La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 11

Contribuciones a la inercia

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

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

www.fbbva.es

CAPÍTULO 11

Contribuciones a la inercia

La inercia total de una tabla de contingencia mide la variación existente en la tabla. Hasta ahora, hemos visto cómo descomponer la inercia en ejes principales, también hemos visto cómo descomponerla en filas o en columnas. Un paso más consiste en descomponer la inercia de filas o columnas en ejes principales. La investigación de estos componentes de la inercia (similar a un análisis de la varianza) desempeña un importante papel en la interpretación del AC. Además, proporciona herramientas de diagnóstico para identificar los puntos que más contribuyen a la definición de los ejes principales. Igualmente, nos permite valorar la calidad de la representación de los puntos.

Contenido

La inercia total mide la variación total de los perfiles

En el capítulo 4, la ecuación (4.7) nos mostró que, geométricamente, la inercia total es una media ponderada de las distancia χ^2 entre los perfiles y el perfil media. Con los perfiles columna obtenemos los mismos resultados. Si sólo existen pequeñas diferencias entre los perfiles y su media, la inercia toma un valor próxi-

Imagen 11.1: *Contribuciones de las filas y las columnas a la inercia, en valores absolutos que sumados dan la inercia total, y en valores relativos en tantos por mil (‰) que sumados dan 1000*

Filas	*Inercia*	*‰ inercia*	Columnas	*Inercia*	*‰ inercia*
Geología	0,01135	137	*A*	0,01551	187
Bioquímica	0,00990	119	*B*	0,00911	110
Química	0,00172	21	*C*	0,00778	94
Zoología	0,01909	230	*D*	0,02877	347
Física	0,01621	196	*E*	0,02171	262
Ingeniería	0,01256	152			
Microbiología	0,00083	10			
Botánica	0,00552	67			
Estadística	0,00102	12			
Matemáticas	0,00466	56			
Total	*0,08288*	*1000*	*Total*	*0,08288*	*1000*

mo a cero; es decir, existe poca variación (lo podemos ver en la imagen 4.2, en el diagrama triangular de arriba a la izquierda). El caso extremo ocurre cuando los perfiles se concentran en unas pocas categorías, y además en categorías distintas en los diferentes perfiles, en tal caso la inercia es grande (imagen 4.2, diagrama triangular de abajo a la derecha). La inercia es una medida de la dispersión de los perfiles en el espacio de perfiles.

Inercia de filas e inercia de columnas

La descomposición de la inercia en sumas de componentes positivos nos permite llevar a cabo un «análisis de inercia» útil para la interpretación de los resultados del AC. De acuerdo con la ecuación (4.7), la contribución de cada fila a la inercia es igual a su masa multiplicada por el cuadrado de su distancia al centroide de las filas, que llamaremos *inercia de filas*. Lo mismo se cumple para las columnas, y así obtendremos la *inercia de columnas*. Para facilitar la interpretación expresamos los componentes de la inercia con relación a la inercia total; se pueden expresar como porcentajes o, mejor, en *tantos por mil* (que indicaremos por ‰). En la tabla de la imagen 11.1 mostramos las inercias de las filas y de las columnas de los datos de la imagen 10.1 sobre la financiación de la investigación científica. Primero como «valores absolutos» y luego en forma relativa en tantos por mil. En nuestra implementación del AC en R utilizamos ampiamente los tantos por mil (lo podemos ver en el apéndice de cálculo, B), ya que nos permite incluir tres dígitos significativos sin utilizar decimales, lo que facilita la lectura de resultados.

Contribuciones grandes y contribuciones pequeñas a la inercia

Podemos ver fácilmente a partir de *«‰ de inercia»* de la imagen 11.1 que Zoología, Física, Ingeniería, Geología y Bioquímica son las filas, por este orden, que más contribuyen a la inercia; mientras que las categorías *D* y *E* son las columnas que más lo hacen. Como pauta general, para decidir qué contribuciones son grandes y cuáles son pequeñas, utilizaremos como valor umbral la media de las contribu-

Disciplina científica	Categoría de financiación					Suma
	A	*B*	*C*	*D*	*E*	
Geología	0	32	16	0	89	*137*
Bioquímica	0	23	4	44	48	*119*
Química	3	12	1	0	5	*21*
Zoología	9	14	11	189	8	*230*
Física	106	11	2	74	3	*196*
Ingeniería	1	11	38	1	102	*152*
Microbiología	2	0	0	3	5	*10*
Botánica	51	4	0	10	2	*67*
Estadística	10	0	0	2	0	*12*
Matemáticas	5	3	22	26	0	*56*
Suma	*187*	*110*	*94*	*347*	*262*	*1000*

Imagen 11.2: *Contribuciones de las celdas a la inercia, expresadas en tantos por mil. La suma de las filas y la de las columnas de esta tabla son idénticas a las inercias de las filas y columnas, expresadas en tantos por mil, de la imagen 11.1*

ciones a la inercia. Así, las contribuciones de las 10 filas suman 1000, su media será de 100, por tanto consideraremos contribuyentes principales las filas con contribuciones mayores del 100‰. Por otro lado, tenemos cinco columnas, lo que da una media de 200‰, por tanto, las columnas *D* y *E* son las que más contribuyen.

Contribuciones de las celdas a la inercia

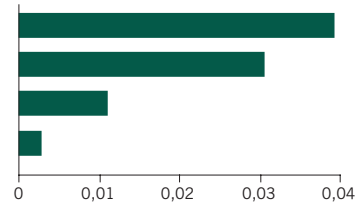
Podemos afinar más en el análisis de las contribuciones a la inercia examinando la contribución de cada celda. Como describimos en el capítulo 4, cada celda de la tabla contribuye con un valor positivo a la inercia total que, de nuevo, podemos expresar en tantos por mil (imagen 11.2). Vemos que las celdas [Zoología, *D*] y [Física, *A*] contribuyen mucho a la inercia; estas dos celdas juntas contribuyen casi al 30% de la inercia total de la tabla (189 + 106 = 295‰, es decir el 0,295 de la inercia total, es decir el 29,5%). Podemos denominar contribuciones ji-cuadrado a las contribuciones de las celdas, ya que sus valores son idénticos a las contribuciones relativas de cada celda al estadístico χ^2. Sumando las filas, o las columnas de la tabla de la imagen 11.2, llegamos a las mismas contribuciones, expresadas en tantos por mil, que vimos en la imagen 11.1.

Descomposición de la inercia en ejes principales

Otra descomposición importante de la inercia es con respecto a los ejes principales. En la página 106, dimos los valores de las inercias de los dos ejes principales de esta tabla de 10 × 5, que tiene cuatro dimensiones. En la imagen 11.3 mostramos los valores de todas las inercias principales, en porcentajes y en forma de diagrama de barras (lo llamaremos *diagrama de descomposición*). Hemos visto que las inercias principales se pueden interpretar por ellas mismas, por ejemplo, como el cuadrado de correlaciones canónicas (cap. 8, pág. 89). Sin embargo, en general, interpretaremos sus valores con relación a la inercia total, en general expresadas en porcentajes y no en tantos por mil.

Imagen 11.3:
Inercias principales de todas las dimensiones de los datos sobre la financiación científica expresadas en valores absolutos, en porcentajes y en porcentajes acumulados, y diagrama de descomposición

Dimensión	*Inercia principal*	*% inercia*	*% acumulado*
1	0,03912	47,2%	47,2%
2	0,03038	36,7%	83,9%
3	0,01087	13,1%	97,0%
4	0,00251	3,0%	100,0%



Componentes de cada inercia principal

Cada inercia principal es por ella misma una inercia, que hemos calculado a partir de las proyecciones de los perfiles fila (o de los perfiles columna) sobre los ejes principales. Por ejemplo, los 10 perfiles fila de los datos sobre la financiación de la investigación científica ocupan *un espacio completo* de dimensionalidad 4, uno menos que el número de columnas. La suma ponderada de los cuadrados de las distancias de los perfiles fila a su centroide es igual a la inercia total: su valor es de 0,08288. El primer eje principal es la recta mínimo-cuadrática más próxima a los perfiles. Este eje pasa por el centroide de las filas que se halla en el *origen*, o punto cero de la representación. Supongamos que proyectamos todos los perfiles fila sobre este eje. En tal caso, la primera inercia principal es la suma ponderada de los cuadrados de las distancias de estas proyecciones al centroide. Es decir, la primera inercia principal, igual a 0,03912, es la inercia de los puntos proyectados sobre el eje principal unidimensional. En la tabla de la imagen 11.4, mostramos las contribuciones de las filas y de las columnas a la primera inercia principal, calculadas a partir de las coordenadas principales de las filas y columnas. Así, vemos que la categoría *D* es la que más contribuye al primer eje, seguida de la *A*, mientras que las restantes categorías contribuyen muy poco. Con relación a las filas, vemos que Zoología (muy asociada con *D*) y Física (muy asociada con *A*) contribuyen en casi el 78% a la inercia del primer eje.

Imagen 11.4:
Contribución de las filas y de las columnas a la primera inercia principal; en valores absolutos, cuya suma es igual a la primera inercia principal, y expresadas de forma relativa en tantos por mil (‰)

Filas	*Inercia*	*‰ inercia*	Columnas	*Inercia*	*‰ inercia*
Geología	0,00062	16	*A*	0,00890	228
Bioquímica	0,00118	30	*B*	0,00260	67
Química	0,00023	6	*C*	0,00265	68
Zoología	0,01616	413	*D*	0,02471	632
Física	0,01426	365	*E*	0,00025	6
Ingeniería	0,00153	39			
Microbiología	0,00001	0			
Botánica	0,00345	88			
Estadística	0,00057	14			
Matemáticas	0,00112	29			
Total	*0,03912*	*1000*	*Total*	*0,03912*	*1000*

Descomposición completa de la inercia en los perfiles en los ejes principales

Podemos repetir lo anterior en todos los ejes principales. Así, en la imagen 11.5 mostramos los valores de la inercia de las filas descompuesta en los cuatro ejes principales (podemos construir una tabla similar para las columnas). Igual que en la imagen 11.4, en la que expresamos la descomposición de la inercia del primer eje principal en valores absolutos y en valores relativos, en tantos por mil, podríamos hacer lo mismo para todos los ejes principales. Así, veríamos que las filas que más contribuyen al eje 2 son Geología, Ingeniería y Bioquímica. El examen de las contribuciones de las filas (o de las columnas) a la inercia de los ejes principales nos proporciona un respaldo numérico a la interpretación de los mapas.

Componentes de la inercia de cada perfil

En la tabla de la imagen 11.5, los totales de las columnas nos dan las inercias principales de los ejes, mientras que los totales de las filas nos dan las inercias de los perfiles (por tanto, los valores de los totales de las filas deben ser iguales a los valores de la primera columna de la imagen 11.1). También podemos expresar las contribuciones a la inercia en términos relativos con relación a la inercia de las filas, como proporciones, en porcentajes o en tantos por mil. Estos resultados nos informarán sobre la inercia de las filas explicada por cada eje. Se trata de una miniversión de lo que hacíamos al determinar el porcentaje de inercia total que explicaba cada eje —aquí hacemos lo mismo pero fila a fila—. En la imagen 11.6 mostramos estos valores relativos en tantos por mil, es decir, el total de cada fila es 1000. Así, por ejemplo, vemos que el eje 2 es el que mejor explica la fila Geología, mientras que el eje 1 es el que mejor explica Física. Por otro lado, ni el eje 1 ni el 2 explican Matemáticas, puesto que su inercia se halla principalmente en la tercera dimensión.

Álgebra de la descomposición de la inercia

En la imagen 11.7 hemos representado gráficamente la descomposición de la inercia al mismo tiempo que introducimos un poco de notación. El punto $\mathbf{a}_i$, representa un perfil en un espacio multidimensional, por ejemplo el *i*-ésimo

	Eje principal				
Disciplina científica	*Eje 1*	*Eje 2*	*Eje 3*	*Eje 4*	*Total*
Geología	0,00062	0,00978	0,00082	0,00013	*0,01135*
Bioquímica	0,00118	0,00754	0,00084	0,00034	*0,00990*
Química	0,00023	0,00088	0,00029	0,00032	*0,00172*
Zoología	0,01616	0,00158	0,00063	0,00073	*0,01909*
Física	0,01426	0,00010	0,00169	0,00016	*0,01621*
Ingeniería	0,00153	0,00941	0,00127	0,00036	*0,01256*
Microbiología	0,00001	0,00056	0,00008	0,00019	*0,00083*
Botánica	0,00345	0,00016	0,00180	0,00011	*0,00552*
Estadística	0,00057	0,00001	0,00042	0,00003	*0,00102*
Matemáticas	0,00112	0,00037	0,00302	0,00015	*0,00466*
Total	*0,03912*	*0,03038*	*0,01087*	*0,00251*	*0,08288*

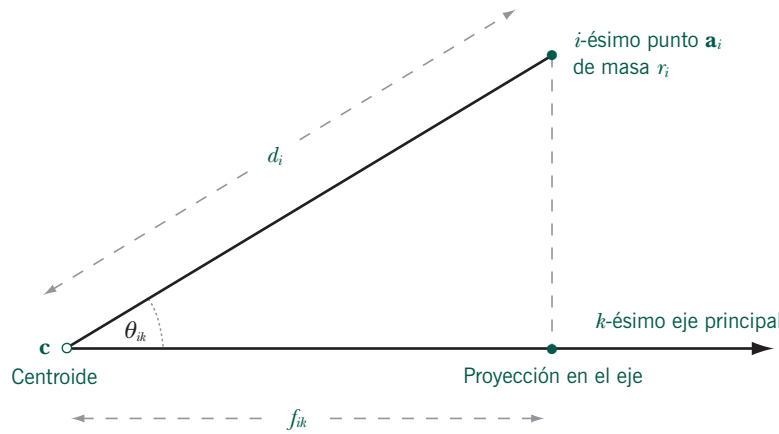
Imagen 11.5:
Descomposición en valores absolutos de la inercia de las filas (disciplinas científicas) en los cuatro ejes principales. La suma de las contribuciones de los ejes a las filas (totales de las filas) son las inercias de las filas de la imagen 11.1. Las sumas de las contribuciones de filas a los ejes (totales de las columnas) con las inercias principales de la imagen 11.3. La suma total de la tabla es la inercia total

Imagen 11.6: *Contribuciones relativas (en ‰) de los ejes principales a la inercia de las filas. En la última fila expresamos las inercias principales también en valores relativos que podemos interpretar como contribuciones relativas medias (comparar estos valores con los de la imagen 11.3)*

	Eje principal				
Disciplina científica	*Eje 1*	*Eje 2*	*Eje 3*	*Eje 4*	*Total*
Geología	55	861	72	11	*1000*
Bioquímica	119	762	85	35	*1000*
Química	134	510	170	186	*1000*
Zoología	846	83	33	38	*1000*
Física	880	6	104	10	*1000*
Ingeniería	121	749	101	28	*1000*
Microbiología	9	671	96	224	*1000*
Botánica	625	29	326	20	*1000*
Estadística	554	7	410	30	*1000*
Matemáticas	240	79	649	33	*1000*
Media	*472*	*367*	*131*	*30*	*1000*

perfil, de masa r_i, a una distancia d_i del perfil fila medio **c**. Por la ecuación (4.7), sabemos que la inercia total es igual a $\sum_i r_i d_i^2$. Hemos simbolizado la coordenada principal de $\mathbf{a}_i$ en el eje principal k por f_{ik}. Por tanto, la inercia en este eje (es decir la inercia principal k-ésima) será $\sum_i r_i f_{ik}^2$, en general simbolizada como λ_k. La contribución relativa de un punto i a la inercia principal del eje k sera $r_i f_{ik}^2$ dividido por λ_k (en la imagen 11.4 damos, en tantos por mil, las contribuciones relativas de los puntos al eje 1). En la imagen 11.5, mostramos los valores absolutos $r_i f_{ik}^2$ de las 10 filas y los 4 ejes principales de los datos sobre la financiación de la investigación científica. Los totales de las columnas de la imagen 11.5 son iguales a λ_k, mientras que los totales de las filas son la suma de las inercias de los ejes en esta fila $r_i d_i^2$. Gracias al teorema de Pitágoras, sabemos que $d_i^2 = \sum_k f_{ik}^2$, por tanto la contribución de los ejes a la inercia de las filas será:

Imagen 11.7: *Representación gráfica de un perfil $\mathbf{a}_i$ en un espacio multidimensional, a una distancia χ^2 d_i del centroide **c**, proyectado en la coordenada f_{ik} sobre el k-ésimo eje principal*



$$\sum_k r_i f_{ik}^2 = r_i d_i^2$$

Por tanto, la contribución relativa del eje k a la inercia del punto i es $r_i f_{ik}^2$ dividido por $r_i d_i^2$ (en la tabla de la imagen 11.6 damos estos valores relativos en tantos por mil).

Las contribuciones relativas como cuadrados de los cosenos de los ángulos

También podemos interpretar geométricamente las contribuciones relativas que hemos mostrado en la imagen 11.6. Dado que la proporción de inercia del punto i explicada por el eje k es $r_i f_{ik}^2 / r_i d_i^2 = (f_{ik}/d_i)^2$, viendo la imagen 11.7, queda claro que este valor es el cuadrado del coseno del ángulo formado por el punto i y el eje k. Supongamos que θ_{ik} sea dicho ángulo, entonces la contribución relativa del eje a la inercia del punto es $\cos^2(\theta_{ik})$. Por ejemplo, la contribución relativa del eje 1 a Física es de 0,880, por tanto, $\cos^2(\theta_{51}) = 0{,}880$, así pues, $\cos(\theta_{51}) = 0{,}938$ y, en consecuencia, el ángulo es $\theta_{51} = 20^\circ$. Este resultado muestra que Física, explicada principalmente por el eje 1, se halla cerca del eje 1, y forma un pequeño ángulo de 20º con dicho eje. En cambio, la contribución relativa del eje 1 a Geología es sólo de 0,055, lo que corresponde a un ángulo de $\theta_{11} = 76^\circ$ entre el eje 1 y Geología, luego esta disciplina científica no se halla cerca de este eje, más bien se encuentra en otras dimensiones del espacio (de hecho, se halla principalmente en el eje 2, lo que podemos deducir por la contribución relativa, de 0,861, de este eje a Geología).

Las contribuciones relativas como correlaciones al cuadrado

Existe todavía otra interpretación de las contribuciones relativas. Podemos interpretar los cosenos de los ángulos entre vectores como coeficientes de correlación. Por tanto, las contribuciones relativas son correlaciones al cuadrado. Así pues, podemos decir que la correlación de Física con el eje 1 es $\sqrt{0{,}880} = 0{,}938$, mientras que la correlación de Geología con dicho eje es de sólo $\sqrt{0{,}055} = 0{,}234$. Si la correlación es 1, el perfil se halla sobre el eje principal, mientras que si la correlación es 0, el perfil es perpendicular al eje principal (forma un ángulo de 90º).

Calidad de la representación en un subespacio

Gracias al teorema de Pitágoras, podemos sumar los cuadrados de los cosenos de los ángulos formados por un perfil y cada uno de los ejes, para obtener una suma de cosenos al cuadrado que relaciona el perfil con el subespacio definido por estos ejes. Así por ejemplo, podemos calcular el ángulo entre un perfil fila y el plano principal a partir de la suma de las contribuciones relativas de los dos ejes principales. Así, en la imagen 11.8 hemos sumado las dos primeras columnas de la tabla de la imagen 11.6. Interpretamos estas sumas como una medida de la *calidad* de la representación de los perfiles en los mapas bidimensionales que vimos en el capítulo 10, de la misma manera que la suma de los dos primeros porcentajes de inercia nos da una medida de la calidad global (o media) de la representación. Además, podemos ver qué perfiles están bien representados y cuáles no. Así, por ejemplo, en la última fila de la imagen 11.8 podemos ver que la calidad global del mapa bidimensional es del 83,9% y que, por tanto, no explicaría el 16,1%

Imagen 11.8:
Calidad de la representación (en tantos por mil) de los perfiles fila en dos dimensiones; solamente para Matemáticas *la inercia explicada es menor del 50%*

Disciplina científica	*Calidad*	Categoría de financiación	*Calidad*
Geología	916	*A*	587
Bioquímica	881	*B*	816
Química	644	*C*	465
Zoología	929	*D*	968
Física	886	*E*	990
Ingeniería	870		
Microbiología	680		
Botánica	654		
Estadística	561		
Matemáticas	319		
Global	*839*	*Global*	*839*

de la inercia de los perfiles. Algunos perfiles no estarán bien representados por hallarse más en el tercer y cuarto ejes que en los dos primeros. Tenemos, por ejemplo, que las Matemáticas están mal representadas, pues dos tercios de su inercia se encuentran fuera del plano. En las imágenes 10.2 y 10.3 el perfil de Matemáticas se parece al de Estadística, pero en este caso la proyección de su posición no es un reflejo preciso de su verdadera posición.

Analogías con el análisis factorial

Este apartado va dirigido, principalmente, a los lectores que conozcan el análisis factorial. Varios elementos del AC tienen elementos análogos a los del análisis factorial.

- El análogo al *coeficiente de carga del factor* es el coseno del ángulo formado por un perfil y un eje, es decir, la raíz cuadrada de la correlación al cuadrado con el signo de la coordenada del perfil. Por ejemplo, a partir de las imágenes 11.1 y 11.4, podemos calcular las correlaciones al cuadrado entre las categorías *A*, *B*, *C*, *D* y *E* con el primer eje principal.

$$A: \frac{0{,}00890}{0{,}01551} = 0{,}574 \qquad B: \frac{0{,}00260}{0{,}00911} = 0{,}286 \qquad C: \frac{0{,}00265}{0{,}00778} = 0{,}341$$

$$D: \frac{0{,}02471}{0{,}02877} = 0{,}859 \qquad E: \frac{0{,}00025}{0{,}02171} = 0{,}012$$

  Con los signos de las coordenadas de las columnas del mapa de la imagen 10.3, los «coeficientes de carga de los factores» serían las raíces cuadradas de los valores anteriores con sus correspondientes signos:

$$A: 0{,}758 \qquad B: 0{,}535 \qquad C: 0{,}584 \qquad D: -0{,}927 \qquad E: -0{,}108$$

- El análogo a la *comunalidad* es la calidad, expresada de 0 a 1. Por ejemplo, en la última columna de la imagen 11.8, mostramos las «comunalidades» de las

cinco categorías de las columnas de la solución bidimensional: 0,587; 0,816; 0,465; 0,968 y 0,990, para A, B, C, D y E, respectivamente.

- El análogo a la *unicidad* es 1 menos la calidad, expresada de 0 a 1. Por ejemplo, en la solución bidimensional, las «unicidades» de las cinco categorías de las columnas serán: 0,413; 0,184; 0,535; 0,032 y 0,010, para A, B, C, D y E, respectivamente.

RESUMEN: Contribuciones a la inercia

1. La inercia (total) de una tabla cuantifica la variación existente en los perfiles fila o en los perfiles columna.
2. Cada una de las filas y cada una de las columnas contribuye a la inercia total. Denominamos a estas contribuciones *inercias de las filas* e *inercias de las columnas*, respectivamente.
3. El AC se lleva a cabo con el objetivo de explicar la máxima inercia posible en el primer eje. El segundo eje explica el máximo de la inercia restante, y así sucesivamente. Por tanto, los ejes principales también descomponen la inercia total; a las inercias de los ejes principales las llamamos *inercias principales*.
4. A su vez, podemos descomponer las inercias principales con relación a las filas (o a las columnas). Tenemos dos posibilidades para expresar la *contribucion* del *k*-ésimo eje a la inercia de las filas (o de las columnas):
   a) con relación a la inercia principal del eje;
   b) con relación a la inercia de la fila o de la columna.
5. La posibilidad (a) nos permite diagnosticar qué filas (o columnas) han tenido un mayor papel en la determinación de la orientación de los ejes principales. Estas contribuciones nos facilitan la interpretación de los ejes principales.
6. La posibilidad (b) nos permite diagnosticar la posición de los perfiles con relación a los ejes y si éstos están bien representados en el mapa. Si están bien representados los podemos interpretar con seguridad, en cambio, si están pobremente representados, debemos interpretar sus posiciones con más cautela. Estos valores de inercia son los cuadrados de cosenos de los ángulos formados por los perfiles y los ejes principales, también los podemos interpretar como correlaciones al cuadrado.
7. La suma de los cuadrados de las correlaciones de un perfil con las dimensiones de un determinado subespacio nos proporciona una medida de la *calidad* de la representación del perfil en dicho subespacio.
8. Las correlaciones de los perfiles con los ejes, y las calidades de la representación del AC equivalen, respectivamente, a los coeficientes de carga de los factores y a las comunalidades del análisis factorial.