Análisis conjunto en el análisis de correspondencia generada por una tabla de contingencia

Eliseo Martínez H.

1. Análisis conjunto

En nuestro artículo anterior, habíamos designado arbitrariamente una variable cualitativa como fila, de tal forma que, si ahora, en nuestra tabla de contingencia, la que era variable columna la designamos como variable fila, el análisis correspondiente para encontrar su mejores proyecciones es *mutatis mutandis*. Dicho de otra forma, realizamos el mismo análisis de correspondencia pero esta vez la tabla de contingencia será la traspuesta.

En cualquier caso, si el análisis se hace por fila debemos calcular los autovalores propios no nulos y distinto de 1 de la matriz $\mathbf{Z}^t\,\mathbf{Z}$, y si el análisis se hace por columnas se debe calcular los autovaloes no nulos y distinto de 1 de la matriz $\mathbf{Z}\,\mathbf{Z}^t$. Sin embargo estas matrices tienen los mismos autovalores propios no nulos y además los autovectores asociados a estos autovalores están relacionados. En efecto, supongamos que $\mathbf{a}_i$ es un vector propio de $\mathbf{Z}^t\,\mathbf{Z}$ asociado al autovalor λ_i, esto es

$$\mathbf{Z}^t\,\mathbf{Z}\,\mathbf{a}_i = \lambda_i \mathbf{a}_i$$

entonces si multiplicamos por $\mathbf{Z}$, tenemos

$$\mathbf{Z}\,\mathbf{Z}^t\,\mathbf{Z}\,\mathbf{a}_i = \lambda_i \mathbf{Z}\,\mathbf{a}_i$$

Y en esta ecuación observamos que el autovector asociado a λ_i de la matriz $\mathbf{Z}\,\mathbf{Z}^t$ es precisamente $\mathbf{Z}\,\mathbf{a}_i$.

De esta forma entonces, calculamos los autovectores (normalmente dos) de la matriz de dimensión más pequeña entre $\mathbf{Z}^t\,\mathbf{Z}$ y $\mathbf{Z}\,\mathbf{Z}^t$ y luego calculamos los autovectores de la otra matriz mediante $\mathbf{Z}\,\mathbf{a}_i$ o $\mathbf{Z}\,\mathbf{b}_i$, según sea el caso. Luego las coordenadas de cada fila se obtienen mediante

$$\mathbf{C}_f = \mathbf{D}_f^{-1/2}\mathbf{Z}\,\mathbf{A}_2 = \mathbf{Y}\mathbf{A}_2$$

donde $\mathbf{A}_2$ tiene en sus columnas los dos vectores propios de $\mathbf{Z}^t\,\mathbf{Z}$.

Las coordenadas de las columnas vienen dadas por

$$\mathbf{C}_c = \mathbf{D}_c^{-1/2}\mathbf{Z}^{\,t}\,\mathbf{B}_2 = \mathbf{Y}^t\mathbf{B}_2$$

donde $\mathbf{B}_2$ tiene en sus columnas los dos vectores propios de $\mathbf{Z}\,\mathbf{Z}^t$, y recordando que

$$\mathbf{Y} = \mathbf{D}_f^{-1}\mathbf{F}\,\mathbf{D}_c^{-1/2}$$

o de otra forma a través de sus entradas

$$y_{i\,j} = \left\{ \frac{f_{i\,j}}{f_{i\,\cdot}\sqrt{f_{\cdot\,j}}} \right\}$$

2. Asignación de puntuaciones

El análisis de correspondencia sirve también para un problema crucial que tienen los invstigadores cuando trabajan con variables cualitativas, en una tabla de contingencia, y, por razones de operabilidad numérica, desean cuantificar estas variables: "ponerles número". Pues bien, esete procedimiento se llama "asignar puntuaciones", y es un problema que lo puede solucionar el análisis de correspondencias.

Nuestro problema será entonces asignar valores numericos, que denotaremos por

$$y_c(1), \ldots, y_c(J)$$

a las columnas de la matriz $\mathbf{F}$ de observaciones.

Vamos a efectuar un análsis para ver las condiciones que deben cumplir las puntuaciones. Supongamos que hemos asignado el puntaje $y_c(1), ..., y_c(J)$ numérico a las columnas, esto generará automáticamente unos valores numéricos a las filas. En efecto, podemos asignar a la fila i el promedio de la variable y_c en esa fila, esto es

$$y_i = \frac{\sum_{j=1}^{J} f_{i\,j}\, y_c(j)}{\sum_{j=1}^{J} f_{i\,j}} = \sum_{j=1}^{J} r_{i\,j}\, y_c(j)$$

donde $r_{i\,j} = f_{i\,j} \,/\, f_{i\,\cdot}$ es la frecuencia relativa condicionada a la fila. El vector generado por todos estos promedio se puede escribir matricialmente como

$$\mathbf{y}_f = \mathbf{R}\,\mathbf{y}_c = \mathbf{D}_f^{-1}\,\mathbf{F}\;\mathbf{y}_c \tag{1}$$

De forma análoga, si damos ahora unas puntuaciones para las filas $\mathbf{y}_c$, las puntuaciones de las columnas se pueden estimar como las medias como las medias de cada columna, de esta manera obtenemos un vector $J \times 1$,

$$\mathbf{y}_c = \mathbf{D}_c^{-1}\mathbf{F}^{\,t}\;\mathbf{y}_f \tag{2}$$

Escribiendo de manera conjunta las ecuaciones (1) y (2) obtenemos

$$\mathbf{y}_f = \mathbf{D}_f^{-1}\,\mathbf{F}\;\mathbf{D}_c^{-1}\mathbf{F}^{\,t}\;\mathbf{y}_f \tag{3}$$

$$\mathbf{y}_c = \mathbf{D}_c^{-1}\mathbf{F}^{\,t}\;\mathbf{D}_f^{-1}\,\mathbf{F}\;\mathbf{y}_c \tag{4}$$

Las matrices $\mathbf{D}_c^{-1}\mathbf{F}^{\,t}$ y $\mathbf{D}_f^{-1}\,\mathbf{F}$ suman uno por filas, puesto que son frecuencias relativas normalizadas por los totales por columnas y filas respectivamente, de modo que las matrices cuadradas $\mathbf{D}_f^{-1}\,\mathbf{F}\;\mathbf{D}_c^{-1}\mathbf{F}^{\,t}$ y $\mathbf{D}_c^{-1}\mathbf{F}^{\,t}\;\mathbf{D}_f^{-1}\,\mathbf{F}$ tienen autovalor 1 (radio espectral) y por lo tanto $\mathbf{y}_f$ y $\mathbf{y}_c$ son los autovectores de estas matrices respectivamente, asociados al autovalor 1. Las soluciones triviales son $\mathbf{y}_f = (1, ..., 1)^t$; $\mathbf{y}_c = (1, ..., 1)^t$ y que ni por asomo parecen buen puntaje de asignación. Para encontrar una solución no trivial al problema vamos a realizar lo siguiente. A la ecuación (3) la multiplicamos por $\lambda\mathbf{D}_f^{1/2}$ y la ecuación (4) la multiplicamos por $\lambda\mathbf{D}_c^{1/2}$ y utilizando las ecuaciones (1) y (2) obtenemos

$$\lambda\,(\mathbf{D}_f^{1/2}\mathbf{y}_f) = \mathbf{D}_f^{-1/2}\,\mathbf{F}\;\mathbf{D}_c^{-1/2}(\mathbf{D}_c^{1/2}\,\mathbf{y}_c) \tag{5}$$

$$\lambda\,(\mathbf{D}_c^{1/2}\mathbf{y}_c) = \mathbf{D}_f^{-1/2}\,\mathbf{F}^{t}\;\mathbf{D}_c^{-1/2}(\mathbf{D}_f^{1/2}\,\mathbf{y}_f) \tag{6}$$

Observemos que exigimos a las puntuaciones que tengan una suerte de normalización con un factor de proporcionalidad.

Para resolver las ecuaciones (5) y (6) hagamos

$$\mathbf{b} = \mathbf{D}_f^{1/2}\,\mathbf{y}_f \;;\; \mathbf{a} = \mathbf{D}_c^{1/2}\,\mathbf{y}_c \;;\; \mathbf{Z} = \mathbf{D}_f^{-1/2}\,\mathbf{F}^t\;\mathbf{D}_c^{-1/2} \tag{6a}$$

nos quedan las ecuaciones

$$\lambda^2\,\mathbf{b} = \mathbf{Z}\mathbf{Z}^t\mathbf{b} \tag{7}$$

$$\lambda^2\mathbf{a} = \mathbf{Z}^t\,\mathbf{Z}\,\mathbf{a} \tag{8}$$

De modo que $\mathbf{b}$ y $\mathbf{a}$ son vectores propios asociados al autovalor λ^2 de las matrices $\mathbf{Z}\mathbf{Z}^t$ y $\mathbf{Z}^t\,\mathbf{Z}$. Los vectores de puntuación mediante la resoluciòn de las ecuaciones anteriores y en virtud de (6a) son

$$\mathbf{y}_f = \mathbf{D}_f^{-1/2}\,\mathbf{b} \tag{9}$$

$$\mathbf{y}_c = \mathbf{D}_c^{-1/2}\,\mathbf{a} \tag{10}$$

Puesto que las matrices $\mathbf{Z}\mathbf{Z}^t$ y $\mathbf{Z}^t\,\mathbf{Z}$ siempre admiten el valor propio 1, tomamos como $\mathbf{a}$ y $\mathbf{b}$ los vectores propios ligados al segundo valor propio menor que la unidad, y de estas matrices obtenemos las puntuaciones óptimas de filas y columnas.

Para obtener una representación gráfica de las filas y columnas. Sustituimos las puntuaciones $\mathbf{y}_c$ de (10) asociadas a las columnas en la ecuación (1), y escribimos

$$y_f\;(\mathbf{a}) = \mathbf{D}_f^{-1}\mathbf{F}\,\mathbf{D}_c^{-1/2}\,\mathbf{a}$$

obteniendo las proyecciones de las filas. Análogamente, sustituimos las puntuaciones $\mathbf{y}_f$ en la ecuación (9), y escribimos

$$y_c\;(\mathbf{b}) = \mathbf{D}_c^{-1}\mathbf{F}^{\,t}\,\mathbf{D}_f^{-1/2}\,\mathbf{b}$$

y encontramos las proyecciones de las columnas.

De esta forma el proceso de asignar puntuaciones a filas y columnas a una tabla de contingencia, es equivalente al problema de encontrar una representación óptima en una dimensión de las filas y columnas de la matriz. El análisis de correspondencia proporciona en la primera coordenada de las filas y columnas una forma consistente de asignar puntuaciones numéricas a las filas y a las columnas de la tabla de contingencia.

Veamos un ejemplo.

3. Ejemplo para asignar puntuaciones[1]

Este ejemplo sirve para asignar "puntuaciones" a un grupo de profsores que han sido evaluados mediante variables cualitativas que, para hacer las cosas sencillas, vamos a suponer que la variable tiene tres tipos de valores: Alta (A), media (M) y baja (B). Se adjunta la

[1] Este ejemplo fue obtenido del libro de Daniel Peña, Análisis de datos multivariantes, Edit. McGrawHill, 2002, pags. 214-215. Se eligió este ejemplo porque los resultados obtenidos en nuestro desarrollo, si bien parecen diferentes en la elección de los vectores propios, al "escalar" las puntuaciones los resultados son coincidentes.

tabla de frecuencia con los resultados

	A	M	B
P_1	2	6	2
P_2	4	4	4
P_3	1	10	4
P_4	7	5	0

Haremos el análisis de estos datos con un programa llamado *correspondencia.mth* (en DERIVE)[2]. La tabla de frecuencias relativas obtenida es

$$\mathbf{F} = \begin{pmatrix} 0.04081632653 & 0.1224489795 & 0.04081632653 \\ 0.08163265306 & 0.08163265306 & 0.08163265306 \\ 0.02040816326 & 0.2040816326 & 0.08163265306 \\ 0.1428571428 & 0.1020408163 & 0 \end{pmatrix}$$

Y además obtenemos la matriz $\mathbf{Z}$,

$$\mathbf{Z} = \begin{pmatrix} 0.1690308509 & 0.3794733192 & 0.2 \\ 0.3086066999 & 0.2309401076 & 0.3651483716 \\ 0.06900655593 & 0.5163977794 & 0.3265986323 \\ 0.5400617248 & 0.2886751345 & 0 \end{pmatrix}$$

Puesto que queremos asignar puntaje a los profesores, debemos calcular los vectores propios de la matriz $\mathbf{Z}\mathbf{Z}^t$, para esto calculamos primero los autovalores, que son

$$\lambda_1 = 1;\ \lambda_2 = 0.1985618768\,;\ \lambda_3 = 0.04900955170\,;\ \lambda_4 = 0$$

y puesto que el mayor autovalor después del 1 es $\lambda_2 = 0.1985618768$, calculamos su autovalor asociado que es

$$\mathbf{b} = \begin{pmatrix} -0.1658205077 & 0.004317520078 & -0.5811018713 & 0.7967468439 \end{pmatrix}$$

Entonces para calcular la puntuación debemos simplemente, en virtud de la ecuación (9), dividir las componentes de este vector por los valores $\sqrt{f_{i\,\cdot}}$, y obtenemos

$$\mathbf{y}_f = \begin{pmatrix} -0.3670593409 & 0.008724524827 & -1.050279005 & 1.610007016 \end{pmatrix}$$

Y estas son las puntuaciones asignadas a los profesores P_1, P_2, P_3 y P_4 respectivamente.

Para trabajar con puntuaciones clásicas, por ejemplo asignar puntajes en una escala del 0 al 10 hacemos la transformación

$$P(x) = \frac{x - \min \mathbf{y}_f}{\max \mathbf{y}_f - \min \mathbf{y}_f} \times 10$$

en el entendido que $\min \mathbf{y}_f$ y $\max \mathbf{y}_f$ corresponden al mínimo y máximo de las componentes del vector $\mathbf{y}_f$ respectivamente. Este escalamiento nos da la puntuación

$$\begin{pmatrix} 2.57 & 3.98 & 0 & 10 \end{pmatrix} \tag{11}$$

De esta manera asignamos puntaje 0.257 al profesor P_1; 0.398 al profesor P_2; 0 puntaje al profesor P_3 y puntaje máximo de 10 al profesor P_4.

A estas alturas debemos hacer una importante aclaración. Este puntaje asignado de

[2] Se encuentra ubicado en el sitio Internet:
http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/correspondencia.mth

manera alguna significa inmediatamente que el mejor puntaje lo obtuvo el profesor P_4 y el profesor P_3 fue el peor evaluado. Este puntaje debe interpretarse como una medida de espaciamiento entre los profesores, de otra forma la interpretación inmediata es que los profesores P_1 y P_2 están más próximos en la evaluación que hicieron los alumnos. Y esto es así toda vez que la puntuación depende del autovector elegido, en efecto, podríamos haber considerado el vector $-\mathbf{b}$ que es igualmente válido y en este caso los puntajes se habrían traspuestos cambiando simétricamente en las componentes. Estos puntajes se entienden a la luz de la propia tabla de contingencia. En efecto, si consideramos que los valores de alta, media y baja significan bien evaluado, medianamente evaluado y mal evaluado, entonces la puntuación entregada en (11) tiene sentido de jerarquía de mejor a peor, toda vez que, el profesor P_4 no tuvo evaluación en el valor de B (baja) como se puede observar en la tabla, y a su vez tuvo concentración mayoritaria en los valores de A y M. El profesor P_3 tiene sus evaluaciones concentradas en los valores de B y M que es notoriamente superior a las evaluaciones en esos valores respecto de los restantes profesores, y los profesores P_1 y P_2 si bien tienen un empate en los valores de B y M, el profesor P_2 le gana al profesor P_1 en la frecuencia de evaluación en A, de tal forma que tenemos indicios para pensar que, aparte de que las puntuaciones propuestas en (11) determinan un espaciamiento o distancia evaluativa, también esta puntuación indica un orden de evaluación que va de mejor a peor en el puntaje de 10 a 0. Este análisis es válido en tanto y en cuanto los valores de A, B y M indiquen la calidad de mejor a peor.

Hemos hecho este análisis toda vez que discrepamos de la puntuación elegida y fundamentada en el libro ya citado de Daniel Peña, páginas 214-215.[3].

[3] En el libro se propone la puntuación $\begin{pmatrix} 10 & 0 & 3.98 & 2.57 \end{pmatrix}$.