La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 3

Masa y centroides

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

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

www.fbbva.es

CAPÍTULO 3

Masas y centroides

Existe una forma equivalente de ver las posiciones de los perfiles en el espacio de los perfiles, que nos será útil para nuestra eventual comprensión e interpretación del AC. Se basa en el concepto de media ponderada o centroide de un conjunto de puntos. En el cálculo usual de la media (no ponderada), todos los puntos tienen la misma masa. Sin embargo, una media ponderada permite asociar diferentes masas a los distintos puntos. Cuando ponderamos los puntos de distinta manera, el centroide no se sitúa exactamente en el centro «geográfico» de la nube de puntos, sino que tiende a situarse cerca de los puntos con mayor masa.

Contenido

Conjunto de datos 2: tipos de lectura y nivel de educación

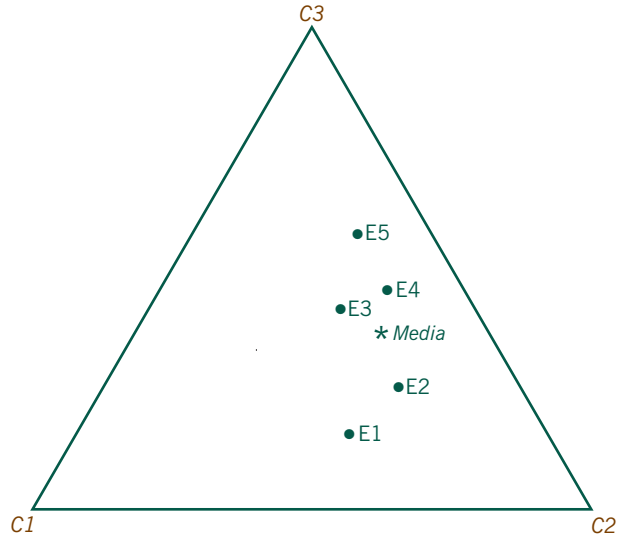
Utilicemos ahora un conjunto de datos habitual en investigación en ciencias sociales, una tabla de contingencia (o «clasificación cruzada») derivada de dos variables obtenidas en una encuesta. La tabla de la imagen 3.1 hace referencia a 312 lectores de un determinado periódico; en particular contiene datos sobre la minuciosidad de los encuestados en la lectura del periódico. Hemos clasificado a los encuestados en tres grupos de lectores: *rápidos*, *minuciosos* y *muy minuciosos*. Hemos cruzado estas categorías de lectura con el nivel de educación de los encuestados: una variable ordinal con cinco categorías que van desde algo de educación

Imagen 3.1:
Tabla del cruce del nivel de educación por tipo de lector, que muestra los perfiles fila y el perfil fila medio entre paréntesis, así como las masas de las filas (derivadas de los totales de las filas)

	Tipo de Lector				
Nivel de educación	*Rápidos C1*	*Minuciosos C2*	*Muy minuciosos C3*	*Total*	*Masas de las filas*
Educación primaria incompleta	5	7	2	*14*	*0,045*
E1	(0,357)	(0,500)	(0,143)		
Educación primaria	18	46	20	*84*	*0,269*
E2	(0,214)	(0,548)	(0,238)		
Educación secundaria incompleta	19	29	39	*87*	*0,279*
E3	(0,218)	(0,333)	(0,448)		
Educación secundaria	12	40	49	*101*	*0,324*
E4	(0,119)	(0,396)	(0,485)		
Educación universitaria incompleta	3	7	16	*26*	*0,083*
E5	(0,115)	(0,269)	(0,615)		
Total	*57*	*129*	*126*	*312*	
Perfil fila medio	*(0,183)*	*(0,413)*	*(0,404)*		

primaria, hasta algo de educación universitaria. En la imagen 3.1 mostramos los recuentos originales, así como los perfiles de los niveles de educación entre paréntesis, es decir, los perfiles fila. En la imagen 3.2 hemos representado gráficamente el diagrama de coordenadas triangular de los perfiles fila, similar al que vimos en el capítulo 2. En esta imagen, los puntos de las esquinas, o los vértices, del triángulo equilátero, representan los tres tipos de lectores (recordemos que cada vértice ocupa la posición de un perfil fila «puro», es decir un perfil completamente

Imagen 3.2:
Representación gráfica de los perfiles fila (nivel de educación) de la imagen 3.1 en coordenadas triangulares, que también indica la posición del perfil fila medio (última fila de la imagen 3.1)



concentrado en una categoría). Por ejemplo, el vértice *muy minuciosos, C3*, representa un perfil fila ficticio [0 0 1], que supuestamente contiene un 100% de lectores *muy minuciosos*.

Los puntos como medias ponderadas

Las posiciones de los niveles de educación en el triángulo las podemos ver también como medias ponderadas. Asignar pesos a los valores de una variable es un concepto bien conocido en estadística. Por ejemplo, supongamos que, en una clase de 26 estudiantes, la media de sus calificaciones calculada sumando las calificaciones de los 26 estudiantes y dividiendo por 26 es 7,5. En realidad, tres estudiantes obtuvieron un 9, siete un 8, y 16 un 7, de manera que podemos calcular de forma equivalente la calificación media asignando un peso de 3/26 a la calificación de 9, un peso de 7/26 a la de 8 y un peso de 16/26 a la de 7, siendo los pesos las frecuencias relativas de cada calificación. Dado que la calificación de 7 tiene más peso que las restantes, el valor de la media ponderada, 7,5, se halla «más cerca» de esta calificación. La media aritmética usual de los valores 7, 8 y 9 es de 8.

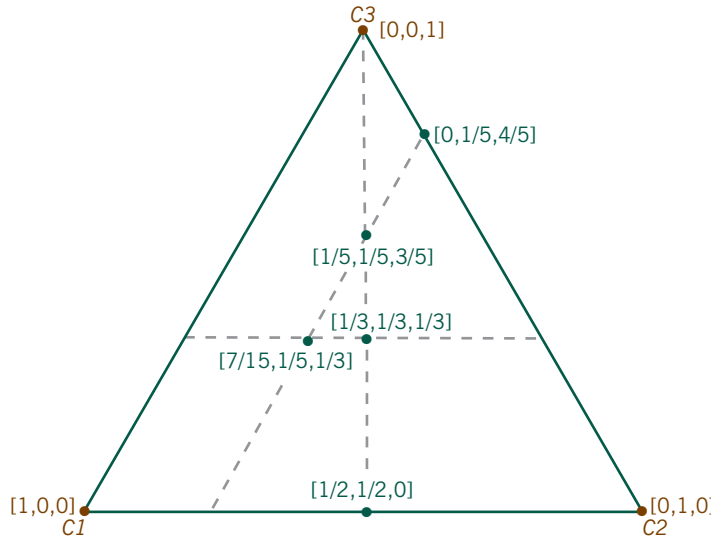
Los valores de los perfiles son los pesos asignados a los vértices

En la última fila de los datos de la imagen 3.1, vemos que a los encuestados de nivel de educación E5 (algo de educación universitaria) les corresponden las frecuencias 3, 7 y 16, es decir, las frecuencias relativas 0,115, 0,269 y 0,615, respectivamente. Imaginemos ahora cuál sería la *posición* media de estos 26 encuestados, si tres casos estuvieran situados en el vértice *rápidos, C1*, del triángulo; siete casos en el vértice *minuciosos, C2*, y 16 casos en el vértice *muy minuciosos, C3*. Es decir, en el espacio de perfiles, más que asociar los pesos con los valores de una variable, asociamos los pesos con las posiciones de los vértices. Hay más casos en la esquina de los *muy minuciosos*, por tanto, cabe esperar que la posición media de E5 se halle más cerca de este vértice, como ocurre en realidad. Por la misma razón, el perfil fila E1 se halla lejos de la esquina *muy minuciosos*, *C3*, ya que tiene muy poco peso (2 de 14, el 0,143) en esta categoría. Es decir, dentro del triángulo, situamos el punto que representa cada perfil fila como un punto medio de los vértices, en el que los valores del perfil —es decir, las frecuencias relativas— son los pesos asignados a los vértices. En consecuencia, podemos considerar los valores de los perfiles no sólo como coordenadas en un espacio multidimensional, sino también como los pesos asignados a los vértices de un símplex. Podemos extender este concepto a perfiles de más dimensiones. Por ejemplo, un perfil con cuatro elementos es también una posición media con relación a los cuatro vértices de un tetraedro tridimensional, que hemos ponderado con los elementos de este perfil.

Cada perfil es una media ponderada, o centroide, de los vértices

Términos alternativos a media ponderada son *centroide* o *baricentro*. En la imagen 3.3, hemos representado gráficamente algunos ejemplos de medias ponderadas en un espacio de perfiles. Por ejemplo, el perfil [1/3 1/3 1/3], que da el mismo peso a los tres vértices, se halla exactamente en el centro del triángulo, equidis-

Imagen 3.3:
Ejemplos de algunos centroides (medias ponderadas) de los vértices de un espacio de coordenadas triangular: los tres valores son los pesos asignados a los vértices (C1, C2, C3)



tante de los vértices, es decir en la posición de la media usual de los tres vértices. El perfil [1/2 1/2 0] se halla a medio camino entre el primer y el segundo vértices, ya que da igual peso a estos dos vértices y un peso igual a cero al tercero. En general podemos expresar la posición de un perfil [*a b c*], para el que se cumple que $a + b + c = 1$, como la de un centroide de los tres vértices, de la manera siguiente:

$$\textit{posición del centroide} = (a \times \text{vértice 1}) + (b \times \text{vértice 2}) + (c \times \text{vértice 3})$$

Por ejemplo, en la imagen 3.2 obtenemos la posición del nivel de educación E5 de la siguiente manera:

$$\text{E5} = (0{,}115 \times \textit{rápidos}) + (0{,}269 \times \textit{minuciosos}) + (0{,}615 \times \textit{muy minuciosos})$$

De forma similar, la posición del perfil medio es también una media ponderada de los vértices:

$$\textit{media} = (0{,}183 \times \textit{rápidos}) + (0{,}413 \times \textit{minuciosos}) + (0{,}404 \times \textit{muy minuciosos})$$

La media se halla más lejos del vértice *rápidos*, ya que su peso en este vértice es menor que el de los otros dos, que tienen aproximadamente los mismos pesos (imagen 3.2).

El perfil medio es también una media ponderada de los perfiles

El perfil medio es un punto especial; como acabamos de ver, no es solamente un centroide de los tres vértices, como cualquier otro perfil, sino que también es un centroide de los cinco perfiles fila, a los que hemos asignado diferentes pesos. Si volvemos de nuevo a la imagen 3.1, vemos que los totales de las filas son distintos: el primer nivel de educación E1 (algo de educación primaria) tiene sólo 14 indivi-

duos, mientras que el nivel de educación E4 (educación secundaria) tiene 101 individuos. En la última columna, titulada «masas de las filas», aparecen estas frecuencias marginales de las filas expresadas con relación al total de la muestra, 312. De la misma manera que contemplamos los perfiles como medias ponderadas de los vértices, podemos ver el perfil fila medio de la imagen 3.2 como una media ponderada de los perfiles, a los que hemos asignado pesos de acuerdo con sus frecuencias marginales; como si hubiera 14 individuos (una proporción de 0,045 de la muestra) en la posición E1, 84 individuos (una proporción de 0,269 de la muestra) en la posición E2, y así sucesivamente. Asignando estos pesos a los cinco perfiles, obtenemos exactamente la posición del perfil fila medio:

$$\textit{Perfil fila medio} = (0{,}045 \times \text{E1}) + (0{,}269 \times \text{E2}) + (0{,}279 \times \text{E3}) + (0{,}324 \times \text{E4}) + (0{,}083 \times \text{E5})$$

Este perfil fila medio se halla en una posición central entre los perfiles fila, pero más cerca de los perfiles con mayor frecuencia.

Las masas de las filas y las masas de las columnas

En el AC, los pesos asignados a los perfiles son tan importantes que les damos un nombre específico: *masas*. En la última columna de la imagen 3.1, se muestran las masas de las filas: 0,045, 0,269, 0,279, 0,324 y 0,083. En el AC preferimos el término «masa», sin embargo, es completamente equivalente a «peso». Preferimos un término alternativo a peso, para diferenciar la ponderación geométrica, de otros tipos de ponderación que podemos encontrar en el AC, como, por ejemplo, los pesos que asignamos a los subgrupos poblacionales en una encuesta. Todo lo referido a perfiles fila y a masas de las filas se puede aplicar de la misma manera a las columnas. En la imagen 3.4 mostramos la misma tabla de contingencia de la

	Tipo de Lector				
Nivel de educación	*Rápidos C1*	*Minuciosos C2*	*Muy minuciosos C3*	*Total*	*Perfil columna medio*
Educación primaria incompleta	5	7	2	*14*	*0,045*
E1	(0,088)	(0,054)	(0,016)		
Educación primaria	18	46	20	*84*	*0,269*
E2	(0,316)	(0,357)	(0,159)		
Educación secundaria incompleta	19	29	39	*87*	*0,279*
E3	(0,333)	(0,225)	(0,310)		
Educación secundaria	12	40	49	*101*	*0,324*
E4	(0,211)	(0,310)	(0,389)		
Educación universitaria incompleta	3	7	16	*26*	*0,083*
E5	(0,053)	(0,054)	(0,127)		
Total	*57*	*129*	*126*	*312*	
Masas de las columnas	*(0,183)*	*(0,413)*	*(0,404)*		

Imagen 3.4:
Tabla del cruce de nivel de educación por tipo de lector, que muestra los perfiles columna y el perfil columna medio entre paréntesis, así como las masas de las filas (obtenidas de los totales de las filas)

imagen 3.1, pero desde la óptica de las columnas. Es decir, se expresan las tres columnas como frecuencias relativas, con respecto al total de las columnas. Así, hemos obtenido tres perfiles con cinco elementos cada uno de ellos. Ahora, los totales de las columnas, con relación al total de la tabla, son las masas de las columnas que asignaremos a los perfiles de las columnas. El perfil columna medio está constituido por los totales de las filas dividido por el total de la tabla. Igual que antes para las filas, podemos expresar el perfil columna medio como una media ponderada de los tres perfiles columna *C1*, *C2* y *C3*:

$$\textit{Perfil columna medio} = (0{,}183 \times C1) + (0{,}413 \times C2) + (0{,}404 \times C3)$$

Fijémonos en que las masas de las filas y las masas de las columnas ejercen dos papeles distintos: como pesos y como elementos de los perfiles medios. En la imagen 3.4, el perfil columna medio está formado por las masas de las filas de la imagen 3.1. Sin embargo, aquí, las masas de las columnas son los elementos de lo que anteriormente era el perfil fila medio.

Interpretación del espacio de perfiles

En este momento, a pesar de que todavía no hemos visto algunos de los conceptos clave del AC, podemos empezar a interpretar la imagen 3.2. Los vértices del triángulo representan «perfiles puros» de tipos de lectores *C1*, *C2* y *C3*, mientras que los niveles de educación están constituidos por «mezclas» de tipos de lectores. Sus posiciones dentro del triángulo dependen de las proporciones de cada una de las categorías anteriores. Fijémonos en los siguientes aspectos de la representación gráfica:

- Dentro del triángulo, el grado de dispersión de los perfiles nos da una idea sobre la variabilidad existente en la tabla de contingencia. Cuanto más cerca se hallen los perfiles del centroide, menor será la variabilidad. En cambio, cuanto más se alejen del centroide, mayor será la variabilidad. El espacio de los perfiles se halla delimitado, de manera que los perfiles más extremos se situarán cerca de los lados del triángulo, o en el caso más extremo en uno de los vértices (por ejemplo, individuos con poca formación se situarían cerca del vértice *C1*). En las tablas de datos sobre ciencias sociales, como las que estamos considerando, dado que, en general, la variabilidad de los valores de los perfiles es relativamente pequeña, los perfiles ocupan sólo una pequeña región del espacio de perfiles, cerca de la media. Por ejemplo, el recorrido de los perfiles del primer elemento (es decir, la categoría de lector *C1*) va solamente de 0,115 a 0,357 (imagen 3.1), mientras que su recorrido potencial va de 0 a 1. En cambio, en datos sobre la investigación en ecología, como veremos más adelante, el recorrido de los valores de los perfiles es mucho mayor, a menudo debido a la presencia de muchos ceros en la tabla; en consecuencia, los perfiles se dispersan mucho más dentro del espacio de perfiles (segundo ejemplo del capítulo 10).

- En la representación, los perfiles se esparcen, en lo que se llamamos «dirección de dispersión», aproximadamente de abajo hacia arriba. Efectivamente, vemos que los cinco perfiles de los niveles de educación se sitúan de abajo hacia arriba en su orden natural (de E1 a E5), de menos a más formación. Arriba, el grupo E5 se halla cerca del vértice *C3*, que representa la categoría de los lectores *muy minuciosos* —ya hemos visto que este grupo contiene la mayor proporción (0,615) de este tipo de lectores—. Abajo, el nivel de educación más bajo, no lejos del borde del triángulo que sabemos que muestra perfiles con cero lectores *C3* (por ejemplo el punto [1/2 1/2 0] de la imagen 3.3 es un ejemplo de uno de estos puntos). La interpretación de esta distribución hay que buscarla en el hecho de que cuando nos movemos, de abajo hacia arriba, en general, los perfiles cambian respecto a la frecuencia relativa de la categoría *C3*, en contraste con la combinación de las categorías *C1* y *C2*. No observamos tendencia particular alguna hacia *C1* o hacia *C2*.

Unión de filas o de columnas

Supongamos que queremos reunir los dos niveles de educación primaria, E1 y E2, de la imagen 3.1, en una nueva fila que llamaremos E1&E2. Existen dos posibilidades de unión: la primera es sumar ambas filas, para obtener una nueva fila de frecuencias [23 53 22] con un total de 98 individuos y de perfil [0,235 0,541 0,224]; la segunda posibilidad es que consideremos al perfil E1&E2 como la media ponderada de los perfiles E1 y E2:

$$[0{,}235\ \ 0{,}541\ \ 0{,}224] = \frac{0{,}045}{0{,}314} \times [0{,}357\ \ 0{,}500\ \ 0{,}143] + \frac{0{,}269}{0{,}314} \times [0{,}214\ \ 0{,}548\ \ 0{,}238]$$

donde las masas de E1 y E2 son 0,045 y 0,269, respectivamente, que sumadas dan 0,314 (fijémonos en que los pesos de esta media ponderada son iguales a 14/98 y 84/98, siendo 14 y 84 los totales de las filas E1 y E2, respectivamente). Geométricamente, el perfil E1&E2 se halla en una línea que une E1 y E2, pero más cerca de E2, como podemos ver en la imagen 3.5. Las distancias de E1 a E1&E2 y de

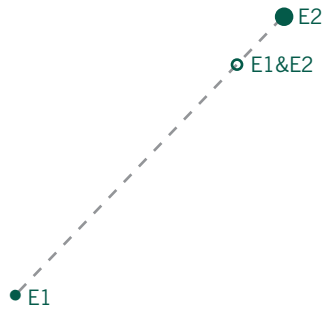


Imagen 3.5:
Ampliación de las posiciones de E1 *y* E2 *en la imagen 3.2, que muestra la posición del punto* E1&E2 *al unir ambas categorías.* E2 *tiene seis veces más masa que* E1*, en consecuencia* E1&E2 *se halla más cerca de* E2*, en un punto que divide el segmento que une* E1 *con* E2 *de acuerdo con la proporción 84:14 = 6:1*

E2 a E1&E2 se hallan en la misma proporción que los totales 84 y 14, es decir de 6 a 1. Por tanto, podemos considerar E1&E2 como el punto de equilibrio de las dos masas situadas en E1 y en E2, con la mayor masa en E2.

Distribuciones equivalentes de filas o de columnas

Supongamos que añadimos una fila a los datos de la imagen 3.1, una categoría de «sin educación reglada» que simbolizaremos por E0, de frecuencias [10 14 4] con relación a los tipos de lectores. El perfil de E0 es idéntico al perfil de E1, ya que las frecuencias de E0 son simplemente el doble de las de E1. Los dos conjuntos de frecuencias son *distribucionalmente equivalentes*. Por tanto, los perfiles de E0 y de E1 se hallan exactamente en el mismo punto del espacio de perfiles. Los podemos unir para dar a este punto una masa igual a la combinación de las masas de los dos perfiles, es decir un punto de frecuencias [15 21 6].

Cambio de masas

Las masas de las filas y las de las columnas son proporcionales a las sumas marginales de la tabla. Si por alguna razón importante tenemos que modificar las masas es fácil transformar la tabla. Por ejemplo, supongamos que queremos que los cinco niveles de educación de la imagen 3.1 tengan masas proporcionales a los tamaños de las poblaciones de procedencia y no proporcionales al tamaño de sus muestras. En tal caso, podemos cambiar los valores de la tabla multiplicando los perfiles de los niveles de educación por el tamaño de sus respectivas poblaciones de origen. Los perfiles de esta nueva tabla serán idénticos a los perfiles originales; sin embargo, las masas de las filas serán proporcionales a los tamaños de las poblaciones. Alternativamente, supongamos que, en vez de ponderar de forma distinta los niveles de educación, como hasta ahora, queremos ponderarlos de igual forma. Para ello podemos tomar la tabla de perfiles fila (o de forma equivalente, de porcentajes en cada fila), como si fuera la tabla original. En esta tabla, la suma total de los elementos de las filas es 1 (o el 100%), es decir, todos los niveles educativos tendrán la misma ponderación.

RESUMEN: Masas y centroides

1. Supongamos que queremos representar gráficamente los perfiles fila, es decir representamos los perfiles fila en el espacio símplex definido por los vértices de las columnas. En tal caso, cada vértice representa una categoría de las columnas: un perfil fila completamente concentrado en esa categoría.
2. Podemos interpretar cada perfil como un centroide (o media ponderada) de los vértices, en el que los pesos son los elementos de su perfil. Por tanto, los perfiles tenderán a hallarse cerca de los vértices para los que tengan valores mayores.
3. Cada perfil fila tiene asociado un peso, llamado *masa*, proporcional a la suma de los elementos de la fila de la tabla original. Podemos obtener el perfil fila medio como el centroide de los perfiles fila, ponderando cada perfil con su correspondiente masa.

4. Todo lo que hemos visto hasta ahora para los perfiles fila, lo podemos aplicar de la misma manera a las columnas de la tabla. En realidad, la mejor manera de pasar del análisis de filas al de columnas es transponer la tabla, es decir, que las columnas sean las filas y viceversa, y hacer lo que hemos visto para las filas.
5. Las filas (o las columnas) formadas sumando las frecuencias de filas (o de columnas) de la tabla tienen un perfil igual a las medias ponderadas de los perfiles de las filas (o de las columnas) que lo componen.
6. Las filas (o las columnas) con perfiles iguales son *distribucionalmente equivalentes*. Las podemos agregar en un solo punto.
7. Podemos modificar las masas de las filas (o de las columnas) para que sean proporcionales a determinados valores, simplemente multiplicando por un factor de escala.