Procesamiento de datos multivariados

Facundo Vecchi - A01283666

25 de octubre de 2022

Carga de datos

```
datos <- read.csv("C:/Users/facun/Documents/GitHub/ai_avanzada_personal/Periodo_2/Modulo_5/Momento_retr
datos_only_nums <- subset(datos, select = -c(X1, X2, X12))
```

Importar librerías

```
library(data.table)
library(MVN)
library(ggplot2)
library(stats)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(mmeln)
```

1. Realice un análisis de normalidad

a) Prueba de normalidad de Mardia y la prueba de Anderson Darling

```
mvn(datos_only_nums,
    subset = NULL,
    mvn = "mardia",
    covariance = FALSE,
    showOutliers = FALSE)
```

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 410.214790601478 7.04198777815398e-23     NO
## 2 Mardia Kurtosis 4.59612555772731 4.30419392238868e-06     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X3        3.6725  <0.001      NO
## 2 Anderson-Darling    X4        0.3496  0.4611      YES
## 3 Anderson-Darling    X5        4.0510  <0.001      NO
## 4 Anderson-Darling    X6        5.4286  <0.001      NO
## 5 Anderson-Darling    X7        0.9253  0.0174      NO
## 6 Anderson-Darling    X8        8.6943  <0.001      NO
## 7 Anderson-Darling    X9        1.9770  <0.001      NO
```

```
## 8 Anderson-Darling    X10       0.6585   0.081      YES
## 9 Anderson-Darling    X11       1.0469  0.0086      NO
##
## $Descriptives
##      n       Mean    Std.Dev Median  Min    Max  25th  75th       Skew
## X3  53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50  0.9679170
## X4  53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40 -0.2458771
## X5  53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60  1.3045868
## X6  53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70  2.4130571
## X7  53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77  0.5986343
## X8  53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00  2.5808773
## X9  53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33  1.0729099
## X10 53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33  0.4645925
## X11 53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70  0.9449951
##       Kurtosis
## X3  -0.4705349
## X4  -0.6239638
## X5   0.6130359
## X6   6.1042185
## X7  -0.6312607
## X8   6.0089455
## X9   0.4060828
## X10 -0.6692490
## X11  0.5733500
```

Como se puede observar en la tabla anterior, solo se encuentran dos variables que se distribuyen individualmente como una normal, las variables X4 y X10. Asimismo podemos observar como las pruebas de Mardia indician que no se pasan las pruebas de kutosis y sesgo.

b) Realiza la prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores

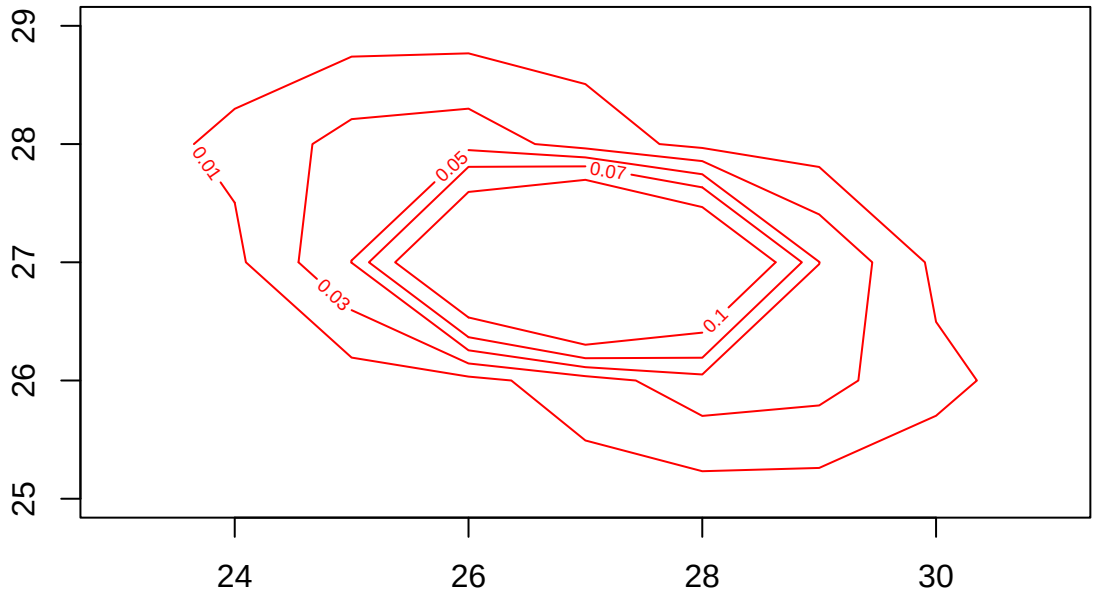
```
datos_subset <- subset(datos_only_nums, select = c(X4, X10))
mvn(datos_subset,
    subset = NULL,
    mvn = "mardia",
    covariance = FALSE,
    showOutliers = FALSE)
```

```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X4        0.3496    0.4611    YES
## 2 Anderson-Darling    X10       0.6585    0.0810    YES
##
## $Descriptives
##      n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew   Kurtosis
## X4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771 -0.6239638
## X10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925 -0.6692490
```

Una vez ya utilizando unicamente las variables que se distribuyen como normales, se puede observar que las pruebas de Mardia indican que si se pasan las pruebas de kutosis y sesgo. Por lo que se puede concluir que las variables X4 y X10 tienen normalidad multivariada.

c) Haz la gráfica de contorno de la normal multivariada obtenida en el inciso B

```
x <- seq(datos_subset[,1])
y <- seq(datos_subset[,2])
mu <- c(mean(x), mean(y))
sigma <- cov(datos_subset)
f <- function(x, y) dmnorm(cbind(x, y), mu, sigma)
z <- outer(x, y, f)
contour(x, y, z,
        col = "red",
        levels = c(0.01,0.03,0.05,0.07,0.1),
        xlim = c(23, 31),
        ylim = c(25, 29))
```



d) Detecta datos atípicos o influyentes en la normal multivariada encontrada en el inciso B

```
mvn(datos_subset,
    subset = NULL ,
    mvn = "mardia",
    covariance = FALSE,
```

```
    showOutliers = TRUE)
```

```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X4        0.3496    0.4611    YES
## 2 Anderson-Darling    X10       0.6585    0.0810    YES
##
## $Descriptives
##      n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew   Kurtosis
## X4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771 -0.6239638
## X10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925 -0.6692490
##
## $multivariateOutliers
## NULL
```

Utilizando la función de R mvn, al pasarle el parametro `showOutliers` como TRUE, nos muestra que no se encuentran datos atípicos o influyentes en la normal multivariada encontrada en el inciso B.

2. Realice un análisis de componentes principales

a) Justifique por qué es adecuado el uso de componentes principales para analizar la base

Este análisis de componentes principales es apropiado para este conjunto de datos ya que el objetivo final es obtener un modelo de regresión para poder predecir la contaminación del mercurio en el agua de los lagos. Trabajar con mas de 10 variables diferentes vuelve complicada la selección de estas para dicho modelo. Es por esto que a través de componentes principales, es apropiado buscar reducir la dimensionalidad de los datos para reducir la complejidad.

b) Realiza el análisis de componentes principales y justifica el número de componentes principales apropiados para reducir la dimensión de la base

```
pca <- princomp(datos_only_nums, cor = TRUE)
summary(pca)
```

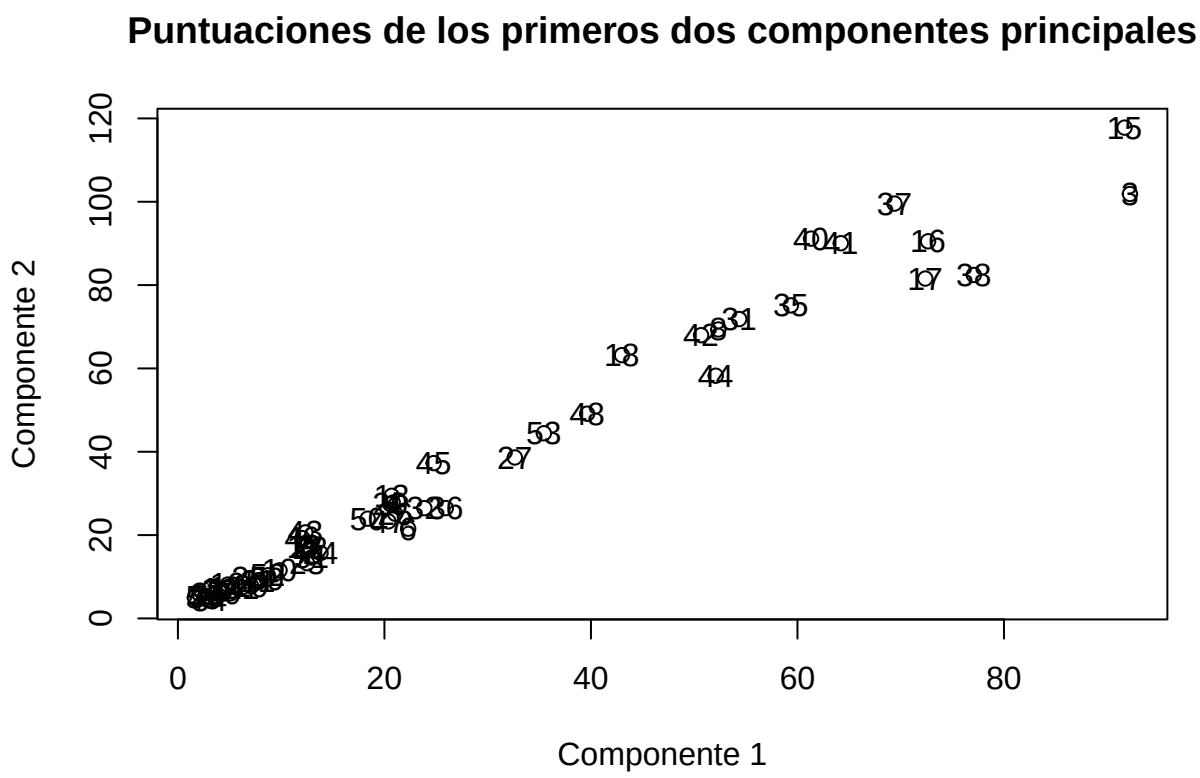
```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.3121220 1.1049470 1.0210443 0.81722905 0.57940716
## Proportion of Variance 0.5939898 0.1356564 0.1158368 0.07420704 0.03730141
## Cumulative Proportion  0.5939898 0.7296462 0.8454831 0.91969010 0.95699151
##                            Comp.6     Comp.7      Comp.8      Comp.9
## Standard deviation     0.45709713 0.32749661 0.228103640 0.137307403
## Proportion of Variance 0.02321531 0.01191711 0.005781252 0.002094814
## Cumulative Proportion  0.98020682 0.99212393 0.997905186 1.000000000
```

Observando las proporciones de la varianza, se puede notar como del componente 1 al 2 hay un gran salto, pero a partir de ahí empiezan a haber saltos mas pequeños en la varianza explicada de cada componente. Esto indica que los componentes 3 a 9, aunque la varianza explicada acumulada de estos es cerca del 30%,

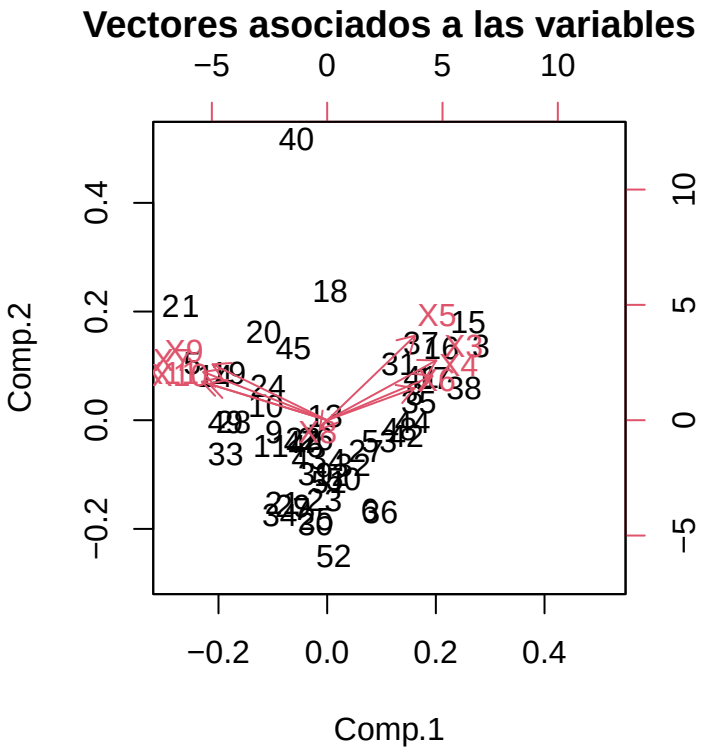
individualmente no aportan mucho. Ya que el propósito de este análisis es reducir la dimensionalidad los mas que se pueda, se utilizaran solo los componentes 1 y 2.

c) Representa en un gráfico los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes

```
pcaS <- as.matrix(datos_only_nums)%*%pca$loadings
plot(pcaS[,1:2],
     type="p",
     main = "Puntuaciones de los primeros dos componentes principales",
     xlab = "Componente 1",
     ylab = "Componente 2")
text(pcaS[,1],pcaS[,2],1:nrow(pcaS))
```




```
biplot(pca, main = "Vectores asociados a las variables")
```



Al graficar los vectores de los componentes se puede observar que que se crearon un total de 3 grupos de variables. El de la izquierda que esta compuesto por las variables X7, X9, X10, X11, el del centro compuesto por X8 y el de la derecha compuesto por X3, X4, X5, X6. Este ultimo grupo esta compuesto por las variables de Alcalinidad, PH, Calcio y Clorofila, que son las variables que se buscan utilizar para un modelo de regresión.

d) Interprete los resultados. Explique brevemente a qué conclusiones llega con su análisis y qué significado tienen los componentes seleccionados en el contexto del problema

```
sum_pca <- summary(pca)
sum_pca
```

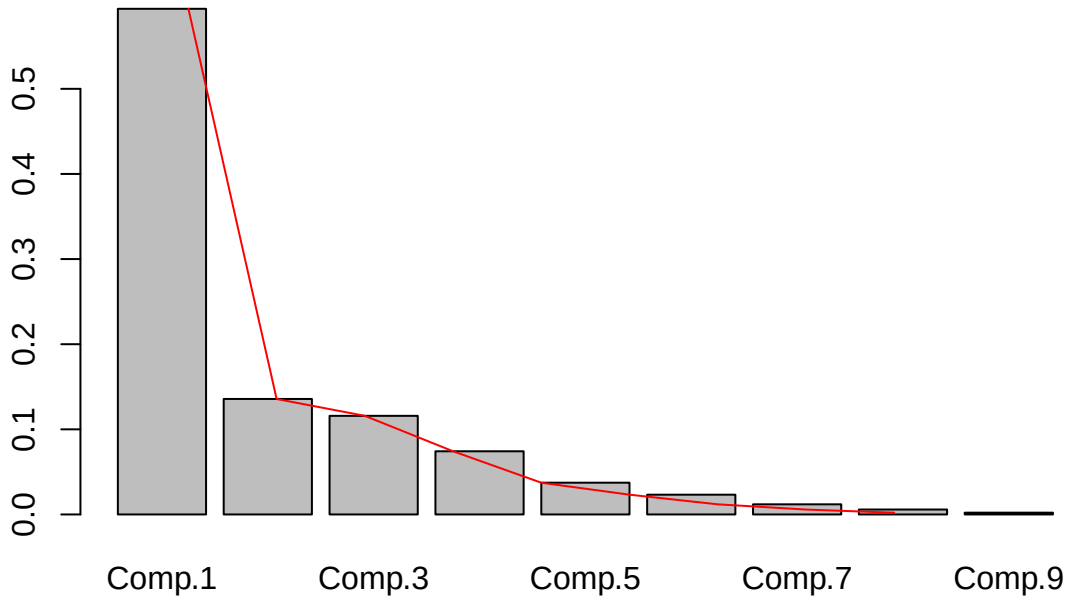
```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.3121220 1.1049470 1.0210443 0.81722905 0.57940716
## Proportion of Variance 0.5939898 0.1356564 0.1158368 0.07420704 0.03730141
## Cumulative Proportion  0.5939898 0.7296462 0.8454831 0.91969010 0.95699151
##                           Comp.6     Comp.7      Comp.8      Comp.9
## Standard deviation     0.45709713 0.32749661 0.228103640 0.137307403
## Proportion of Variance 0.02321531 0.01191711 0.005781252 0.002094814
## Cumulative Proportion  0.98020682 0.99212393 0.997905186 1.000000000
```

```
pca$loadings
```

```
##
```

```
## Loadings:
##     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## X3   0.351  0.403         0.304         0.284  0.726
## X4   0.339  0.298        -0.232 -0.826        -0.223
## X5   0.283  0.569         0.374  0.328 -0.298 -0.488 -0.141
## X6   0.281  0.215        -0.831  0.395         0.111
## X7  -0.399  0.325                                       0.850
## X8                 0.970                0.150 -0.140
## X9  -0.369  0.376 -0.117 -0.114  0.106  0.489 -0.224  0.528 -0.340
## X10 -0.380  0.244  0.162        -0.165 -0.711  0.307  0.212 -0.311
## X11 -0.403  0.259                       0.223        -0.803 -0.248
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
## Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

```
proportions <- pca$sdev^2/sum(pca$sdev^2)
barplot(proportions)
lines(proportions, col = "red")
```



Con este análisis de componentes principales se concluyo que se puede reducir la dimensionalidad de los datos a dos, utilizando los componente 1 y 2 de este análisis. En esta grafica se puede notar que del componente 2 en adelante hay una gran diferencia en la variabilidad explicada por cada componente. Indicando que es viable solo utilizar los componentes 1 y 2

3. Conclusion general

a) ¿Se qué forma te ayuda este nuevo análisis a contestar la pregunta principal del estudio?

Con este analisis podemos detectar cuales son las variables que explican mejor la varianza de los datos para despues utilizar estas variables para predecir la concentracion de mercurio en los lagos.

b) ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?

Los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida son las variables X3, X4, X5, y X6 que son las variables que se podrian utilizar como variables independientes para predecir la concentracion de mercurio en los lagos.

c) ¿En qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas?

En que la seleccion de variables para realizar predicciones u otros tipos de estudios.

d) ¿Cómo te ayudan los componentes principales a abordar este problema?

Al reducir la dimensionalidad de los datos, vuelve mas simple el problema, ya que se mantiene la mayor parte de la varianza de los datos y se reduce la cantidad de variables a utilizar para realizar predicciones u otros tipos de estudios.