Tarea 2, Reconocimiento de Patrones

Francisco Javier Peralta Ramírez

1. Verifica que para la entropía de Shannon:

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

es decir, la información mutua es simétrica en X y Y.
Tomando en cuenta la definición de la entropía de Shannon como:

$$H(X) = E[-logP(X)]$$

Y recordamos el teorema de Bayes

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

$$\begin{aligned}
H(X) - E_Y[H(X|Y)] &= H(X) - E_Y E_X(-log(P(X|Y))) \\
&= H(X) - E_Y E_X(log[P(Y|X)P(X)/P(Y)]) \\
&= H(X) - E_Y E_X(log[P(Y)] - log[P(Y|X)] - log[P(X)]) \\
&= H(X) - E_Y[E_X log[P(Y)]) - E_X(log[P(Y|X)]) - E_X(log[P(X)])] \\
&= H(X) - E_Y(log[P(Y)] + E_Y E_X(log[P(Y|X)]) - E_Y H(X) \\
&= H(X) + H(Y) - H(Y|X) - H(X) \\
&= H(Y) - H(Y|X)
\end{aligned}$$

2. Verifica que 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 $|\alpha 2| = 1, \alpha 1 = 0$

Recordamos la definición de Kurtosis $Kurt(X) = E(X - EX)^4 - 3Var(X)^2$, y las propiedades de la varianza $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$ cuando X y Y son independientes.

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

Derivamos con respecto α_1 y α_2 y recordamos que $E(X - EX) = 0$

$$\begin{aligned}
\frac{\delta}{\delta\alpha_1} &= 4\alpha_1^3 E(X - EX)^4 + 12\alpha_1^2\alpha_2 E[(X - EX)^3(Y - EY)] + 6\alpha_1\alpha_2^2 E[(X - EX)^2(Y - EY)^2] \\
&+ 4\alpha_2^3 E[(X - EX)(Y - EY)^3] - 12[\alpha_1^2 Var(X) + \alpha_2^2 Var(Y)]\alpha_1 Var(X) \\
&= 4\alpha_1^3[E(X - EX)^4 - 3Var(X)^2] + 12\alpha_1^2\alpha_2 E(X - EX)^3 E(Y - EY) \\
&+ 6\alpha_1\alpha_2^2[E(X - EX)^2 E(Y - EY)^2 - 2Var(Y)Var(X)] + 4\alpha_2^3 E(X - EX)E(Y - EY)^3 \\
&= 4\alpha_1^3 Kurt(X) + 6\alpha_1\alpha_2^2[Var(X)Var(Y) - 2Var(X)Var(Y)] \\
&= 4\alpha_1^3 Kurt(X) - 12\alpha_1\alpha_2^2 Var(X)Var(Y)
\end{aligned}$$

$$
\begin{aligned}
\frac{\delta}{\delta\alpha_2} &= 4\alpha_1^3 E(Y-EY)^4 + 12\alpha_1^2\alpha_2 E[(Y-EY)^3(X-EX)] + 6\alpha_2\alpha_1^2 E[(Y-EY)^2(X-EX)^2] \\
&\quad + 4\alpha_1^3 E[(Y-EY)(X-EX)^3] - 12[\alpha_2^2 Var(Y) + \alpha_1^2 Var(X)]\alpha_2 Var(Y) \\
&= 4\alpha_2^3[E(Y-EY)^4 - 3Var(Y)^2] + 12\alpha_2^2\alpha_1 E(Y-EY)^3 E(X-EX) \\
&\quad + 6\alpha_2\alpha_1^2[E(Y-EY)^2 E(X-EX)^2 - 2Var(X)Var(Y)] + 4\alpha_1^3 E(Y-EY)E(X-EX)^3 \\
&= 4\alpha_2^3 Kurt(Y) + 6\alpha_2\alpha_1^2[Var(Y)Var(X) - 2Var(Y)Var(X)] \\
&= 4\alpha_2^3 Kurt(Y) - 12\alpha_2\alpha_1^2 Var(Y)Var(X)
\end{aligned}
$$

Igualamos para encontrar la solución, llamemos $Kurt(X)$, $Kurt(Y)$ como $\boldsymbol{K}$

$$
\begin{aligned}
4\alpha_1^3\boldsymbol{K} - 12\alpha_1\alpha_2^2 Var(X)Var(Y) &= 4\alpha_2^3\boldsymbol{K} - 12\alpha_2\alpha_1^2 Var(X)Var(Y) \\
4\alpha_1^3\alpha_2^3\boldsymbol{K} - 12\alpha_1\alpha_2^5 Var(X)Var(Y) &= 4\alpha_2^3\alpha_1^3\boldsymbol{K} - 12\alpha_2\alpha_1^5 Var(X)Var(Y) \\
\alpha_1\alpha_2^5 &= \alpha_2\alpha_1^5 \\
\alpha_1\alpha_2(\alpha_2^4 - \alpha_1^4) &= 0
\end{aligned}
$$

Obtenemos los puntos (1, 0), (0, 1), (-1, 0), (0, -1), (1/2, 1/2), (-1/2, 1/2), (1/2, -1/2) y (-1/2, -1/2). Para asegurar si son minimos o máximos tomamos las segundas derivadas parciales

$$
\begin{aligned}
\frac{\delta}{\delta\alpha_1^2} &= 12\alpha_1^2 Kurt(X) - 12\alpha_2^2 Var(X)Var(Y) \\
\frac{\delta}{\delta\alpha_1\alpha_2} &= -24\alpha_1\alpha_2 Var(X)Var(Y) \\
\frac{\delta}{\delta\alpha_2^2} &= 12\alpha_2^2 Kurt(Y) - 12\alpha_1^2 Var(X)Var(Y)
\end{aligned}
$$

Evaluamos la matriz Hessiana en cada punto

$$\nabla^2 Kurt(0X+Y)) = \begin{pmatrix} -12Var(X)Var(Y) & 0 \\ 0 & 12Kurt(Y) \end{pmatrix} \rightarrow definida\ neg \rightarrow max$$

$$\nabla^2 Kurt(0X-Y)) = \begin{pmatrix} -12Var(X)Var(Y) & 0 \\ 0 & 12Kurt(Y) \end{pmatrix} \rightarrow definida\ neg \rightarrow max$$

$$\nabla^2 Kurt(X+0Y)) = \begin{pmatrix} 12Kurt(X) & 0 \\ 0 & -12Var(X)Var(Y) \end{pmatrix} \rightarrow definida\ neg \rightarrow max$$

$$\nabla^2 Kurt(-X+0Y)) = \begin{pmatrix} 12Kurt(X) & 0 \\ 0 & -12Var(X)Var(Y) \end{pmatrix} \rightarrow definida\ neg \rightarrow max$$

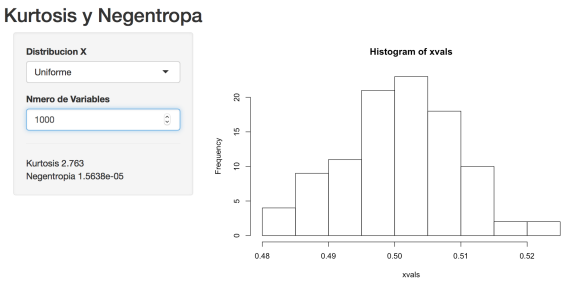
$$\nabla^2 Kurt(X/2+Y/2)) = \begin{pmatrix} 3[Kurt(X) - Var(X)Var(Y)] & -6Var(X)Var(Y) \\ -6Var(X)Var(Y) & 3[Kurt(Y) - Var(X)Var(Y)] \end{pmatrix} \rightarrow min\ o\ silla$$

3. Tomando una suma de n variables aleatorias de cierta distribución, podemos ver el comportamiento de su Kurtosis y Negentropía. Dado que la Negentropía y la Kurtosis miden la "no gaussianindad", esperaríamos ver que ambos valores decrementan conforme se incrementa el número de variables. Se creó una app en *Shiny* y se utilizaron los paquetes de *moments* y *entropy*. La función *kurtosis* regresa el valor $Kurt_N = \frac{E(X-EX)^4}{Var(X)^2}$ por lo que el mínimo se encuentra en $Kurt_N = 3$.

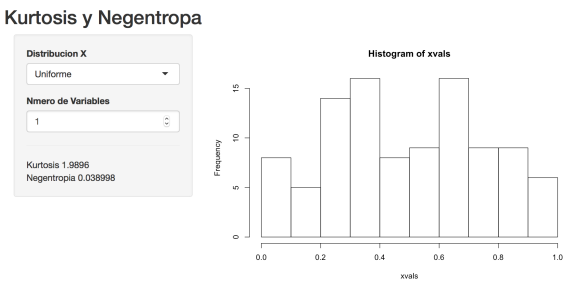
   Los resultados que se pueden ver es que la Negentropía decrementa hasta llegar casi a cero, si el número de variables aleatorias fuerna infinito este llegaría por completo a cero. Lo mismo pasa con la Kurtosis, pero el valor se acerca cada vez más a 3. Esto es independiente de la distribución original, se probó con una distribución uniforme, exponencíal y normal. La normal al ya ser normal tiene valores más cercanos a los minimos.

```
library(moments)
library(entropy)
dists <- c('Uniforme', 'Normal', 'Exponential')
ui <- fluidPage(
  titlePanel('Kurtosis y Negentropa'),
  sidebarPanel(
    selectInput('xcol', 'Distribucion X', dists),
    numericInput("nvars", "N mero de Variables", 1, min=1, max=1000, step=1),

    tags$hr(),
    textOutput('kurt'),
    textOutput('negentropia')
    ),
  mainPanel(
    plotOutput('plot1')
    )
  )
server <- function(input, output) {
  nv = 100
  xdist <- reactive({
    y <- seq(0, 0, length=nv)
    switch( input$xcol,
      "Normal" = {
        for (i in 1:input$nvars)
          y = y + rnorm(nv)
      },
      "Exponential" = {
        for (i in 1:input$nvars)
          y = y + rexp(nv)
      },
      "Uniforme" = {
        for (i in 1:input$nvars)
          y = y + runif(nv)
      }
      )
    return(y/input$nvars)
  })
  output$plot1 <- renderPlot({
    xvals <-xdist()
    hist(xvals, freq = TRUE, breaks=nv/10)
  })
  output$kurt <- renderText({
    xvals <-xdist()
    sprintf("Kurtosis %0.5g", kurtosis(xvals))
  })
  output$negentropia <- renderText({
    xvals <-xdist()
    v = var(xvals)
    m = mean(xvals)
    n = rnorm(nv, mean=m, sd=sqrt(v))
    e = entropy(n) - entropy(xvals)
    sprintf("Negentropia %0.5g", e)
  })
}
shinyApp(ui = ui, server = server)
```



(a) Suma de 1000 v.a Unif



(b) Una v.a Unif

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 $H(Y) = H(X) + log(|det(M)|)$

$$H(Y) = E[-logP(Y)] = E[-logP(MX)]$$

Si Y es una transformación de X, $Y = h(X)$

$$f_Y(y) = \frac{f(x)\delta x}{|h'(x)|\delta x} = \frac{f_X(h^{-1}(y))}{|h'(h^{-1}(y))|}$$

La densidad de $Y = (y_1, y_2, \cdots, y_n)$ puede ser calculada como

$$f_Y(y_1, y_2, \cdots, y_n) = \frac{1}{|\frac{\delta(y_1,\cdots,y_n)}{\delta(x_1,\cdots,x_n)}|} f_X(h^{-1}(x_1, x_2, \cdots, x_n))$$

En este caso, como Y es una transformación lineal de X.

$$f_Y(Y) = f_Y(y_1, y_2, \cdots, y_n) = \frac{1}{|det\, M|} f_X(M^{-1}(x_1, x_2, \cdots, x_n))$$

Y tomando en cuenta $f_Y(y_1, \cdots, y_n)\delta(y1, \cdots, y_n) = fX(x1, \cdots, x_n)\delta(x1, \cdots, x_n)$
Con estas propiedades en cuenta, observamos

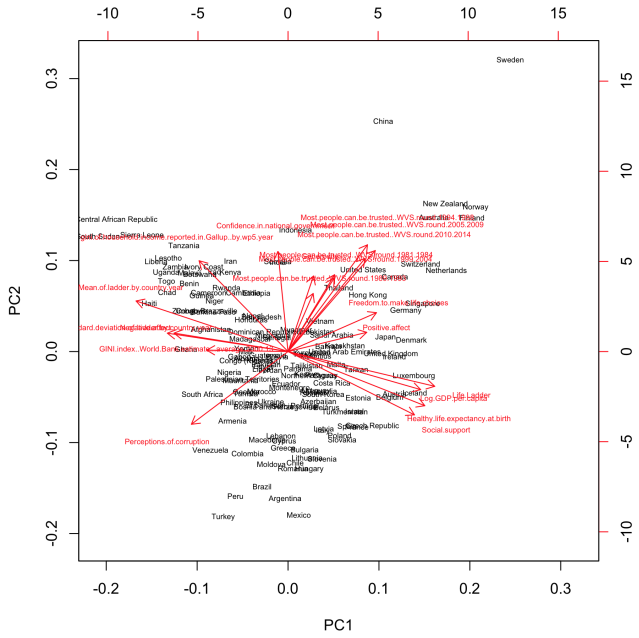
$$\begin{aligned}
H(Y) &= E[-logP(Y)] = E[-logP(MX)] \\
H(Y) &= -\int log[f_Y(Y)] f_Y(Y)\delta y \\
&= -\int log\left(\frac{f_X(M^{-1}Y)}{|det\, M|}\right)\frac{f_X(M^{-1}Y)}{|det\, M|}\delta y \\
&= -\int (log[f_X(M^{-1}Y)] - log|det\, M|)\frac{f_X(M^{-1}Y)}{|det\, M|}\delta y \\
&= -\int log[f_X(X)] f_X(X)\delta x + \int log|det\, M| f_X(X)\delta x \\
&= H(X) + log|det\, M| \int f_X(X)\delta x = H(X) + log(|det\, M|)
\end{aligned}$$

5. Tomando los datos del world report happiness de la ONU del 2017. Para los valores no existentes, llenamos con el promedio sobre la columna del valor y eliminando las columnas que sólo tienen 'NA'

   Primero hacemos análisis de componente preincipal. Notamos de forma rápida que PCA probablemente no es el mejor camino, ya que requerimos de muchos compoentes para llegar a proporción de varianza aceptable (≥ 0.95)

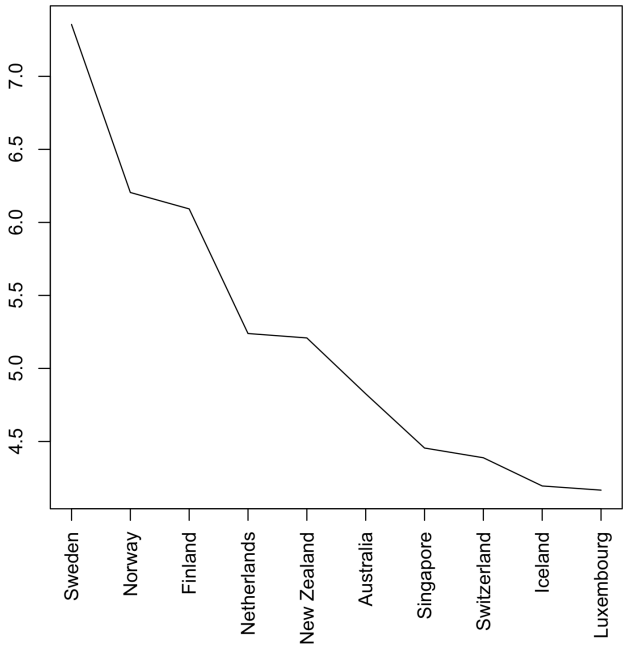
```
> summary(p)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5334 1.7413 1.4562 1.14987 1.06414 0.93976 0.86740
Proportion of Variance 0.3209 0.1516 0.1060 0.06611 0.05662 0.04416 0.03762
Cumulative Proportion  0.3209 0.4725 0.5785 0.64465 0.70127 0.74543 0.78305
                          PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.81791 0.81196 0.7416 0.68680 0.61006 0.54157 0.5348
Proportion of Variance 0.03345 0.03296 0.0275 0.02358 0.01861 0.01466 0.0143
Cumulative Proportion  0.81650 0.84946 0.8770 0.90054 0.91915 0.93381 0.9481
                          PC15    PC16    PC17    PC18    PC19    PC20
Standard deviation     0.50164 0.48404 0.47581 0.40610 0.36823 0.15777
Proportion of Variance 0.01258 0.01171 0.01132 0.00825 0.00678 0.00124
Cumulative Proportion  0.96070 0.97241 0.98373 0.99198 0.99876 1.00000
```

   Aprovechando que tenemos estos resultados podemos graficar el biplot sobre los dos primeros componentes principales, aun que estos sólo cuentan con el 47 % de la varianza total, quizá podemos comenzar a ver ciertos patrones.



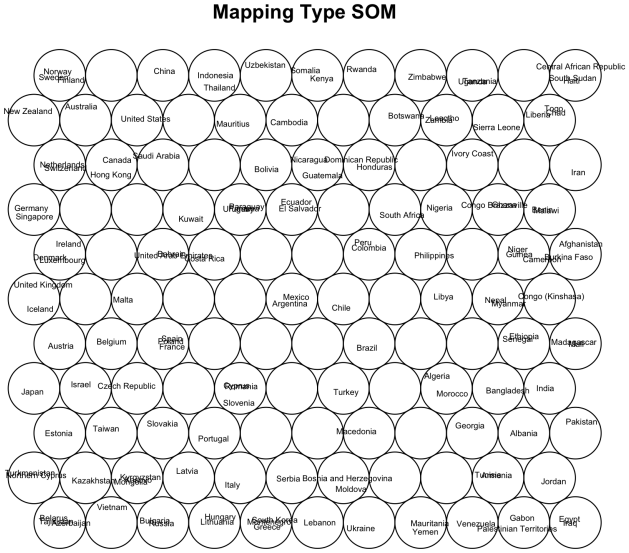
   En general parecería la grágfica tiene algo de sentido, pero se puede interpretar que paises como China tienen niveles de felicidad muy altos, lo cual no va del todo con nuestra percepción del país. Por otra parte se ve que muchos paises latino americanos están en la parte baja de la gráfica, y considerando la situación es varios de estos paises es de esperarse. Al mismo tiempo podemos ver que algunas variables tienen sentido en como afectan a la felicidad. Altos niveles de percepción de corrupción están ligados a una posición baja en la gráfica mientras que confianza en el govierno y libertad de elecciones de vida van de la mano con una posición alta en la gráfica.

   Podemos ordenar los datos sobre el primer componente princial y graficar sólamente la primera componente principal vs pais. Esto nos daría una idea de los paises más felices.



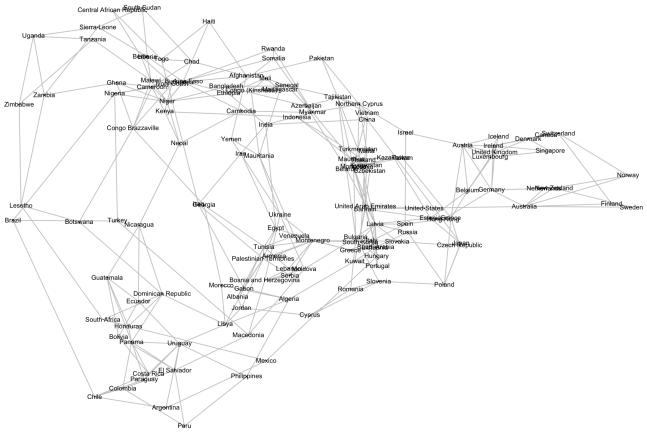
Dado que los primeros dos componentes principales sólo represetan el 45 % de la varianza no es muy bueno tratar de interpretar los resultados.

Usando Self-Organizing Maps, podríamos tener una mejor idea de la similitud entre los datos. SOM nos permite reducir la dimensión creando *cubetas* donde ponemos los datos más similares entre si, y estas *cubetas* están cercanas dependiendo de la similitud entre ellas. Puede haber varias *cubetas* vacias, lo que nos permite ver más claramemte separaciones en los datos. Con SOM no podemos obtener información de que paises son más felices, pero sí sobre su similitud.



Podemos ver que hay grupos de paises erupeos, latinos y africanos. Interesantemente SOM al igual que PCA pone muy cerca a Noruega y USA de China. Esto otra vez contradice la precepsión que se tiene de los paises, pero nos indica que posiblemente son más similares, en cuanto a felicadad de los ciudadanos, a lo que pensabamos. Es bastante interesante que Alemania y Singapore se encuentren en el mismo bloque aun cuando los paises son económicamente hablando, completamente diferentes. Otra vez, podemos probar con otro método de reducción de dimensión para comprobar los resultados y tener una aún mejor idea de lo que está pasando.

Con ISOMAP se genra un grafo donde sólo están conectados los datos más similates entre si. Generamos la gráfica y podemos ver que nuevamente los paieses que esperaríamos que esten cercanos, lo están. Y esta vez China aparece lejano a Noruega y Suecia, lo que nos indica que no es del todo tan similar como los otros métodos nos podrían hacer pensar.



Como vimos, cada método nos dio una visualización muy diferente que se pueden interpretar de muchas formas. Es dificil saber cual es más correcta, pero los datos que se mantuvieron similares nos indican que estos verdaderamente son similares. Pasar los datos por multiples métodos nos puede ayudar a verificar patrones que creemos observar y al repetirse una tras otra vez nos aclaran que posiblemente existe una relación.

```
require(kohonen)
# Load Data
data = read.xls('data.xlsx', sheet=1)
data2016 = data[data[, 'year'] == 2016,][-1][-1][-1]
rownames(data2016) <- data[data[, 'year'] == 2016,][,1]
for(i in 1:ncol(data2016)){
  data2016[is.na(data2016[,i]), i] <- mean(data2016[,i], na.rm = TRUE)
}
data2016 <- data2016[, colSums(is.na(data2016)) != nrow(data2016)]
# PCA
p <- prcomp(data2016, scale=TRUE) #remove country, country, year
# plot(p$x[,1], p$x[,2], ylab="PC2", xlab='PC1')
biplot(p, cex=0.5)
px = p$x[order(p$x[,1], decreasing=TRUE),]
par(oma=c(2,2,2,2))
plot(head(px[,1],10), type='l', xaxt='n', ann=FALSE)
axis(1,at=1:10,
  labels=head(rownames(px), 10), las=2, cex=0.2)

data2016s <- scale(data2016)
# SOM
som1 <- som(data2016s, grid = somgrid(xdim = 11, ydim=11, topo="hexagonal"))
# plot(som1, type="dist.neighbours")
# plot(som1, type="codes")
plot(som1, type = "mapping", main = "Mapping Type SOM", labels=rownames(data2016s),data=data2
# ISOMAP
# dis <- vegdist(data2016s)

# ord <- isomap(dis, k = 3)
# pl <- plot(ord, main="isomap k=3", pch=rownames(data2016s))
dis <- dist(data2016s, method = "euclidean")
ord <- isomap(dis, k = 3)
pl <- plot(ord, main="isomap k=3", pch='.')
a <- data.frame(pl$sites)
text(a[,1], a[,2], labels=rownames(data2016s), cex=0.5)
```