La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 12

Puntos adicionales

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

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

www.fbbva.es

CAPÍTULO 12

Puntos adicionales

Con frecuencia ocurre que tenemos filas y/o columnas de datos que no hemos considerado inicialmente, pero que, sin embargo, nos pueden ser útiles para interpretar características que hayamos descubierto en los datos originales. Siempre que tenga sentido comparar los perfiles de estas nuevas filas (o columnas) con los de las filas (o columnas) de la matriz de datos originales que configuraron el mapa, tendremos la posibilidad de añadirlos en el mapa. Llamamos *puntos adicionales* o *suplementarios* a las filas o columnas que añadimos en un mapa preexistente.

Contenido

Puntos activos

Hasta ahora hemos utilizado todas las filas y columnas de una determinada tabla de datos para configurar los ejes principales y, en consecuencia, el mapa: son las filas y columnas *activas* del análisis. No todos los puntos activos tienen la misma fuerza de atracción sobre los ejes principales. Esta fuerza de atracción depende de la posición del punto y de su masa. Los perfiles alejados de la media «influyen» principalmente en la orientación del mapa, mientras que los perfiles con mayor masa tienen más «fuerza de atracción» sobre los ejes.

Imagen 12.1: *Frecuencias de las categorías de financiación de 796 investigadores (imagen 10.1), con una columna adicional Y, correspondiente a una nueva categoría de «nuevos investigadores prometedores», una fila adicional correspondiente a los investigadores que trabajan en museos, y una nueva fila que contiene la suma de las frecuencias de* Estadística *y* Matemáticas, *etiquetada como Ciencias matemáticas*

	Categoría de financiación					
Disciplina científica	*A*	*B*	*C*	*D*	*E*	*Y*
Geología	3	19	39	14	10	*0*
Bioquímica	1	2	13	1	12	*1*
Química	6	25	49	21	29	*0*
Zoología	3	15	41	35	26	*0*
Física	10	22	47	9	26	*1*
Ingeniería	3	11	25	15	34	*1*
Microbiología	1	6	14	5	11	*1*
Botánica	0	12	34	17	23	*1*
Estadística	2	5	11	4	7	*0*
Matemáticas	2	11	37	8	20	*1*
Museos	*4*	*12*	*11*	*19*	*7*	
Ciencias matemáticas	*4*	*16*	*48*	*12*	*27*	

Definición de puntos adicionales

Sin embargo, hay situaciones en las que deseamos visualizar las proyecciones de determinados puntos que no queremos que intervengan en el cálculo de la configuración del mapa. Queremos que la configuración del mapa se ajuste sólo a los puntos activos. Lo más simple es considerarlos como puntos que tienen una posición en el mapa, pero no tienen masa. Es decir, son puntos que no contribuyen a la inercia, y, por tanto, no influyen en la configuración de los ejes principales. Los llamamos *puntos adicionales*, *pasivos* o *suplementarios* y así los distinguimos de los puntos activos que sí tienen masa. Existen tres situaciones en las que las filas o las columnas adicionales pueden ser útiles. Vamos a ilustrar cada una de ellas en el contexto de los datos sobre la financiación de la investigación científica que vimos en capítulos precedentes. En la imagen 12.1 mostramos una versión ampliada de los mencionados datos. Hemos añadido:

1. Una fila adicional, etiquetada como *Museos*, que contiene las frecuencias de investigadores que trabajan en museos (a diferencia de los restantes que trabajan en universidades).
2. Una columna adicional, etiquetada como *Y*. Se trata de una categoría especial de financiación para investigadores jóvenes, una categoría que se acaba de introducir en el sistema de financiación.
3. Y otra fila, etiquetada como *Ciencias matemáticas,* que es la suma de Estadística y Matemáticas.

Primer caso: un punto intrínsicamente diferente a los restantes

El estudio del que derivan estos datos estaba, en principio, centrado en investigadores universitarios. Sin embargo, los investigadores de los museos tienen niveles similares y obtienen recursos de las mismas organizaciones financieras, por tanto, las frecuencias de 53 investigadores que trabajan en museos están clasificados en

Imagen 12.2:
Mapa simétrico de los datos de la imagen 12.1 (la podemos comparar con la imagen 10.2) que, además, muestra la posición de la columna adicional Y, y las filas adicionales Museos y Ciencias matemáticas

las mismas categorías de financiación. Aunque sea necesario considerar a los investigadores que trabajan en los museos de forma separada de los que trabajan en universidades, sigue teniendo interés visualizar los perfiles de los primeros en el «espacio» de los investigadores universitarios. Lo podemos hacer declarando la fila *Museos* como un punto adicional. De esta manera, su perfil no participará en la configuración de los ejes principales. De todas maneras, su perfil puede proyectarse sobre el mapa. En la imagen 12.2 mostramos un mapa simétrico, como el de la imagen 10.2, pero con el punto adicional *Museos,* abajo a la izquierda del mapa. Este punto no contribuye a la inercia principal, pero permite examinar las contribuciones relativas de los ejes a este punto (es decir, los cosenos o correlaciones al cuadrado). Así, podemos ver que este punto queda bastante bien representado en el mapa, ya que la contribución relativa de los ejes es superior al 50%. Su posición indica que relativamente pocos de los investigadores que trabajan en

museos han visto rechazados sus proyectos, pese a que el nivel de financiación de sus proyectos es bajo. A un conjunto de datos activo le podemos añadir distintos tipos de información adicional. Información que podemos obtener del mismo estudio, como es el caso de los *Museos* que acabamos de ver. O información que proceda de estudios similares, por ejemplo, con el objetivo de seguir la evolución, en función del tiempo, de las posiciones de las disciplinas científicas con relación a las categorías de financiación. Así, podríamos añadir en el mapa filas adicionales correspondientes a los datos de una tabla de frecuencias similar obtenida de un estudio anterior sobre la financiación de los investigadores científicos. Otro ejemplo sería añadir en el mapa perfiles objetivo para las distintas disciplinas científicas. De esta manera podríamos valorar lo lejos que quedan las posiciones actuales de las posiciones objetivo. Este concepto de «punto ideal» se utiliza frecuentemente en estudios sobre el posicionamiento de productos en investigaciones de mercados.

Segundo caso: una observación atípica de poca masa

Supongamos que se acaba de introducir una categoría de financiación adicional *Y,* de la que hasta el momento se han beneficiado muy pocos investigadores; de hecho, sólo seis investigadores de seis disciplinas distintas. Esto significa que el perfil de esta columna es muy poco común: seis valores del perfil son iguales a $\frac{1}{6}$ = 0,167, y los restantes valores son 0. Ningún otro perfil columna tiene el más mínimo parecido con éste, por tanto, hay que esperar que su posición en el espacio multidimensional sea inusual. Efectivamente, como se aprecia en la imagen 12.2, este punto es una *observación atípica*. Si lo hubiéramos incluido en el análisis como punto activo, hubiera contribuido mucho su configuración. No sería una situación deseable, ya que la columna *Y* está constituida sólo por seis individuos —por tanto, aparte de una razón sustantiva, existe una razón de tipo técnico que nos aconseja considerar este punto como punto adicional—. En este caso particular, si incluyéramos a *Y* como punto activo, y a pesar de que su masa representa menos del 1% de la masa de las columnas, la inercia total de la tabla pasaría de 0,0829 a 0,0920, un incremento del 11%. Además, como vemos en la imagen 12.3, cambiaría sustancialmente la configuración del mapa: se produciría una rotación de 30º en comparación con el resultado anterior. La inclusión de *Y* ha hecho girar los ejes. Debemos ponernos en guardia ante este tipo de observaciones atípicas de poca masa que contribuyen mucho a la inercia de la solución. En algunos casos extremos, las observaciones atípicas dominan tanto el mapa que los contrastes más interesantes entre categorías con mayores frecuencias quedan completamente enmascarados. Si declaramos las observaciones atípicas puntos adicionales, podemos seguir visualizando sus posiciones, sin que influyan en la configuración final del espacio. Otra posibilidad es combinar, siempre que tenga sentido, las filas (o las columnas de poca masa) con otras filas (o columnas). Así, si tuviéramos una disciplina adicional, como por ejemplo «Infor-

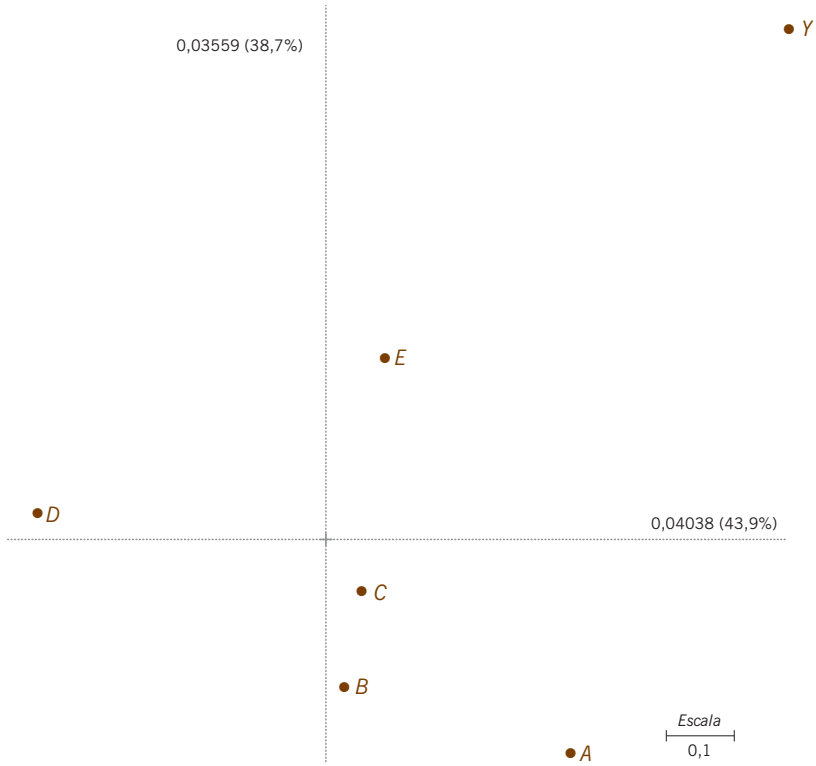

Imagen 12.3:
Mapa del AC de las columnas de la imagen 12.1 en el que hemos incluido Y como punto activo. Para facilitar las comparaciones, hemos expresado los ejes de los mapas de las imágenes 12.2 y 12.3 en la misma escala. Sin embargo, en la imagen 12.3 hemos rotado los ejes 30º respecto a los de la imagen 12.2

mática», con muy pocos investigadores y un posible perfil extraño, podríamos combinar esta disciplina con otra disciplina similar, como por ejemplo Matemáticas o Ingeniería. De todas formas, a pesar de lo que acabamos de comentar, es oportuno que señalemos que, en general, las observaciones atípicas de poca masa, no son un problema serio en el AC. Su influencia viene determinada por su masa multiplicada por una distancia al cuadrado, por tanto, la poca masa disminuye su influencia. El problema real es que las observaciones atípicas se hallen muy lejos de los restantes puntos (volveremos a este tema en el capítulo 13, cuando veamos distintas posibilidades de elección de las escalas de los mapas).

Tercer caso: representación de grupos o subdivisiones de puntos

También podemos utilizar los puntos adicionales para representar grupos de categorías o subdivisiones de una categoría. Por ejemplo, la fila adicional *Ciencias matemáticas* de la imagen 12.1, corresponde a la suma de las frecuencias de Matemáticas y Estadística, dos disciplinas que a menudo se agrupan. El perfil de esta nueva fila es el centroide de las dos filas que lo componen, ponderadas con sus respectivas masas. Dado que en Matemáticas y en Estadística hay 78 y 29 investigadores, respectivamente, el perfil de *Ciencias matemáticas* será:

$$\text{perfil de } \textit{Ciencias matemáticas} = \frac{78}{107} \times \text{perfil de Matemáticas} + \frac{29}{107} \times \text{perfil de Estadística}$$

de manera que el perfil de *Ciencias matemáticas* se parecerá más al perfil de Matemáticas que al de Estadística. Geométricamente, esto significa que el punto que representa el perfil de *Ciencias matemáticas* se halla en la línea que une Matemáticas con Estadística, pero más cerca de Matemáticas (compárese con la imagen 12.2). Vamos a representar *Ciencias matemáticas*, en la imagen 12.2, como un punto adicional. No lo podemos considerar un punto activo, ya que si así lo hiciéramos resultaría que en el análisis habríamos considerado dos veces los 107 investigadores de ambas disciplinas. En mapas de AC ya existentes representaremos de la misma manera las subdivisiones de categorías. Así, por ejemplo, supongamos que disponemos de datos que nos permiten subdividir Ingeniería en distintas ramas, como eléctrica, mecánica, civil, etc. Consideraremos estas nuevas filas como adicionales e investigaremos las posiciones de sus perfiles en el mapa. Igual que ocurría con *Ciencias matemáticas*, el punto activo Ingeniería es el centroide de las distintas ramas de la ingeniería que hemos considerado.

Cálculo de las posiciones de los puntos adicionales

Hasta ahora, hemos descrito los puntos adicionales como perfiles adicionales que proyectamos en un mapa calculado anteriormente. Una manera alternativa de obtener las posiciones de los puntos adicionales es situarlos con relación a los vértices de un mapa asimétrico. Así, en el capítulo 3 vimos que las posiciones de los perfiles fila resultaban del cálculo de la media ponderada de los vértices columna, ponderados con los elementos de los perfiles. Podemos situar los puntos adicionales exactamente de la misma forma. Una vez determinadas las posiciones de los ejes principales de los perfiles fila, conocemos, en cada eje principal, las posiciones de las coordenadas de los vértices que representan las columnas, es decir, las coordenadas estándares de las columnas. A partir de este momento, podemos situar los puntos adicionales calculando sus centroides con relación a los vértices ponderando con los elementos de sus perfiles. Por ejemplo, para calcular la posición del punto adicional *Museos*:

$$\text{posición de } \textit{Museos} = \frac{4}{53} \times \text{vértice } A + \frac{12}{53} \times \text{vértice } B + \ldots \text{ etc.}$$

es decir, calculamos la media ponderada de las coordenadas estándares de las columnas en cada uno de los ejes principales.

Contribuciones de los ejes a los puntos adicionales

Dado que los puntos adicionales tienen masa cero, su inercia es cero, por lo que no contribuyen a las inercias principales. Sin embargo, sigue siendo válida la interpretación de las contribuciones relativas de los ejes, en términos de los ángulos formados entre perfiles y ejes. Ello nos permite determinar si los puntos adicionales están bien representados. En el espacio bidimensional, las contribu-

ciones relativas de los ejes y las calidades de la representación de los tres puntos adicionales descritos anteriormente son las siguientes:

	Contribución relativa		*Calidad en dos dimensiones*
PUNTOS ADICIONALES	*Eje 1*	*Eje 2*	
Museos	225	331	556
Ciencias matemáticas	493	66	559
Y	4	587	641

Estos valores describen la bondad de la representación de los puntos adicionales. Por ejemplo, el coseno al cuadrado del ángulo del punto adicional *Y* con el primer eje es de 0,054, y con el segundo eje es de 0,587. Por tanto, la calidad de su representación en el plano es de 0,054 + 0,587 = 0,641, es decir, el 64,1% de su posición se halla en el mencionado plano, mientras que el 35,9% de su posición se halla en las restantes dimensiones. También podemos decir que la correlación de *Y* con el plano es $\sqrt{0,641} = 0,801$.

Los vértices son puntos adicionales

En realidad, ya nos encontramos anteriormente con puntos adicionales, ya que en el cálculo del mapa no tenemos en cuenta las posiciones de los vértices. Los vértices son puntos que proyectamos en los mapas con el objetivo de facilitar su interpretación; no intervienen en su configuración. Esta consideración nos sugiere una forma alternativa de cálculo de las posiciones de los vértices: en primer lugar, aumentaremos el número de filas en tantas filas como columnas tienen los datos; cada una de estas nuevas filas contendrá sólo un 1 y los restantes valores serán ceros. En cada una de las filas, el 1 se hallará en una columna distinta (imagen 12.4); en segundo lugar, declaramos las nuevas filas puntos adicionales. Las posiciones de estas filas adicionales son idénticas a las de los vértices de las columnas, es decir, sus coordenadas serán las coordenadas estándares de las columnas.

Variables categóricas adicionales y variables binarias

No debemos confundir el ejemplo de la columna adicional *Y* o el cálculo de las posiciones de vértices como filas adicionales de la imagen 12.4, con la codificación en «variables binarias», un tema que trataremos en detalle cuando, en los últimos capítulos, lleguemos al análisis de correspondencias múltiple. Supongamos, por

CATEGORÍA DE FINANCIACIÓN	*A*	*B*	*C*	*D*	*E*
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1

Imagen 12.4:
Agregación de filas adicionales a la tabla de la imagen 12.1: sus posiciones son idénticas a la de los vértices de las columnas

ejemplo, que clasificamos las disciplinas científicas en «Ciencias naturales» *(CN)* y en «Ciencias biológicas» *(CB)*, el último grupo incluiría Bioquímica, Zoología, Microbiología y Botánica, mientras que el primero contendría las restantes disciplinas científicas. En AC, una manera estándar de codificar estos datos sería con un par de variables binarias, *CN* y *CB*, es decir, como variables que toman los valores cero o uno. Así, para Geología (una ciencia natural), los valores de las variables binarias serían $CN = 1$ y $CB = 0$, mientras que para Bioquímica (una ciencia biológica), los valores serían $CN = 0$ y $CB = 1$, y así sucesivamente. Podríamos estar tentados en introducir estas variables binarias como columnas en la tabla y representarlas como puntos adicionales; sin embargo, esto no sería correcto. No estamos ante variables que expresen recuentos, a diferencia de la variable *Y*, que también tomaba los valores 0 y 1. En este caso, los valores de *Y* son verdaderos recuentos, que podían haber tomado otros valores enteros. La manera correcta de representar la información *CN*/*CB* sería mediante un par de filas, de manera similar a como representamos anteriormente *Ciencias matemáticas*. Sumaríamos las frecuencias de las filas *CN* y añadiríamos en la tabla una nueva fila que llamaríamos *CN*; haríamos lo mismo con las filas *CB*. De esta manera, los puntos *CN* y *CB* serían medias ponderadas de los puntos que representan los dos conjuntos de disciplinas científicas (en el capítulo 18 se retoma este tema).

Variables continuas adicionales

La información adicional en forma de variables continuas también requiere una especial consideración. Supongamos que tuviéramos información complementaria sobre las disciplinas científicas. Por ejemplo, el factor de impacto medio de los artículos publicados por investigadores de estas disciplinas en revistas internacionales. Podríamos situar esta información en una columna de datos y, dado que todos los valores son números positivos, podríamos estar tentados de representar el perfil de esta columna en forma de punto adicional. No obstante, debemos recordar que los perfiles columna representan números positivos que expresan proporciones de un total —no valores originales—, que además deben tener sentido en el contexto del estudio. ¿Qué haríamos si, por ejemplo, los datos expresaran cambios de la media del factor de impacto durante un determinado período, de manera que algunos valores fueran positivos y otros negativos? Es obvio que expresar estos valores con relación a su suma no tendría sentido alguno. En esta situación, podríamos representar esta variable continua de forma completamente distinta: por regresión. Veremos este tema con mucho más detalle en los capítulos 13 y 14, así como en el capítulo 24 cuando tratemos el análisis de correspondencias canónico, que consiste en una combinación del AC y de la regresión. Por el momento, únicamente pretendíamos alertar al lector sobre este problema.

RESUMEN: Puntos adicionales

1. Llamamos *puntos activos* a las filas y a las columnas de una tabla analizada por AC. Son los puntos que determinan la orientación de los ejes principales y, por tanto, contribuyen a la construcción de los mapas de baja dimensionalidad. Las filas y las columnas activas las proyectamos sobre el mapa.

2. Los puntos *adicionales* (o *pasivos)* son filas o columnas de la tabla que no han participado en la configuración del mapa, pero que tienen verdaderos perfiles. Son puntos que existen en los espacios completos de perfiles fila o de perfiles columna. Los podemos proyectar sobre un mapa de baja dimensionalidad con el objetivo interpretar sus posiciones con relación a los puntos activos.
3. Dado que los puntos adicionales tienen masa cero, el resultado de todos los cálculos en los que intervenga su masa será cero, como por ejemplo la inercia de los puntos o la contribución de los puntos a la inercia los ejes.
4. A pesar de que los puntos adicionales no contribuyen a la solución del AC, podemos calcular las contribuciones de los ejes principales (en términos de coseno o de correlación al cuadrado). Dichas contribuciones nos permiten valorar si los puntos adicionales se hallan bien representados en el mapa.
5. Hay que estar en guardia ante las observaciones atípicas de poca masa, cuya presencia en el análisis puede tener una gran influencia en la solución. Si es así, o bien los consideramos puntos adicionales o bien los combinamos —siempre que tenga sentido— con otras filas (u otras columnas).
6. Podemos crear una variable categórica adicional, por ejemplo una columna, para agrupar filas de acuerdo con las categorías de dicha variable. A continuación hallamos las frecuencias de dichas categorías y luego las añadimos como filas adicionales en la tabla.
7. Hay que ir con cuidado cuando añadamos una variable continua como punto adicional: sus valores no pueden ser negativos, además sus perfiles deben tener sentido en el contexto de los datos.