Esta evidencia sugiere que aplicar el ACP es adecuado para estos datos.

Para confirmar el supuesto anterior se puede recurrir al *índice de Kaiser-Meyer-Olkin* o *medida de adecuación muestral KMO*. Este índice trata de saber si podemos factorizar las variables originales de forma eficiente.

El punto de partida es la matriz de correlaciones entre las variables observadas. Las variables pueden estar más o menos correlacionadas, pero la correlación entre dos de ellas puede estar influenciada por las otras. Así pues, se utiliza la correlación parcial para medir la relación entre dos variables eliminando el efecto del resto. El *índice KMO* compara los valores de las correlaciones entre las variables y sus correlaciones parciales. Si el *índice KMO* está próximo a 1, el ACP se puede hacer. Si, por el contrario, el índice es bajo (próximo a 0), el ACP no será relevante. Kaiser establece la clasificación siguiente:

$0.90 \geq$ KMO	Excelente
$0.80 \leq$ KMO < 0.90	Bueno
$0.70 \leq$ KMO < 0.80	Aceptable
$0.60 \leq$ KMO < 0.70	Mediocre
$0.50 \leq$ KMO < 0.60	Malo
KMO > 0.50	Inaceptable

Como se aprecia en la tabla, el tener un índice de KMO por arriba de 0.70 significará que el análisis a realizar es adecuado. Si KMO es menor a 0.6, hay que entrar a considerar cambiar de variables o de técnica. La franja restante (0.60 – 0.70) queda a criterio del investigador si después de analizar los datos de la muestra es adecuado realizar un ACP.

```
install.packages("psych")      # instalación de la biblioteca que contiene el cálculo del
                               # índice KMO
library(psych)
KMO(Gorriones)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = Gorriones)
## Overall MSA =  0.83
## MSA for each item =
##                   Peso           Longitud ala
##                   0.82                   0.82
## longitud pico y cabeza    longitud del humero
##                   0.86                   0.79
##      longitud esternón
##                   0.87
```

El valor del índice KMO (ver *Overall MSA*) para los datos es cercano al valor de 1. Si se atiende a la escala proporcionada anteriormente estaría en el rango de “Bueno”. Si se analiza el índice para cada una de las variables, también presentan valores altos. Se concluye que es pertinente aplicar el ACP a estos datos. La prueba anterior también proporciona las medidas por variable de forma que se puede detectar aquellas que no están relacionadas con las demás.

1.2 Verificar que los datos son adecuados

El siguiente paso consiste en la obtención de los valores y vectores propios de la matriz de covarianzas muestral o de la matriz de coeficientes de correlación que se obtienen a partir de la matriz de datos. En el caso utilizado como ejemplo se decide utilizar la matriz de correlación. Para el cálculo de los componentes principales se utiliza la función `prcomp()`.

```
Gorriones_pca <- prcomp(Gorriones, center = TRUE, scale = TRUE)
# center = TRUE y scale = TRUE para utilizar la matriz de correlación

round(Gorriones_pca$center,3)    # Despliega la media de las variables originales
```

```
##                   Peso           Longitud ala
##                157.980                241.327
## longitud pico y cabeza    longitud del humero
##                 31.459                 18.469
##      longitud esternón
##                 20.827
```

```
round(Gorriones_pca$scale,3)      # Despliega la desviación estándar de las variables
                                  # originales
```

```
##                   Peso           Longitud ala
##                  3.654                  5.068
## longitud pico y cabeza    longitud del humero
##                  0.795                  0.564
##      longitud esternón
##                  0.991
```

```
summary(Gorriones_pca) # Muestra la desviación estándar, el porcentaje (en decimales) de
                       # varianza explicada, y el porcentaje (en decimales) de la varianza
                       # acumulada para cada uno de los componentes principales
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5
## Standard deviation     1.9016 0.7290 0.62163 0.54915 0.40562
## Proportion of Variance 0.7232 0.1063 0.07728 0.06031 0.03291
## Cumulative Proportion  0.7232 0.8295 0.90678 0.96709 1.00000
```

Se aprecia que el primer componente principal concentra el 72.32% de la varianza total, el segundo componentes contiene 10.63% mientras que el ultimo solo explica 3.29%. La varianza asociada a cada factor (el cuadrado de las desviaciones estándar) viene expresada por su valor propio o raíz característica de la matriz de coeficientes de correlación (en este caso) o de la matriz de covarianzas.

```
round(Gorriones_pca$sdev^2,3)
```

```
##  [1] 3.616 0.532 0.386 0.302 0.165
```

```
round(Gorriones_pca$rotation,4)
```

```
##                          PC1     PC2     PC3     PC4     PC5
## Peso                  0.4518 -0.0507  0.6905 -0.4204  0.3739
## Longitud ala          0.4617  0.2996  0.3405  0.5479 -0.5301
## longitud pico y cabeza 0.4505  0.3246 -0.4545 -0.6063 -0.3428
## longitud del humero    0.4707  0.1847 -0.4109  0.3883  0.6517
## longitud esternón      0.3977 -0.8765 -0.1785  0.0689 -0.1924
```

Los otros elementos importantes en el ACP son los *eigenvectores* asociados para cada uno de los *eigenvalores*. Cada columna representa una combinación lineal (loadings o cargas) de las variables originales que proporcionan los componentes principales. Así el primer componente se obtiene con la siguiente combinación:

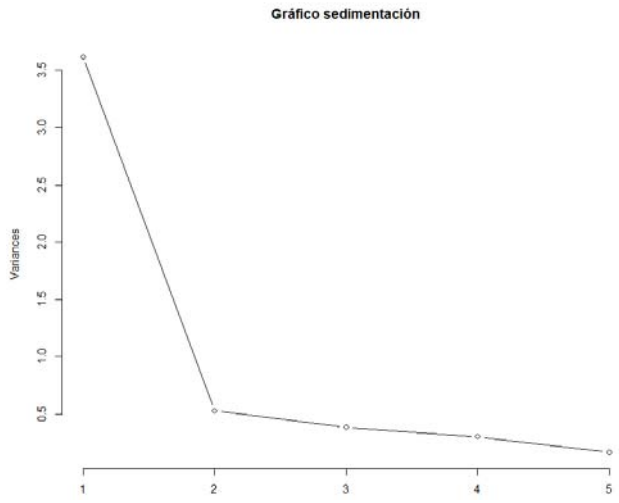
$$Z1 = 0.4518(peso) + 0.4617(longitud\ ala) + 0.4505(longitud\ pico\ y\ cabeza) + 0.4707(longitud\ humero) + 0.3977(longitud\ esternón)$$

NOTA: El signo de las columnas que conforman a los *eigenvectores* es arbitrario y puede diferir entre programas que computaciones para el ACP. Lo que representan es el sentido de la aportación que tiene cada variable con respecto al componente. En el caso del primer componente las variables contribuyen en el mismo sentido. Al analizar las cargas del segundo componente se observa que la primera y la última variable tienen una contribución en sentido contrario al de las restantes tres variables.

1.3 Determinar el número de Componentes Principales

El número de componentes principales a retener queda a juicio del investigador. Es posible representar el gráfico de sedimentación (*scree plot)* de los *eigenvalores* y considerar alguno de los criterios mencionados en el material de apoyo de la presente unidad.

```
plot(Gorriones_pca, type = "line", main = "Gráfico sedimentación")
```



Como se menciona en el material de apoyo si se emplea la técnica de Kaiser, únicamente el primer componente debe ser analizado. Este componente explica cerca del 72% del total de la variabilidad. Si se decide el explicar un cierto porcentaje de la variabalidad, e.g. 80%, se debe analizar la varianza acumulada para cada uno de los componentes. En este caso al incluir los dos primeros componentes se estaría explicando cerca del 83%.

1.4 Análisis e interpretación de resultados

La última etapa del ACP consiste en la interpretación de los resultados tanto numéricos como gráficos. Si se considera el caso de utilizar los dos primeros componentes principales. Como el ejemplo lo indica, un grafica de dispersión nos permitirá resumir la información contenida en las observaciones mediante los componentes seleccionados.

```
# Se obtiene los scores de cada observación en los componentes que se deciden analizar
(Z1 <- round(scale(as.matrix(Gorriones)) %*% Gorriones_pca$rotation[,1],4))
```

```
##         [,1]
##  [1,]  0.0643
##  [2,] -2.1803
##  [3,] -1.1456
##  [4,] -2.3111
##  [5,] -0.2950
##  [6,]  1.9163
##  [7,] -1.0504
##  [8,]  0.4385
##  [9,]  2.6915
## [10,]  0.1857
## [11,]  0.3711
## [12,]  0.2677
## [13,]  2.3592
## [14,]  0.7146
## [15,] -1.3943
## [16,] -1.5587
## [17,]  0.5483
## [18,] -1.6577
## [19,] -1.7767
## [20,]  2.1761
## [21,] -0.4574
## [22,] -0.9651
## [23,] -0.6581
## [24,]  1.5841
## [25,] -3.7168
## [26,]  2.1236
## [27,] -1.3289
## [28,]  1.7233
## [29,]  3.9943
## [30,] -3.7142
## [31,]  0.1484
## [32,]  1.1951
```

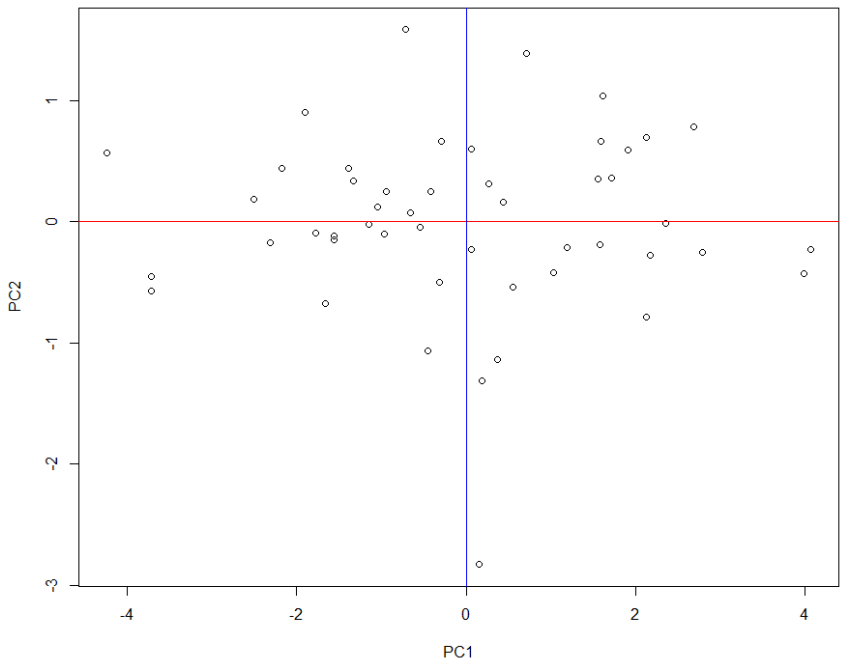
```
## [33,]  1.0299
## [34,] -0.7148
## [35,] -0.3175
## [36,]  2.7963
## [37,] -4.2403
## [38,] -0.5419
## [39,] -1.9057
## [40,]  4.0714
## [41,]  0.0628
## [42,] -0.9383
## [43,] -0.4228
## [44,]  1.5868
## [45,] -2.5090
## [46,]  1.6188
## [47,] -1.5590
## [48,]  1.5570
## [49,]  2.1342
```

```
(Z2 <- round(scale(as.matrix(Gorriones)) %*% Gorriones_pca$rotation[,2],4))
```

```
##          [,1]
##  [1,]  0.6008
##  [2,]  0.4423
##  [3,] -0.0193
##  [4,] -0.1720
##  [5,]  0.6652
##  [6,]  0.5953
##  [7,]  0.1198
##  [8,]  0.1640
##  [9,]  0.7823
## [10,] -1.3137
## [11,] -1.1384
## [12,]  0.3147
## [13,] -0.0111
## [14,]  1.3887
## [15,]  0.4430
## [16,] -0.1447
## [17,] -0.5402
## [18,] -0.6724
## [19,] -0.0945
## [20,] -0.2747
## [21,] -1.0614
## [22,] -0.1030
## [23,]  0.0778
## [24,] -0.1864
## [25,] -0.4496
## [26,] -0.7883
## [27,]  0.3389
## [28,]  0.3636
## [29,] -0.4311
## [30,] -0.5715
## [31,] -2.8304
## [32,] -0.2098
## [33,] -0.4201
```

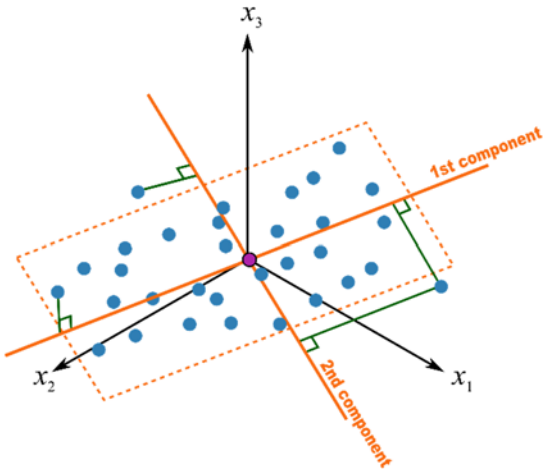
```
## [34,]  1.5868
## [35,] -0.4967
## [36,] -0.2528
## [37,]  0.5667
## [38,] -0.0430
## [39,]  0.9021
## [40,] -0.2309
## [41,] -0.2264
## [42,]  0.2474
## [43,]  0.2487
## [44,]  0.6668
## [45,]  0.1899
## [46,]  1.0431
## [47,] -0.1177
## [48,]  0.3546
## [49,]  0.6975
```

```
plot(z1,z2, xlab="PC1", ylab="PC2")
abline(0,0,col="red")
abline(0,90,col="blue")
```

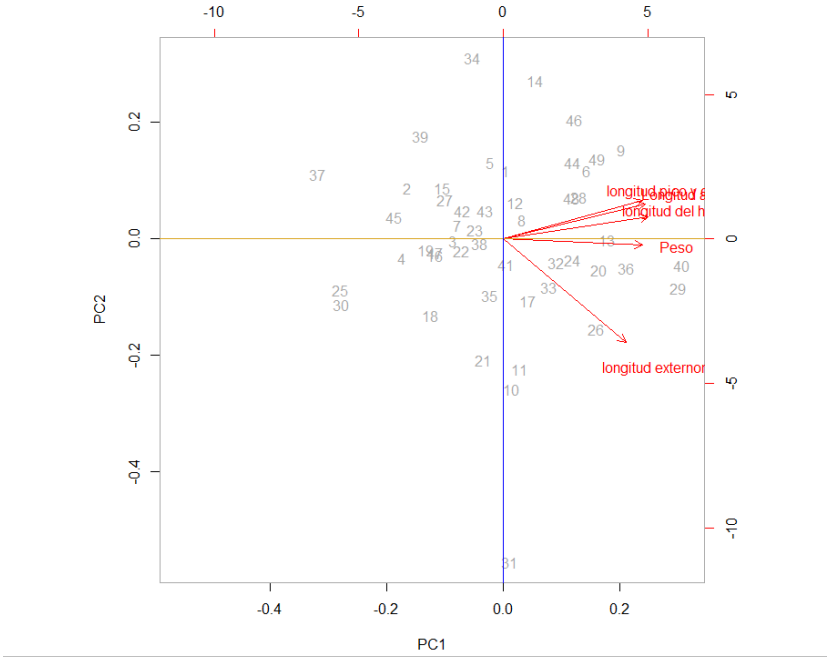


Otra forma de representar las proyecciones de cada observación y las variables al mismo tiempo es mediante el *gráfico biplot*. Las variables son representadas mediante vectores (flechas) mientras que las observaciones aparecen similar al gráfico de dispersión anterior. Como se muestra en la figura lo que busca este tipo de representación gráfica es visualizar la estructura de los datos proyectándolos en un plano fácil de interpretar. Para lograr dicha representación es necesario reducir la dimensionalidad de la base de datos utilizando los *eigenvectores* asociados a los componentes seleccionados.

Al realizar un ACP, la dimensión está dado por el número de variables que se consideran en el estudio, en el caso del ejemplo son 5 las variables medidas por lo que su dimensión es de 5. El análisis de los *eigenvalores* determinó que los dos primeros componentes son retenidos para el estudio, esto indica que la dimensión pasó de 5 a 2. Al realizar la reducción se está dispuesto a sacrificar un porcentaje de información (variablidad) de los datos.




```
biplot(Gorriones_pca, choices = c(1,2), col = c("darkgray", "red"))
# El argumento choice = c ( , ) se emplea para indicar los componentes principales a graficar

abline(h = 0,col="goldenrod")
abline(v = 0,col="blue")
```



En el grafico anterior se resumen muchos aspectos. El tamaño de la flecha indica la importancia que tiene la variable para explicar la variación contenida entre las componentes principales que se grafican, mientras más larga mayor es la cantidad de variabilidad explicada para esta variable por los componentes que se grafican. El ángulo que separa a dos vectores indica que tan correlacionadas están esas variables, mientras menos separación exista más fuerte es la correlación entre ellas. Un ángulo entre las variables de 90° indica que su correlación es igual a 0 y un ángulo mayor a 90° entre las fechas indica correlación negativa. Al analizar el gráfico se observa que las variables *longitud del humero*, *longitud pico* y *longitud ala* tienen una fuerte correlación positiva. La correlación entre las variables *longitud pico* y *longitud esternón* es muy bajo dado que el ángulo entre sus flechas es muy cercano a 90°. La variable *longitud esternón* es la que presenta un tamaño mayor lo que significa que esta variable explica una mayor cantidad de la variabilidad representada por la primera y segunda variable.

La posición del vector en el plano también aporta la siguiente información: todas las variables tienen coeficientes positivos en el eje correspondiente con al primer componte principal; las variables *longitud del humero*, *longitud pico* y *longitud ala* son positivos sus coeficientes en el eje correspondiente al segundo componente principal mientras que *longitud esternón* y *peso* su coeficiente es negativo. Lo anterior se debe a los valores obtenidos en los *eigenvectores* correspondientes a cada componente.

También es posible conocer la importancia que tiene cada observación en cada una de las variables. Por ejemplo, si se alarga la flecha que representa la variable *longitud esternón* y proyectamos perpendicularmente sobre ella cada uno de los puntos que representan las observaciones se obtiene que el gorrión hembra 31 tiene un valor mayor en esta variable, seguida de las observaciones 10, 11, 26, 29 y 40. La dirección a la que apunta la flecha indica hacia donde crecen los valores de dicha variable, por lo que las observaciones proyectadas en el eje más alejadas en la dirección de la flecha son la que tienen los valores mayores.

Si se observa la distribución de las observaciones en el plano se puede establecer conclusiones sobre qué tan similares son entre ellas. A una menor distancia entre dos puntos, mayor es la similitud entre ellos. Por ejemplo, los gorriones hembras 25 y 30 son similares al igual que las observaciones 48 y 28, caso contrario entre el gorrión 37 y 40, ambos tiene un perfil diferente ya que tiene una gran separación al revisar el eje que corresponde al primer componente el cual explica la mayor cantidad de variabilidad de los datos originales. La explicación de esta diferencia se sustenta en la orientación que tiene los vectores que representan las variables. Al tener todos los vectores orientados hacia el lado positivo del eje del primer componente, las observaciones que se encuentran más pegados al extremo derecho son aquellas que presentan mediciones en las variables estudiadas más grandes, contrario a aquellas que se ubican en el extremo izquierdo del primer eje principal. Esto puede comprobarse al revisar los valores originales de dichas observaciones. Por lo tanto, el primer componente es denominado eje de *tamaño*.

Ahora bien, que pasa al analizar el eje del segundo componente principal. Al observa la dispersión de los gorriones en el *grafico biplot* se aprecia que las observaciones 5 y 35 están casi al mismo nivel con respecto al primer eje principal, es decir tienen el mismo tamaño, entonces ¿qué los hace diferente?. El interés estará ahora centrado en encontrar cuales son las características que los hacen diferentes

mediante el análisis del eje que representa al segundo componente principal. Como ya se describió anteriormente al examinar las representaciones vectoriales de las variables se identifica que 3 de ellas están ubicadas en el extremo positivo del eje mientras que *peso* y *longitud esternón* están en el extremo negativo. Lo anterior describe las características adicionales al tamaño que diferencian a cada una de las observaciones, es decir describen la *forma* que tiene el gorrión. Gráficamente se identifica que el *peso* no contribuye mucho a realizar este análisis detallado debido a que su vector es casi horizontal. Regresando al ejemplo, se inferir que la observación 5 tiene longitud más pequeña del esternón que la observación 35 y que al menos una de las otras longitudes medidas es mayor que de las medidas en la observación 35.

2. Medidas corporales de gorriones hembras

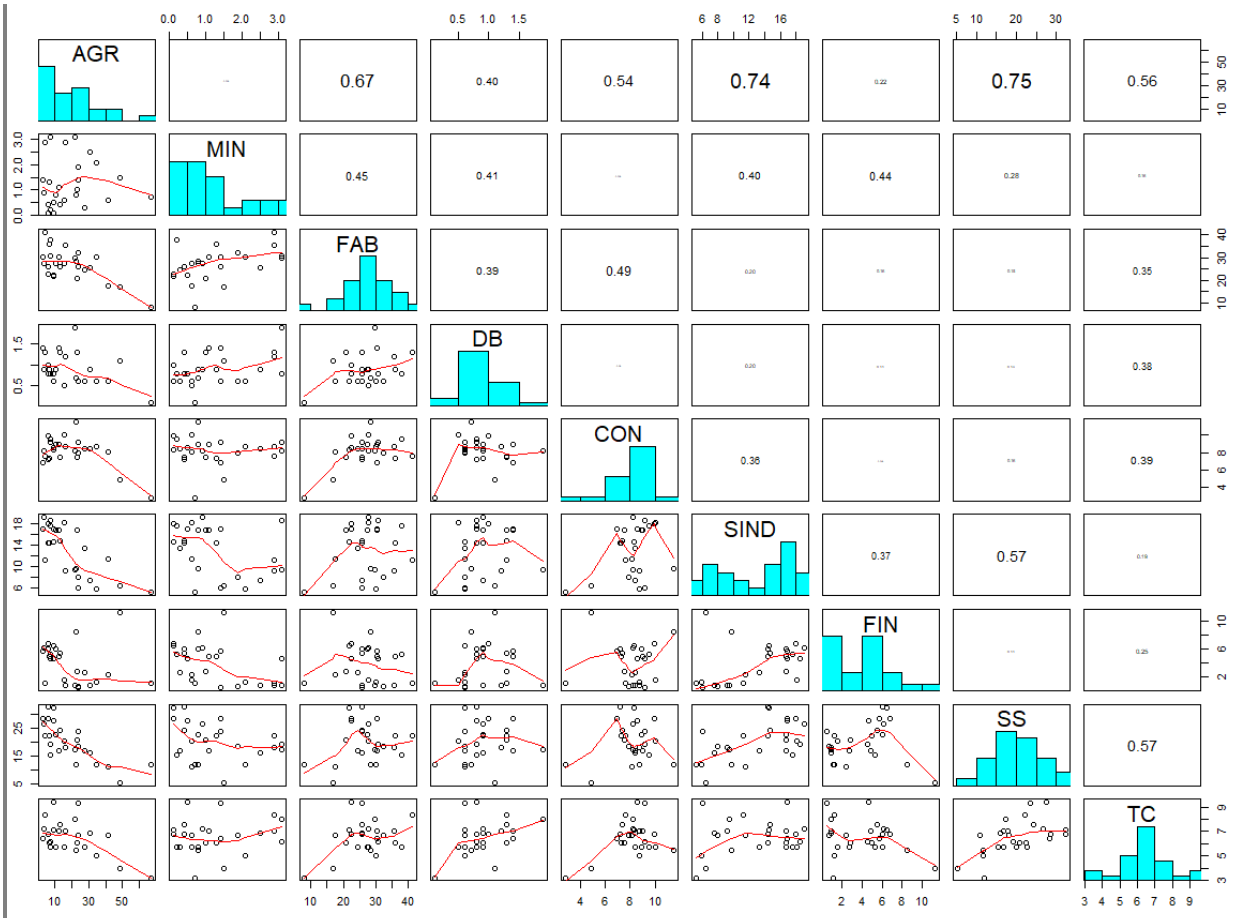
2.1 Verificar que los datos son adecuados

Obtención de la matriz de correlación

```
library("Hmisc")
rcorr(as.matrix(paises_europa[,-1]))
```

```
##         AGR   MIN   FAB    DB   CON  SIND   FIN    SS    TC
## AGR    1.00  0.04 -0.67 -0.40 -0.54 -0.74 -0.22 -0.75 -0.56
## MIN    0.04  1.00  0.45  0.41 -0.03 -0.40 -0.44 -0.28  0.16
## FAB   -0.67  0.45  1.00  0.39  0.49  0.20 -0.16  0.15  0.35
## DB    -0.40  0.41  0.39  1.00  0.06  0.20  0.11  0.13  0.38
## CON   -0.54 -0.03  0.49  0.06  1.00  0.36  0.02  0.16  0.39
## SIND  -0.74 -0.40  0.20  0.20  0.36  1.00  0.37  0.57  0.19
## FIN   -0.22 -0.44 -0.16  0.11  0.02  0.37  1.00  0.11 -0.25
## SS    -0.75 -0.28  0.15  0.13  0.16  0.57  0.11  1.00  0.57
## TC    -0.56  0.16  0.35  0.38  0.39  0.19 -0.25  0.57  1.00

## n= 26


## P
##      AGR    MIN    FAB    DB     CON    SIND   FIN    SS     TC
## AGR         0.8622 0.0002 0.0429 0.0046 0.0000 0.2805 0.0000 0.0026
## MIN  0.8622        0.0227 0.0399 0.9012 0.0449 0.0235 0.1643 0.4448
## FAB  0.0002 0.0227        0.0519 0.0102 0.3179 0.4472 0.4521 0.0790
## DB   0.0429 0.0399 0.0519        0.7713 0.3226 0.5932 0.5190 0.0589
## CON  0.0046 0.9012 0.0102 0.7713        0.0742 0.9371 0.4401 0.0504
## SIND 0.0000 0.0449 0.3179 0.3226 0.0742        0.0663 0.0023 0.3589
## FIN  0.2805 0.0235 0.4472 0.5932 0.9371 0.0663        0.6007 0.2259
## SS   0.0000 0.1643 0.4521 0.5190 0.4401 0.0023 0.6007        0.0025
## TC   0.0026 0.4448 0.0790 0.0589 0.0504 0.3589 0.2259 0.0025
```

```
par(fig=c(0,1,0,1))      # Diagrama de dispersión e histogramas
pairs(paises_europa[,-1], lower.panel=panel.smooth, upper.panel=panel.cor,
      diag.panel=panel.hist)
```



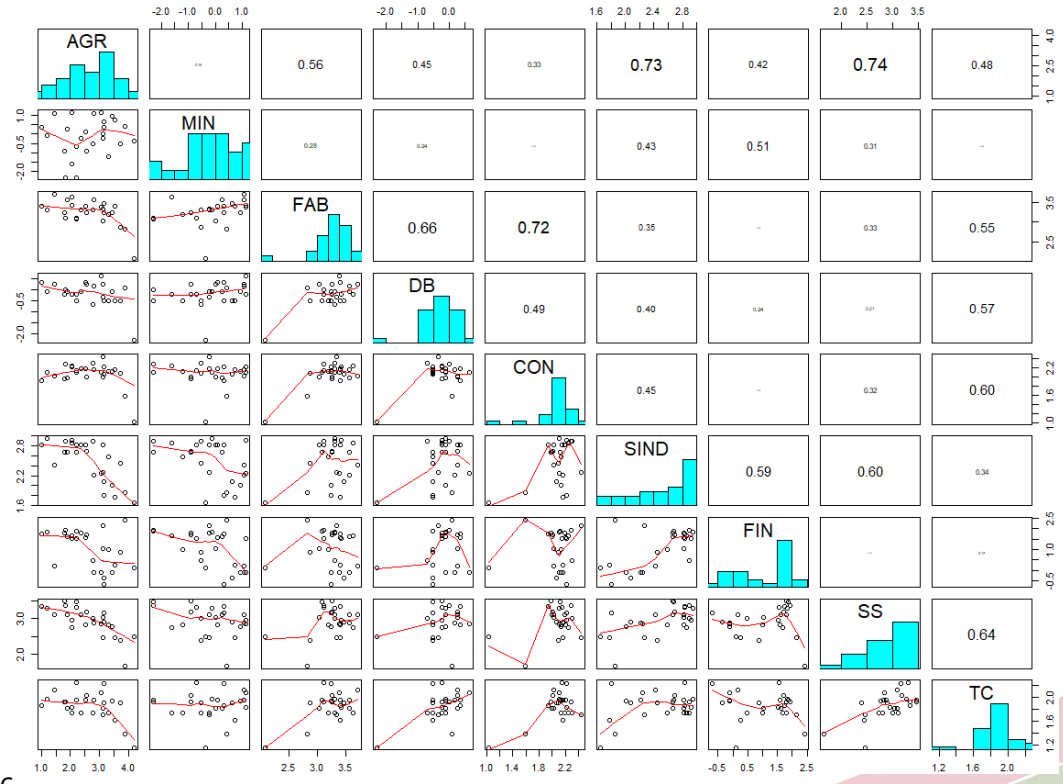

```
KMO(paises_europa[,-1])

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = paises_europa[,-1]
## Overall MSA =  0.13
## MSA for each item =
##  AGR  MIN  FAB   DB  CON SIND  FIN   SS   TC
## 0.24 0.10 0.14 0.10 0.10 0.15 0.06 0.15 0.14
```

Se aprecia que los datos no son adecuados para aplicar la técnica de ACP, los histogramas de la mayoría de las variables no se asemejan a una distribución normal. Se procede a realizar una transformación de los datos aplicando el logaritmo natural a toda la base. Estos valores son importados a R y se ejecutan los comandos anteriores.

```
rcorr(as.matrix(paises_europa_LN[,-1]))
```

```
##         AGR   MIN   FAB    DB   CON  SIND   FIN    SS    TC
## AGR    1.00  0.16 -0.56 -0.45 -0.33 -0.73 -0.42 -0.74 -0.48
## MIN    0.16  1.00  0.28  0.24 -0.09 -0.43 -0.51 -0.31  0.04
## FAB   -0.56  0.28  1.00  0.66  0.72  0.35 -0.03  0.33  0.55
## DB    -0.45  0.24  0.66  1.00  0.49  0.40  0.24  0.21  0.57
## CON   -0.33 -0.09  0.72  0.49  1.00  0.45  0.08  0.32  0.60
## SIND  -0.73 -0.43  0.35  0.40  0.45  1.00  0.59  0.60  0.34
## FIN   -0.42 -0.51 -0.03  0.24  0.08  0.59  1.00  0.07 -0.17
## SS    -0.74 -0.31  0.33  0.21  0.32  0.60  0.07  1.00  0.64
## TC    -0.48  0.04  0.55  0.57  0.60  0.34 -0.17  0.64  1.00

## n= 26

## P
##       AGR    MIN    FAB    DB     CON    SIND   FIN    SS     TC
## AGR          0.4260 0.0029 0.0205 0.0986 0.0000 0.0317 0.0000 0.0139
## MIN   0.4260        0.1664 0.2456 0.6710 0.0299 0.0081 0.1243 0.8379
## FAB   0.0029 0.1664        0.0003 0.0000 0.0818 0.8889 0.1030 0.0038
## DB    0.0205 0.2456 0.0003        0.0101 0.0429 0.2447 0.2971 0.0026
## CON   0.0986 0.6710 0.0000 0.0101        0.0219 0.7025 0.1056 0.0013
## SIND  0.0000 0.0299 0.0818 0.0429 0.0219        0.0015 0.0012 0.0923
## FIN   0.0317 0.0081 0.8889 0.2447 0.7025 0.0015        0.7294 0.4200
## SS    0.0000 0.1243 0.1030 0.2971 0.1056 0.0012 0.7294        0.0004
## TC    0.0139 0.8379 0.0038 0.0026 0.0013 0.0923 0.4200 0.0004
```

```
par(fig=c(0,1,0,1))    # Diagrama de dispersión e histogramas
pairs(paises_europa_LN[,-1], lower.panel=panel.smooth, upper.panel=panel.cor,
      diag.panel=panel.hist)
```



```r
KMO(paises_europa[,-1])
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = paises_europa_LN[, -1])
## Overall MSA =  0.56
## MSA for each item =
##  AGR  MIN  FAB   DB  CON SIND  FIN   SS   TC
## 0.55 0.55 0.52 0.56 0.51 0.79 0.37 0.65 0.60
```

A partir de los valores anteriores se puede decidir la eliminación de la variable que presenta el valor *MSA* más bajo (*FIN*) y volver a correr el análisis. Par efectos de ejemplo se procederá a realizar el análisis con los datos anteriores.

2.2 Obtención de los Componentes Principales

```r
# Se efectua el ACP
paises_europa_LN_pca <- prcomp(paises_europa_LN[,-1], center = TRUE, scale = TRUE)

#despliega los eigenvalores
(eigenvalores <- round(paises_europa_LN_pca$sdev^2,3))
```

```
## [1] 4.100 2.061 1.088 0.755 0.463 0.235 0.157 0.095 0.045
```

```r
round(paises_europa_LN_pca$rotation,4) # eigenvectores asociados
```

```
##          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
## AGR   0.4145  0.1490 -0.0467  0.4714 -0.2730  0.2910 -0.1532 -0.2615 -0.5749
## MIN   0.0838  0.5714 -0.2067 -0.4818  0.1311  0.3956  0.4266 -0.1782 -0.0644
## FAB  -0.3677  0.3416 -0.1956  0.0514  0.4843 -0.3439 -0.2726  0.1355 -0.5116
## DB   -0.3429  0.2353 -0.4330 -0.1214 -0.5688 -0.0404 -0.4448 -0.2291  0.2219
## CON  -0.3557  0.1804 -0.0881  0.6809  0.2269  0.1422  0.2928 -0.2769  0.3680
## SIND -0.3871 -0.3272 -0.0838 -0.0398  0.1064  0.7593 -0.2057  0.2937 -0.1346
## FIN  -0.1646 -0.5071 -0.5199 -0.0488 -0.1266 -0.1945  0.4930 -0.1938 -0.3295
## SS   -0.3612 -0.1257  0.5680 -0.2220  0.0132  0.0141 -0.0599 -0.6605 -0.2039
## TC   -0.3667  0.2582  0.3488  0.1068 -0.5208 -0.0640  0.3843  0.4383 -0.2252
```

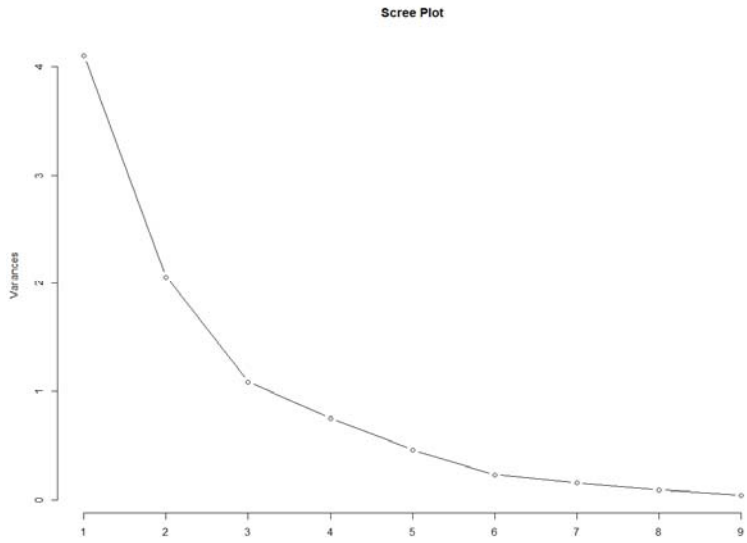
```r
summary(paises_europa_LN_pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7    PC8     PC9
## Standard deviation     2.0248 1.4356 1.0431 0.86903 0.68051 0.48459 0.39626 0.3089 0.21327
## Proportion of Variance 0.4555 0.2290 0.1209 0.08391 0.05146 0.02609 0.01745 0.0106 0.00505
## Cumulative Proportion  0.4555 0.6845 0.8054 0.88935 0.94081 0.96690 0.98435 0.9950 1.00000
```

2.3 Determinar el número de Componentes Principales

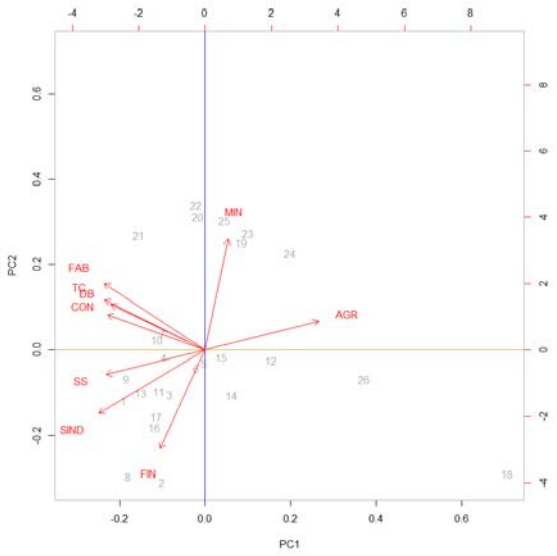
Para determinar el número de componentes a retener en el estudio se puede aplicar el criterio de Kaiser, lo que dejaría a los tres primeros componentes. Los tres primeros componentes principales explican más del 80% de la variabilidad en los datos. Si se aplica la gráfica de sedimentación se obtiene que el cambio ocurre en el tercer componente por lo que solo se considerarían los dos primeros que representan un 68% de la variabilidad en los datos. Se continúa trabajando con tres componentes.

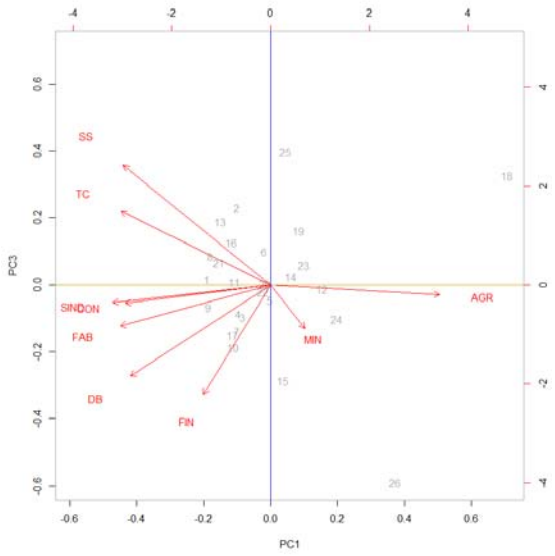
```
plot(paises_europa_LN_pca, type="lines", main="Scree Plot")
```

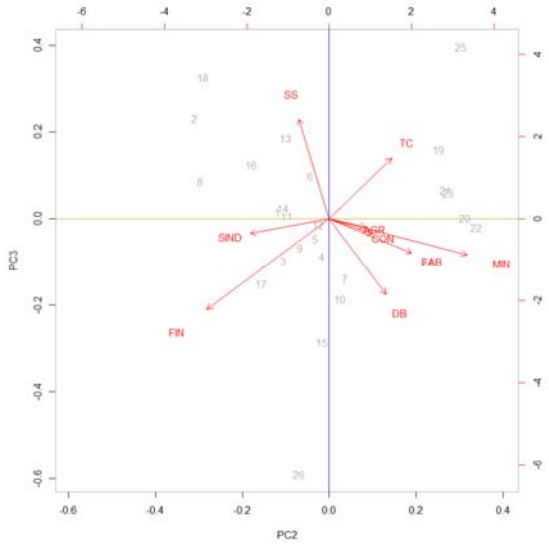


2.4 Análisis e interpretación de resultados

Al considerar tres componentes se reduce la dimensión de los datos al pasar de una dimensión 9 a una dimensión 3. En este caso es posible ya sea emplear gráficos de dispersión 2D o 3D para estudiar el comportamiento de los datos en las dimensiones seleccionadas. En el caso de las *gráficas biplot*, como únicamente muestra dos dimensiones de forma simultánea, se debe realizar 3 gráficos que representen las combinaciones entre los 3 componentes retenidos.


PC2
PC1
MIN
FAB
TC
DB
CON
AGR
SS
SIND
FIN


PC3
PC1
SS
TC
SIND
CON
FAB
AGR
MIN
DB
FIN


PC3
PC2
SS
TC
SIND
AGR
CON
FAB
MIN
DB
FIN