

## Tarea 2 - Reconocimiento estadístico de patrones

Erick Salvador Alvarez Valencia

CIMAT A.C.,  
`erick.alvarez@cimat.mx`

**Resumen** En el presente reporte se hablará sobre la resolución de los ejercicios de la primer segunda de reconocimiento de patrones, de la misma manera se presentará el código de los mismos y los resultados que fueron obtenidos.

### 1. Problema 1

Verifica que la entropía de Shannon:

$$H(X) - E_Y(H(Y|X)) = H(Y) - E_X(H(X|Y))$$

osea, la información mutua es simétrica en  $X$  y  $Y$ .

Trabajamos con el lado izquierdo de la ecuación.

$$\begin{aligned}
H(X) - E_Y(H(Y|X)) &= - \sum_{x \in X} P(X=x) \log(P(X=x)) - \sum_{y \in Y} P(Y=y) H(X|Y=y) \\
&= - \sum_{x \in X} P(X=x) \log(P(X=x)) - \sum_{y \in Y} P(Y=y) \left[ - \sum_{x \in X} P(X|Y=y) \log(P(X|Y=y)) \right] \\
&= - \sum_{x \in X} P(Y=y) \log(P(X=x)) + \sum_{y \in Y} P(Y=y) \sum_{x \in X} \frac{P(X,Y)}{P(Y=y)} \log\left(\frac{P(X,Y)}{P(Y=y)}\right) \\
&= - \sum_{x \in X} P(Y=y) \log(P(X=x)) + \sum_{y \in Y} \sum_{x \in X} P(X,Y) [\log(P(X,Y)) - \log(P(Y=y))] \\
&= - \sum_{x \in X} P(Y=y) \log(P(X=x)) + \sum_{y \in Y} \sum_{x \in X} P(X,Y) \log(P(X,Y)) - \sum_{y \in Y} \sum_{x \in X} P(X,Y) \log(P(Y=y)) \\
&= - \sum_{x \in X} P(Y=y) \log(P(X=x)) + \sum_{y \in Y} \sum_{x \in X} P(X,Y) \log(P(X,Y)) - \sum_{y \in Y} P(Y=y) \log(P(Y=y)) \\
&= - \sum_{y \in Y} P(Y=y) \log(P(Y=y)) + \sum_{y \in Y} \sum_{x \in X} P(X,Y) \log(P(X,Y)) - \sum_{x \in X} P(X=x) \log(P(X=x)) \\
&= - \sum_{y \in Y} P(Y=y) \log(P(Y=y)) + \sum_{y \in Y} \sum_{x \in X} P(X,Y) \log(P(X,Y)) - \sum_{x \in X} \sum_{y \in Y} P(X,Y) \log(P(X=x)) \\
&= - \sum_{y \in Y} P(Y=y) \log(P(Y=y)) + \sum_{y \in Y} \sum_{x \in X} P(X,Y) [\log(P(X,Y)) - \log(P(X=x))] \\
&= - \sum_{y \in Y} P(Y=y) \log(P(Y=y)) - \sum_{y \in Y} P(X=x) \left[ - \sum_{x \in X} \frac{P(X,Y)}{P(X=x)} \log\left(\frac{P(X,Y)}{P(X=x)}\right) \right] \\
&= H(Y) - \sum_{x \in X} P(X=x) H(Y|X=x) \\
&= H(Y) - E_X(H(X|Y)) \\
&\quad (1)
\end{aligned}$$

## 2. Problema 2

Verifica la propiedad de la Kurtosis de la página 10 del archivo febr19.pdf.

Si  $X$  y  $Y$  son independientes,  $Kurt(X) = Kurt(Y)$ ,  $||\alpha|| = 1$ :

$$Kurt(\alpha_1 X + \alpha_2 Y)$$

es máxima en  $|\alpha_1| = 1$ ,  $\alpha_2 = 0$  ó  $|\alpha_2| = 1$ ,  $|\alpha_1| = 0$ .

Se parte de la definición dada de la Kurtosis:

$$Kurt(Z) = E(Z - E(Z))^4 - 3Var(Z)^2 \quad (2)$$

Hacemos la sustitución:  $Z = \alpha_1 X + \alpha_2 Y$ , de la misma forma recordemos que  $Var(Z) = E(Z - E(Z))^2$ .

$$\begin{aligned}
Kurt(Z) &= E(\alpha_1 X + \alpha_2 Y - E(\alpha_1 X + \alpha_2 Y))^4 - 3(E[\alpha_1 X + \alpha_2 Y - E(\alpha_1 X + \alpha_2 Y)]^2)^2 \\
&= E(\alpha_1(X - E(X)) + \alpha_2(Y - E(Y)))^4 - 3(E[\alpha_1(X - E(X)) + \alpha_2(Y - E(Y))]^2)^2 \\
&= E[\alpha_1^4(X - E(X))^4 + 4\alpha_1^3(X - E(X))^3\alpha_2(Y - E(Y)) + 6\alpha_1^2(X - E(X))^2\alpha_2^2(Y - E(Y))^2 \\
&\quad + 4\alpha_1(X - E(X))\alpha_2^3(Y - E(Y))^3 + \alpha_2^4(Y - E(Y))^4] - 3(E[\alpha_1^2(X - E(X))^2 \\
&\quad + 2\alpha_1(X - E(X))\alpha_2(Y - E(Y)) + \alpha_2^2(Y - E(Y))^2])^2 \\
&= \alpha_1^4 E[X - E(X)]^4 + 4\alpha_1^3 \alpha_2 E[(X - E(X))^3(Y - E(Y))] + 6\alpha_1^2 \alpha_2^2 E[(X - E(X))^2(Y - E(Y))^2] \\
&\quad + 4\alpha_1 \alpha_2^3 E[(X - E(X))(Y - E(Y))^3] + \alpha_2^4 E[Y - E(Y)]^4 - 3(\alpha_1^2 E[X - E(X)]^2 \\
&\quad + 2\alpha_1 \alpha_2 E[(X - E(X))(Y - E(Y))] + \alpha_2^2 E[Y - E(Y)]^2)^2 \\
&\quad (3)
\end{aligned}$$

En la última expresión notemos que se tiene  $2\alpha_1 \alpha_2 E[(X - E(X))(Y - E(Y))]$  y sabemos que  $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$  pero como  $X$  y  $Y$  son independientes esa covarianza es cero.

$$\begin{aligned}
&= \alpha_1^4 E[X - E(X)]^4 + 4\alpha_1^3 \alpha_2 E[(X - E(X))^3(Y - E(Y))] + 6\alpha_1^2 \alpha_2^2 E[(X - E(X))^2(Y - E(Y))^2] \\
&\quad + 4\alpha_1 \alpha_2^3 E[(X - E(X))(Y - E(Y))^3] + \alpha_2^4 E[Y - E(Y)]^4 - 3(\alpha_1^4 [E(X - E(X))^2]^2 \\
&\quad + 2\alpha_1^2 \alpha_2^2 E(X - E(X))^2 E(Y - E(Y))^2 + \alpha_2^4 [E(Y - E(Y))^2]^2) \\
&= \alpha_1^4 E[X - E(X)]^4 + 4\alpha_1^3 \alpha_2 E[(X - E(X))^3(Y - E(Y))] + 6\alpha_1^2 \alpha_2^2 E[(X - E(X))^2(Y - E(Y))^2] \\
&\quad + 4\alpha_1 \alpha_2^3 E[(X - E(X))(Y - E(Y))^3] + \alpha_2^4 E[Y - E(Y)]^4 - 3\alpha_1^4 [E(X - E(X))^2]^2 \\
&\quad - 6\alpha_1^2 \alpha_2^2 E[(X - E(X))^2(Y - E(Y))^2] - 3\alpha_2^4 [E(Y - E(Y))^2]^2 \\
&\quad (4)
\end{aligned}$$

Aquí podemos eliminar los términos  $6\alpha_1^2 \alpha_2^2 E[(X - E(X))^2(Y - E(Y))^2] - 6\alpha_1^2 \alpha_2^2 E[(X - E(X))^2(Y - E(Y))^2]$  y agrupar otros de tal forma que se formen dos expresiones de Kurtosis.

$$\begin{aligned}
&= \alpha_1^4(E[X - E(X)]^4 - 3[E(X - E(X))^2]^2) + \alpha_2^4(E[Y - E(Y)]^4 - 3[E(Y - E(Y))^2]^2) \\
&\quad + 4\alpha_1^3\alpha_2 E[(X - E(X))^3(Y - E(Y))] + 4\alpha_1\alpha_2^3 E[(X - E(X))(Y - E(Y))^3] \\
&\quad = \alpha_1^4 Kurt(X) + \alpha_2^4 Kurt(Y) + 4\alpha_1^3\alpha_2 E[(X - E(X))^3(Y - E(Y))] \\
&\quad \quad + 4\alpha_1\alpha_2^3 E[(X - E(X))(Y - E(Y))^3] \\
&\quad = \alpha_1^4 Kurt(X) + \alpha_2^4 Kurt(Y) + 4\alpha_1^3\alpha_2 E[(X^3 - 3X^2E(X) + 3XE(X)^2 \\
&\quad - E(X)^3)(Y - E(Y))] + 4\alpha_1\alpha_2^3 E[(X - E(X))(Y^3 - 3Y^2E(Y) + 3YE(Y)^2 - E(Y)^3)] \\
&\quad = \alpha_1^4 Kurt(X) + \alpha_2^4 Kurt(Y) + 4\alpha_1^3\alpha_2 E[X^3Y - X^3E(Y) - 3X^2YE(X) \\
&\quad + 3X^2E(X)E(Y) + 3XYE(X)^2 - 3XE(X)^2E(Y) - E(X)^3Y + E(X)^3E(Y)] \\
&\quad + 4\alpha_1\alpha_2^3 E[XY^3 - 3XY^2E(Y) + 3XYE(Y)^2 - XE(Y)^3 - E(X)Y^3 \\
&\quad + 3Y^2E(X)E(Y) - 3YE(X)E(Y)^2 + E(X)E(Y)^3] \\
&\quad = \alpha_1^4 Kurt(X) + \alpha_2^4 Kurt(Y)
\end{aligned} \tag{5}$$

Hemos llegado a que  $Kurt(\alpha_1 X + \alpha_2 Y) = \alpha_1^4 Kurt(X) + \alpha_2^4 Kurt(Y)$ . Ahora para maximizar lo anterior con respecto a  $\alpha_1, \alpha_2$  respetando la restricción antes mencionada se usarán los multiplicadores de Lagrange. Definimos la función de Lagrange como:

$$\mathcal{L}(\alpha_1, \alpha_2, \lambda) = \alpha_1^4 Kurt(X) + \alpha_2^4 Kurt(Y) - \lambda(\alpha_1^2 + \alpha_2^2 - 1) \tag{6}$$

El gradiente de la función  $\mathcal{L}$  es:

$$\nabla \mathcal{L} = [4\alpha_1^3 Kurt(X) - 2\lambda\alpha_1, 4\alpha_2^3 Kurt(Y) - 2\lambda\alpha_2, -\alpha_1^2 - \alpha_2^2 + 1]^T \tag{7}$$

Iguálamos a cero y despejamos.

$$\begin{aligned}
4Kurt(X)\alpha_1^3 - 2\lambda\alpha_1 &= 0 \\
\alpha_1(4Kurt(X)\alpha_1^2 - 2\lambda) &= 0
\end{aligned} \tag{8}$$

$$\alpha_1 = 0 \text{ ó } 4Kurt(X)\alpha_1^2 - 2\lambda = 0$$

Asumimos  $\alpha_1 = 0$  y sustituimos en la restricción.

$$\begin{aligned}
-2\alpha_2^2 + 1 &= 0 \\
\alpha_2^2 &= 1 \\
|\alpha_2| &= 1
\end{aligned} \tag{9}$$

Al final nos queda  $\alpha_1 = 0$  y  $|\alpha_2| = 1$ . Asumimos  $4Kurt(X)\alpha_1^2 - 2\lambda = 0$ .

$$\begin{aligned} 4Kurt(X)\alpha_1^2 - 2\lambda &= 0 \\ \lambda &= 2Kurt(X)\alpha_1^2 \end{aligned} \quad (10)$$

Sustituir 10 en la segunda ecuación del gradiente.

$$\begin{aligned} 4Kurt(Y)\alpha_2^3 - 2\alpha_2(2Kurt(X)\alpha_1^2) &= 0 \\ 4Kurt(Y)\alpha_2^3 - 4\alpha_2\alpha_1^2Kurt(Y) &= 0 \\ 4Kurt(Y)\alpha_2(\alpha_2^2 - \alpha_1^2) &= 0 \end{aligned} \quad (11)$$

$4Kurt(Y)\alpha_2 = 0$  ó  $\alpha_2^2 - \alpha_1^2$   
Asumimos  $\alpha_2 = 0$  y sustituimos en la restricción.

$$\begin{aligned} -2\alpha_1^2 + 1 &= 0 \\ \alpha_1^2 &= 1 \\ |\alpha_1| &= 1 \end{aligned} \quad (12)$$

Al final nos queda  $\alpha_2 = 0$  y  $|\alpha_1| = 1$ .  
Asumimos  $\alpha_2^2 - \alpha_1^2 = 0$ .

$$\alpha_2^2 - \alpha_1^2 = 0 \quad (13)$$

Sustituimos la ecuación 13 en la restricción y obtenemos:

$$\begin{aligned} -\alpha_1^2 - \alpha_1^2 + 1 &= 0 \\ \alpha_1 &= \pm\sqrt{\frac{1}{2}} \end{aligned} \quad (14)$$

ó

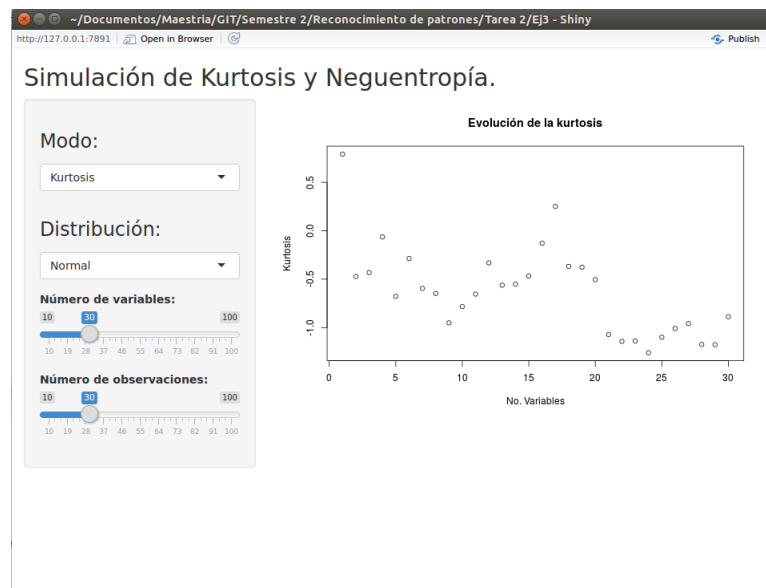
$$\begin{aligned} -\alpha_2^2 - \alpha_2^2 + 1 &= 0 \\ \alpha_2 &= \pm\sqrt{\frac{1}{2}} \end{aligned} \quad (15)$$

Haciendo un análisis análogo vamos a llegar a que tenemos los puntos críticos:  $\alpha_1 = 0$ ,  $|\alpha_2| = 1$  y los mencionados en la ecuación 14.  
Ahora si sustituimos los 3 puntos críticos en la ecuación  $f(X) = (\alpha_1^4 + \alpha_2^4)Kurt(X)$  (Esto porque  $Kurt(X) = Kurt(Y)$ ) vemos que tanto los puntos  $\alpha_1 = 0$ ,  $|\alpha_2| = 1$  y  $|\alpha_1| = 1$ ,  $\alpha_2 = 0$  nos dan como resultado la Kurtosis, mientras que  $\alpha_1 = \alpha_2 = \pm\sqrt{\frac{1}{2}}$  nos da  $\frac{1}{2}Kurt(X)$  lo cual es menor que con los puntos críticos anteriores. Por lo que se ve que los primeros dos puntos críticos son maximizadores.

### 3. Problema 3

Ilustra con algunas simulaciones cómo cambia la kurtosis y negentropía de una suma de v.a. independientes de una cierta distribución cuando se aumenta el número de variables. Considera diferentes distribuciones. Construye una página usando la herramienta shiny que permite jugar con estas simulaciones. Ver <https://shiny.rstudio.com/>, no es necesario usar Rstudio.

Para este ejercicio se usó la herramienta shiny para generar una página web como se ve a continuación:



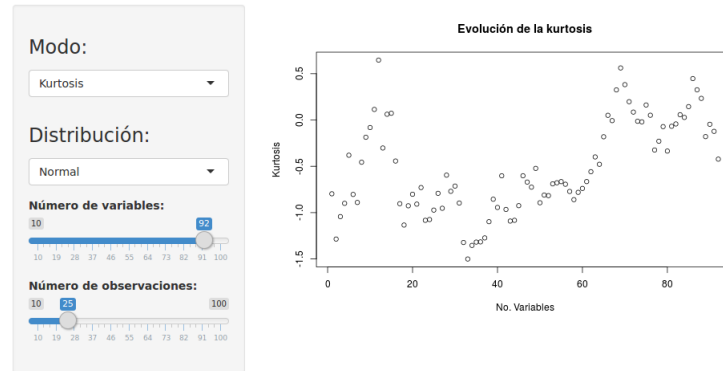
(a) Figura 1. Página web creada con shiny.

En la cual se le pide al usuario introducir un método, una distribución, el número de variables y el número de observaciones para generar un plot ya sea la Kurtosis o la Neguentropía de un conjunto de datos definido por los parámetros antes mencionados.

La página trabaja de la siguiente manera: Se crea un vector vacío con el tamaño de número de observaciones y se suma  $n$  veces con vectores de la distribución indicada por el usuario, por cada suma se calcula la kurtosis o la neguentropía y se almacena en un vector el cual se va a graficar.

Ahora se hablará sobre los resultados observados, primeramente para la Kurtosis.

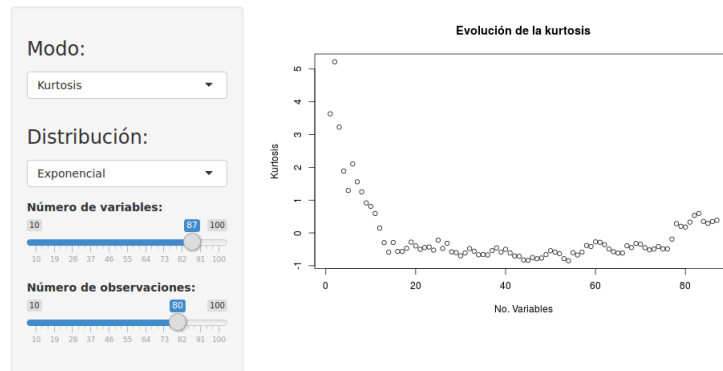
### Simulación de Kurtosis y Neguentropía.



(b) Figura 2. Gráfica de la kurtosis de una distribución normal.

En la Figura 2. se puede ver la kurtosis generada por una distribución normal. Ahora sabemos que la kurtosis mide la no gaussianidad por lo que  $Kurt(X) = 0$  ó  $Kurt_N(X) = 3$  para una distribución normal. En base a lo anterior se puede distinguir que la negentropía oscila muy cerca del cero, y entre más se van sumando vectores normales, esta se acerca cada vez más y más a cero.

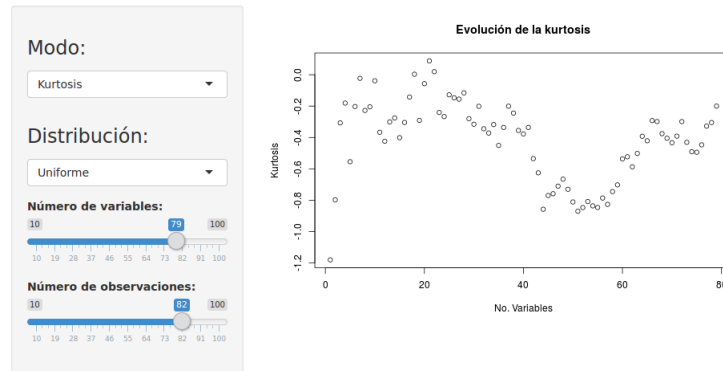
### Simulación de Kurtosis y Neguentropía.



(c) Figura 3. Gráfica de la kurtosis de una distribución exponencial.

En la Figura 3. vemos que para la distribución exponencial su Kurtosis primeramente es alta y a medida que se va sumando los vectores va descendiendo ya que los datos van pareciéndose más y más a una normal.

### Simulación de Kurtosis y Neguentropía.

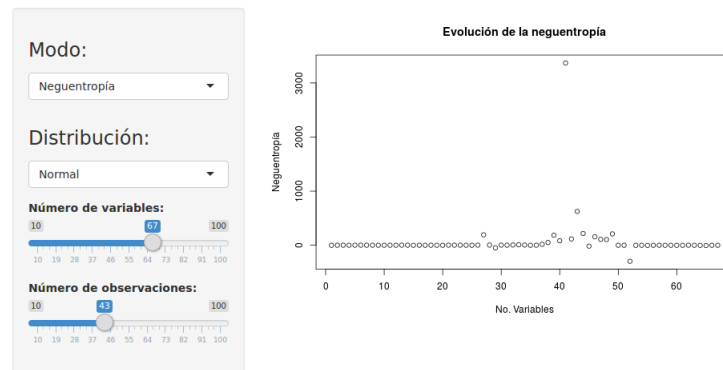


(d) Figura 4. Gráfica de la kurtosis de una distribución uniforme.

Para la Figura 4. que muestra los datos de una distribución uniforme vemos que la Kurtosis oscila por debajo del cero debido a lo mismo que en la distribución exponencial pero una vez va creciendo la suma, la Kurtosis se acerca cada vez más a cero.

Ahora se mostrarán los resultados producidos por la neguentropía.

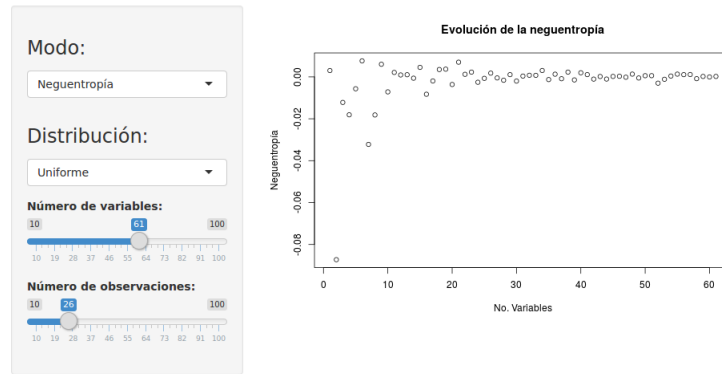
### Simulación de Kurtosis y Neguentropía.



(e) Figura 5. Gráfica de la neguentropía de una distribución uniforme.

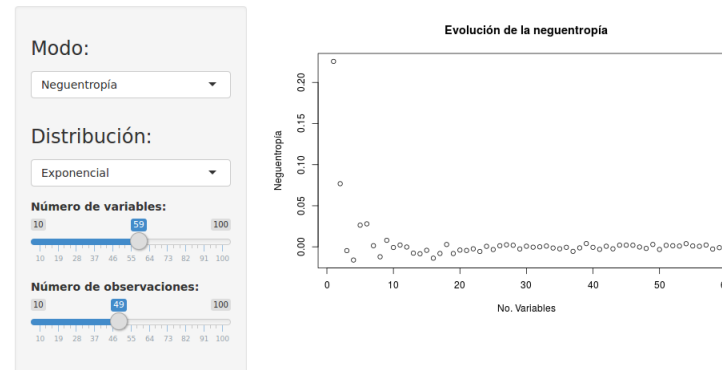


### Simulación de Kurtosis y Neguentropía.



(f) Figura 6. Gráfica de la neguentropía de una distribución uniforme.

### Simulación de Kurtosis y Neguentropía.



(g) Figura 7. Gráfica de la neguentropía de una distribución exponencial.

En las Figuras 5, 6 y 7 se puede apreciar las gráficas de la neguentropía de las 3 distribuciones antes mencionadas, se sabe que esta característica mide la similitud que tiene una distribución con la normal, con valores muy cercanos a cero cuando esto ocurre. Por lo que en la primera gráfica se puede apreciar que la neguentropía siempre está casi a cero ya que los datos son normales. En las últimas dos gráficas vemos que al principio la neguentropía oscila en valores un poco lejanos a cero, pero a medida que se va haciendo la suma de vectores aleatorios y los datos se parecen más a una normal, la neguentropía se va acercando más y más a cero.

A continuación se muestra el código que se hizo para crear la página. Como comentario se requiere instalar los paquetes: *shiny*, *e1071* y *entropy* para correr el programa. Este mismo código se adjuntará con el reporte.

```
library("shiny")
library("e1071")
library("entropy")

ui <- fluidPage(

  # App title ——
  titlePanel("Simulacion_de_Kurtosis_y_Neguentropia."),

  # Sidebar layout with input and output definitions ——
  sidebarLayout(

    # Sidebar panel for inputs ——
    sidebarPanel(
      selectInput("selectMode", label = h3("Modo:"),
                  choices = list("Kurtosis" = 1, "Neguentropia" = 2),
                  selected = 1),

      selectInput("selectDist", label = h3("Distribucion:"),
                  choices = list("Normal" = 1, "Uniforme" = 2, "Exponencial" = 3),
                  selected = 1),

      sliderInput(inputId = "noVars",
                  label = "Numero_de_variables:",
                  min = 10,
                  max = 100,
                  value = 30),

      sliderInput(inputId = "noObs",
                  label = "Numero_de_observaciones:",
                  min = 10,
                  max = 100,
                  value = 30)

    ),

    # Main panel for displaying outputs ——
    mainPanel(

      # Output: Histogram ——
      plotOutput(outputId = "kurtPlot")

    )

  )

)
```

```
)
  )
)
```

```
server <- function(input, output) {
```

```
  output$kurtPlot <- renderPlot({
```

```
    mode <- as.integer(input$selectMode)
    noVars <- as.integer(input$noVars)
    noObs <- as.integer(input$noObs)
    dist <- as.integer(input$selectDist)
```

```
    ind <- seq(1, noVars, by = 1)
    x <- rep(0, noObs)
    y <- rep(0, noVars)
```

```
    if(mode == 1) {
```

```
      for(n in 1:noVars) {
        if(dist == 1) {
          x <- x + rnorm(noObs)
        }
        else if(dist == 2) {
          x <- x + runif(noObs)
        }
        else {
          x <- x + rexp(noObs)
        }
      }
```

```
      y[n] <- kurtosis(x)
    }
```

```
    plot(ind, y, xlab = "No. Variables", ylab = "Kurtosis", title("Evolucion de Kurtosis"))
  }
```

```
  else {
```

```
    for(n in 1:noVars) {
      if(dist == 1) {
        x <- x + rnorm(noObs)
      }
      else if(dist == 2) {
        x <- x + runif(noObs)
      }
      else {
        x <- x + rexp(noObs)
      }
    }
```

```

    }

    aux <- rnorm(noObs, mean = mean(x), sd = sd(x))
    y[n] = entropy(aux) - entropy(x)
  }

  plot(ind, y, xlab = "No. Variables", ylab = "Neguentropia", title("Evoluci
})

})
}
shinyApp(ui = ui, server = server)

```

#### 4. Problema 4

Supongamos que  $Y$  es una transformación lineal de  $X$ , es decir, existe una matrix  $M$  tal que  $Y = MX$ . Si  $X$  y  $Y$  son v.a. continuas y  $M$  invertible, verifica que  $H(Y) = H(X) + \log(|\det(M)|)$

Tenemos que la entropía de  $Y$  se define como:

$$H(Y) = - \int_I f_Y \log(f_Y) dy \quad (16)$$

Con  $I = (-\infty, \infty)$ . Ahora como  $Y = MX$  podemos verlo como un vector  $Y = [Y_1, Y_2, \dots, Y_M]^T$ . De la misma forma  $X$  se define como  $X = M^{-1}Y$ , esto anterior existe porque  $M$  es invertible. Por lo tanto  $X = [X_1, X_2, \dots, X_m]^T$ .

Se tiene la propiedad siguiente:

$$f_Y(Y_1, Y_2, \dots, Y_m) = \frac{1}{|\frac{\partial(Y_1, Y_2, \dots, Y_m)}{\partial(X_1, X_2, \dots, X_m)}|} f_X(h^{-1}(Y_1, Y_2, \dots, Y_m)) \quad (17)$$

De lo anterior podemos deducir que, como  $Y$  es una transformación lineal entonces su Jacobiano es  $M$ .

Ahora denotamos la entropía de  $Y$  en términos de  $X$  con la propiedad antes mencionada.

$$H(Y) = - \int_I \frac{1}{|\det(M)|} f_X(M^{-1}(Y_1, Y_2, \dots, Y_m)) \log\left(\frac{f_X(M^{-1}(Y_1, Y_2, \dots, Y_m))}{|\det(M)|}\right) dy \quad (18)$$

Haciendo  $[X_1, X_2, \dots, X_m]^T = M^{-1}[Y_1, Y_2, \dots, Y_m]^T$  tenemos que  $dy = |\det(M)|dx$  por lo que:

$$\begin{aligned}
H(Y) &= - \int_I f_X \log\left(\frac{f_X}{|det(M)|}\right) dx \\
H(Y) &= - \int_I f_X (\log(f_X) - \log(|det(M)|)) dx \\
H(Y) &= - \int_I f_X \log(f_X) + \int_I f_X \log(|det(M)|) dx \\
H(Y) &= H(X) + \log(|det(M)|) dx
\end{aligned} \tag{19}$$

## 5. Problema 5

En <http://worldhappiness.report/> se puede consultar el reporte de 2017 de la Organización de las Naciones Unidas (UN) sobre la felicidad en diferentes países. Toma los datos que se pueden acceder en esta página donde dice Chapter 2. Online Data de 2016. En Chapter 2. Statistical Appendix se definen los diferentes indicadores. Usa las herramientas que vimos en clase como ISOMAP, SOM y PCA de R para encontrar visualizaciones informativas de este conjunto de datos. El resultado debe ser un reporte de unas 3-4 páginas. No se trata de generar el máximo número de gráficas sino incluir solamente las que son informativas y discutir las.

Los datos corresponden a un nivel de medición sobre la felicidad de las personas de un país, aquí podemos observar varias características que influyen en este comportamiento, como son: esperanza de vida, libertad de toma de decisiones, percepción de corrupción, etc.

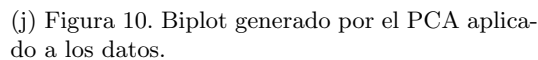
Primeramente se importaron los datos del `xlsx` y se hizo el filtrado de quitar los datos que no correspondían al año 2016, una vez esto se notó que había muchas columnas que tenían pocos datos, incluso algunas que no tenían. Aquí se pudieron haber tomado muchos caminos, uno de ellos rellenar los campos sin datos con las medias de la columna correspondiente. Otro camino a seguir y de hecho es el que se tomó, quitar las columnas que tienen más del 90 % de espacios vacíos ya que vemos que estas columnas no aportan mucho a la información general. Finalmente se quitaron las columnas relacionadas al número de año de los datos y a la que indica el nombre de los países.

Ahora, lo primero que se hizo fue PCA, esto para ver qué componentes aportaban más variabilidad a los datos, o dicho de otra forma: Qué características hacen que las personas sean más felices y cuáles no.

A continuación mostramos los primeros resultados de aplicar PCA a los datos filtrados.

(h) Figura 8. Resumen del PCA aplicado a los datos.

(i) Figura 9. Resumen del PCA aplicado a los datos.



En las Figuras 8, 9 y 10 mostradas podemos ver los resultados que nos generó el PCA. Lo primero que notamos es que el primer componente principal aporta casi toda la variabilidad de los datos, y dicha variabilidad corresponde a la

característica: *Esperanza de vida al nacer*, de la misma forma podemos ver que hay características que aportan muy poca variabilidad, tales como: La desviación estándar media de la percepción de la felicidad, etc. Esto anterior nos sugiere que dicha característica es la causante de que más personas sean felices, aunque viendo en general los datos se puede apreciar que hay más características posiblemente importantes en la felicidad de las personas tales como la libertad de toma de decisiones o el PIB de su país, por lo que podemos suponer que el PCA está siendo sesgado por la esperanza de vida. Para ello volveremos a aplicar el método usando los datos normalizados con la matriz de correlación.

A continuación mostramos los resultados del PCA normalizado.

```

Loadings:
               Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9  Comp.10
Life.Ladder      -0.366      -0.217 -0.230      -0.111 -0.269  0.160  0.564
Log.GDP.per.capita -0.347      -0.248 -0.294  0.230  0.242  0.101  0.160  0.130
Social.support    -0.334      -0.272  0.119 -0.197  0.397      -0.545 -0.398 -0.372
Healthy.life.expectancy.at.birth -0.348  0.134 -0.136 -0.336      -0.452  0.207 -0.401
Freedom.to.make.life.choices -0.216 -0.445 -0.136  0.201  0.200 -0.547  0.526 -0.147      -0.226
Generosity        -0.466  0.357 -0.474 -0.610      0.184
Perceptions.of.corruption  0.215  0.335 -0.420  0.104 -0.598 -0.183  0.111      0.451 -0.148
Positive.affect    -0.210 -0.429 -0.315  0.308 -0.220 -0.100 -0.561  0.415 -0.153
Negative.affect     0.291      -0.226 -0.587  0.157 -0.408 -0.383 -0.248 -0.187 -0.283
Standard.deviation.of.ladder.by.country.year  0.295 -0.168 -0.552 -0.130      0.159  0.324  0.118 -0.233  0.403
Standard.deviation.Mean.of.ladder.by.country.year  0.378 -0.121 -0.117      0.230  0.210  0.251 -0.256 -0.197
gini.of.household.income.reported.in.Gallup..by.wp5.year  0.243 -0.460      0.217  0.416 -0.185 -0.258  0.607 -0.154

               Comp.11  Comp.12
Life.Ladder      -0.186  0.537
Log.GDP.per.capita  0.740 -0.109
Social.support
Healthy.life.expectancy.at.birth -0.543
Freedom.to.make.life.choices
Generosity
Perceptions.of.corruption  0.137
Positive.affect
Negative.affect
Standard.deviation.of.ladder.by.country.year -0.233 -0.383
Standard.deviation.Mean.of.ladder.by.country.year  0.127  0.734
gini.of.household.income.reported.in.Gallup..by.wp5.year

               Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9  Comp.10  Comp.11  Comp.12
SS Loadings      1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var   0.083  0.083  0.083  0.083  0.083  0.083  0.083  0.083  0.083  0.083  0.083  0.083
Cumulative Var   0.083  0.167  0.250  0.333  0.417  0.500  0.583  0.667  0.750  0.833  0.917  1.000

```

(k) Figura 11. Resumen del PCA normalizado aplicado a los datos.

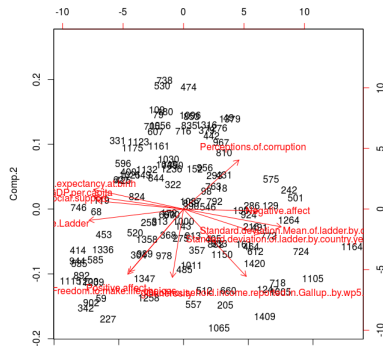
```

> summary(PCA)
Importance of components:
               Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9  Comp.10
Standard deviation  2.4452888  1.3849737  1.00047438  0.85170530  0.8482299  0.6856725  0.62841911  0.52034474  0.47281863  0.41365685
Proportion of Variance  0.4982865  0.1598460  0.08341242  0.06045016  0.0588322  0.0391789  0.03290921  0.02256322  0.01862979  0.01425933
Cumulative Proportion  0.4982865  0.6581325  0.74154489  0.80199505  0.8608272  0.9008061  0.93291536  0.95547858  0.97410837  0.98836770

               Comp.11  Comp.12
Standard deviation  0.33792183  0.159362501
Proportion of Variance  0.00951593  0.002116367
Cumulative Proportion  0.99788363  1.000000000

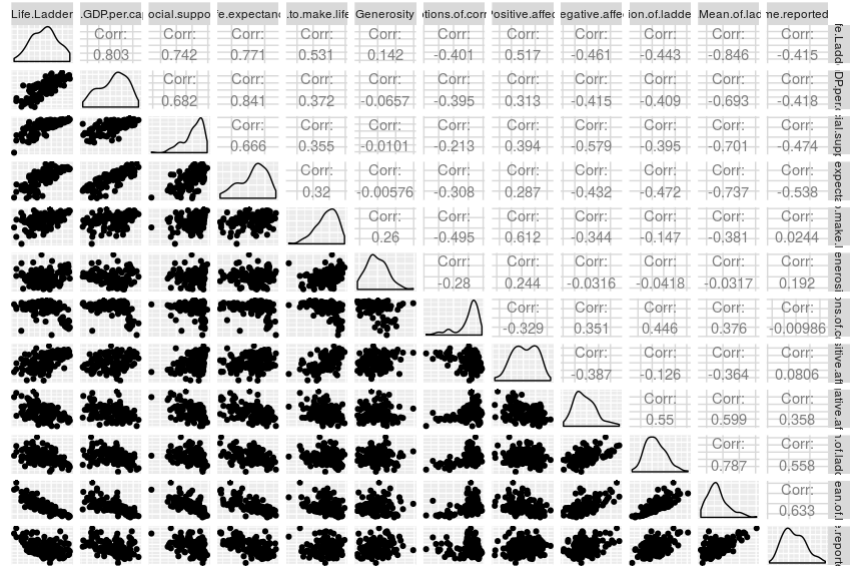
```

(l) Figura 12. Resumen del PCA normalizado aplicado a los datos.



(m) Figura 13. Biplot generado por el PCA normalizado aplicado a los datos.

En las tres Figuras anteriores podemos ver que usando la matriz de correlación en lugar de la de covarianza, la variabilidad de los componentes no se ve sesgada hacia la característica de la esperanza de vida al nacer. Sino que hay muchas más características que influyen en la felicidad de las personas. Este comportamiento se puede ver fácilmente si hacemos una gráfica de GGPairs la cual analiza la correlación de todas las características usadas.



(n) Figura 14. GGPairs de los datos.

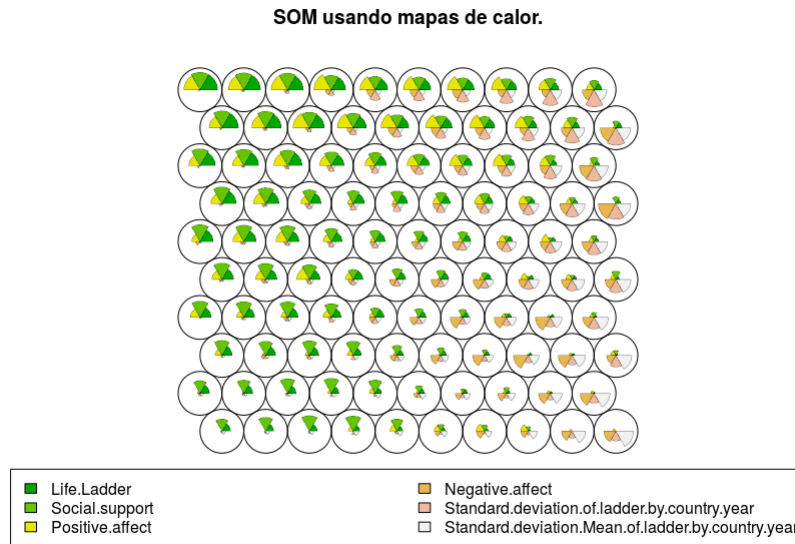
En la Fig 14. podemos ver que mucha correlación entre las variables de los datos. Se puede notar que varias de las características como el afecto negativo



o la corrupción generan correlación negativa, osea menos felicidad, y por el contrario, la generosidad o el afecto positivo generan correlación positiva que ayuda al aumento de la felicidad.

Otro tipo de análisis que se realizó a los datos fue generar (SOM) o Self-Organizing Maps los cuales son una técnica no supervisada que agrupa los datos según alguna relación entre sus características. La representación que se usó fue el mapa de códigos, los cuales se componen de valores normalizados de las variables originales utilizadas para generar el SOM. El vector de peso de cada nodo es representativo / similar de las muestras mapeadas a ese nodo. Al visualizar los vectores de ponderación en el mapa, podemos ver patrones en la distribución de muestras y variables.

A continuación mostramos el SOM generado a los datos, para el cual se usó como característica en común la distancia euclideana.



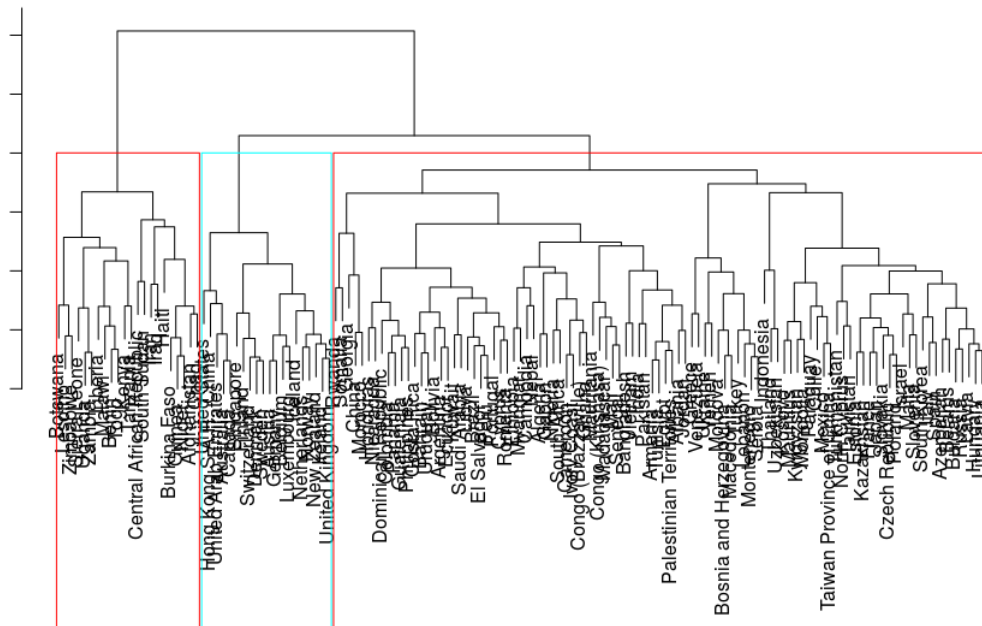
(ñ) Figura 15. GGPairs de los datos.

En la Figura 15 podemos notar varias características marcadas, como son la percepción de la felicidad, ayuda social, afecto positivo, etc. Algo que se notó fue que al realizar varias veces este método siempre se resaltan estas características, aunque no estén en la misma posición, por lo cual podemos notar que esta técnica nos sugiere que mucha de la información está relacionada con estas características.

Como comentario, el Life Ladder o la percepción de la felicidad es una medida que se genera encuestando a las personas sobre qué tan felices creen que son en su vida, por lo que esta característica podría verse como subjetiva aunque

si es una encuesta aplicada a muchas personas referente a varias clases sociales si podemos llegar a resultados más cuerdos ya que no estamos sesgando la información a posiblemente una clase social. Viendo esta característica en los datos, podemos notar que Noruega es el país donde las personas se consideran más felices, por el contrario Afghanistan es el país menos feliz. Otro comentario es que se trató de plotear un SOM como Heat Map resaltando alguna característica en especial pero no se llegó a ninguna conclusión para ello.

Un último método que se aplicó fue el dendograma, el cual es un algoritmo no supervisado que genera un árbol usando alguna distancia de la información. A continuación se muestra el resultado de un dendograma con 3 grupos generados.



(o) Figura 16. Dendograma de los datos.

Para el dendograma mostrado en la Figura 16. se usó la distancia euclídeana, se puede ver que denota los grupos donde podríamos considerar los países más felices o menos felices desde el punto económico, ya que en un grupo encontramos a países como el Reino Unido, Suecia, Estados Unidos, etc. y en otro grupo denotamos a países como África Central, Haití, Nigueria, etc. En el tercer grupo encontramos al resto de los países, los cuales probablemente los veamos como ni tan felices pero ni tan infelices. Se generaron más dendogramas con diferente número de grupos para ver si se encontraba alguna característica

relevante pero no fue el caso.

A continuación se muestra el código que se usó para generar todas las gráficas mostradas anteriormente.

Como comentario se requiere instalar los paquetes: *gdata*, *GGally* y *kohonen* para correr el programa. Este mismo código se adjuntará con el reporte.

```
setwd("Directorio_donde_se_encuentran_los_datos")
library("gdata")
library("GGally")
library("kohonen")

data <- read.xls("data.xlsx")
f_data <- data[data$year == 2016, ]
f_data[[1]] <- NULL
f_data[[2]] <- NULL

sn <- f_data[, colSums(is.na(f_data)) < (1 - 0.9) * nrow(f_data)]
sn <- sn[, -1]
print(f_data)

#Hacer PCA
PCA <- princomp(~., sn, na.action = na.exclude, cor = T)
loadings(PCA)
summary(PCA)
biplot(PCA)
ggpairs(sn, diag = list(continuous = "density"), axisLabels = "none")

#SOM
data_train_matrix <- data.matrix(sn)
som_grid <- somgrid(xdim = 10, ydim = 10, topo = "hexagonal")
som_model <- supersom(data_train_matrix,
                      grid=som_grid,
                      rlen=100,
                      alpha=c(0.05, 0.01),
                      keep.data = TRUE )

plot(som_model, type = "codes", main = "SOM_usando_mapas_de_calor.")
plot(som_model, type = "property", property = getCodes(som_model)[,4], main=colnames(som_model$codes))

#Dendograma
rownames(sn) <- f_data$country
scale = scale(sn)
d = dist(scale, method = "euclidean") # definimos la distancia
fit = hclust(d) # definimos el metodo para hacer clusters
```

```
# Generamos el dendograma  
plot(fit , main="")  
rect.hclust(fit , 3, border=rainbow(2))
```