Análisis de Correspondencias

Introducción

El análisis de correspondencias es una técnica descriptiva para representar tablas de contingencia. Los datos de partida para el análisis de correspondencias es una matriz $\mathbf{X}$ de dimensiones $n \times k$ que representa las frecuencias absolutas observadas en una tabla de contingencia de dos variables, donde la primera se representa por filas y tiene n categorías y la segunda por columnas y tiene k categorías. Por ejemplo, clasificamos un conjunto de estudiantes en n posibles procedencias geográficas y k posibles opiniones respecto a la docencia. En general, el elemento x_{ij} de la matriz $\mathbf{X}$ representa la frecuencia absoluta observada en la casilla (i, j) de la tabla de contingencia.

La metodología la desarrolló Benzecri, a principios de los años 60 del siglo XX en la Universidad de Renner (Francia). En esencia. es un tipo especial de análisis de componentes principales pero realizado sobre una tabla de contingencia y usando una distancia euclídea ponderada llamada *chi-cuadrado*.

Ejemplo: supongamos 400 tiendas de discos repartidas entre los países de la U.E. Se clasifica a los compradores en 3 categorías distintas: Jóvenes, Edad Media, Mayores, y a los tipos de música en 5 tipos:

A = Música disco

B = Rock'n'roll y música americana

C = Pop y música inglesa (melódicas)

D = Jazz y música autóctona

E = Clásica

Así, se tienen dos variables categóricas: *Compradores* y *Música*:

	Jov	Med	May	Total
A	70	0	0	70
B	45	45	0	90
C	30	30	30	90
D	0	80	20	100
E	35	5	10	50
Total	180	160	60	400

Cada uno de los entrevistados sólo valora un tipo de música, es decir, aparece en solo una de las casillas de la tabla.

Se puede definir el *perfil* de una tabla como el conjunto de las frecuencias de una fila o de una columna dividido entre el total de la fila o columna.

Por ejemplo, las frecuencias del tipo de música **B** son $\begin{pmatrix} 0{,}5 & 0{,}5 & 0 \end{pmatrix}$ ya que teníamos $\begin{pmatrix} 45 & 45 & 0 \end{pmatrix}$ y el total es 90. Así se obtiene:

	Jov	Med	May
A	1	0	0
B	0.5	0.5	0
C	0.33	0.33	0.33
D	0	0.8	0.2
E	0.7	0.1	0.2
Total	0.45	0.40	0.15

Se puede ver que un 45 % de compradores es Joven, 40 % Medianos y 15 % Mayores. También se puede ver por tipos de música, por ejemplo en el tipo **E** el reparto de edades difiere de la media: (70 % frente a 45 % en Jóvenes, 10 % frente a 40 % en Medianos).

Si se considera el análisis de las columnas, es decir, en vez de centrarnos en los tipos de música, nos centramos en las edades, se pueden considerar también perfiles columna. Así, por ejemplo de los 160 compradores en el caso de los de mediana edad, un 50 % compra el tipo de música D en vez del porcentaje general del 25 %. Es decir, con la tabla completa:

	Jov	Med	Jub	Total
A	0.39	0	0	0.175
B	0.25	0.28	0	0.225
C	0.17	0.19	0.50	0.225
D	0	0.50	0.33	0.25
E	0.19	0.03	0.17	0.125

Se pueden establecer visualmente relaciones entre los porcentajes de las categorías, tanto por filas como por columnas y representar las categorías de las filas según un espacio tridimensional determinado por las tres categorías de grupos de edad. Esto se denomina representación *baricéntrica*.

En el caso del ejemplo se puede hacer la representación dado el escaso número de categorías, pero se hace necesario encontrar un sistema de representación que disminuya el número de dimensiones mediante proyecciones. Una forma de hacerlo es usar las técnicas de multidimensional scaling.

Esencialmente, el análisis de correspondencias se puede considerar una aplicación del multidimensional scaling usando una distancia específica que se puede usar para datos categóricos. Dicha distancia se denomina distancia *chi cuadrado*.

Independencia

Si el hecho de que aparezca o se presente una categoría junto con otra no es ni más ni menos probable de que se presenten las dos categorías por separado, se dice que las variables son independientes y, en general, se dice que la tabla es *homogénea*.

Así, dadas dos variables aleatorias X e Y, son independientes si

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

para todo i, j.

En el caso de una tabla de contingencia, si se aproxima la probabilidad de que sucedan x_i e y_j como la frecuencia relativa en un experimento con N tiradas totales (regla de

Laplace), entonces:

$$\begin{aligned} p_{ij} &= \frac{n_{ij}}{n_{..}} \\ p_{i\cdot} &= \frac{n_{i.}}{n_{..}} \\ p_{\cdot j} &= \frac{n_{.j}}{n_{..}} \end{aligned}$$

Así, si

$$P(X = x_i, Y = y_j) = p_{ij} = p_{i\cdot} \times p_{\cdot j}$$

para todo i, j, las variables X e Y son independientes y la tabla es homogénea. Si es cierta la hipótesis de independencia esperaremos encontrar E_{ij} objetos dentro de la casilla (i, j)-ésima, donde

$$E_{ij} = n_{..}p_{ij} = n_{..}p_{i\cdot}p_{\cdot j} = \frac{n_{i.}n_{.j}}{n_{..}}$$

Si no vemos que ocurra así en la tabla, se rechaza la hipótesis de independencia.

Es necesario definir un contraste o test que me mida las *distancias* entre lo que uno observa y lo que esperaría si se cumple la hipótesis nula de independencia. La forma tradicional de hacerlo es mediante un contraste de la chi cuadrado, en el que se define el estadístico como

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}}\right)^2}{\frac{n_{i.}n_{.j}}{n_{..}}}.$$

Habitualmente se usa este contraste de independencia en tablas de contingencia.

Distancia chi cuadrado

En general, una tabla de contingencia donde hay r filas y c columnas se puede escribir como

		Columnas				
		1	2	$\cdots$	c	
Filas	1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	$n_{\cdot\cdot}$

A partir de aquí se pueden construir las tablas de proporciones de filas y columnas:

		Columnas			
		1	2	$\cdots$	c
Filas	1	$p_{11} = \frac{n_{11}}{n_{1\cdot}}$	$p_{12} = \frac{n_{12}}{n_{1\cdot}}$	$\cdots$	$p_{1c} = \frac{n_{1c}}{n_{1\cdot}}$
	2	$p_{21} = \frac{n_{21}}{n_{2\cdot}}$	$p_{22} = \frac{n_{22}}{n_{2\cdot}}$	$\cdots$	$p_{2c} = \frac{n_{2c}}{n_{2\cdot}}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	$p_{r1} = \frac{n_{r1}}{n_{r\cdot}}$	$p_{r2} = \frac{n_{r2}}{n_{r\cdot}}$	$\cdots$	$p_{rc} = \frac{n_{rc}}{n_{r\cdot}}$

		Columnas			
		1	2	$\cdots$	c
Filas	1	$q_{11} = \frac{n_{11}}{n_{\cdot 1}}$	$q_{12} = \frac{n_{12}}{n_{\cdot 2}}$	$\cdots$	$q_{1c} = \frac{n_{1c}}{n_{\cdot c}}$
	2	$q_{21} = \frac{n_{21}}{n_{\cdot 1}}$	$q_{22} = \frac{n_{22}}{n_{\cdot 2}}$	$\cdots$	$q_{2c} = \frac{n_{2c}}{n_{\cdot c}}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	$q_{r1} = \frac{n_{r1}}{n_{\cdot 1}}$	$q_{r2} = \frac{n_{r2}}{n_{\cdot 2}}$	$\cdots$	$q_{rc} = \frac{n_{rc}}{n_{\cdot c}}$

La distancia chi cuadrado entre las columnas i y j se define, entonces, como

$$d_{ij}^{\text{col}} = \sum_{k=1}^{r} \frac{1}{p_{k\cdot}} \left(p_{ki} - p_{kj}\right)^2$$

donde

$$p_{k\cdot} = \frac{n_{k\cdot}}{n_{\cdot\cdot}}$$

Se denominan tablas de perfiles fila y perfiles columna

La distancia chi cuadrado se puede considerar como una distancia euclídea ponderada basada en las proporciones de las columnas. Será igual a cero si las dos columnas tienen los mismos valores para esas proporciones. Si observamos que las diferencias al cuadrado anteriores se multiplican o ponderan mediante el factor $\frac{1}{p_{k\cdot}}$, de modo que categorías de

la variable que está en la columna con pocos valores tienen una mayor influencia en el cálculo de la distancia que las categoría comunes.

Se puede definir una distancia similar entre dos filas i y j

$$d_{ij}^{\text{fil}} = \sum_{k=1}^{c} \frac{1}{q_{\cdot k}} \left(q_{ik} - q_{jk}\right)^2$$

donde

$$q_{\cdot k} = \frac{n_{\cdot k}}{n_{\cdot\cdot}}$$

La distancia chi cuadrado cumple la propiedad de *equivalencia distribucional*:

Si dos categorías de los perfiles fila tienen el mismo valor de perfil, entonces al agruparlas en una única categoría no se modifican las distancias entre el resto de categorías de la tabla que forman las columnas. Lo mismo se puede decir en cuanto a las columnas: si se juntan o separan columnas, esto no afecta a las distancias entre los perfiles fila.

En muchas ocasiones se habla del concepto de *masa* de una fila o una columna de una tabla de contingencia. Esto es simplemente la proporción de observaciones de la fila (o columna) respecto al total de observaciones ($n_{i\cdot}/n_{\cdot\cdot}$)

El perfil medio de las filas (la fila *media* de perfiles) es el centroide de los perfiles fila cuando se calcula la media ponderando cada perfil por su masa. Todo esto mismo, obviamente, se puede considerar para las columnas.

A la expresión $\frac{X^2}{N}$ se denomina *inercia total* de la tabla de contingencia. Se puede interpretar como la media ponderada de las distancias chi cuadrado entre los perfiles fila y su perfil medio. O, alternativamente, se puede definir del mismo modo para los perfiles columna.

Reducción de dimensiones

En general, los perfiles están situados en espacios de altas dimensiones de modo que no se pueden observar directamente. Se pueden determinar subespacios de dimensión menor al número mínimo entre filas y columnas menos uno, donde se puede aproximar la posición

original de los perfiles. La calidad de representación en subespacios de dimensión menor se mide en porcentajes de inercia con respecto a la total.

El cálculo matemático de los subespacios se basa en minimizar las sumas de las distancias entre los perfiles y el subespacio, ponderadas por las masas de los puntos. Es decir, se calcula por el método de los mínimos cuadrados ponderados. Se pueden proyectar perfiles fila y perfiles columna de modo equivalente en el subespacio extraído.

Una manera de hacer lo anterior es mediante una aplicación directa del multidimensional scaling (*MDS*) en cada matriz de distancias (por filas o por columnas). Luego, se consideran y se dibujan las dos primeras coordenadas para las categorías de las filas y de las columnas en la misma gráfica etiquetadas de modo conveniente para que se puedan distinguir ambas variables.

Cuando las coordenadas de las categorías de ambas variables son grandes y positivas se deduce una asociación positiva entre las columnas y las filas correspondientes. Del mismo modo se razona en el caso de coordenadas negativas. La conclusión es que los valores de la tabla n_{ij} son mayores que los esperados bajo la hipótesis de independencia entre ambas variables.

Cuando las coordenadas de las categorías de ambas variables son grandes en valor absoluto, pero tienen signos opuestos las filas y columnas correspondientes tienen asociación negativa; así los valores de la tabla n_{ij} son menores que los esperados bajo la hipótesis de independencia entre ambas variables.

Finalmente, cuando el producto de las coordenadas está próximo a 0, la asociación entre las variables es baja, de modo que n_{ij} se encuentra cerca del valor esperado bajo la hipótesis de independencia.

Ejemplo

Supongamos la tabla de contingencia siguiente (de Everitt):

	Grupos de Edades					
	1	**2**	**3**	**4**	**5**	
Sin Pareja	21 (*0.68*)	21 (*0.64*)	14 (*0.58*)	13(*0.42*)	8(*0.40*)	77(*0.55*)
Con Pareja (no sexo)	8 (*0.26*)	9 (*0.27*)	6 (*0.25*)	8 (*0.26*)	2 (*0.10*)	33(*0.24*)
Con Pareja (sí sexo)	2 (*0.06*)	3 (*0.09*)	4 (*0.17*)	10 (*0.32*)	10 (*0.50*)	29(*0.21*)
Total	31(*0.22*)	33(*0.24*)	24(*0.17*)	31(*0.22*)	20(*0.15*)	139

En esta tabla se trata de ver la influencia de la edad en relación a las relaciones personales. Se pueden calcular las distancias chi cuadrado entre los elementos de la tabla. Por ejemplo, la distancia entre al columna 1 y la 2 es:

$$d_{12}^{\text{col}} = \sqrt{\frac{(0{,}68-0{,}64)^2}{0{,}55} + \frac{(0{,}26-0{,}27)^2}{0{,}24} + \frac{(0{,}06-0{,}09)^2}{0{,}21}} = 0{,}09$$

Esta distancia es similar a la distancia euclídea habitual pero difiere en que se divide cada término entre la proporción media correspondiente. De este modo se compensan los diferentes niveles de ocurrencia de las categorías. En el ejemplo la matriz de distancias entre las columnas es

$$D^{\text{col}} = \begin{pmatrix} 0{,}00 & 0{,}09 & 0{,}26 & 0{,}66 & 1{,}07 \\ 0{,}09 & 0{,}00 & 0{,}19 & 0{,}59 & 1{,}01 \\ 0{,}26 & 0{,}19 & 0{,}00 & 0{,}41 & 0{,}83 \\ 0{,}66 & 0{,}59 & 0{,}41 & 0{,}00 & 0{,}51 \\ 1{,}07 & 1{,}01 & 0{,}83 & 0{,}51 & 0{,}00 \end{pmatrix}$$

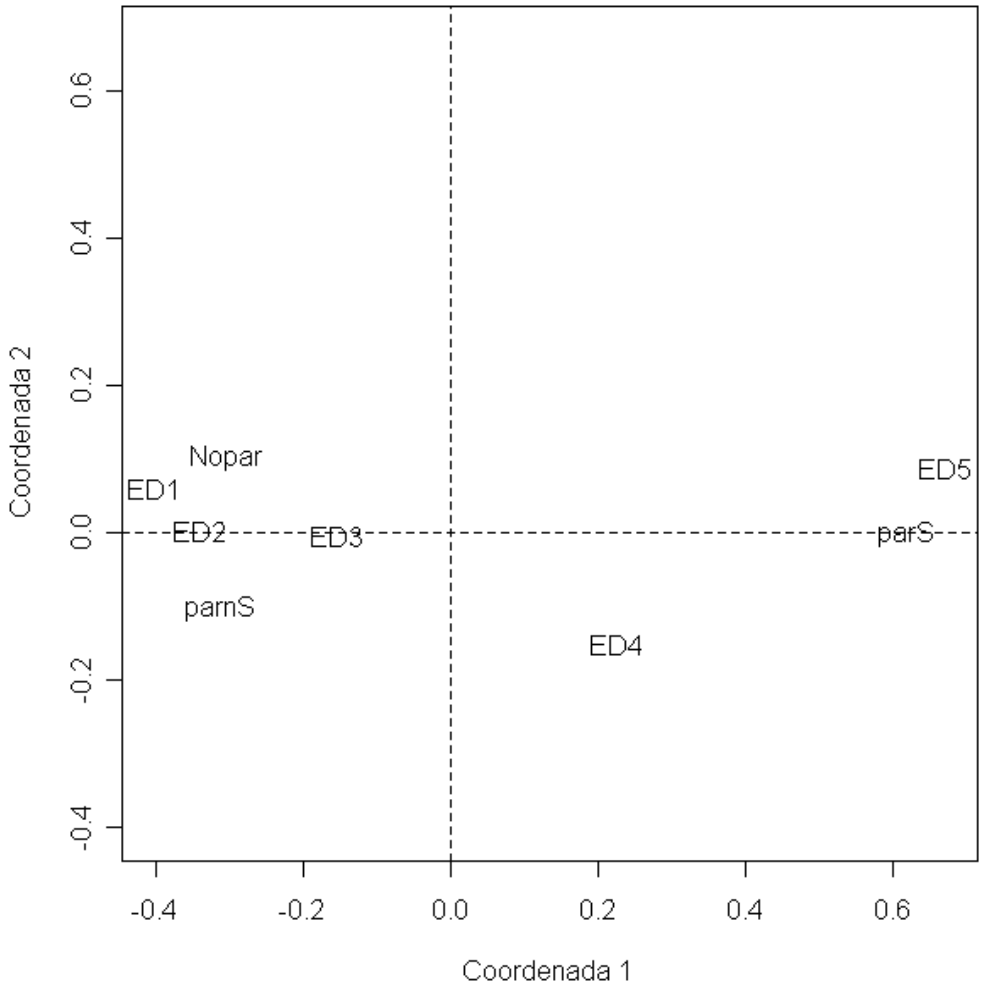
La matriz de distancias entre filas es

$$D^{\text{fil}} = \begin{pmatrix} 0{,}00 & 0{,}21 & 0{,}93 \\ 0{,}21 & 0{,}00 & 0{,}93 \\ 0{,}93 & 0{,}93 & 0{,}00 \end{pmatrix}$$

Aplicamos un multidimensional scaling (*MDS*) clásico a cada una de las matrices de distancias, obteniéndose las coordenadas respectivas de las categorías. Estas se dibujan posteriormente en un gráfico con las etiquetas correspondientes.

Análisis de Correspondencias básico con R

```
# Se introduce la tabla
sex<-matrix(c(21,21,14,13,8,8,9,6,8,2,2,3,4,10,10),ncol=5,byrow=TRUE)

# Se calculan los porcentajes
ncol<-5
nrow<-3
n<-sum(sex)
rtot<-apply(sex,1,sum)
ctot<-apply(sex,2,sum)

xrtot<-cbind(rtot,rtot,rtot,rtot,rtot)
xctot<-rbind(ctot,ctot,ctot)

xrtot<-sex/xrtot
xctot<-sex/xctot

rdot<-rtot/n
cdot<-ctot/n

# Se calculan las matrices de distancias entre columnas
dcols<-matrix(0,ncol,ncol)
for(i in 1:ncol){
       for(j in 1:ncol){d<-0
            for(k in 1:nrow) d<-d+(xctot[k,i]-xctot[k,j])^2/rdot[k]
            dcols[i,j]<-sqrt(d)}}

# Se calculan las matrices de distancias entre filas
drows<-matrix(0,nrow,nrow)
for(i in 1:nrow){
       for(j in 1:nrow){d<-0
            for(k in 1:ncol) d<-d+(xrtot[i,k]-xrtot[j,k])^2/cdot[k]
            drows[i,j]<-sqrt(d)}}


# Se aplica el MDS metrico
r1<-cmdscale(dcols,eig=TRUE)
r1$points
r1$eig
c1<-cmdscale(drows,eig=TRUE)
c1$points
c1$eig
xrtot

# Se dibujan las coordenadas en un dos dimensiones
par(pty="s")
plot(r1$points,xlim=range(r1$points[,1],c1$points[,1]),ylim=range(r1$p
oints[,1],c1$points[,1]),type="n",
xlab="Coordenada 1",ylab="Coordenada 2",lwd=2)
text(r1$points,labels=c("ED1","ED2","ED3","ED4","ED5"),lwd=2)
text(c1$points,labels=c("Nopar","parnS","parS"),lwd=4)

abline(h=0,lty=2)
abline(v=0,lty=2)
```


Coordenada 2
0.6
0.4
0.2
0.0
-0.2
-0.4
Nopar
ED1
ED2
ED3
ED5
parS
parnS
ED4
-0.4
-0.2
0.0
0.2
0.4
0.6
Coordenada 1