Capítulo 5

Anál. Múltiple de Correspondencias

5.1. INTRODUCCIÓN

El Análisis Múltiple de Correspondencias es una técnica que estudia las relaciones entre las categorías de Q variables cualitativas a partir de una muestra de n individuos.

Si existe alguna variable cuantitativa, ésta se puede transformar en cualitativa, dividiendo su rango en intervalos homogéneos. Cada intervalo será una categoría de la nueva variable cualitativa.

5.2. TABLAS DE DATOS

Tabla de datos inicial:

Ind.\ Var.	v_1	v_2	$\cdots$	v_q	$\cdots$	v_Q
1	v_{11}	v_{12}	$\cdots$	v_{1q}	$\cdots$	v_{1Q}
2	v_{21}	v_{22}	$\cdots$	v_{2q}	$\cdots$	v_{2Q}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	v_{i1}	v_{i2}	$\cdots$	v_{iq}	$\cdots$	v_{iQ}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	v_{n1}	v_{n2}	$\cdots$	v_{nq}	$\cdots$	v_{nQ}

Notación:

n núm. individuos, $\mathcal{I} = \{1, 2, \ldots, n\}$ índices de los individuos
Q núm. variables, $\mathcal{Q} = \{1, 2, \ldots, Q\}$ índices de las variables
p_q número de categorías de la variable v_q, $q \in \mathcal{Q}$.

Matriz disyuntiva: $K_{IJ} = (k_{ij})$
Esta matriz se construye de la forma siguiente:

a) Para cada variable v_q, se construyen tantas variables dummy como categorías de v_q, es decir, se construyen p_q variables binarias que ocupan p_q columnas en la matriz K_{IJ}.

b) Sea $p = \sum_{q \in \mathcal{Q}} p_q$ el total de categorías de las Q variables. Reenumeramos las categorías desde 1 hasta p y denotamos por $\mathcal{J} = \{1, \ldots, p\}$ al conjunto de todas las categorías.

La matriz disyuntiva K_{IJ} es

Ind.\ Categ.	1	2	$\cdots$	p	Total
1	k_{11}	k_{12}	$\cdots$	k_{1p}	
2	k_{21}	k_{22}	$\cdots$	k_{2p}	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	k_{n1}	k_{n2}	$\cdots$	k_{np}	
Total	k_{+1}	k_{+2}	$\cdots$	k_{+p}	

donde, para el individuo $i \in \mathcal{I}$ y la categoría $j \in \mathcal{J}$,

$$k_{ij} = \begin{cases} 1 & \text{si el individuo } i \text{ está en la categoría } j \\ 0 & \text{en otro caso} \end{cases}$$

Llamamos $\mathcal{J}_q$ al conjunto de índices de $\mathcal{J}$ asociados a las categorías de la variable v_q, de manera que $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2 \cup \cdots \cup \mathcal{J}_Q$.

Ejemplo 5.1 Se dispone de los datos de $n = 5$ individuos en $Q = 2$ variables cualitativas v_1 y v_2, cada una con $p_1 = p_2 = 3$ categorías.

Ind.\ Var.	v_1	v_2
1	1	2
2	1	3
3	2	1
4	1	3
5	3	2

En este caso

$$\mathcal{J}_1 = \{\ldots \qquad \}, \quad \mathcal{J}_2 = \{\ldots \qquad \}.$$

Construir la matriz disyuntiva K_{IJ}:

	$\mathcal{J}_1$			$\mathcal{J}_2$			
	—	—	—	—	—	—	
Ind.\Cat.	1	2	3	4	5	6	Total
1							
2							
3							
4							
5							

Propiedades:

(a) $\displaystyle\sum_{j \in \mathcal{J}_q} k_{ij} = \qquad .$

(b) $k_{i+} = \qquad .$

(c) $\displaystyle k_{+j} = \sum_{i \in \mathcal{I}} k_{ij}.$

(d) $\displaystyle\sum_{j \in \mathcal{J}_q} k_{+j} = \qquad .$

(e) $k_{++} = \qquad .$

Tabla de contingencia múltiple o matriz de Burt:

$$B_{JJ} = K'_{IJ} K_{IJ},$$

El elemento en la fila j y la columna ℓ de $B_{JJ} = (b_{j\ell})$ es

$$b_{j\ell} = \sum_{i \in \mathcal{I}} k_{ij} k_{i\ell}, \quad \forall j, \ell \in \mathcal{J}\,.$$

Ejemplo 5.2 Para la matriz inicial dada en el Ejemplo 5.1, la matriz de Burt es:

$$B_{JJ} = K'_{IJ} K_{IJ} = \left(\begin{array}{ccccc} 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{array}\right) \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right)$$

$$= \left(\begin{array}{c|c} & \\ & \\ & \\ \hline & \\ & \\ & \end{array}\right).$$

Propiedades:

(a) $\forall j, \ell \in \mathcal{J}_q, \quad b_{j\ell} = \begin{cases} 0, & \text{si } j \neq \ell \\ \quad\quad , & \text{si } j = \ell\,. \end{cases}$

(ningún individuo pertenece a dos categorías distintas j y ℓ de la misma variable)

(b) Si j y ℓ no son categorías de la misma variable, entonces $b_{j\ell}$ es el total de individuos en las categorías j y ℓ.

(c) $b_{j+} = ...$

(d) $b_{++} = ...$

Estas propiedades nos muestran la forma de tabla de contingencia múltiple de la matriz de Burt,

	v_1	v_2	$\cdots$	v_Q
v_1	k_{+1} $\ddots$ k_{+p_1}	Tabla de contingencia entre v_1 y v_2	$\cdots$	$\cdots$
v_2	Tabla de contingencia entre v_2 y v_1	$k_{+(p_1+1)}$ $\ddots$ $k_{+(p_1+p_2)}$		
$\vdots$	$\vdots$		$\ddots$	
v_Q	$\vdots$			$\ddots$

5.3. ANÁLISIS

El Análisis de Correspondencias Múltiple consiste en aplicar Análisis de Correspondencias Simple a la matriz disyuntiva K_{IJ}.

Proposición 5.1 El Análisis de Correspondencias Simple sobre la matriz disyuntiva K_{IJ} es equivalente al Análisis de Correspondencias Simple sobre la matriz de Burt B_{JJ}.

El ACS sobre la matriz disyuntiva K_{IJ} consiste en:

(1) Construir la tabla de frecuencias relativas. En este caso,

$$
\begin{aligned}
f_{ij} &= ... \\
f_{i+} &= ... \\
f_{+j} &= ...
\end{aligned}
$$

(2) Construir la tabla de perfiles fila R_i, $i = 1, \ldots, n$, y de perfiles columna C_j, $j = 1, \ldots, p$.

(3) Construir la tabla de puntos fila transformados R_i^* , $i = 1, \ldots, n$, o de perfiles columna transformados C_j^*, $j = 1, \ldots, p$.

(4) Obtener la matriz de puntos fila centrados X_i, $i = 1, \ldots, n$, o de puntos columna centrados Y_j, $j = 1, \ldots, p$.

(5) Calcular la matriz de inercia de los puntos fila $T = X' D_{\mathcal{I}} X$ o la de los puntos columna $S = Y D_{\mathcal{J}} Y'$. Calcular valores y vectores propios.

(6) Calcular las coordenadas de los puntos fila (individuos) y de los puntos columna (categorías) en los ejes principales, $G_{i\alpha}$ y $H_{j\alpha}$.

Proposición 5.2 La coordenada de un individuo i en un eje α, $G_{i\alpha}$, es el promedio de las coordenadas $H_{j\alpha}$ de las categorías j a las que pertenece el individuo i, afectada del factor $1/\sqrt{\lambda_\alpha}$.

Denotamos j_q categoría de la variable v_q a la que el individuo i pertenece;

$$G_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha} Q} \sum_{q \in \mathcal{Q}} H_{j_q \alpha}.$$

Proposición 5.3 El centro de gravedad de las coordenadas de las categorías de una variable es igual al centro de gravedad de las coordenadas de toda la nube de categorías.

$$\frac{\displaystyle\sum_{j\in\mathcal{J}_q} f_{+j}\, H_{j\alpha}}{\displaystyle\sum_{j\in\mathcal{J}_q} f_{+j}} = 0\,.$$

5.4. INERCIA DE LA NUBE DE CATEGORÍAS

Nube de puntos de las categorías:

C_j punto columna asociado a la categoría j
f_{+j} peso de C_j

$$N(\mathcal{J}) = \{(C_j, f_{+j}),\ j = 1, 2, \ldots, p\}\,.$$

Inercia de un punto columna:

$$I(C_j) = f_{+j}\, d_e^2(Y_j, O)\,,$$

donde $d_e^2(Y_j, O)$ viene dada por

$$d_e^2(Y_j, O) = \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{+j}\sqrt{f_{i+}}} - \sqrt{f_{i+}} \right)^2 .$$

Proposición 5.4 Llamamos $d_j = k_{+j}/n$ a la proporción de individuos que pertenecen a la categoría j. Se cumple

$$d_e^2(Y_j, O) = \frac{1 - d_j}{d_j}.$$

Por tanto, la inercia del punto columna (categoría) C_j es

$$I(C_j) = \frac{1 - d_j}{Q}.$$

La contribución de una categoría j de la nube $N(\mathcal{J})$ a la inercia total es mayor cuanto menor sea su frecuencia. En consecuencia, es aconsejable eliminar las categorías con frecuencias muy bajas, o unirlas a categorías próximas.

Proposición 5.5 La inercia de la nube de categorías de una varible v_q es

$$I[N(\mathcal{J}_q)] = \sum_{j \in \mathcal{J}_q} I(C_j) = \frac{p_q - 1}{Q}.$$

Como la contribución de una variable v_q a la inercia total de $N(\mathcal{J})$ aumenta con su número de categorías p_q, cada variable debería tener el mismo, o un número parecido de categorías.

Prueba. Efectivamente,

$$I[N(\mathcal{J}_q)] = \sum_{j \in \mathcal{J}_q} I(C_j) = \sum_{j \in \mathcal{J}_q} \frac{1-d_j}{Q} = \frac{1}{Q} \sum_{j \in \mathcal{J}_q} (1-d_j)$$

$$= \frac{1}{Q} \left(p_q - \sum_{j \in \mathcal{J}_q} d_j \right) = \frac{1}{Q} \left(p_q - \frac{1}{n} \sum_{j \in \mathcal{J}_q} k_{+j} \right) = \frac{p_q - 1}{Q} . \square$$

Proposición 5.6 La inercia total de la nube de categorías $N(\mathcal{J})$ es

$$I[N(\mathcal{J})] = \sum_{q \in \mathcal{Q}} I[N(\mathcal{J}_q)] = \sum_{q \in \mathcal{Q}} \frac{p_q - 1}{Q} = \frac{p - Q}{Q} .$$

5.5. CONTRIBUCIONES ABSOLUTAS Y RELATIVAS

Contribución absoluta de una categoría a la inercia de un eje:

La inercia del eje α es

$$I_\alpha = \sum_{j=1}^{p} f_{+j} (H_{j\alpha})^2 = \lambda_\alpha .$$

La contribución absoluta de la categoría j a la inercia del eje α es

$$ca_\alpha(j) = \frac{f_{+j} (H_{j\alpha})^2}{\lambda_\alpha}$$

Contribución absoluta de una variable a la inercia de un eje:

La contribución absoluta de una variable v_q a la inercia del eje α es la suma de las contribuciones de sus categorías

$$ca_\alpha(\mathcal{J}_q) = \sum_{j\in\mathcal{J}_q} ca_\alpha(j) = \sum_{j\in\mathcal{J}_q} \frac{f_{+j}(H_{j\alpha})^2}{\lambda_\alpha}\,.$$

Contribución relativa de un eje a la inercia de una categoría:

La inercia de la categoría j es

$$I(C_j) = f_{+j}\, d_e^2(Y_j, O),$$

donde la distancia de Y_j al origen es

$$d_e^2(Y_j, O) = \sum_{\alpha=1}^{r} (H_{j\alpha})^2 = \frac{1-d_j}{d_j}\,.$$

Por tanto, la contribución relativa del eje α a la inercia de la categoría j es

$$cr_\alpha(j) = \frac{(H_{j\alpha})^2}{d_e^2(Y_j, O)} = \frac{d_j(H_{j\alpha})^2}{1-d_j}\,.$$

Contribución relativa de un eje a la inercia de una variable:

La inercia de una variable v_q es

$$I[N(\mathcal{J}_q)] = \sum_{j\in\mathcal{J}_q} I(C_j) = \sum_{j\in\mathcal{J}_q} f_{+j} \sum_{\alpha=1}^{r} (H_{j\alpha})^2 = \frac{p_q-1}{Q}.$$

Por tanto, la contribución relativa del eje α a la inercia de la variable v_q es

$$cr_\alpha(\mathcal{J}_q) = \frac{\sum\limits_{j\in\mathcal{J}_q} f_{+j}(H_{j\alpha})^2}{I[N(\mathcal{J}_q)]} = \frac{\sum\limits_{j\in\mathcal{J}_q} f_{+j}(H_{j\alpha})^2}{(p_q-1)/Q}\,.$$

Una vez se ha seleccionado el número de ejes, la interpretación de los resultados se hace a partir de las contribuciones absolutas y relativas.

5.6. INTERPRETACIÓN

Las contribuciones absolutas ayudan a dar una interpretación a los ejes principales. En primer lugar, se buscan las variables que más contribuyen a la construcción de los ejes:

La contribución absoluta media de las variables a la inercia del eje α es

$$\begin{aligned}\frac{1}{Q}\sum_{q\in\mathcal{Q}} ca_\alpha(\mathcal{J}_q) &= \frac{1}{Q\lambda_\alpha}\sum_{q\in\mathcal{Q}}\sum_{j\in\mathcal{J}_q} f_{+j}\left(H_{j\alpha}\right)^2\\ &= \frac{1}{Q\lambda_\alpha}\sum_{j\in\mathcal{J}} f_{+j}\left(H_{j\alpha}\right)^2 = \frac{1}{Q}\,.\end{aligned}$$

Una variable v_q aporta suficiente inercia al eje α si

$$ca_\alpha(\mathcal{J}_q) > 1/Q.$$

Para cada variable significativa, observamos sus categorías. Una categoría se considera significativa si su contribución absoluta es mayor que la contribución absoluta media de las categorías de dicha variable.

Las contribuciones relativas nos dirán lo bien representadas que están las categorías por los ejes.

Se considera que una categoría está bien representada en el espacio de los ejes principales si la suma de sus contribuciones relativas con los ejes es de al menos un 60 %. Las categorías que no estén bien representadas no se pueden interpretar.

Se considera que una categoría está bien representada por un eje concreto si su contribución relativa con ese eje es de al menos un 30 %. Solo se puede interpretar una categoría en relación a un eje si está bien representada con ese eje.