La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 15

Agrupación de filas o 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 15

Agrupación de filas o de columnas

Hasta ahora hemos transformado matrices de datos en mapas en los que representamos filas y columnas como puntos en un espacio continuo, en general un espacio bidimensional. Una forma alternativa de representar una estructura consiste en realizar un análisis de grupos de los perfiles de filas o de columnas. Esta aproximación tiene muchas similitudes con el AC. En ambos análisis descomponemos la inercia de los perfiles, en grupos en el análisis de grupos y en ejes continuos en el AC. El análisis de grupos aplicado a tablas de contingencia conlleva poder disponer de una prueba estadística que nos permite contrastar si existen diferencias entre grupos de filas o de columnas.

Contenido

Agrupación de filas o de columnas

La idea de agrupar objetos es omnipresente en el análisis de datos. La agrupación puede venir dada por una determinada clasificación, o por algún criterio que agrupe objetos similares. En primer lugar, tratemos una agrupación establecida de acuerdo con una determinada variable categórica que clasifique las filas o las

Imagen 15.1: *Frecuencias de las categorías de financiación para 796 investigadores agrupados en cuatro categorías según disciplinas científicas*

	Categoría de financiación					
Disciplina científica	*A*	*B*	*C*	*D*	*E*	*Suma*
Geol/Fis/Est/Mat	17	57	134	35	63	*306*
Bioq/Quim	7	27	62	22	41	*159*
Zool/Micr/Biol	4	33	89	57	60	*243*
Ing	3	11	25	15	34	*88*
Suma	*31*	*128*	*310*	*129*	*198*	*796*

columnas de una tabla. Consideremos otra vez el ejemplo sobre la financiación de la investigación científica, y supongamos que existe una agrupación predeterminada de las disciplinas científicas en cuatro grupos, según las facultades de una determinada universidad: {Geología, Física, Estadística, Matemáticas}, {Bioquímica, Química}, {Zoología, Microbiología, Botánica} e {Ingeniería}. Como apuntamos en el capítulo 12, cuando definimos una variable categórica sobre las filas, como en este ejemplo, cada categoría de esta variable categórica define una fila adicional en la tabla que reúne las frecuencias afectadas por dicha categoría. Así, las diez filas de la imagen 10.1 se convierten en cuatro filas que corresponden a los cuatro grupos que mostramos en la imagen 15.1. El AC de los datos originales de la imagen 10.1 tenían una inercia total de 0,08288, mientras que si realizamos el AC de los datos de la imagen 15.1, la inercia total es de 0,04386. Cuando reunimos puntos, se produce una pérdida de inercia, de la misma manera que cuando dividimos filas o columnas, siguiendo algún criterio de subclasificación, se produce un incremento de inercia.

Inercia inter e intra grupos

La inercia de la tabla de grupos de la imagen 15.1 corresponde a la *inercia intergrupos*, ya que mide la variabilidad entre los cuatro grupos de filas de la tabla. Llamamos *inercia intragrupos*, a la diferencia entre la inercia total, 0,08288 y la inercia intergrupos, 0,04386. Esta diferencia mide la variabilidad que se pierde, dentro de los grupos, cuando unimos filas en grupos. Esta descomposición de la inercia es un resultado clásico del análisis de la varianza, en general aplicado a una sola variable, aunque también se puede aplicar a datos multivariantes. En AC, la inercia total de las filas viene dada por $\sum_i r_i d_i^2$ (fómula 4.7), siendo d_i la distancia χ^2 entre $\mathbf{a}_i$ y **c,** donde $\mathbf{a}_i$, es una fila que tiene asociada una masa r_i y **c** es el perfil fila medio (centroide) igual a las masas de las columnas. La inercia intergrupos se calcula de forma similar, mediante la ecuación $\sum_g \bar{r}_g \bar{d}_g^2$ que se aplica a las filas agrupadas, siendo $\bar{\mathbf{a}}_g$ los perfiles de las filas resultantes de la agrupación, donde $g = 1,..., G$ indica el grupo (aquí $G = 4$), $\bar{r}_g$ es la masa del g-ésimo grupo, resultante de la suma de las masas de los miembros del grupo. Los perfiles $\bar{\mathbf{a}}_g$ siguen teniendo el centroide en **c**, $\bar{d}_g$ son las distancias χ^2 al centroide. La inercia de cada grupo g a su propio centroide $\bar{\mathbf{a}}_g$ la calculamos mediante la expresión $\sum_{i \in g} r_i d_{ig}^2$ don-

Grupo	*Definición*	*Componente*	*Porcentaje sobre cada parte*	*Porcentaje sobre el total*
Inercia intergrupos				
Geol/Fis/Est/Mat	$\overline{r}_1\overline{d}_1^2$	0,01482	33,8%	17,9%
Bioq/Quim	$\overline{r}_2\overline{d}_2^2$	0,00099	2,3%	1,2%
Zool/Micr/Biol	$\overline{r}_3\overline{d}_3^2$	0,01548	35,3%	18,7%
Ing	$\overline{r}_4\overline{d}_4^2$	0,01256	28,6%	15,2%
Total	$\sum_g \overline{r}_g\overline{d}_g^2$	0,04386	100,0%	52,9%
Inercia intragrupos				
Geol/Fis/Est/Mat	$\sum_{i\in1} r_i d_{i1}^2$	0,01842	47,2%	22,2%
Bioq/Quim	$\sum_{i\in2} r_i d_{i2}^2$	0,01064	27,3%	12,8%
Zool/Micr/Biol	$\sum_{i\in3} r_i d_{i3}^2$	0,00996	25,5%	12,0%
Ing	$\sum_{i\in4} r_i d_{i4}^2$	0	0%	0%
Total	$\sum_g \sum_{i\in g} r_i d_{ig}^2$	0,03902	100,0%	47,1%

Imagen 15.2: *Descomposición de la inercia inter e intragrupos, que muestra los valores absolutos expresados como porcentajes con relación a la inercia de cada parte, y con relación a la inercia total. Las sumas de la inercia total intergrupos y la inercia total intragrupos es la inercia total, 0,08288 de la tabla de original (imagen 10.1)*

de d_{ig} es la distancia χ^2 de cada perfil i del grupo g al centroide $\overline{\mathbf{a}}_g$. Sumando estos valores para los cuatro grupos obtenemos la inercia intragrupos. Por tanto, la descomposición de la inercia es:

$$\text{inercia total} = \text{inercia intergrupos} + \text{inercia intragrupos}$$

$$\sum_i r_i d_i^2 = \sum_g \overline{r}_g \overline{d}_g^2 + \sum_g \sum_{i\in g} r_i d_{ig}^2 \qquad (15.1)$$

$$0{,}08288 = 0{,}04386 + 0{,}03902$$

Cálculo de la inercia dentro de un grupo

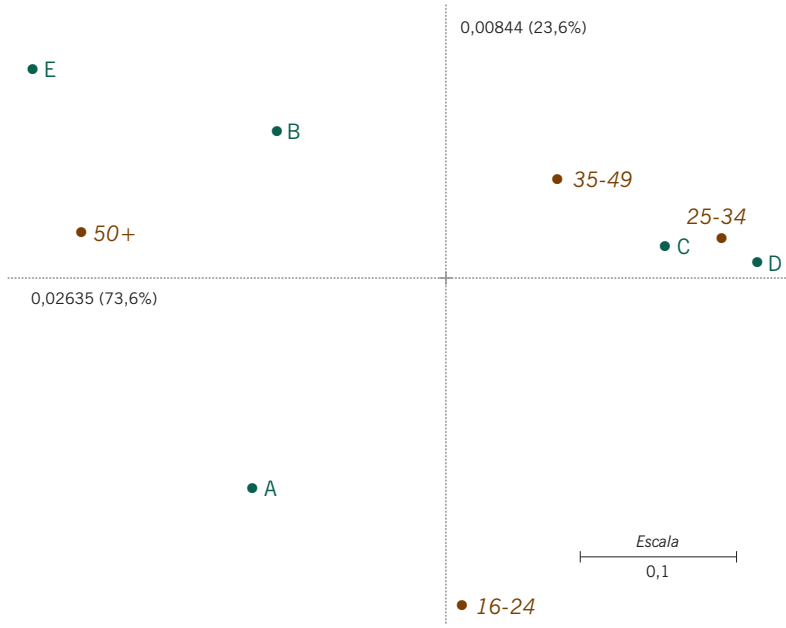
Según lo que acabamos de ver, la inercia intragrupos es igual a 0,03902, pero ¿cuál es la contribución de cada uno de los cuatro grupos? Lo podemos calcular directamente, recordando que, en todos los cálculos de distancias χ^2, debemos utilizar los mismos valores de **c**. Sin embargo, una manera más rápida de hallar esta contribución es aplicar el AC a las matrices resultantes de ir formando, uno a uno, los distintos grupos. Por ejemplo, si formamos el primer grupo reuniendo Geología, Física, Estadística y Matemáticas y analizamos este grupo juntamente con las restantes filas, sin reunir, es decir en total siete filas, la inercia total es 0,06446. Comparando este valor con el valor de la inercia total de los datos originales, 0,08288, la disminución de 0,01842, que corresponde a la inercia intragrupos perdida al formar este grupo. Si ahora reunimos Bioquímica y Química y llevamos a cabo de nuevo el AC de seis grupos, en esta ocasión el valor de la inercia total disminuye hasta 0,05382. Por tanto, la inercia intragrupos atribuible a este grupo es la diferencia, 0,06446 – 0,05382 = 0,01064, y así sucesivamente. En la imagen 15.2 mostramos la descomposición completa de la inercia en valores absolutos y en porcentajes. Fijémonos en que la inercia intragrupos del grupo formado por una sola fila, Ingeniería, es 0.

Conjunto de datos 8: distribución de edades en tiendas de comida

Hasta ahora, hemos realizado la agrupación de filas partiendo de información disponible. Vamos a considerar ahora la posibilidad de formar grupos utilizando un determinado criterio de análisis de grupos. Para ilustrar los cálculos vamos a utilizar una pequeña matriz de datos obtenida de una muestra real de compradores en cinco tiendas de comida distintas. En la imagen 15.3, mostramos la tabla de contingencia de 5 × 4 que hemos obtenido cruzando los datos según tiendas y grupos de edad. El estadístico χ^2 de esta tabla es 25,06, al que le corresponde un valor p de 0,015. Por tanto, existe una asociación significativa entre la edad y la elección de la tienda. Junto con la tabla mostramos el mapa simétrico del AC. Un investigador de mercados estaría interesado en conocer dónde se halla esta asociación significativa. Por ejemplo, estaría interesado en

Imagen 15.3: *Combinación de tiendas de comida con grupos de edad de una muestra de 700 consumidores, y mapa simétrico del AC, que explica el 97,2% de la inercia total de 0,03580*

	Grupo de edad (años)				
Tienda de comida	*16-24*	*25-34*	*35-49*	*50+*	*Suma*
A	37	39	45	64	*185*
B	13	23	33	38	*107*
C	33	69	67	56	*225*
D	16	31	34	22	*103*
E	8	16	21	35	*80*
Suma	*107*	*178*	*200*	*215*	*700*



saber qué tiendas o grupos de tiendas tienen un perfil de edad significativamente distinto de los otros. Observamos que el mayor contraste se halla entre el grupo de mayor edad a la izquierda y el segundo grupo más joven a la derecha (imagen 15.3). La tienda E es la que muestra una mayor asociación al mencionado grupo de mayor edad, mientras que las tiendas C y D tienden más hacia los grupos más jóvenes. El eje vertical contrasta el grupo de edad más joven con los otros. Vemos también que la tienda A se halla hacia el grupo de edad más joven, separada de las restantes tiendas.

Algoritmo de agrupación

Vamos a agrupar las filas y las columnas utilizando un algoritmo de agrupación que trata (al mismo tiempo) de maximizar la inercia intergrupos y de minimizar la inercia intragrupos. En la imagen 15.4 ilustramos, para las filas, dicho algorit-

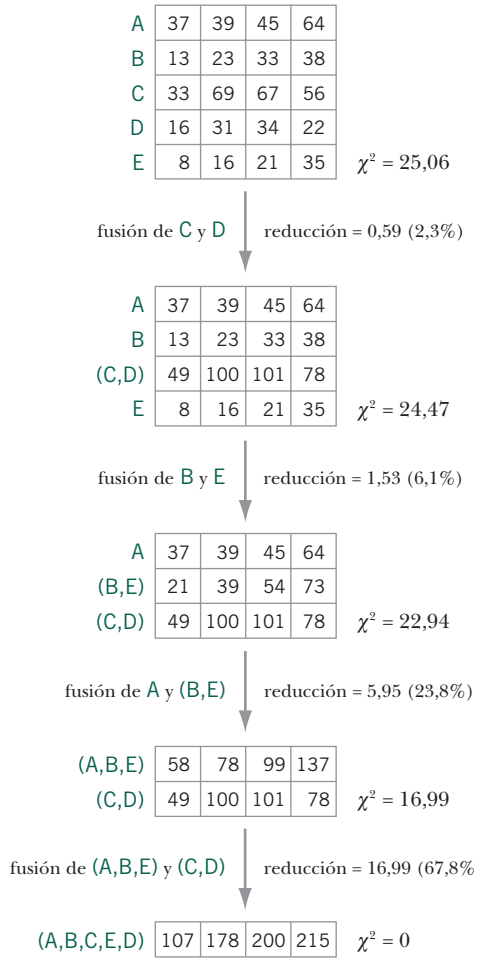


Imagen 15.4:
Pasos en la agrupación de las filas de la imagen 15.1: en cada paso se reúnen las dos filas que conducen a una menor reducción del valor del estadístico χ^2 o, de forma equivalente, a una menor reducción de la inercia intergrupos (para pasar de valores χ^2 a inercia, dividimos por el tamaño de la muestra, N = 700)

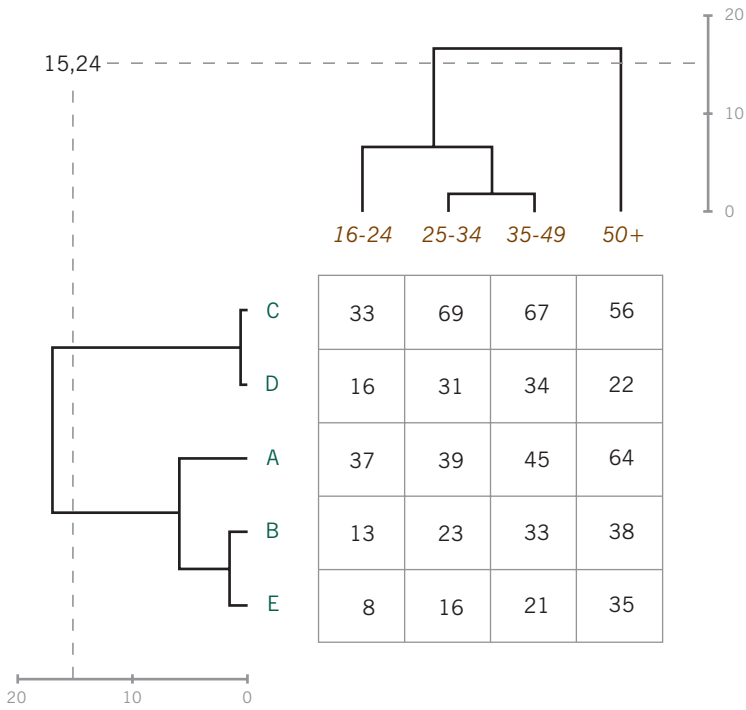
mo de agrupación. Al principio del proceso todas las filas están separadas entre sí, la inercia total intergrupos es igual a la inercia total de la tabla. Cualquier fusión de filas reducirá la inercia intergrupos. Por tanto, el primer paso consiste en identificar qué pares de filas (tiendas) se pueden reunir reduciendo al mínimo la pérdida de inercia. En este sentido, las filas más parecidas son las tiendas C y D. Cuando fusionamos estas dos filas para formar una nueva fila, etiquetada como (C,D), la inercia de la tabla de 4 × 4 resultante se reduce en 0,00084, alcanzándose el valor 0,03496, o en términos de χ^2 la reducción es de 0,59, llegándose al valor de 25,06 (en la imagen 15.4 mostramos los valores χ^2, son los valores de la inercia multiplicados por el tamaño de la muestra: $\chi^2 = 0{,}03496 \times 700 = 25{,}06$). En términos porcentuales, la reducción es del 2,3% tanto en la inercia como en el valor de χ^2. Luego, repetimos el procedimiento para hallar, en la nueva tabla, las filas, que son más parecidas en el sentido mencionado anteriormente. Son las tiendas B y E, lo que conduce a una disminución adicional de χ^2 de 1,53 (6,1%). Ahora la tabla tendrá tres filas etiquetadas como A, (B,E) y (C,D). Repetimos el procedimiento y vemos que la menor reducción se produce cuando la tienda A se une al par (B,E) para formar una nueva fila etiquetada como (A,B,E), χ^2 se reduce en 5,95 unidades adicionales (23,8%). Finalmente, se reúnen las dos filas (A,B,E) y (C,D) para formar una sola fila, que consta de las sumas marginales de las columnas de la tabla original, para la cual el valor de χ^2 es cero. Por tanto la reducción final es de 16,99 (67,8%), que es la inercia de la penúltima tabla de la imagen 15.4. Podemos repetir el procedimiento de la misma manera para las columnas de la tabla.

Representación en árbol de las agrupaciones

Podemos representar gráficamente la reunión sucesiva de filas, llamada *agrupación jerárquica*, como un *árbol binario* o *dendrograma* (se muestra en la imagen 15.5 junto con una agrupación jerárquica similar de las columnas). Fijémonos en que, generalmente, la ordenación original de filas y columnas exige modificaciones para adaptarlas a las representaciones en árbol. En este ejemplo sólo hemos reordenado las filas. Podemos ver en el árbol que las tiendas C y D han sido las primeras en fusionarse. Llamamos *nudo* al punto en el que ocurre esta unión, correspondiendo al mismo una determinada reducción del valor del estadístico χ^2.

Descomposición de la inercia (o del estadístico χ^2)

La descomposición del estadístico χ^2 hasta llegar a cero es la siguiente: 25,06 = 16,99 + 5,95 + 1,53 + 0,59. Dividiendo por 700, el tamaño de la muestra, obtenemos la correspondiente descomposición de la inercia: 0,03580 = 0,02427 + 0,00851 + 0,00218 + 0,00084. Si expresamos las dos descomposiciones anteriores como porcentajes, obtenemos los mismos valores: 67,8%, 23,8%, 6,1% y 2,3%. Hemos seguido un procedimiento de agrupación similar para las columnas, en esta ocasión, la descomposición de la inercia que nos señalan los nudos es la siguiente: 0,03580 = 0,02383 + 0,00938 + 0,00259, que en forma de porcentajes es: 66,6%, 26,2% y 7,2%.



C	33	69	67	56
D	16	31	34	22
A	37	39	45	64
B	13	23	33	38
E	8	16	21	35

Imagen 15.5:
Estructuras en árbol que representan la agrupación jerárquica de filas y de columnas. La agrupación se expresa en términos de χ^2, podemos convertirlo en inercias dividiendo por el tamaño de la muestra, 700. Indicamos el nivel crítico de χ^2, 15,24 (de filas y de columnas)

Decidiendo sobre la agrupación

En un análisis de grupos de este tipo es útil inspeccionar los árboles para decidir el número final de grupos con el que finalmente nos quedamos. Por ejemplo, si nos fijamos en la agrupación de las filas, vemos que existe una gran diferencia entre los dos grupos de tiendas (C,D) y (A,B,E), que viene indicada por el alto valor del nudo en el que se unen estos dos grupos. La descomposición de la inercia nos indica que la formación de estos dos grupos explica el 67,8% de la inercia de las filas. Si separamos la tienda A, como tercer grupo, entonces se explica un 23,8% de inercia adicional. Es decir, con estos tres grupos se explica el 91,6% de la inercia total. Así pues en este tipo de análisis de grupos, interpretamos los porcentajes de inercia asociados con los nudos igual que los porcentajes de inercia de los ejes principales en AC. La decisión sobre hasta qué porcentaje tenemos que llegar para detener la formación de grupos es, en general, informal, se basa en la secuencia de porcentajes y en la interpretación sustantiva de cada nudo o eje principal.

Contraste de hipótesis sobre los grupos de filas o columnas

El estadístico χ^2 de la tabla de contingencia original es significativo (p = 0,015); por tanto, en algún lugar de la tabla tienen que existir diferencias significativas entre perfiles. Precisar, desde un punto de vista estadístico, qué perfiles son significativamente distintos no es sencillo, ya que podemos contrastar la significación de muchos grupos de tiendas. Además, debemos tener en cuenta que, cuando queremos hacer muchas pruebas con unos mismos datos, es necesario ajustar

el nivel de significación. A todo ello hay que añadir que, en este caso, las agrupaciones de tiendas como por ejemplo la C con la D y de la B con la E, las sugieren los mismos datos, no se han establecido como hipótesis antes de la obtención de los datos.

Comparaciones múltiples

Estamos sobre la delicada línea existente entre el análisis exploratorio y el análisis confirmatorio de datos. Intentamos sacar conclusiones a partir de unos datos que hemos obtenido de manera exploratoria sin hipótesis establecidas *a priori*. Afortunadamente, se ha desarrollado un área de la estadística especialmente para este tipo de situaciones, son las llamadas *comparaciones múltiples*. Esta aproximación se utiliza más en el análisis de experimentos cuando se quieren comparar varios «tratamientos» entre sí, que en el análisis de experimentos clásicos sencillo en los que un tratamiento se compara con un control. El procedimiento de comparaciones múltiples permite que cualquier tratamiento (o grupos de tratamientos) se pueda contrastar con cualquier otro. Las decisiones estadísticas se pueden realizar a un nivel de significación preestablecido para proteger todas estas pruebas del llamado «error tipo I», es decir, de obtener un resultado que se deba completamente al azar.

Comparaciones múltiples para tablas de contingencia

Igual que en el caso de los diferentes tratamientos en una situación experimental, podríamos querer contrastar las diferencias entre cualquier par de filas de la tabla original o en las diferencias entre cualquier par de grupos de filas. Si sólo hiciéramos una prueba, calcularíamos la tabla reducida conteniendo dos filas (o grupos) y haríamos una prueba χ^2 de forma habitual. El procedimiento de comparación múltiple desarrollada para esta situación permite contrastar diferencias entre dos filas (o grupos de filas) cualquiera. Para ello, en primer lugar, calculamos el valor del estadístico χ^2 de la tabla reducida; a continuación, para conocer si la prueba es significativa o no, comparamos el valor del estadístico χ^2 calculado con el correspondiente valor crítico de la tabla que aparece en la imagen A.1 (pág. 277) que hemos incluido en el apéndice teórico, A. En esta tabla damos los valores críticos para tablas de contingencia de distinto tamaño a un nivel de significación del 5%. Así por ejemplo, a nuestra tabla de 5×4 le corresponde un valor crítico de 15,24. Por tanto, si el estadístico χ^2 es mayor que 15,24, decimos que las dos filas (o grupos de filas) son significativamente distintas.

Límites del valor de χ^2 para agrupaciones significativas

Podemos utilizar el valor crítico de la prueba de comparaciones múltiples para cualquier grupo de filas o de columnas de la tabla, en particular, para separar de forma estadísticamente significativa los grupos que mostramos en la agrupación jerárquica de la imagen 15.5. Así, con relación a los grupos de edad, vemos que el único contraste estadísticamente significativo se produce entre el grupo de mayor edad (50 o más años) y el resto de grupos; con relación a las tiendas de comida, la diferencia estadística se halla entre el grupo (A,B,E) y el grupo (C,D). Por

tanto, la separación observada en el segundo eje de la imagen 15.3 puede ser debida a la variabilidad aleatoria de los datos observados, ya que no existen diferencias significativas entre el grupo de edad más joven y los restantes grupos. Además, en el segundo eje, la distinción entre el grupo de edad *16-24* y el grupo *35-49*, también es difícil de justificar desde un punto de vista estadístico. Todo ello no significa que no podamos inspeccionar la información original en forma de mapa bidimensional como el la imagen 15.3 (prescindiendo de las consideraciones sobre la significación estadística, pues la información mostrada por los datos siempre es útil). En el capítulo 25, utilizaremos estos mismos valores críticos para llevar a cabo una prueba de significación sobre las inercias principales de una tabla de contingencia.

Agrupación de Ward

El algoritmo de agrupación que hemos descrito en este capítulo es un caso especial de la *agrupación de Ward*. En este tipo de agrupación, los grupos se reúnen según un criterio de distancia mínima que tienen en cuenta los pesos de los puntos que se agrupan. Por tanto, en vez de considerar en cada paso sólo la reducción de χ^2 (o de la inercia), utilizamos las distancias χ^2 entre perfiles y sus masas asociadas. Por ejemplo, la «distancia» entre dos grupos de filas g y h es:

$$\frac{\bar{r}_g \bar{r}_h}{\bar{r}_g + \bar{r}_h} \left\| \bar{\mathbf{a}}_g - \bar{\mathbf{a}}_h \right\|_c^2 \qquad (15.2)$$

donde $\bar{r}_g$ y $\bar{r}_h$ son las masas de los respectivos grupos, y $\left\| \bar{\mathbf{a}}_g - \bar{\mathbf{a}}_h \right\|_c$, la distancia χ^2 entre los perfiles de los grupos:

RESUMEN: Agrupación de filas o de columnas

1. El análisis de grupos de filas o de columnas, consistente en la fusión de filas (o columnas) similares en grupos discretos, proporciona una alternativa al examen de la estructura de los datos.
2. Los resultados de la agrupación se pueden representar gráficamente mediante una estructura en árbol *(dendrograma o árbol binario)*, en el que los nudos indican las uniones sucesivas de las filas (o columnas).
3. La inercia total (o de forma equivalente el estadístico χ^2) de la tabla se reduce lo menos posible en cada nivel sucesivo de agrupación de las filas (o columnas). Este procedimiento de *agrupación de Ward* proporciona una descomposición de la inercia con relación a los nudos del árbol, análogo a la descomposición de inercia con relación a los ejes principales en el análisis de correspondencias.
4. Gracias al procedimiento de *comparaciones múltiples,* podemos contrastar la significación de la inercia explicada por cada nudo, lo que nos permite hacer afirmaciones estadísticas sobre las diferencias intergrupos de filas (o de columnas). Esta prueba sólo se puede aplicar a verdaderas tablas de contingencia.