Estadística Multivariada

Tarea sobre MDS

Instrucciones:

- Subir la tarea a la plataforma en un zip que contenga un archivo pdf con las respuestas documentadas y el código
- Fecha límite de entrega: 14 de mayo de 2021,23:00

1. Frecuentemente en las aplicaciones nos encontramos con una variable categórica nominal con k estados excluyentes medida sobre una muestra de $n = n_1 + \cdots + n_g$ 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 $n_i = n_{i1} + \cdots + n_{ik}$ 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:

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

La siguiente tabla contiene las proporciones génicas (observadas) de los grupos sanguíneos correspondientes a 10 poblaciones.

	Población	Grupo A	Grupo AB	Grupo B	Grupo O
1	Francesa	0.21	0.06	0.06	0.67
2	Checa	0.25	0.04	0.14	0.57
3	Germanica	0.22	0.06	0.08	0.64
4	Vasca	0.19	0.04	0.02	0.75
5	China	0.18	0.00	0.15	0.67
6	Ainu	0.23	0.00	0.28	0.49
7	Esquimal	0.30	0.00	0.06	0.64
8	Afromericana USA	0.10	0.06	0.13	0.71
9	Española	0.27	0.04	0.06	0.63
10	Egipcia	0.21	0.05	0.20	0.54

   (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? Grafica 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?

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"

   Existen muchos coeficientes de similaridad, pero los de Sokal-Michener y de Jacard son especialmente interesantes porque dan lugar a una configuración euclidiana. Se definen como

$$Sokal - Michener : s_{ik} = \frac{a+d}{p}, \qquad Jacard : s_{ik} = \frac{a}{a+b+c}$$

   donde p es el número de variables observadas. Aplicando uno de estos coeficientes a un conjunto de n individuos se obtiene una matriz de similaridades $\boldsymbol{S} = \{s_{ij}\}_{n\times n}$. Utilizando la siguiente transformación podemos convertir la matriz de similaridades a una matriz de distancias

$$\boldsymbol{D}^2 = 2(\mathbf{1}_n\mathbf{1}_n' - \boldsymbol{S})$$

   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

3. Sea O un conjunto de n individuos cuya matriz de distancias euclidianas es $\boldsymbol{D}$ y cuya representación en coordenadas principales es $\boldsymbol{X}$. 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 $\boldsymbol{d} = (\delta^2_{n+1,1}, ..., \delta^2_{n+1,n})'$ 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

$$x_{n+1} = \frac{1}{2}\boldsymbol{\Lambda^{-1}X'(b-d)}$$

donde $\boldsymbol{b} = \boldsymbol{diag}(\boldsymbol{B}) = (b_{11}, ..., b_{nn})'$, $\boldsymbol{B} = \boldsymbol{XX'} = \boldsymbol{U\Lambda U'}$ y $\boldsymbol{U}$ es una matriz ortogonal. La ecuacion anterior se conoce como formula de interpolación de Gower.

Para los datos del ejercicio 2

(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