Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Estadística Multivariada
Tarea 4

Marcelo Alberto Sanchez Zaragoza

14 de mayo de 2021

1. Problema 1

Frecuentemente en las aplicaciones nos encontramos con una variable categórica nominal con k estados excluyentes medida sobre una muestra de n $= n_1 + ... + n_g$ individuos provenientes de g poblaciones. Se desea obtener una medida de disimilaridad entre estas poblaciones. En estas condiciones, el vector de frecuencias de cada población $n_i = (n_{i1}, ..., n_{ik})$, para $i = 1, ..., g$, tiene una distribución conjunta multinomial con parámetros (n_i, p_i), donde $ni = n_{i1} + ... + n_{ik}$ y $p_i = (p_{i1}, ..., p_{ik})$. Una medida de disimilaridad es la distancia de Bhattacharyya, conocida en genética como

distancia Cavalli-Sforza, cuya expresión es:

$$d_{ij}^2 = arccos(\sum_{l=1}^{k} \sqrt{p_{il}p_{jl}})$$

a) Obtenga las distancias de Bhattacharyya entre estas poblaciones.

b) Construye una configuración MDS de las poblaciones mediante la solución clasica (coordenadas principales), utilizando la matriz de distancias Bhattacharyya.

c) ¿Cual la dimensión adecuada de la representación euclidiana?, ¿cuál es el porcentaje de la variabilidad explicada por las dos primeras coordenadas principales? Grafíca las poblaciones con las dos primeras coordenadas.

d) Construye una configuración MDS de las poblaciones utilizando el enfoque de mínimos cuadrados considerando la matriz de distancias Bhattacharyy, tomando como solución inicial la solución clásica y considerando las transformaciones de tipo razón, intervalo y ordinal para las disimilaridades. Compara los resultados obtenidos en cada modelo y justifica la dimensionalidad adecuada de representación y grafica las dos primeras dimensiones.

e) Compara las configuraciones MDS obtenidas con el enfoque clásico y de mínimos cuadrados, ¿existen diferencias? ¿Cuáles son las conclusiones?

Solución

Inciso a)

Las distancias Bhattacharyya se presentan en el código que anexamos, la matriz lleva por nombre M.

Inciso b)

Al realizar la configuración MDS tenemos los siguientes valores:

```
Francesa          0.20895842  0.01098544
Checa            -0.02491951  0.12723205
Germanica         0.16539227  0.05980017
Vasca             0.23554673 -0.14702890
China            -0.22782060 -0.11843395
Ainu             -0.38252414  0.01583101
Esquimal         -0.08502023 -0.29832442
Afromericana USA  0.05743718  0.17807845
Española          0.15082488 -0.05661571
Egipcia          -0.09787501  0.22847586
```

La representación se muestra en la figura 1.1.

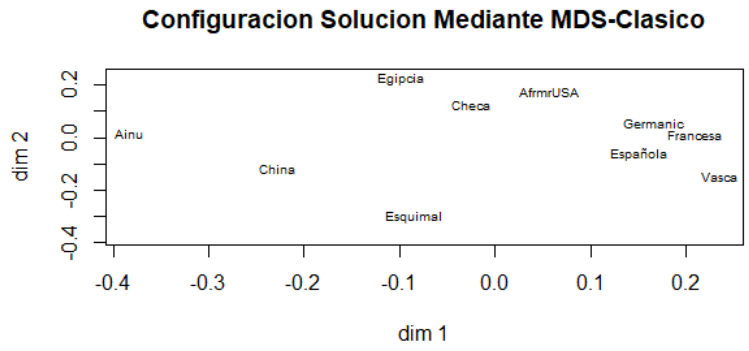

Figura 1.1: Configuración MDS-Clasico

Inciso c)

Una vez que encontramos las distancias euclidianas de las poblaciones procedemos a encontrar la configuración mediante MDS en dos dimensiones

ya que nuestra intención es representar las proximidades entre las poblaciones, en este caso contamos con 10 por lo que es más sencillo observa estas poblaciones en un gráfico de dos dimensiones. Al realizar dicha configuración el procentaje de la variabilidad explicada por las dos primeras coordenadas principales es de: 95.54 %. Observando este porcentaje podemos observar que dos coordenadas principales nos ayudan mucho en la representación.
Las coordenada que resultaron al utilizar una matriz con distancias euclidianas fueron las siguientes:

Francesa	0.070559013	-0.00772619
Checa	-0.062027716	-0.03568507
Germanica	0.034307898	-0.01570367
Vasca	0.155478974	0.01217982
China	0.006792157	0.05275680
Ainu	-0.216418097	0.01780887
Esquimal	0.033280242	-0.09730726
Afromericaba USA	0.063759384	0.12777442
Española	0.033288768	-0.07139795
Egipcia	-0.119020622	0.01730024

En la figura 1.2 se muestra la representación que resulto de ocupar una matriz con distancias euclidianas.

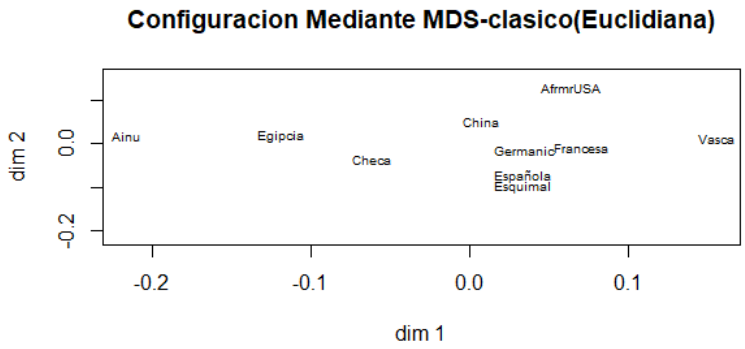

Figura 1.2: Configuración MDS-Clasico

Inciso d)
En este inciso nos piden un nuevo enfoque, con minimos cuadrados encontrar las configuraciones. Nos vamos a apoyar del valor que nos resulte en STRESS y en el diagrama de Sherpard para argumentar la dimensionalidad, para realizar los tres casos se hizo uso de la matriz de distancias Bhattacharyy.
Para la primer transformación razón, tenemos en la figura 1.3 el diagrama de Sherpard.

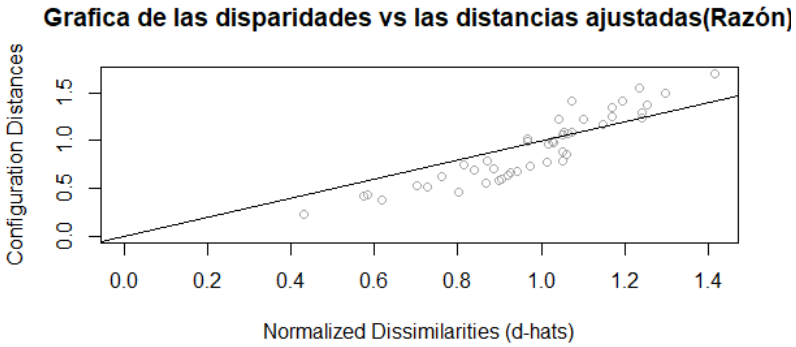

Figura 1.3: Diagrama de Sherpard MC-Razón

Si prestamos atención con este diagrama se busca encontrar que los datos esten lo más pegados a la recta de 45 grados que se dibuja y en este caso no lo hacen muy bien nuestros datos. El valor del Stress es de 0.193833. Para la transformación de invervalo se muestra su diagrama de Sherpard en la figura 1.4:

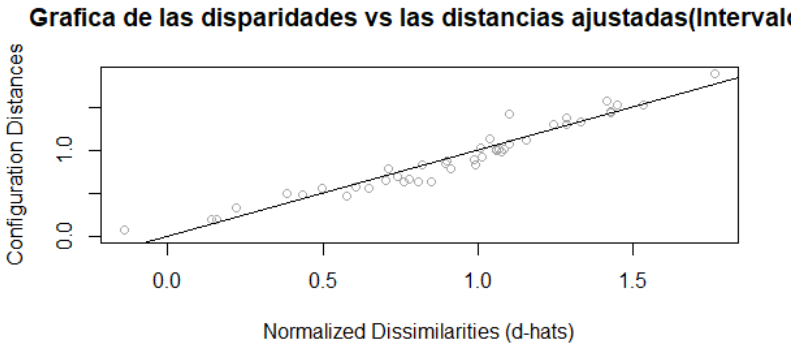

Figura 1.4: Diagrama de Sherpard MC-Intervalo

Se observa que los valores estan más cercanos a la recta de 45 grados, el valor de Stress es de: 0.102636.
Para la utima, la transformación ordinal su diagrama de Sherpard se presenta en la figura 1.5:

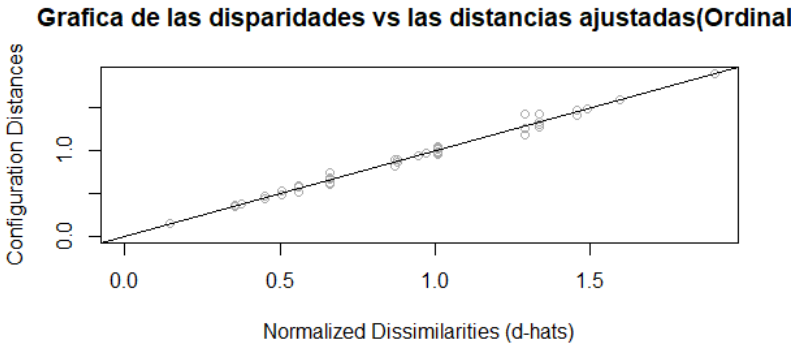

Figura 1.5: Diagrama de Sherpard MC-Ordinal

Con un valor de Stress de 0.04223, en este caso observamos el valor del Stress muy pequeño, en clase se observo un criterio por investigadores que sugiere que si el valor de Stress para una dimensionalidad dada es menor de 0.05, la representación es muy buena. Se oberva que en la figura 1.5 que nuestros datos estan sobre la recta de 45 grados por que para esta transformación habra una buena representación de los datos.
En la figura 1.6 se presenta la configuración con razón en las dos coordenadas principales.

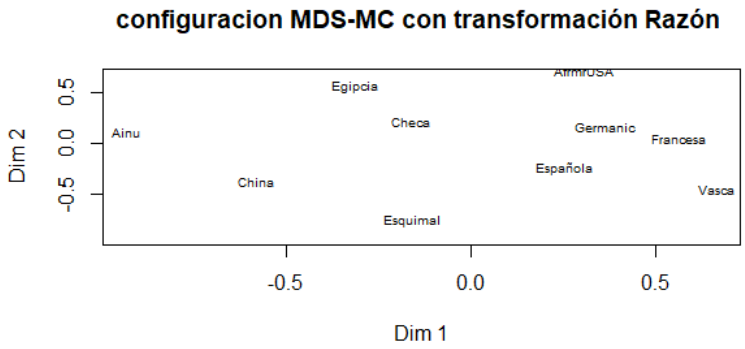

Figura 1.6: Configuración MC-Razón

En la figura 1.7 se presenta la configuración con Intervalo en las dos coordenadas principales.

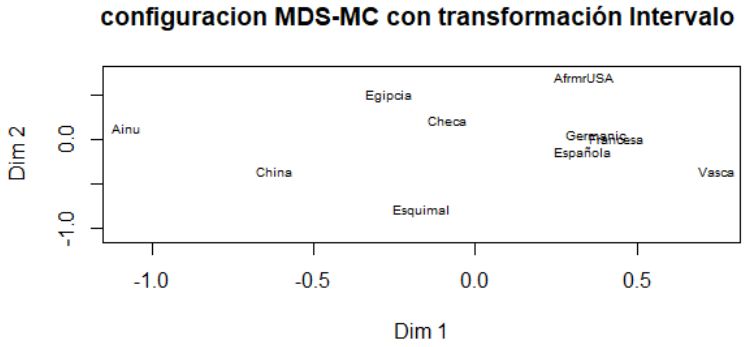

Figura 1.7: Configuración MC-Intervalo

En la figura 1.8 se presenta la configuración con Ordinal en las dos coor-

denadas principales.

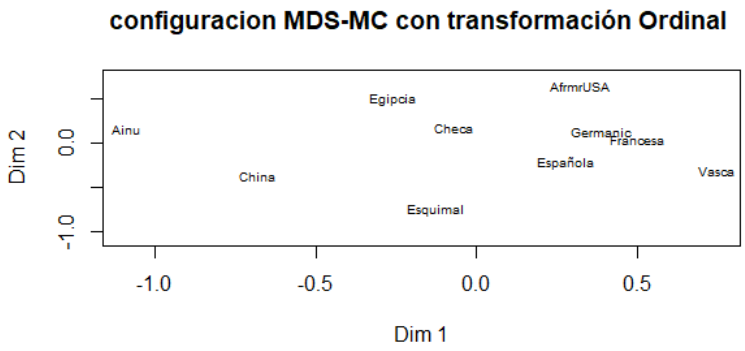

Figura 1.8: Configuración MC-Ordinal

La dimensión adecuada para las dos primeras transformaciones sugiere que sea mayor ya que el valor del Stress y el diagrama no fueron muy bueno, el diagrama no nos dio tantos problemas pero el valor si, basado en el criterio antes mencionado, para la transformación Ordinal si obtuvimos una buena representación en las dos coordenadas princiaples, basada en los resultados del Stress y el diagrama de Sherpard.

Inciso e)

Al observar las representaciones que nos resultaron de tomar el enfoque clasico con las dos matrices de distancias y el enfoque de minimos cuadrados con las tranformaciones, observamos que entre la figura 1.1 y figura 1.2 hay mucha diferencia ya que si recordamos la intención de ocupar el método de MDS, es encontrar una representación de las distancias en un espacio de baja dimensión y al menos para nuestro ejercicio esperariamos encontrar cierta relación entre nuestras poblaciones, es decir, no representarlas al cien por ciento como en un mapa. En este caso observamos

que la representación de la figura 1.1 observamos que si nos agrupa a las poblaciones europeas(Germanica, Francesa, Española y Vasca) por lo que tenemos una representación adecuada. En el caso cuando ocupamos la matriz de distancias euclidianas nos pega mucho las poblaciones Esquimal y Española, puede existir cierta similitud pero no la que estamos buscando en la representación por lo que no es de mucha ayuda.
Cuando ocupamos el enfoque de minimos cuadrados observamos que las representaciones se van poco a poco acercando a la figura 1.1, de las transformaciones la mejor o bueno la que se acerco más a la figura 1.1 fue cuando ocupamos la transformación Ordinal y tambien observamos que el diagrama de Sherpard nos dibujo un gráfico adecuado al criterio y tambien el valor del Stress ya que de acuerdo al criterio que mencionamos para tener una representación adecuada se desea tener un valor menor a 0.05 el cual si se consiguio.
En conclusión podemos decir que las figuras 1.1 y 1.8 fueron las mejor y se observa la similitud entre ellas.

2. Problema 2

En muchas situaciones las variables que se observan sobre un conjunto de individuos son de naturaleza binaria. En estos casos para poder disponer de una matriz de distancias entre individuos se utilizan coeficientes de similaridad. El coeficiente de similaridad entre el individuo i y el individuo j, s_{ij}, se calcula a partir de las frecuencias:

- a = numero de variables con respuesta 1 en ambos individuos.
- b = numero de variables con respuesta 0 en el primer individuo y con respuesta 1 en el segundo individuo.
- c = numero de variables con respuesta 1 en el primer individuo y con respuesta 0 en el segundo individuo.
- d = numero de variables con respuesta 0 en ambos individuos.

Se considera el siguiente conjunto de 6 individuos formado por 5 animales, león, jirafa, vaca, oveja, gato doméstico, junto con el hombre. Se miden 6 variables binarias sobre estos individuos: $X_1 =$ tiene cola, $X_2 =$ es salvaje, $X_3 =$ tiene el cuello largo, $X_4 =$ es animal de granja, $X_5 =$ es carnivoro y $X_6 =$ camina sobre 4 patas.

a) Obtenga la matriz de datos.

b) Calcule los coeficientes de similaridad de Sokal-Michener y de Jacard para cada par de individuos y obtenga las matrices de distancias asociadas.

Solución

Inciso a)
Como nos piden la matriz de datos la presentamos

$$Datos = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Donde en cada renglón colocamos los datos de: León, Jirafa, Vaca, Oveja, Gato-domestico y Humano. En la siguiente tabla se presenta con las etiquetas:

`Leon:`	1	1	0	0	1	1
`Jirafa:`	1	1	1	0	0	1
`Vaca`	1	0	0	1	0	1
`Oveja`	1	0	0	1	0	1
`Gato-doms`	1	0	0	0	1	1
`Humano`	0	0	0	0	1	0

Inciso b)
Presentamos las matrices correspondiente a los coeficientes de similaridad de Sokal-Michener y Jacard:

$$Sokal - Michener = \begin{pmatrix} 1{,}00 & 0{,}67 & 0{,}50 & 0{,}50 & 0{,}83 & 0{,}50 \\ 0{,}67 & 1{,}00 & 0{,}50 & 0{,}50 & 0{,}50 & 0{,}167 \\ 0{,}50 & 0{,}50 & 1{,}00 & 1{,}00 & 0{,}67 & 0{,}333 \\ 0{,}50 & 0{,}50 & 1{,}00 & 1{,}00 & 0{,}67 & 0{,}33 \\ 0{,}83 & 0{,}50 & 0{,}67 & 0{,}67 & 1{,}00 & 0{,}667 \\ 0{,}50 & 0{,}167 & 0{,}33 & 0{,}33 & 0{,}67 & 1{,}00 \end{pmatrix}$$

$$Jacard = \begin{pmatrix} 1{,}00 & 0{,}6 & 0{,}4 & 0{,}4 & 0{,}75 & 0{,}25 \\ 0{,}60 & 1{,}0 & 0{,}4 & 0{,}4 & 0{,}40 & 0{,}00 \\ 0{,}40 & 0{,}4 & 1{,}0 & 1{,}0 & 0{,}50 & 0{,}00 \\ 0{,}40 & 0{,}4 & 1{,}0 & 1{,}0 & 0{,}50 & 0{,}00 \\ 0{,}75 & 0{,}4 & 0{,}5 & 0{,}5 & 1{,}00 & 0{,}33 \\ 0{,}25 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}33 & 1{,}00 \end{pmatrix}$$

Las matrices de distancias asociadas son las siguientes:

$$D_{sokal} = \begin{pmatrix} 0{,}00 & 0{,}67 & 1{,}00 & 1{,}00 & 0{,}33 & 1{,}00 \\ 0{,}67 & 0{,}00 & 1{,}00 & 1{,}00 & 1{,}00 & 1{,}67 \\ 1{,}00 & 1{,}00 & 0{,}00 & 0{,}00 & 0{,}667 & 1{,}33 \\ 1{,}00 & 1{,}00 & 0{,}00 & 0{,}00 & 0{,}67 & 1{,}33 \\ 0{,}33 & 1{,}00 & 0{,}67 & 0{,}67 & 0{,}00 & 0{,}67 \\ 1{,}00 & 1{,}667 & 1{,}33 & 1{,}33 & 0{,}67 & 0{,}00 \end{pmatrix}$$

$$D_{jacard} = \begin{pmatrix} 0{,}0 & 0{,}8 & 1{,}2 & 1{,}2 & 0{,}5 & 1{,}50 \\ 0{,}8 & 0{,}0 & 1{,}2 & 1{,}2 & 1{,}2 & 2{,}00 \\ 1{,}2 & 1{,}2 & 0{,}0 & 0{,}0 & 1{,}0 & 2{,}00 \\ 1{,}2 & 1{,}2 & 0{,}0 & 0{,}0 & 1{,}0 & 2{,}00 \\ 0{,}5 & 1{,}2 & 1{,}0 & 1{,}0 & 0{,}0 & 1{,}33 \\ 1{,}5 & 2{,}0 & 2{,}0 & 2{,}0 & 1{,}3 & 0{,}00 \end{pmatrix}$$

3. Problema 3

Sea O un conjunto de n individuos cuya matriz de distancias euclidianas es D y cuya representación en coordenadas principales es X. Se desean

obtener las coordenadas de un nuevo individuo al que llamaremos individuo $n+1$, del cual se conocen los cuadrados de sus distancias a los n individuos del conjunto O. Si $d = (\delta^2_{n+1,1}, ..., \delta^2_{n+1,n})^t$ es el vector columna que contiene las distancias al cuadrado del individuo $n+1$ a los restantes n individuos, se puede probar que las coordenadas principales del individuo $n+1$ están dadas por:

$$x_{n+1} = \frac{1}{2}\Delta^{-1}X^t(b-d)$$

a) Obtenga una representación en coordenadas principales utilizando la matriz de distancias calculada a partir del coeficiente de similaridad de Sokal-Michener.

b) Sin volver a recalcular las coordenadas principales, añada el elefante al conjunto de animales y obtenga sus coordenadas principales.

Solución

Inciso a)
La representación que encontramos se presenta en la figura 3.1:

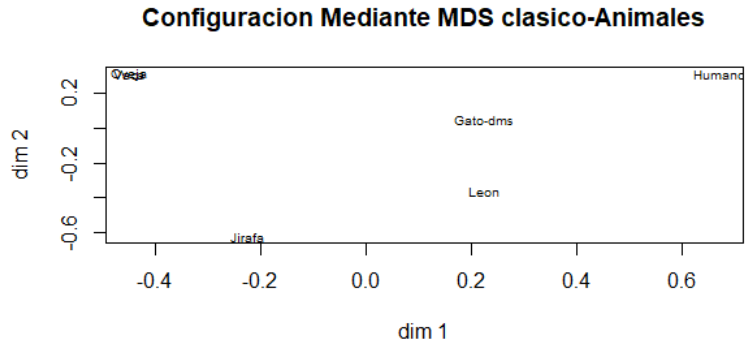

Figura 3.1: Configuración MDS-Animales

Obsevamos que en el gráfico nos pone muy pegados los animales oveja y vaca, esto puede ser resultado de que sus valores en la matriz de datos son iguales. La jirafa por un lado se separa mucho de los demás ya que el cuello largo es el que tiene peso, mientras que con el gato domenstico y el león lo pone sobre la misma recta vectical pero un factor que cambia en ellos es ser salvaje, mientras que el humano lo manda muy lejor de todo ellos ya que tomamos como unico valor el que sea carnivoro. La representación podria ser mejor si tomaramos más variables porque es claro que la Oveja y Vaca son diferentes pero al menos para las variables que tomamos lo son. Los valores de la representación se presentan a continuación:

```
Leon         0.2236068 -0.35823182
Jirafa      -0.2236068 -0.61643071
Vaca        -0.4472136  0.30821536
Oveja       -0.4472136  0.30821536
Gato-domes   0.2236068  0.05001647
Humano       0.6708204  0.30821536
```

Inciso b)

Al realiza las operaciones que nos mostraron en el problema llegamos a que el valor de las coordenadas del elefante son las siguientes: -0,14907, -0.3582. Al colocarlo en la figura 3.1 tenemos como resultado la figura 3.2

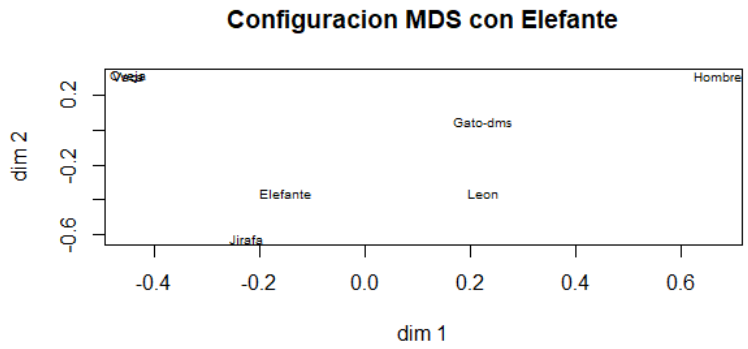

Figura 3.2: Configuración MDS-Elefante

El elefante lo coloca cercano a la jirafa ya que coinciden en valores pero solo cambian en la variable que toma en cuenta el cuello largo. Una variable extra que se podria tomar podria ser las orejas o la trompa grande.
Todas las operaciones realizadas se realizaron con ayuda de un programa el cual se anexa al trabajo.