La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 20

Propiedades del escalado óptimo del análisis de correspondencias múltiples

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

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

www.fbbva.es

CAPÍTULO 20

Propiedades del escalado óptimo del ACM

En los capítulos 7 y 8 vimos que existen distintas definiciones de AC, así como diversas maneras de abordar este método. En este libro hemos enfatizado la aproximación geométrica de Benzécri que lleva a la visualización de datos. En los capítulos 18 y 19 quedó claro que el paso del AC simple, de dos variables, a las formas multivariadas no es directo, especialmente si tratamos de generalizar la interpretación geométrica. Una aproximación alternativa al caso multivariado, con exactamente el mismo aparato matemático que el ACM, consiste en acercarse a este método como una manera de cuantificar datos categóricos, generalizando de esta manera las ideas sobre escalado óptimo que vimos en el capítulo 7. También aquí veremos que existen distintas formas de presentar el ACM como una técnica de escalado. El estudio de estas aproximaciones alternativas enriquecerán nuestra comprensión sobre las propiedades de este método. En la literatura, la aproximación al ACM como un método de escalado óptimo se llama *análisis de homogeneidad*.

Contenido

Conjunto de datos 11: actitudes hacia la ciencia y el medio ambiente 205
La cuantificación de categorías como objetivo 206
El ACM como el análisis de componentes principales de la matiz binaria 206
Maximización de la correlación entre ítems 207
ACM del ejemplo de la opinión sobre la ciencia 208
Correlaciones individuales al cuadrado 209
Pérdida de homogeneidad 210
Geometría de la función de pérdida en el análisis de homogeneidad 210
Fiabilidad y alfa de Cronbach 212
RESUMEN: Propiedades del escalado óptimo del ACM 213

Conjunto de datos 11: actitudes hacia la ciencia y el medio ambiente

Este conjunto de datos lo hemos obtenido de la encuesta multinacional del ISSP de 1993 sobre el medio ambiente. Nos centraremos en $Q = 4$ preguntas sobre la opinión de la gente acerca del papel de la ciencia. Se preguntó a los encuestados si estaban o no de acuerdo con las afirmaciones siguientes:

A Creemos demasiado en la ciencia, y demasiado poco en los sentimientos y en la fe.

B En general, la ciencia moderna provoca más daños que beneficios.

C Cualquier cambio que el hombre cause en la naturaleza, a pesar de que tenga un base científica, es probable que empeore las cosas.

D La ciencia moderna solucionará los problemas medioambientales sin modificar sustancialmente nuestro modo de vida.

Tenemos cinco posibles respuestas categóricas:

1 Muy de acuerdo.

2 Algo de acuerdo.

3 Ni de acuerdo ni en desacuerdo.

4 Algo en desacuerdo.

5 Muy en desacuerdo.

Para simplificar solamente hemos utilizado datos de Alemania Occidental. Hemos omitido los casos con valores perdidos en cualquiera de las cuatro preguntas, lo que nos ha llevado a una muestra de $N = 871$. (Estos datos se hallan incluidos en nuestro paquete **ca** para R, que se ofrece en el apéndice B.)

La cuantificación de categorías como objetivo

En el capítulo 7 definíamos el AC como un método de cuantificación de las categorías de la variable columna que nos lleva a la mayor diferenciación, o discriminación, posible entre las categorías de la variable fila, o viceversa. Es lo que llamaríamos definición «asimétrica», ya que las filas y las columnas desempeñan papeles distintos en la definición, lo que también se refleja en los resultados. Así, expresamos los resultados de las columnas en coordenadas estándares, mientras que los de las filas los expresamos en coordenadas principales. En el capítulo 8 definimos el AC de forma «simétrica» como un método de cuantificación de las categorías que nos lleva a la mayor correlación entre filas y columnas. En esta definición, el papel de filas y columnas es el mismo. Esta cuantificación de las categorías no incluye ningún concepto geométrico específico; en concreto, no hace mención alguna a un espacio en el que podamos imaginar situados los datos, lo que, por el contrario, es muy importante en la aproximación geométrica para poder medir la inercia total y los porcentajes de inercia en los subespacios de baja dimensionalidad.

El ACM como el análisis de componentes principales de la matiz binaria

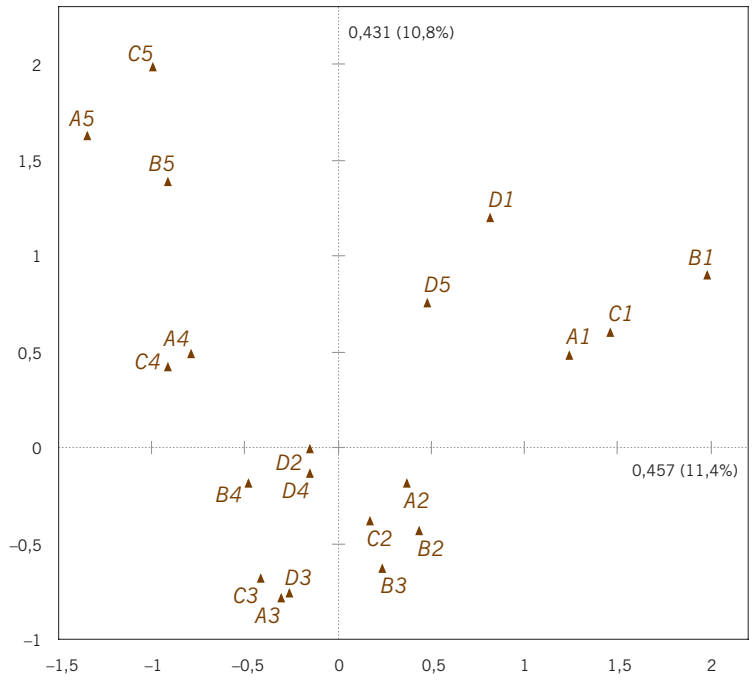
La metodología de cuantificación de categorías utilizada en el AC asimétrico sobre la matriz binaria se parece mucho al análisis de componentes principales (ACP). En general, aplicamos el ACP a datos derivados de una escala continua, no obstante, el ACP está muy relacionado con la teoría y el cálculo del AC (en realidad, podríamos decir que el AC es una variante del ACP aplicado a datos categóricos). En el ACP,

dado un conjunto de datos, en el que las filas son los casos y las columnas las variables (m variables, $x_1,..., x_m$), asignamos a las columnas unos coeficientes $\alpha_1,...,\alpha_m$ (que tendremos que estimar) que conducen a combinaciones lineales para las filas (casos) de la forma $\alpha_1 x_1 + ... + \alpha_m x_m$, las *puntuaciones*. Calculamos los coeficientes de manera que se maximice la varianza de las puntuaciones de las filas. Como vimos anteriormente, para hallar la solución tenemos que definir unas condiciones de identificación. En el ACP, estas condiciones son, en general, que la suma de los cuadrados de los coeficientes sea 1: $\sum_j \alpha_j^2 = 1$. Aplicar estas ideas a la matriz binaria, que sólo consta de ceros y unos, y asignar coeficientes $\alpha_1,..., \alpha_J$ a las variables binarias, para calcular luego las combinaciones lineales de las filas, simplemente significa sumar los coeficientes α (es decir los valores de la escala) de cada caso. Por tanto, la maximización de la varianza de los casos recuerda el procedimiento de escalado óptimo que vimos en el capítulo 7 (maximización de la discriminación entre filas). De hecho se trata de un concepto casi idéntico, con la excepción de las condiciones de identificación. En el escalado óptimo, las condiciones de identificación serían que la varianza ponderada (inercia) de los coeficientes (no la simple suma de cuadrados) fuera 1: $\sum_j c_j \alpha_j^2 = 1$. Aquí las c_j son las masas de las columnas, es decir, la suma de las columnas de la matriz binaria divididas por la suma total NQ de la matriz binaria; así, para cada variable categórica, la suma de los c_j es $1/Q$. Por tanto, con este cambio en las condiciones de identificación, podríamos llamar al ACM, ACP de datos categóricos, que maximiza la varianza de los casos. Los coeficientes son las coordenadas estándares de las categorías de las columnas, mientras que las coordenadas principales del ACM de los casos son las medias de los valores de éstos. Es decir $1/Q$ veces la suma de lo que hemos llamado antes «puntuaciones». La primera dimensión del ACM maximiza la varianza (primera inercia principal), la segunda dimensión maximiza la varianza con la condición de que las puntuaciones no estén correlacionadas con las de la primera dimensión y así sucesivamente.

Maximización de la correlación entre ítems

El *análisis de homogeneidad*, visto como una técnica de escalado óptimo del ACM, se contempla, habitualmente, como una generalización de la correlación según se expuso en el capítulo 8. En concreto, vimos la ecuación (8.1) como una manera alternativa de optimizar la correlación entre dos variables categóricas, que podemos fácilmente generalizar a más de dos variables. Para ilustrar este hecho utilizaremos una notación correspondiente a cuatro variables, sin embargo, podemos extenderlo fácilmente a Q variables con cualquier número de categorías (en nuestro ejemplo $Q = 4$, y el número total de categorías es $J = 20$). Supongamos que las cuatro variables toman los valores (desconocidos) de a_1 a a_5, de b_1 a b_5, de c_1 a c_5 y de d_1 a d_5. Asignaremos a los encuestados cuatro de estos valores a_i, b_j, c_k y d_l de acuerdo con sus respuestas, y de esta manera cuantificaremos las respuestas de toda la muestra, que simbolizamos como a, b, c y d (es decir, a indica todas las 871 respuesta cuantificadas a la pregunta A, etc.). Cada encuestado tendrá como puntuación la suma estos valores, $a_i + b_j + c_k + d_l$. Simbolizaremos las puntuaciones

Imagen 20.1:
Mapa del ACM (versión matriz binomial) sobre la actitud hacia la ciencia, que muestra los puntos correspondientes a las categorías en coordenadas principales. Dado que las inercias principales difieren sólo ligeramente (e incluso menos en forma de raíces cuadradas), en ambos ejes, las coordenadas principales presentan casi la misma contracción que las coordenadas estándares



de toda la muestra como $a + b + c + d$. En este contexto, llamamos *ítems* a las variables, *puntuaciones de los ítems* a los valores de a a d, y *suma de puntuaciones* a la suma $a + b + c + d$. Expresaremos, el criterio de búsqueda de los valores óptimos de la escala, como la maximización de la media de las correlaciones al cuadrado entre las puntuaciones de los ítems y la suma de puntuaciones:

$$\text{correlaciones al cuadrado} = \frac{1}{4}[\text{cor}^2(a, a+b+c+d) + \text{cor}^2(b, a+b+c+d) + \text{cor}^2(c, a+b+c+d) + \text{cor}^2(d, a+b+c+d)] \quad (20.1)$$

(Comparemos con el caso de dos variables en (8.1).) De nuevo necesitamos las condiciones de identificación. Es conveniente aplicar a la suma de puntuaciones la condición de media 0 y varianza 1: media $(a + b + c + d) = 0$, $\text{var}(a + b + c + d) = 1$. Obtenemos, de forma exacta, la solución a este problema de maximización, con las coordenadas estándares de las categorías de los ítems en el primer eje principal del ACM, la media de las correlaciones al cuadrado (20.1) maximizada es exactamente la primera inercia principal (del ACM de la matriz binaria).

ACM del ejemplo de la opinión sobre la ciencia

En el mapa bidimensional de la matriz binomial de la imagen 20.1 vemos, de nuevo, porcentajes de inercia muy bajos (los porcentajes basados en las inercias ajustadas son el 44,9% y el 34,2%, respectivamente). Sin embargo, en este caso, dado

Categorías	Preguntas				
	A	*B*	*C*	*D*	*Suma*
1 «Muy de acuerdo»	115	174	203	25	*518*
2 «Algo de acuerdo»	28	21	6	3	*57*
3 «Ni de acuerdo ni en desacuerdo»	12	7	22	9	*49*
4 «Algo en desacuerdo»	69	41	80	3	*194*
5 «Muy en desacuerdo»	55	74	32	22	*182*
Suma	*279*	*317*	*343*	*62*	*1000*

Imagen 20.2:
Contribuciones a la inercia en tantos por mil (‰) del primer eje principal (versión matriz binaria) de los datos sobre ciencia y medio ambiente

que los valores de las inercias principales son medias de correlaciones al cuadrado, debemos ignorar los porcentajes, ya que los valores de las inercias principales tienen interés *per se*. El valor máximo de (20.1) es 0,457. La segunda inercia principal, 0,431, se halla buscando un nuevo conjunto de valores que nos lleven a unas puntuaciones que no estén correlacionadas con los que se obtuvieron anteriormente, y que además maximicen (20.1); este valor máximo es 0,431. Y continuaríamos de esta manera para hallar los resultados de los restantes ejes, siempre no correlacionados con los hallados anteriormente. En el mapa de la imagen 20.1, vemos que las preguntas *A*, *B* y *C* presentan una distribución muy similar, como una cuña en forma de herradura, que va de profundos desacuerdos, a la izquierda, a fuertes acuerdos, a la derecha. Sin embargo, la pregunta *D* sigue una trayectoria completamente distinta con los dos valores extremos muy próximos. Las primeras tres preguntas presentaban un redactado negativo hacia la ciencia, mientras que la pregunta *D* tenía un redactado mucho más positivo; por tanto, habríamos esperado que *D5* se hallara hacia *A1*, *B1* y *C1*, y *D1* se hallara al lado de *A5*, *B5* y *C5*. Sin embargo, el hecho de que *D1* y *D5* se hallen tan cerca y dentro de la herradura indica que ambas están asociadas con respuestas extremas de las restantes tres preguntas: la explicación más plausible es que algunos encuestados hayan interpretado mal el cambio de sentido del redactado de la cuarta pregunta.

Correlaciones individuales al cuadrado

También es interesante conocer los valores de cada una de las correlaciones al cuadrado que componen (20.1). Podemos obtener estos valores directamente sumando la contribución de cada pregunta a la inercia del primer eje principal. Habitualmente, los resultados del ACM proporcionan esta información en forma de proporciones o en tanto por mil. En la imagen 20.2 detallamos estos valores en esta última forma para ilustrar cómo recuperar estas correlaciones. Las preguntas de *A* a *D* contribuyen, en las proporciones 0,279, 0,317, 0,343 y 0,062 de la inercia principal de 0,457. Dado que 0,457 es la media de las cuatro correlaciones al cuadrado, las correlaciones al cuadrado y, en consecuencia, las correlaciones son:

$$A: 0{,}279 \times 0{,}457 \times 4 = 0{,}510 \quad \text{correlación} = \sqrt{0{,}510} = 0{,}714$$
$$B: 0{,}317 \times 0{,}457 \times 4 = 0{,}579 \quad \text{correlación} = \sqrt{0{,}579} = 0{,}761$$
$$C: 0{,}343 \times 0{,}457 \times 4 = 0{,}627 \quad \text{correlación} = \sqrt{0{,}627} = 0{,}792$$
$$D: 0{,}062 \times 0{,}457 \times 4 = 0{,}113 \quad \text{correlación} = \sqrt{0{,}113} = 0{,}337$$

Estos cálculos muestran el bajo valor de la correlación de la pregunta *D* con relación a la puntuación total. Fijémonos en que, a pesar de que el ACM de la matriz binomial era la peor, desde un punto de vista geométrico habitual de distancias χ^2, inercia total, etc., las inercias principales y las contribuciones a las inercias principales tienen una interpretación muy interesante por sí mismas. En el *análisis de homogeneidad*, que teóricamente es equivalente al ACM de la matriz binaria, pero que interpreta el método desde el punto de vista de cuantificación de las categorías, denominamos *valores de discriminación* a las correlaciones al cuadrado 0,510; 0,579; 0,627 y 0,113.

Pérdida de homogeneidad

El análisis de homogeneidad generaliza la función objetivo (8.3) a muchas variables. Utilizando la notación anterior para el ejemplo que nos ocupa con cuatro variables, calcularíamos la puntuación media $\frac{1}{4}(a_i + b_j + c_k + d_l)$ de las puntuaciones de los ítems de cada encuestado y luego calcularíamos la varianza del encuestado dentro de su grupo de respuestas cuantificadas:

$$\begin{aligned}\text{varianza (para un caso)} = \frac{1}{4}\Big(& [a_i - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \\ & + [b_j - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \\ & + [c_k - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \\ & + [d_l - \frac{1}{4}(a_i + b_j + c_k + d_l)]^2 \Big)\end{aligned} \qquad (20.2)$$

A continuación, calcularíamos la *pérdida de homogeneidad* como la media de todos estos valores con relación a los *N* casos. El objetivo es minimizar esta pérdida. De nuevo el ACM (versión matriz binaria) resuelve este problema. La minimización de la pérdida es 1 menos la primera inercia principal, es decir, 1 – 0,457 = 0,543. Minimizar la pérdida es equivalente a maximizar la correlación que hemos definido anteriormente.

Geometría de la función de pérdida en el análisis de homogeneidad

El objetivo de minimizar la pérdida de homogeneidad tiene una interpretación geométrica muy atractiva, muy relacionada con la definición de distancia entre filas y columnas del AC que vimos en el capítulo 7. En realidad, el cálculo de la pérdida de homogeneidad es exactamente igual al cálculo de la distancia ponderada (7.6), aplicada a la matriz binaria. En la imagen 20.3 mostramos el mapa asimétrico de ACM de todos los *N* = 871 encuestados (en coordenadas principales) y las

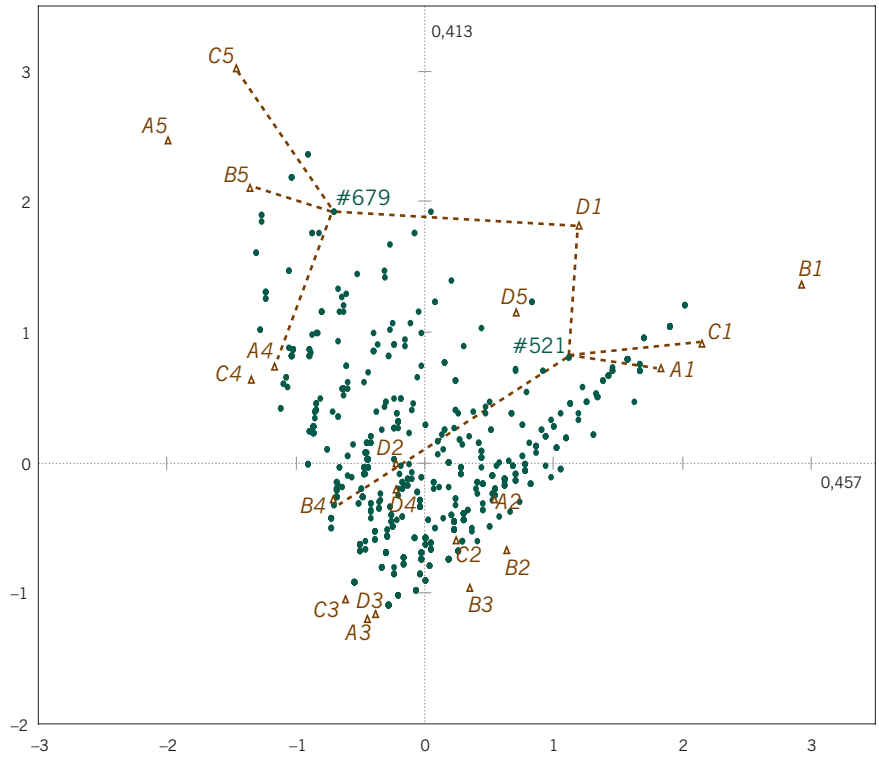


Imagen 20.3:
Mapa asimétrico (versión matriz binaria) de la opinión sobre la ciencia, que muestra los encuestados en coordenadas principales y las categorías en coordenadas estándares. Cada encuestado se halla en la media de sus cuatro respuestas. El ACM minimiza la suma de las distancias al cuadrado entre los puntos correspondientes a los individuos y sus respuestas

$J = 20$ categorías (en coordenadas estándares). Esto significa que los encuestados se hallan en los centroides de las categorías, siendo los pesos los valores relativos de las filas de la matriz binaria. Cada encuestado tiene un perfil que consta de ceros, y valores de $\frac{1}{4}$ en las posiciones correspondientes a las cuatro respuestas. Por tanto, el punto correspondiente a cada encuestado se halla en la posición de la media ordinaria de sus respuestas. En el mapa (imagen 20.3) hemos etiquetado los encuestados #679 y #521. Las respuestas del individuo #679 son: (*A4*, *B5*, *C5*, *D1*). Es decir, está en desacuerdo con las tres primeras preguntas y de acuerdo con la cuarta —en el mapa hemos unido mediante trazo discontinuo este individuo, situado a la izquierda, con las categorías correspondientes a sus respuestas—. Se trata de una fuerte y consistente opinión a favor de la ciencia. En contraste, las respuestas del individuo #521 son más variadas: (*A1*, *B4*, *C1*, *D1*). Este último individuo opina que creemos demasiado en la ciencia y que la interferencia de los humanos en la naturaleza empeorará las cosas. Sin embargo, también está muy de acuerdo en que la ciencia solucionará nuestros problemas medioambientales, al mismo tiempo que opina que la ciencia hace más daño que bien. Este tipo de respuestas explica el hecho de que *D1* se haya acercado hacia el centro del mapa, entre las dos opiniones extremas. Cada encuestado se halla en la media de sus

cuatro respuestas. Para cualquier configuración de respuestas categóricas, los encuestados se hallarán en la posición media. El mapa que mostramos en la imagen 20.3 es óptimo en el sentido de que las líneas que unen los encuestados con las categorías son las más cortas posibles (en términos de sumas de distancias al cuadrado). Llamamos *diagrama de estrellas* a los diagramas resultantes de unir los puntos correspondientes a los individuos con los de sus respuestas. Podríamos decir que el objetivo del ACM es la obtención de diagramas de estrellas con las menores distancias entre los individuos y sus respuestas en el sentido mínimo-cuadrático. El número de uniones entre los puntos correspondientes a los N encuestados y los correspondientes a sus Q respuestas es NQ. La pérdida de homogeneidad es la media de los cuadrados de las uniones (por ejemplo, en (20.2) donde $Q = 4$, dividimos la suma de los cuatro cuadrados por 4; para los N individuos dividimos la suma de cuadrados por $4N$). Por tanto, la media de la suma de las uniones al cuadrado en la primera dimensión es 1 – 0,457 = 0,513 y en la segunda dimensión es 1 – 0,413 = 0,587. Por el teorema de Pitágoras, la media de la suma de las uniones al cuadrado en el mapa bidimensional de la imagen 20.3 es 0,513 + 0,587 = 1,100.

Fiabilidad y alfa de Cronbach

En el ejemplo que nos ocupa con datos sobre la ciencia y el medio ambiente, vimos que la pregunta *D* no está muy correlacionada con las restantes (pág. 210). En este contexto, si hubiéramos querido obtener un indicador global de la opinión de la gente sobre la ciencia, hubiéramos dicho que estos resultados nos muestran que la pregunta *D* empeoraba la *fiabilidad* de la puntuación total, y que lo mejor habría sido eliminarla. En teoría de fiabilidad suponemos que las Q preguntas o ítems miden una estructura subyacente. La *alfa de Cronbach* es una medida estándar de fiabilidad definida como:

$$\alpha = \frac{Q}{Q-1}\left(1 - \frac{\sum_q s_q^2}{s^2}\right) \tag{20.3}$$

donde s_q^2 es la varianza de la puntuación del ítem q-ésimo, $q = 1,..., Q$ (por ejemplo, las varianzas de *a, b, c* y *d*) y s^2 es la varianza de la suma de las puntuaciones media (por ejemplo, la varianza de $(a + b + c + d)$). Aplicando esta definición a la primera dimensión del resultado del ACM, vemos que la alfa de Cronbach se reduce a:

$$\alpha = \frac{Q}{Q-1}\left(1 - \frac{1}{Q\lambda_1}\right) \tag{20.4}$$

donde λ_1 es la primera inercia principal de la matriz binaria. Por tanto, cuanto mayor sea la inercia principal, mayor será la fiabilidad. Utilizando $Q = 4$ y $\lambda_1 = 0,4574$ (cuatro dígitos significativos para aumentar un poco la exactitud) obtenemos:

$$\alpha = \frac{4}{3}\left(1 - \frac{1}{4 \times 0,4574}\right) = 0,605$$

Una vez visto el comportamiento de la pregunta D, una posibilidad es eliminarla y hacer de nuevo los cálculos con las tres preguntas que están altamente intercorrelacionadas. No mostramos aquí estos resultados, solamente señalaremos que la primera inercia principal correspondiente a estas tres variables es $\lambda_1 = 0{,}6018$, con un incremento de fiabilidad hasta $\alpha = 0{,}669$ (utilizando (20.4), con $Q = 3$). Como comentario final es interesante que nos demos cuenta de que la media de las correlaciones al cuadrado de un conjunto de variables aleatorias, cuyas correlaciones dos a dos no son iguales a cero, es igual a $1/Q$, lo que corresponde a una alfa de Cronbach igual a 0. El valor $1/Q$ es exactamente el umbral que hemos utilizado en (19.7) para ajustar las inercias principales (valores propios), y es también la inercia principal media del ACM de la matriz binaria que hemos mencionado en el capítulo 18.

RESUMEN: Propiedades del escalado óptimo del ACM

1. En el contexto de dos variables, definimos el escalado óptimo como la búsqueda de valores numéricos para las categorías de una variable que lleven a la separación máxima de los grupos definidos por la otra variable. Este problema es equivalente a hallar los valores numéricos, para las categorías, que conduzcan a la máxima correlación entre las variables fila y las variables columna.
2. En un contexto multivariado, el escalado óptimo consiste en la búsqueda de valores numéricos para las categorías de todas las variables que optimicen la correlación entre las variables y su suma (o promedio). En concreto, maximizamos la media de las correlaciones al cuadrado de los valores numéricos de cada variable, *puntuaciones de los ítems*, con su suma (o promedio), es decir, su *puntuación*.
3. De forma equivalente, el *análisis de homogeneidad* consiste en minimizar el promedio de las varianzas de las puntuaciones de los ítems de cada encuestado de la muestra.
4. En general, la aproximación al ACM como un escalado óptimo, ejemplificado por el análisis de homogeneidad, es el mejor marco para la interpretación de los resultados del ACM de una matriz binaria. Es mejor interpretar las inercias principales y su descomposición como correlaciones que de forma geométrica, tal como hacíamos en el AC simple.
5. La primera inercia principal del ACM, versión binaria, presenta una relación monotónica con la fiabilidad expresada con la alfa de Cronbach: cuanto mayor sea la inercia principal, mayor será la fiabilidad.
6. Dado que las coordenadas estándares de la matriz binaria, de la matriz de Burt y de la forma ajustada son idénticas, podemos aplicar las propiedades de escalado óptimo a las tres versiones del ACM.