La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 8

Simetría entre el análisis de filas y el de columnas

Primera edición: julio 2008
ISBN: 978-84-96515-71-0

Traducción: Jordi Comas Angelet
Revisión: Carles M. Cuadras Avellana

www.fbbva.es

CAPÍTULO 8

Simetría entre el análisis de filas y el de columnas

En todos los ejemplos y análisis que hemos mostrado hasta ahora, nos hemos centrado en el análisis de las filas de una tabla. Hemos visualizado e interpretado las posiciones de los perfiles fila, utilizando las columnas como puntos de referencia, es el «análisis de filas». Sin embargo, podemos aplicar el análisis anterior de forma completamente simétrica a las columnas de la tabla. Lo podemos ver como una transposición de la tabla, en la que intercambiamos las filas por las columnas y viceversa, para a continuación repetir los procedimientos que hemos descrito del capítulo 2 al 7. En este capítulo, veremos que los análisis de filas y columnas están muy relacionados. En realidad, si llevamos a cabo el análisis de filas, también estamos efectuando el análisis de columnas y viceversa. Por tanto, podemos ver el AC como un análisis simultáneo de las filas y de las columnas de una tabla.

Contenido

Resumen del análisis de las filas

Consideremos de nuevo los datos de la tabla de la imagen 6.1 sobre la autopercepción de la salud. En el capítulo 6 hicimos el análisis de filas de estos datos porque queríamos representar los perfiles de los grupos de edad con relación a las categorías de salud. Estos siete perfiles se hallaban en un espacio tetradimensional, delimitado por los cinco vértices que representan los perfiles unidad extremos de cada una de las categorías de salud. Llegamos a la conclusión que la mayor parte de la variación espacial de los perfiles se producía en una recta (imagen 6.3). Finalmente, proyectamos e interpretamos las proyecciones de los perfiles y de los cinco vértices sobre la mencionada recta (imagen 6.5).

Análisis de columnas: los valores de los perfiles tienen una interpretación simétrica

Consideremos ahora la posibilidad de analizar los perfiles columna de la tabla de la imagen 6.1. Es decir, los perfiles de las categorías de salud con relación a los grupos de edad que mostramos en la tabla de la imagen 8.1. Para cada categoría de salud, los perfiles columna proporcionan porcentajes de individuos con relación a los grupos de edad. Por ejemplo, en la categoría *mala* salud, el 4,3% de los individuos tiene de 16 a 24 años, el 8,5% de 25 a 34 años, y así sucesivamente. A pesar de que la tabla de perfiles columna tiene un aspecto completamente distinto que el de la tabla de perfiles fila de la imagen 6.2, cuando nos fijamos en valores concretos y los comparamos con sus medias, vemos que contienen la misma información (en el capítulo 2, con los datos sobre mis viajes, ya nos dimos cuenta de ello). Consideremos, por ejemplo, el 23,7% de la columna *mala* del grupo de edad de 65 a 74 años. Comparemos este valor con el porcentaje de individuos de ese grupo de edad para el total de la muestra, que podemos encontrar en la última columna: el 11,2%. Llegamos a la conclusión que, en el grupo de edad de 65 a 74 años, algo más del doble de los encuestados manifiestan que su salud es *mala* en comparación con la media global de ese grupo de edad (el cociente es 23,7/11,2 = 2,1). Si nos fijamos ahora en la misma celda de la tabla de la imagen 6.2, vemos que el 13,7% del grupo de 65 a 74 años manifiesta que su salud es *mala*, mientras que esta proproción en el total de la muestra es del 6,5% (última fila de la tabla de la imagen 6.2). De nuevo, llegamos a la conclusión de que, en este grupo de edad, algo más del doble de los

Imagen 8.1:
Perfiles de la columna de las categorías de salud con relación a los grupos de edad, expresados como porcentajes

Grupo de edad	*Muy buena*	*Buena*	*Regular*	*Mala*	*Muy mala*	*Media*
16–24	29,7	22,3	11,2	4,3	5,8	*19,2*
25–34	26,9	22,8	11,0	8,5	5,8	*19,4*
35–44	18,0	18,6	12,1	9,9	7,8	*16,2*
45–54	11,0	13,2	15,8	12,1	15,5	*13,5*
55–64	6,5	11,7	20,5	25,6	29,1	*14,3*
65–74	5,4	7,5	19,0	23,7	19,4	*11,2*
75+	2,4	3,8	10,5	15,9	16,5	*6,2*
Suma	*100*	*100*	*100*	*100*	*100*	*100*

encuestados manifiesta que su salud es *mala*, en comparación con la media global de esta categoría de salud, el cociente es idéntico: 13,7/6,5 = 2,1.

Análisis de columnas: la misma inercia total

En el capítulo 4 vimos que la inercia total de los perfiles columna es igual a la inercia total de los perfiles fila; ambos cálculos son tan sólo maneras alternativas de expresar la misma fórmula: el estadístico χ^2 dividido por el tamaño de la muestra. Para los datos sobre la autopercepción de la salud, la inercia total es de 0,1404.

Análisis de columnas: igual dimensionalidad

Los perfiles columna definen una nube de cinco puntos, cada uno de ellos con siete componentes, que deben hallarse en un espacio de seis dimensiones, ya que la suma de sus componentes es 1. Sin embargo, los cinco puntos no llegan a ocupar las seis dimensiones de este espacio, pues sólo ocupan cuatro. Podemos percibir este hecho de forma intuitiva, si tenemos en cuenta que dos puntos se hallan exactamente en una recta unidimensional, tres puntos en un plano bidimensional, cuatro puntos en un espacio tridimensional y, por tanto, cinco puntos se hallarán en un espacio tetradimensional. En consecuencia, aunque los perfiles fila y los perfiles columna se encuentran en espacios distintos, la dimensionalidad de estas dos nubes de puntos es la misma, en este caso, de cuatro. Se trata de la primera coincidencia geométrica entre el análisis de los perfiles fila y el de los perfiles columna. Pronto veremos muchas más similitudes.

Análisis de columnas: la misma aproximación para reducir la dimensionalidad

Consideremos todavía los perfiles de las cinco categorías de salud en un espacio tetradimensional. Nos planteamos ahora las mismas preguntas que antes: ¿se pueden representar, de forma aproximada, estos puntos en un subespacio de pocas dimensiones?, ¿cuál es la calidad de esa aproximación? Haciendo el mismo tipo de cálculos matemáticos que vimos en el capítulo 6, llegamos a que podemos representar los perfiles columna en un espacio unidimensional y que la calidad de la representación es del 97,3%, exactamente el mismo porcentaje que hallamos con los perfiles fila. Estamos ante una segunda coincidencia geométrica entre los dos análisis.

Análisis de columnas: los mismos valores de coordenadas pero redimensionados

En el mapa de la imagen 8.2 representamos las proyecciones de los perfiles columna sobre la recta que mejor se ajusta a los perfiles. Vemos que las categorías de salud se sitúan exactamente en el mismo orden que los vértices en el mapa de la imagen 6.5. Aunque los valores de sus coordenadas no son iguales, sus posiciones relativas son idénticas. Comparando las posiciones de las categorías de salud

Imagen 8.2:
Mapa unidimensional óptimo de los perfiles de las categorías de salud

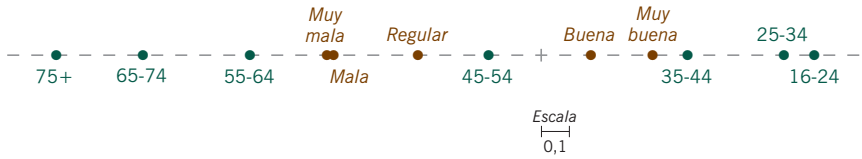

Imagen 8.3:
El mismo mapa de la imagen 8.2, que muestra las posiciones de las proyecciones de los vértices de los grupos de edad

del mapa de la imagen 8.2 con la escala del mapa de la imagen 6.5, vemos que las coordenadas de los perfiles son una versión encogida, o contraída, de las posiciones de los vértices. Pronto interpretaremos de forma específica este «factor de contracción». Sin embargo, profundicemos un poco más. En el mapa de la imagen 8.3 representamos las proyecciones de los siete vértices que corresponden a los grupos de edad sobre la misma recta. La comparación de las posiciones de los vértices aquí con la de los perfiles de los grupos de edad en el mapa de la imagen 6.5 (o con las posiciones del mapa de la imagen 6.3 en una escala mayor) pone de manifiesto el mismo fenómeno, pero para las filas; en el mapa de la imagen 6.5, las posiciones de los perfiles fila con relación a la recta que mejor se ajusta son una versión encogida de las posiciones de los vértices de los grupos de edad proyectados sobre la recta, que mejor se ajusta a los perfiles de las categorías de salud del mapa de la imagen 8.2. Es decir, en el análisis de columnas, las posiciones de los vértices fila son una expansión de las posiciones de los perfiles fila del análisis de filas. Estamos ante la tercera, y más importante, coincidencia geométrica entre los dos análisis.

Ejes e inercias principales

En estos análisis, la recta que mejor se ajusta se denomina *eje principal*. En los próximos capítulos veremos que existen otros ejes principales. Por ello, de forma más precisa, llamaremos a esta recta «primer eje principal». Hemos visto que tanto en el análisis de filas, como en el análisis de columnas la inercia total es de 0,1404 y que, en ambos casos, el porcentaje de inercia explicada por el primer eje es del 97,3%. También en ambos casos, el valor concreto de la inercia explicada por el primer eje es de 0,1366, por tanto, el porcentaje de inercia explicada es igual a $100 \times 0{,}1366/0{,}1404 = 97{,}3\%$. Llamamos *inercia principal* a la inercia explicada por un eje principal (de 0,1366 en este caso). En este ejemplo se trata de la primera inercia principal ya que nos referimos a la inercia del primer eje principal. La inercia principal también recibe el nombre de *valor propio*, ya que se puede calcular como un valor propio de una matriz cuadrada simétrica.

El factor de escala es la raíz cuadrada de la inercia principal

Parece pues, que tenemos que hacer un solo análisis: de filas o de columnas. Los resultados de uno de ellos se pueden obtener de los resultados del otro. Sin embargo, ¿cuál es exactamente la relación entre ambos? Dicho de otra manera, ¿cuál es el factor de escala que nos permite pasar de las posiciones de los vértices de un análisis a las posiciones de los perfiles del otro? Pues bien, este factor de escala es igual a la raíz cuadrada de la inercia principal. Así, en este ejemplo,

Categoría de la salud	*Coordenadas de perfiles*
Muy buena	0,423
Buena	0,198
Regular	–0,439
Mala	–0,755
Muy mala	–0,767

Grupos de edad	*Coordenadas de vértices*
16–24	1,004
25–34	0,893
35–44	0,538
45–54	–0,192
55–64	–1,070
65–74	–1,463
75+	–1,782

Imagen 8.4:
Valores de las coordenadas de los puntos del mapa de la imagen 8.2, es decir las coordenadas de los perfiles columna y de los vértices de las filas en el primer eje principal de los perfiles columna (compárese con las tablas de la imagen 7.1)

es $\sqrt{0{,}1366} = 0{,}3696$. En consecuencia, para pasar de los vértices fila de la imagen 8.3, a los perfiles fila de los mapas de las imágenes 6.3 o 6.5, simplemente multiplicamos los valores de las coordenadas por 0,3696, es decir, algo más de un tercio. A la inversa, para pasar de los perfiles columna del mapa de la imagen 8.3 a los vértices columna del mapa de la imagen 6.5, multiplicamos los valores de las coordenadas por el inverso de este valor, concretamente por $1/0{,}3696 = 2{,}706$. En las tablas de las imágenes 7.1 y 8.4 se muestran todos los valores numéricos de las coordenadas de los perfiles y de las coordenadas de los vértices. Comparando los valores de ambas imágenes llegamos a la expresión:

$$\text{Coordenada del perfil} = \text{coordenadas del vértice} \times \sqrt{\text{inercia principal}}$$

Fijémonos que en los mapas de las imágenes 6.5 y 8.3, los perfiles están más juntos que los vértices. El factor de escala es una medida directa de lo apretados que están los perfiles «interiores» en comparación con los vértices «exteriores». En este caso, un factor de escala de 0,3696 indica que la dispersión de los perfiles es aproximadamente un tercio de la de los vértices. Al final del capítulo 4 interpretamos la inercia total como una medida de la dispersión de los perfiles en relación a los vértices exteriores (imagen 4.2). Las inercias principales (o sus raíces cuadradas) son también medidas de dispersión, pero se refieren a los ejes principales de forma individual, no al espacio de perfiles en su conjunto. Cuanto mayor sea la inercia principal, y en consecuencia cuanto mayor sea el factor de escala, mayor será la dispersión de los perfiles con relación a los vértices en el eje principal. En consecuencia, es obvio que la inercia principal no puede ser mayor que 1, pues los perfiles deben hallarse en el «interior» de sus correspondientes vértices.

La correlación como una interpretación de la inercia principal

La raíz cuadrada de la inercia principal, que ya hemos señalado que siempre toma un valor menor de 1, tiene otra interpretación como coeficiente de correlación. En general, los coeficientes de correlación se calculan entre pares de medidas, como por ejemplo la correlación entre los ingresos y la edad. En el caso que nos ocupa, para cada encuestado tenemos dos observaciones —el grupo de

edad y la categoría de la salud—, pero se trata de observaciones categóricas, no de medidas. Podemos calcular el coeficiente de correlación entre estas dos variables recurriendo a los códigos enteros que utilizamos anteriormente por defecto, es decir de 1 a 7 para los grupos de edad y de 1 a 5 para las categorías de salud. En tal caso obtenemos una correlación de 0,3456. Utilizando otros valores, obtendríamos otras correlaciones. Por tanto, nos podemos plantear las siguientes preguntas: para obtener la máxima correlación, ¿qué valores debemos utilizar para los grupos de edad?, ¿y para las categorías de salud? Llamamos *correlación canónica* a la correlación máxima que obtenemos de esta manera. En este ejemplo, la correlación canónica es de 0,3696, exactamente la raíz cuadrada de la inercia, es decir, el factor de escala que vincula el análisis de filas con el de columnas. Los valores numéricos de los grupos de edad y de las categorías de salud que dan la máxima correlación son precisamente los valores de las coordenadas de los grupos de edad y de las categorías de salud en el eje principal del AC que aparecen en las imágenes 7.1 y 8.4 y que representamos gráficamente en las imágenes 6.3, 6.5, 8.2 y 8.3. Podemos utilizar las coordenadas de los perfiles o las coordenadas de los vértices, ya que la correlación no se ve afectada ni por un cambio de origen ni por redimensionamiento de las escalas. Sin embargo, en general, utilizamos escalas estandarizadas de media 0 y varianza 1.

Representación gráfica de la correlación

Es habitual mostrar gráficamente la correlación entre dos variables en un diagrama de dispersión de los casos, como por ejemplo los grupos de edad (eje *y*) y categorías de la salud (eje *x*). En este diagrama de dispersión tenemos 6371 casos, sin embargo, sólo aparecen 7 valores en el eje *y*, y 5 en el *x*. Por tanto, en el diagrama de dispersión tenemos sólo $7 \times 5 = 35$ posibles puntos (imagen 8.5). En cada punto hallamos todos los casos de la categoría de la salud y del grupo de edad de la celda correspondiente de la tabla de contingencia original (imagen 6.1). Aquí hemos representado los puntos como cuadrados de área proporcional a la frecuencia de la celda. En este diagrama de dispersión, la correlación canónica de los 6371 individuos, es igual a la correlación de Pearson. Cuando decimos que la correlación canónica es óptima queremos decir que no existen otros valores de las categorías de las filas y de las categorías de las columnas que proporcionen un coeficiente de correlación mayor. Obtendríamos una correlación canónica igual a 1 cuando todos los puntos se hallaran en una recta, en este caso significaría que cada grupo de edad está asociado sólo con una categoría de salud (los perfiles serían todos perfiles unidad, es decir, vértices).

Coordenadas principales y coordenadas estándares

En esta etapa es conveniente introducir algo de terminología para evitar tener que repetir continuamente las expresiones «coordenadas de las posiciones de los vértices» y «coordenadas de las posiciones de los perfiles». Hemos estandarizado las primeras para que tengan media 0 y varianza 1, y las llamaremos *coordenadas*

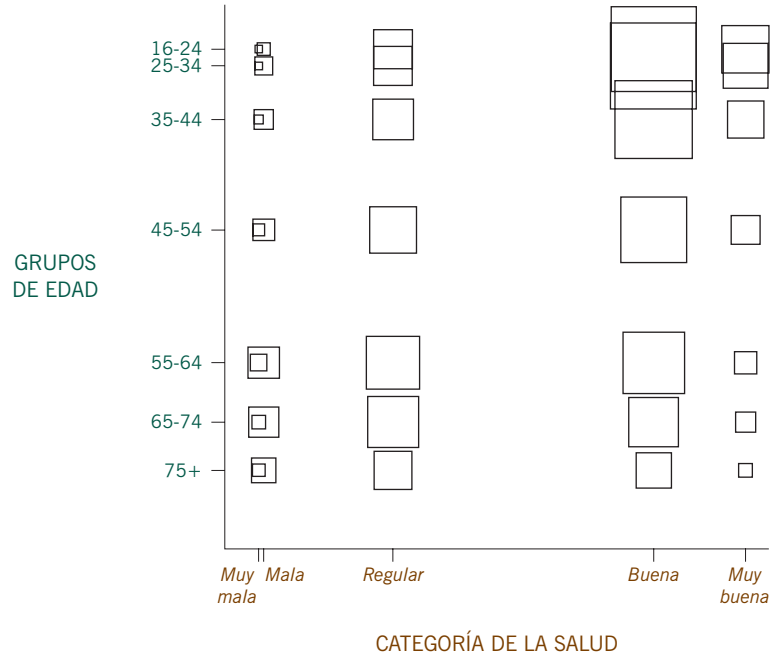

Imagen 8.5:
Diagrama de dispersión de los valores que maximizan la correlación entre las categorías de salud y los grupos de edad; los cuadrados correspondientes a cada combinación de valores tienen un área proporcional al número de individuos. La correlación es igual a 0,3456

estándares; las segundas son las coordenadas de los perfiles en los ejes principales, y las denominaremos *coordenadas principales*. Por ejemplo, en las tablas de la imagen 8.4, la primera columna de resultados numéricos contiene las coordenadas principales de las categorías de la salud, mientras que la segunda columna contiene las coordenadas estándar de los grupos de edad. En ambos casos son coordenadas en el primer eje principal del AC. Veremos que en capítulos posteriores tendremos, en general, más de un eje principal.

Maximización del cuadrado de las correlaciones con la media

Vamos a ver otra manera de calcular la correlación en el AC. Así, a cada uno de los 6371 individuos del ejemplo sobre la encuesta de salud le asignemos un par de valores: uno (a_i) para el grupo de edad y otro (h_j) para la categoría de salud. Como antes, estos valores son desconocidos, sin embargo, vamos a definir un criterio de optimización que nos permita determinarlos. Supongamos que cada individuo tiene una puntuación igual a la suma de los dos valores $a_i + h_j$. Por ejemplo, alguien del grupo de edad de 25 a 34 años con *muy buena* salud (segundo grupo de edad y primera categoría de salud) tendría una puntuación igual a $a_2 + h_1$. Supongamos ahora que indicamos la correlación entre todos los pares de valores $\{a_i, a_i + h_j\}$ como cor$(a, a + h)$, donde a y h indican los 6371 valores de la muestra. De forma similar, indicamos la correlación entre todos los pares $\{h_j, a_i + h_j\}$ como cor$(h, a + h)$. Habría que buscar unas escalas que optimizaran estas dos correlaciones. Se puede demostrar que la primera dimensión del AC

proporciona unos valores que son óptimos en el sentido de que maximizan la media de los cuadrados de estas correlaciones:

$$\text{media de los cuadrados de las correlaciones} = \frac{1}{2}[\text{cor}^2(a, a+h) + \text{cor}^2(h, a+h)] \quad (8.1)$$

Dado que para cualquier par de variables estandarizadas X e Y, la $\text{cor}(X, X+Y) = \sqrt{[1+\text{cor}(X,Y)]/2}$, en (8.1) la media de los cuadrados de las correlaciones será igual a:

$$\text{media de los cuadrados de las correlaciones} = \frac{1+\text{cor}(a,h)}{2} \quad (8.2)$$

En consecuencia, cuando con el AC maximizamos la $\text{cor}(a,h)$, es decir, obtenemos la correlación canónica (8.2), también maximizamos (8.1). Este resultado nos será útil más tarde ya que lo podemos generalizar fácilmente a más de dos variables, como veremos en el capítulo 20.

Minimización de la pérdida de homogeneidad entre variables

Utilizando la notación anterior, podemos ver otro criterio de optimización que también nos conduce a los resultados del AC. En primer lugar, en vez de calcular las sumas de los valores de cada individuo, calculemos la media de estos valores, $\frac{1}{2}(a_i + h_j)$. A continuación calculemos las diferencias entre los valores de la edad y de salud de cada individuo con su media: $a_i - \frac{1}{2}(a_i + h_j)$ y $h_j - \frac{1}{2}(a_i + h_j)$. Una medida de la similitud entre los valores de edad y los de la salud de cada individuo es la media de la suma de cuadrados de estas dos diferencias, lo que nos lleva a una medida de la varianza de los valores a_i y h_j:

$$\text{varianza (de un caso)} = \frac{1}{2}\left([a_i - \frac{1}{2}(a_i + h_j)]^2 + [h_j - \frac{1}{2}(a_i + h_j)]^2\right) \quad (8.3)$$

Sin embargo, en este contexto, preferimos el término *homogeneidad* porque si los valores a_i y h_j fueran iguales, su varianza sería cero; a un individuo con esta combinación de categorías le llamamos individuo *homogéneo*. Un término alternativo a homogeneidad es el de *consistencia interna*. Calculando la media de los valores (8.3) para todos los individuos de la muestra, obtenemos un valor llamado *pérdida de homogeneidad* (en la página 69, se usa este término en el mismo sentido). Si todos los valores de edad coincidieran con los de salud, la pérdida de homogeneidad sería cero, es decir la muestra sería completamente homogénea (o internamente consistente). El objetivo del AC es hallar una escala de valores que minimice esta pérdida. Una vez más, los valores que minimizan la pérdida de homogeneidad coinciden con las coordenadas de la edad y de salud de la primera dimensión del AC. Como veremos en el capítulo 20, podemos fácilmente extender esta definición a más de dos variables.

RESUMEN:
Simetría entre el análisis de filas y el de columnas

1. Todo lo que hemos hecho en el análisis de filas lo podemos aplicar de forma completamente simétrica a las columnas, como si repitiéramos todas las operaciones en la tabla transpuesta.
2. Con el análisis de columnas visualizamos los perfiles de las columnas y los vértices de las filas en el subespacio de representación óptimo de los perfiles de las columnas.
3. El (primer) *eje principal* de perfiles es la recta, o dimensión, que mejor se ajusta y la (primera) *inercia principal* es la inercia explicada por esta dimensión.
4. Las *coordenadas principales* son las posiciones de las coordenadas de los perfiles en un eje principal, y las *coordenadas estándares* son las posiciones de las coordenadas de los vértices en un eje principal.
5. Los dos análisis son equivalentes en el sentido de que tienen la misma inercia total, la misma dimensionalidad y la misma descomposición de la inercia total en inercias de los ejes principales.
6. Además, en ambos análisis, los perfiles y los vértices están íntimamente relacionados de la siguiente manera: en un eje principal, las posiciones de los perfiles (en coordenadas principales) tienen exactamente las mismas posiciones relativas que los correspondientes vértices (en coordenadas estándares) en el otro análisis, pero con valores contraídos. El factor de escala implicado es exactamente la raíz cuadrada de la inercia principal de ese eje.
7. Este factor de escala también se puede interpretar como una *correlación canónica*, especialmente cuando nos referimos al primer eje principal. Se trata de la máxima correlación que podemos obtener con las variables fila y las variables columna como resultado de la asignación de valores numéricos a las categorías de estas variables.