UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Escalado multidimensional (MDS)

Francesc Carmona

7 de mayo de 2019

1. La matriz adjunta indica las similitudes[1] entre siete animales según un experto:

	A	B	C	D	E	F	G
A	0	7	5	9	5	7	9
B	7	0	4	6	4	6	7
C	5	4	0	3	4	5	6
D	9	6	3	0	3	2	2
E	5	4	4	3	0	5	4
F	7	6	5	2	5	0	4
G	9	7	6	2	4	4	0

   a) Construir la matriz $\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}^{(2)}\mathbf{H}$, donde $\mathbf{D}^{(2)}$ es la matriz de distancias al cuadrado y $\mathbf{H}$ es la matriz de centrado, y calcular sus valores propios. Observar si la matriz de distancias es euclídea.[2]

   b) Obtener la representación con las dos primeras coordenadas principales e indicar el grado de bondad de esta representación.

   Se puede hacer a partir de la descomposición de la matriz $\mathbf{B}$ o con la función

   `cmdscale(`*dist*`, eig=T)`.

2. (∗) Poner un ejemplo para comprobar que el escalado multidimensional clásico aplicado a las distancias euclídeas calculadas sobre una matriz de datos multivariantes $\mathbf{X}$ es equivalente a la solución que se obtiene por el análisis de componentes principales de la matriz de covarianzas de $\mathbf{X}$.

3. (∗) 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. En estas condiciones, el vector de frecuencias de la i-ésima población $\mathbf{n}_i = (n_{i1}, \ldots, n_{ik})$ tiene una distribución conjunta multinomial con parámetros $(n_i, \mathbf{p}_i)$, donde $n_i = n_{i1} + \cdots + n_{ik}$ y $\mathbf{p}_i = (p_{i1}, \ldots, p_{ik})$ son las probabilidades de un individuo de la población i tales que $\sum_{j=1}^{k} p_{ij} = 1$.

   En genética, una medida de disimilaridad, entre otras muchas en estas circunstancias, es la distancia de Bhattacharyya cuya expresión es

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

   Si definimos el coeficiente de Bhattacharyya entre dos poblaciones como

$$\cos\varphi = \sum_{l=1}^{k}\sqrt{p_{il}p_{jl}}$$

[1] Aquí la palabra "similitudes" se utiliza coloquialmente y no es correcta según la definición estadística. Se ha utilizado porque estaba en el ejercicio original. En realidad es una matriz de distancias.

[2] El paquete `ade4` dispone de una función `is.euclid()`.

Tabla 1: Proporciones génicas entre 10 poblaciones.

	Población	grupo A	grupo B	grupo AB	grupo 0
1	francesa	0.21	0.06	0.06	0.67
2	checa	0.25	0.04	0.14	0.57
3	germánica	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	negra 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

esta distancia está relacionada en genética con las distancias de Cavalli-Sforza

$$d_{\text{C-Sf}} = \sqrt{2(1-\cos\varphi)}$$

y la distancia de Nei [3]

$$d_{\text{Nei}} = -\ln\cos\varphi$$

La tabla 1 contiene las proporciones génicas observadas de los grupos sanguíneos correspondientes a 10 poblaciones distintas y excluyentes.

a) Obtener las distancias de Bhattacharyya según la fórmula 1.

b) Representar estas poblaciones con las dos primeras coordenadas principales. ¿Se observa algún tipo de agrupación?

c) ¿Es ésta una distancia euclídea? ¿Cuál es la dimensión de la representación euclídea? Determinar el porcentaje de variabilidad explicado por las dos primeras coordenadas principales.

4. Si las variables observadas sobre un conjunto de individuos son del tipo binario, es necesario disponer de una matriz de distancias basadas en los coeficientes de similaridad que se obtienen de las tablas del tipo

	1	0
1	a	b
0	c	d

donde a es el número de variables con respuesta 1 en ambos individuos, d es el número de variables con respuesta 0 en ambos individuos, etc.

Entre los muchos coeficientes de similaridad destacan dos:

$$\text{Sokal y Michener: } s_{ij} = \frac{a+d}{p} \qquad \text{Jaccard: } s_{ij} = \frac{a}{p-d}$$

donde $p = a+b+c+d$ es el número de variables binarias observadas.

El coeficiente de Sokal y Michener[4] se utiliza con variables binarias simétricas, es decir, aquellas en las que los valores 1 y 0 son intercambiables. Uno no es más importante que el otro, como el sexo. El coeficiente de Jaccard, en cambio, se utiliza con variables binarias asimétricas, es decir, aquellas en las que la presencia de la característica (el 1) es importante.

[3] Fuera de la genética, ésta es la que se conoce como distancia de Bhattacharyya.
[4] También llamado coeficiente de acoplamiento simple.

Aplicando uno de estos coeficientes a un conjunto de individuos se obtiene una matriz de similaridades (s_{ij}) que se puede transformar en una distancia

$$d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij} = 2(1 - s_{ij})$$

Dada la siguiente matriz de datos

A	1	1	0	0	1	1
B	1	1	1	0	0	1
C	1	0	0	1	0	1
D	0	0	0	0	1	0

con las observaciones de 6 variables binarias sobre 4 individuos,

a) Calcular los coeficientes de Sokal y Michener para cada par de individuos y obtener la matriz de distancias asociada.

b) Lo mismo que en el apartado anterior pero con el coeficiente de Jaccard.

5. En muchos análisis multivariantes se dispone de un conjunto de variables mixto, es decir, unas variables son del tipo cuantitativo y otras cualitativo (nominales, ordinales o incluso binarias). En estos casos es necesario disponer de una distancia como la de Gower que tiene en cuenta el tipo de variable.

   El cuadrado de la distancia de Gower entre dos filas de datos se define como $d_{ij}^2 = 1 - s_{ij}$, donde s_{ij} es el coeficiente de similaridad

   $$s_{ij} = \frac{\sum_{h=1}^{p_1}(1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

   p_1 es el número de variables cuantitativas, p_2 es el número de variables binarias, p_3 el número de variables cualitativas (no binarias), a el número de coincidencias $(1, 1)$ en las variables binarias, d el número de coincidencias $(0, 0)$ en las variables binarias, α el número de coincidencias en las variables cualitativas (no binarias) y G_h es el rango (o recorrido) de la h-ésima variable cuantitativa.

   a) Calcular la distancia de Gower entre los datos de 18 tipos de flores de la base de datos flower del paquete cluster de **R**.

   Probar la función `daisy()`[5] del paquete cluster.

   b) Realizar un escalado multidimensional con esta distancia y representar las flores en un gráfico de dispersión de dos dimensiones.

6. Consideremos los datos sobre cráneos de varones egipcios de cinco épocas históricas que se pueden bajar[6] desde la página del libro de Everitt (2005). Los podremos cargar directamente en **R** con la siguientes instrucciones:

```
skulls <- source("/(path)/chap5skulls.dat")$value
str(skulls)
attach(skulls)
```

   Donde el *path* debe ser la dirección a la carpeta donde hemos dejado el archivo una vez descomprimido. Así tendremos la base de datos `skulls` con cinco variables. La primera variable es el factor EPOCH y las otras cuatro son las medidas biométricas estudiadas del cráneo.

   a) En primer lugar podemos realizar un MANOVA para contrastar la diferencia de medias entre los niveles del factor o poblaciones. No entraremos aquí en la comprobación de las hipótesis de normalidad y de igualdad de las matrices de covarianzas.

[5] Considerar atentamente si las variables binarias son simétricas o asimétricas.
[6] También se pueden obtener con el *data.frame* `skulls` del paquete `HSAUR`.

b) (∗∗) Comprobaremos que el test anterior rechaza la igualdad de medias y, por lo tanto, justifica un **análisis canónico de poblaciones**.
Utilizar la función `candisc()` del paquete `candisc` o consultar

`http://erre-que-erre-paco.blogspot.com/2010/03/analisis-canonico-de-poblaciones.html`

c) Calcular las distancias de Mahalanobis entre cada pareja de épocas.
Para ello, considerar la matriz de covarianzas común

$$\hat{\mathbf{S}} = \frac{29\hat{\mathbf{S}}_1 + 29\hat{\mathbf{S}}_2 + 29\hat{\mathbf{S}}_3 + 29\hat{\mathbf{S}}_4 + 29\hat{\mathbf{S}}_5}{145}$$

donde $\hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_5$ son las matrices de covarianzas en cada época.

d) Realizar un escalado multidimensional con la matriz de distancias obtenida en el apartado anterior y representar las cinco épocas con las dos coordenadas principales.

7. (∗) **Escalado multidimensional no métrico**

En los problemas de escalado no métrico se parte de una matriz de similaridades, disimilaridades o distancias (δ_{ij}) entre objetos o individuos que se ha obtenido por la estimación directa de un juez o grupo de expertos, o por estimación por rangos o rangos por pares, es decir, sin una métrica a partir de unas variables observadas.

La distancia así obtenida se supone que está relacionada con la verdadera distancia pero de una manera compleja. Es decir, seguramente los expertos utilizan en sus valoraciones ciertas variables o dimensiones, además de elementos de error y variabilidad personal. Así, la verdadera distancia se puede considerar una función monótona creciente de las distancias dadas

$$\hat{d}_{ij} = \varphi(\delta_{ij})$$

Escalado multidimensional isotónico

Una versión de escalado multidimensional no métrico se basa en una vieja idea de Kruscal y Shepard (1960). Dada una disimilaridad (δ_{ij}), se trata de elegir una configuración de puntos con su distancia euclídea (d_{ij}) que minimize

$$\text{stress}^2 = \sum_{i<j} [d_{ij} - \varphi(\delta_{ij})]^2 / \sum_{i<j} d_{ij}^2$$

para ambas, la configuración y la función creciente φ. Así, la localización, la rotación, la reflexión y la escala de la configuración de puntos son todas indeterminadas.

En **R** tenemos la función `isoMDS()` del paquete `MASS` en la que se ha implementado un algoritmo de optimización, lógicamente, un poco lento.

Aplicar este escalado isotónico a la matriz de distancias del ejercicio 1.