La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del apéndice E

Epílogo

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

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

www.fbbva.es

Epílogo

En este libro hemos presentado el análisis de correspondencias (AC) como un método versátil para la visualización de datos, aplicable a una amplia variedad de situaciones. Este epílogo tiene como objeto avanzar algo más en el análisis de algunos aspectos de este método que aparecen con frecuencia en discusiones sobre AC, así como aportar algunas consideraciones personales.

Mapas simétricos

La interpretación de los mapas simétricos, aunque es una opción más de los mapas de AC, sigue siendo uno de los aspectos más controvertidos de este método. Este tipo de mapas expresan tanto las filas como las columnas en coordenadas principales; es decir, a pesar de que las proyecciones de los perfiles fila y los perfiles columna ocupan espacios distintos, mostramos sus proyecciones en un mismo mapa. Hemos visto (por ejemplo, en los capítulos 9 y 10) que la diferencia entre los mapas simétricos y los asimétricos (en los que todos los puntos se hallan en el mismo espacio) es el factor de escala de los ejes principales, la raíz cuadrada de sus respectivas inercias principales. Por tanto, las direcciones, indicadas por los puntos en coordenadas principales y por sus homólogos en coordenadas estándares, son casi iguales cuando las raíces cuadradas de las inercias principales no son muy distintas; así, podemos ver un ejemplo en el mapa de la imagen 13.4 en la que los ejes del biplot, que pasan a través de los vértices, casi coinciden con los puntos correspondientes a los perfiles. En tales casos, la forma de interpretar los mapas simétricos y los asimétricos como si fueran biplots es válida. Sin embargo, si las raíces cuadradas de las inercias principales son muy distintas, al interpretar los mapas simétricos como si fueran un biplot pueden aparecer problemas; lo podemos ver, por ejemplo, en las diferentes direcciones definidas por las categorías de fumadores en los mapas de las imágenes 9.2 y 9.5. Aun así, como se pone de manifiesto en el artículo de Gabriel que mencionamos a continuación, la distorsión que se produce al interpretar los mapas simétricos como si fueran verdaderos biplots, no es demasiado grande.

- Gabriel K.R. «Goodness of Fit of Biplots and Correspondence Analysis». *Biometrika* 89 (2002): 423-436.

Esto significa que el debate sobre las diferencias de escala es más bien un tema académico. Toda la discusión que ha generado este tema tiene poco interés cuando se trata de aplicar el AC. En mi opinión, el mapa simétrico sigue sien-

do, por defecto, el mejor mapa. De hecho, es la opción que aparece por defecto en nuestro paquete **ca** para R. Si interpretamos de forma asimétrica la matriz de datos, en la que la filas representen «unidades observacionales» (como, por ejemplo, individuos en estudios sociales, localidades de muestreo en ecología o en arqueología, o textos en lingüística, etc.) y las columnas representen «variables» (como, por ejemplo, las respuestas categóricas en sociología, las especies en ecología, los artefactos en arqueología, o los indicadores de estilo en lingüística, etc.), el biplot estándar del AC es una buena alternativa. Representa de forma óptima las distancias entre unidades y permite una interpretación tipo biplot válida de las unidades proyectadas sobre las direcciones de las variables. Además, las longitudes de los vectores (variables) tienen una interpretación clara.

«No puedes comerte un pastel y, al mismo tiempo, conservarlo»

Desgraciadamente, en el contexto que nos ocupa, se cumple este dicho inglés. Podemos decir lo mismo de la expresión: «En la vida, no lo puedes tener todo». Sería maravilloso que en un solo mapa pudiéramos representar de forma óptima e interpretar los tres elementos siguientes:

1. Las distancias entre perfiles fila.
2. Las distancias entre perfiles columna.
3. Los productos escalares entre filas y columnas, que reconstruyen los datos originales (es decir, el biplot).

Sin embargo, la realidad es que, al mismo tiempo y como máximo, podemos tener representados óptimamente sólo dos de los tres elementos anteriores. Los mapas simétricos representan óptimamente las distancias ji-cuadrado entre los perfiles fila y entre los perfiles columna. Por tanto, podemos interpretar las distancias entre filas y las distancias entre columnas (es decir, se cumplen los puntos 1 y 2). No podemos interpretar de forma óptima las relaciones entre filas y columnas. Sin embargo, teniendo en cuenta las observaciones del párrafo anterior, las podemos interpretar con una seguridad razonable. En los mapas asimétricos representamos de forma óptima, por ejemplo, los perfiles fila, mientras que los vértices columna proporcionan los perfiles extremos como puntos de referencia. Sus proyecciones sobre los ejes del biplot nos permiten interpretar de forma óptima las relaciones entre filas y columnas (es decir, se cumplen los puntos 1 y 3). Los biplots estándares del AC son una variante de los mapas asimétricos que muestran, por ejemplo, los perfiles fila, al mismo tiempo que acercan los vértices columna, multiplicando por la raíz cuadrada de sus masas, para mejorar la representación conjunta (es decir, se cumplen 1 y 3). En este último biplot, podemos relacionar las proyecciones de los vectores columna sobre los ejes del biplot con sus contribuciones a los ejes principales (capítulo 13).

Normalización simétrica en SPSS

Aparte del programa libre R, y del programa comercial XLSTAT que hemos descrito en el apéndice de cálculo, todavía no hemos comentado nada sobre otros softwares que incluyen el AC. Entre estos programas encontramos Minitab, Stata, Statistica, SPAD, SAS y SPSS. Dado que SPSS es ampliamente utilizado, es conveniente que hagamos algunos comentarios sobre esta opción. En el módulo *Categories* del programa de AC del SPSS, se proporciona un biplot llamado *symmetrical normalization* que no hemos visto en este libro. Podríamos confundir dicho biplot con el mapa simétrico que sí hemos descrito. Sin embargo, no se trata de lo mismo, ya que el primero presenta las coordenadas estándares multiplicadas por las raíces cuadradas de los valores singulares (es decir, la raíz cuarta de las inercias principales) y no por los valores singulares. Dicho de otro modo —con relación a los pasos (A.8) y (A.9) del algoritmo básico de cálculo del AC que vimos en la página 267—, este procedimiento calcula $\mathbf{\Phi D}_\alpha^{\frac{1}{2}}$ y $\mathbf{\Gamma D}_\alpha^{\frac{1}{2}}$ en vez de $\mathbf{\Phi D}_\alpha$ y $\mathbf{\Gamma D}_\alpha$ como en los mapas simétricos. Por tanto, la «normalización simétrica» del SPSS proporciona una representación óptima de los productos escalares, pero no proporciona una representación óptima de distancias, ya que ni filas ni columnas se expresan en coordenadas principales. Por tanto, esta representación gráfica proporciona sólo uno de los tres elementos mencionados anteriormente (se cumple 3, pero ni 1 ni 2). A pesar de que la diferencia entre esta representación gráfica y el mapa simétrico es sólo un tema de factores de escala en los dos ejes —que en la mayoría de casos son difícilmente distinguibles para un observador no experimentado—, no recomendamos la utilización de este mapa ya que no aporta beneficio alguno (en realidad representa una pérdida) con relación a las otras opciones existentes. Si las inercias principales de los dos ejes son similares, entonces, como vimos anteriormente, las posiciones relativas de los puntos en la «normalización simétrica» son prácticamente idénticas a las del mapa simétrico. Sin embargo, es preferible el mapa asimétrico ya que representa las distancias ji-cuadrado en su verdadera escala. El mapa con «normalización simétrica» lo denominamos *symmetric biplot*, y es una de las posibilidades de nuestro paquete **ca** de R. Para obtenerlo escribiremos: `map="symbiplot"` (págs. 304-305). Curiosamente, en las últimas versiones de SPSS no era posible representar un mapa simétrico, una de las representaciones gráficas más populares entre los investigadores franceses. Sigue siendo imposible en las últimas versiones del programa obtener un mapa conjunto de filas y columnas en coordenadas principales. La mejor opción es seleccionar la normalización «principal», que proporciona los valores numéricos de las coordenadas principales de filas y columnas. Sin embargo, el programa siempre rechaza el realizar un mapa conjunto con estos datos, prefiere mapas separados. A no ser que los datos originales del usuario se hallen en formato SPSS, como decíamos, no recomendamos el programa del AC de SPSS. Sin embargo, dentro del módulo *Categories*, resultan muy útiles para ciencias sociales el programa de optimización de escalas para análisis de correspondencias múltiples (llamado, en versiones anteriores, análisis de homogeneidad) y el de análisis de componentes principales no lineal (CatPCA).

Categorías poco frecuentes

El efecto de las categorías poco frecuentes sobre la distancia χ^2 y sobre el resultado del AC es también un tema que ha generado mucha discusión, especialmente entre los investigadores en ecología, casi siempre sin justificación. Por ejemplo, según C.R. Rao, «la distancia ji-cuadrado que utiliza proporciones marginales en el denominador otorga al medir las afinidades entre perfiles, demasiada importancia a las categorías con bajas frecuencias» (en pág. 42 del siguiente artículo):

- Rao C.R. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Quëstiió* 19 (1995): 23-63. Disponible en Internet en:

  www.idescat.net/sort/questiio/questiiopdf/19.1,2,3.1.radhakrishna.pdf

Sin embargo, la realidad es que en AC ponderamos cada categoría proporcionalmente a su masa, lo que reduce el papel de las categorías de baja frecuencia. Lo podemos ver de forma muy simple analizando las contribuciones numéricas de las distintas categorías a los ejes principales. Así, podemos constatar que las categorías poco frecuentes tienen, en general, poca influencia sobre la solución hallada; es decir, la solución sería casi la misma si eliminásemos estas categorías del análisis.

Consideremos, a título ilustrativo los datos sobre abundancia de especies del capítulo 10 (pág. 109) con los que calculamos la abundancia relativa de las 10 especies más frecuentes y la de las 10 menos abundantes, y lo comparamos con sus contribuciones relativas, en porcentaje, a los dos primeros ejes del mapa de AC de la imagen 10.5. Los resultados son los siguientes:

		Contribución a los ejes	
Especies	*Abundancia relativa*	Eje 1	Eje 2
10 más abundantes	74,6%	77,3%	89,3%
10 menos abundantes	0,4%	0,8%	0,5%

Estos cálculos ilustran que las especies poco frecuentes no contribuyen demasiado a la solución bidimensional, pues las contribuciones se hallan mucho más en la línea con las abundancias de cada grupo de especies. Según nuestra experiencia, sólo de vez en cuando, las categorías poco frecuentes contribuyen de forma excesiva a los ejes principales. En tales casos, debemos eliminarlas o combinarlas con otras categorías. Esta situación se da en estudios sociológicos, en los que las categorías de baja frecuencia, como los valores perdidos, coinciden en el mismo grupo de encuestados. Estas categorías pueden dominar la solución del ACM, a menudo definiendo el primer eje. Lo vimos, en los mapas de las imágenes 18.2 y 18.5. Podemos rectificar esta situación mediante un análisis de subgrupos o

combinando, de forma razonable, las respuestas correspondientes a categorías de baja frecuencia con otras similares. En ecología se produciría una situación análoga cuando determinadas especies poco frecuentes se hallaran simultáneamente en las mismas muestras. Sin embargo, no se trata de una situación común; en general, las especies poco frecuentes ocurren de forma aleatoria en distintas muestras.

Las categorías de baja frecuencia son, a menudo, observaciones atípicas

A menudo, las filas y columnas con frecuencias bajas son observaciones atípicas con extraños perfiles. Probablemente por este motivo llaman la atención y dan la impresión de que pueden afectar, de forma importante, al análisis. Sin embargo, como hemos dicho, tienen en general poca influencia sobre la solución del AC debido a su escasa masa. Además, según hemos mostrado en el capítulo 13 y mencionamos anteriormente, el biplot estándar del AC podría solucionar este problema, ya que «acerca» estos puntos a razón de la raíz cuadrada de sus masas, lo que en la práctica implica una eliminación de las observaciones atípicas de poca frecuencia. Ello también constituye una ilustración gráfica de su escaso efecto sobre la configuración de los ejes principales.

La distancia χ^2 es una distancia de Mahalanobis

Este apartado es algo técnico, aunque resulta útil para que el lector formado estadísticamente pueda comprender que la distancia ji-cuadrado, aparte de ser la clave de todas las propiedades del AC, es una distancia estadística apropiada. Matricialmente, podemos expresar la distancia euclídea ponderada como:

$$\text{distancia euclídea ponderada} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\mathbf{D}_w(\mathbf{x}-\mathbf{y})} \qquad \text{(E.1)}$$

donde $\boldsymbol{x}$ e $\boldsymbol{y}$ son vectores con elementos x_j y y_j, $j = 1, ..., J$, $^{\mathsf{T}}$ indica la transposición de una matriz o de un vector, y $\mathbf{D}_w$ es la matriz digonal que contiene los factores de ponderación w_j. Podemos suponer que las filas de una tabla de contingencia corresponden a una variable aleatoria *multinomial*. La distribución multinomial es una generalización de la distribución binomial. Constituye un modelo para la descripción del comportamiento de datos muestreados de poblaciones con probabilidades p_j, $j = 1, ..., J$ para cada uno de los J grupos. Por ejemplo, los tres tipos de lectores del capítulo 3 (tabla de la imagen 3.1). A partir de la hipótesis nula de que hemos muestreado los datos en la misma población, los cinco niveles educativos de este conjunto de datos serían muestras multinomiales de la población con probabilidades p_1, p_2, p_3 en las que las estimaciones de p_j de los tres grupos son los elementos del perfil medio $\hat{p}_1 = c_1 = 0{,}183$, $\hat{p}_2 = c_2 = 0{,}413$ y $\hat{p}_3 = c_3 = 0{,}404$ (última fila de la tabla de la imagen 3.1). La *distancia de Mahalanobis* es la distancia clásica utilizada para datos multivariantes agrupados. Se basa en la inversa de la matriz de covarianzas de las variables:

$$\text{distancia de Mahalanobis} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\,\Sigma^{-1}(\mathbf{x}-\mathbf{y})} \qquad \text{(E.2)}$$

excepto por el hecho de que implica una matriz cuadrada completa de pesos Σ^{-1}, y no una matriz diagonal, tiene el aspecto de una distancia euclídea ponderada (E.1). Para una distribución multinomial, la matriz de covarianzas Σ tiene una forma simple. Por ejemplo, en nuestro caso trinomial $J = 3$ (los resultados serían similares para cualquier número de grupos):

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{bmatrix} = \mathbf{D}_p - \mathbf{p}\mathbf{p}^\top \qquad \text{(E.3)}$$

donde $\mathbf{p}$ es el vector de las p_j y $\mathbf{D}_p$ la correspondiente matriz diagonal. Estimamos (E.3) sustituyendo las probabilidades p_j por sus estimaciones c_j. No es posible invertir la matriz de covarianzas Σ de la forma habitual, ya que se trata de una matriz singular. Por tanto no podemos hallar una matriz Σ^{-1} tal que $\Sigma\Sigma^{-1} = \mathbf{I}$. Una manera de sortear este problema es eliminar una de las categorías y seguir con sólo $J - 1$ categorías. No obstante, calquiera que sea la categoría que se omita, la distancia de Mahalanobis será la misma. Una aproximación alternativa más elegante, completamente equivalente pero que utiliza las J categorías, consiste en utilizar la *generalización inversa*, simbolizada como Σ^-, que tiene la propiedad de que $\Sigma\Sigma^-\Sigma = \Sigma$ (la *inversa de Moore-Penrose*). La inversa generalizada de Moore-Penrose de (E.3) es igual a:

$$\Sigma^- = \begin{bmatrix} 1/p_1 & 0 & 0 \\ 0 & 1/p_2 & 0 \\ 0 & 0 & 1/p_3 \end{bmatrix} = \mathbf{D}_p^{-1} \qquad \text{(E.4)}$$

Es decir, la distancia χ^2 estima de forma exacta la distancia de Mahalanobis (E.2). Aquí la situación es similar a la del análisis discriminante lineal: para maximizar la discriminación entre grupos, suponemos que los grupos tienen matrices de covarianzas iguales, lo que en el caso multinomial equivale a asumir el modelo de independencia y que los vectores se hallan en un espacio de Mahalanobis, que equivale a un espacio χ^2.

Rotación de las soluciones

En este libro no hemos visto nada sobre rotaciones debido a que raramente se justifican o se necesitan en AC. Debemos tener en cuenta que el espacio de perfiles no es un espacio de vectores real ilimitado, es un espacio delimitado por puntos unidad o vértices, que definen un simplex en un espacio multidimensional. La idea de alinear los puntos de las distintas categorías en ejes que formen ángulos rectos no tiene, en nuestro contexto, el mismo significado que en el análisis factorial en que los ángulos rectos indican que las correlaciones entre variables son cero (recordemos que en AC, la suma de los elementos del perfil es 1; por tanto, la posición de un determinado punto viene determinada por las de los restantes puntos). Las rotaciones pueden ser apropiadas en algunos contextos como el

ACM y en ACP no lineal (que no hemos visto en este libro) cuando analizamos varias variables simultáneamente. Por ejemplo, en ACM ocurre con frecuencia que los puntos correspondientes a las no respuestas se hallan juntos —mostrando así una elevada asociación dentro del conjunto de datos— y que, sin embargo, su posición no coincida con ningún eje principal. En tal caso podría tener interés hacer girar los ejes para separar el efecto de los puntos de no respuesta de los restantes. De todas formas, podemos solucionar mejor este problema haciendo un análisis de subgrupos (capítulo 21), que permite ignorar las no respuestas y concentrar el analisis en las respuestas sustantivas. En cualquier caso, si queremos llevar a cabo una rotación, deberemos tener en cuenta las masas de las categorías. Una posibilidad podría ser una versión ponderada de la rotación varimax del análisis factorial cuyo (para el caso de las columnas) criterio de maximización sería:

$$\sum_j \sum_k c_j^2 \left(\tilde{y}_{jk}^2 - \frac{1}{J} \sum_{j'} \tilde{y}_{j'k}^2 \right)^2 \qquad \text{(E.5)}$$

donde $\tilde{y}_{jk}$ es la coordenada estándar rotada, es decir, el *(j,k)*-ésimo elemento de $\tilde{\mathbf{Y}} = \mathbf{YQ}$, siendo $\mathbf{Q}$ una matriz ortogonal de rotación. Fijémonos en que las masas c_j se hallan al cuadrado ya que la función objetivo implica la cuarta potencia de las coordenadas. Dado que $c_j \tilde{y}_{jk}^2 = (c_j^{\frac{1}{2}} \tilde{y}_{jk})^2$, sugerimos una alternativa casi idéntica, que deriva de una pequeña modificación del usual criterio varimax: llevar a cabo una rotación (sin ponderar) con las coordenadas estándares recalibrabas $c_j^{\frac{1}{2}} y_{jk}$, que son exactamente las mismas utilizadas en el biplot estándar del AC. Es decir, rotar la solución para concentrar (o, concretar, en terminología del análisis factorial) las contribuciones de las categorías sobre los ejes rotados.

AC y modelización

En el capítulo 13 vimos el AC en K^* dimensiones como una descomposición que se puede expresar de la siguiente manera [véanse (13.4) y (A.14) en el apéndice teórico]:

$$p_{ij} = r_i c_j + r_i c_j \left(\sum_{k=1}^{K^*} \sqrt{\lambda_k} \phi_{ik} \gamma_{jk} \right) + e_{ij} \quad i = 1,\ldots,I; \quad j = 1,\ldots,J \qquad \text{(E.6)}$$

Obtenemos la solución del AC minimizando la suma ponderada de los cuadrados de los residuos e_{ij}. La primera parte de la descomposición, $r_i c_j$, es el valor esperado según el modelo de independencia, de manera que la segunda parte explica las desviaciones del modelo de independencia como la suma de K^* términos bilineales (esta parte bilineal tienen una interpretación geométrica en K^* dimensiones, lo que constituye la mayor parte del tema de este libro). Sin embargo, podemos sustituir el modelo de independencia por cualquier otro modelo a elección del usuario. Por ejemplo, en el artículo que mencionamos a continuación, los autores consideran para tablas de contingencia, modelos log-lineales, así que utilizan

el AC como una manera para explorar la esructura de las posibles desviaciones del modelo log-lineal.

- Van der Heijden P.G.M., A. de Falguerolles y J. de Leeuw. J. «A Combined Approach To Contingency Table Analysis and log-Linear Analysis (with Discussion)». *Applied Statistics* 38 (1989): 249-292.

También podemos utilizar esta estrategia en tablas de contingencia de múltiples entradas, utilizando una modelización de las tablas de contingencia que primero tenga en cuenta los efectos principales y determinadas interacciones para, a continuación, calcular los residuos del modelo para analizarlos mediante AC. Sin embargo, dado que los datos ya se han centrado con relación al modelo, no se trata de una aplicación directa del AC. Por tanto, al realizar el AC no debemos llevar a cabo el centrado, y en el ajuste de mínimos cuadrados ponderado debemos utilizar los valores marginales originales de la tabla.

AC y mapas espectrales

El análisis de correspondencias presenta una gran afinidad con los *mapas espectrales*, un método desarrollado originalmente por Paul Lewi en los años setenta y que en el desarrollo de nuevos medicamentos se ha utilizado ampliamente en el análisis biológico de espectros de actividad. Una referencia reciente es:

- Lewi P.J. «Analysis of Contingency Tables». En: B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi y J. Smeyers-Verbeke (eds.). *Handbook of Chemometrics and Qualimetrics*: *Part B*. Amsterdam: Elsevier, 1998: 161-206.

En los mapas espectrales trabajamos con los logaritmos de los valores de la tabla. Sin embargo, llevamos a cabo la ponderación de filas y de columnas como en el AC —utilizamos las masas de filas y de columnas de la tabla original—. Antes de realizar la DVS, llevamos a cabo un centrado con relación a las medias ponderadas de filas y columnas, como en el AC. Si la inercia de los datos es baja, el mapa espectral y el mapa del AC son casi iguales. La diferencia entre los dos métodos es más acusada para inercias mayores. En los mapas espectrales representamos los cocientes de los logaritmos de los datos, lo que hace que este procedimiento tenga propiedades para el diagnóstico del modelo muy interesantes. Además de cumplir el principio de equivalencia distribucional (pág. 60), es *subcomposicionalmente coherente*. Es decir, los cocientes entre valores permanecen constantes aunque se eliminen filas o columnas del análisis. Una propiedad que refuerza este tipo de análisis; pues nos permite analizar con seguridad grupos de filas o de columnas. Por el contrario, en el AC cuando analizamos subgrupos los perfiles y las distancias se ven afectados. Es decir, el AC no es subcomposicionalmente coherente. De ahí la necesidad de desarrollar el AC de subgrupos que vimos en el capítulo 21. Para más detalles y referencias, podemos consultar el documento de trabajo aceptado en el Journal of Classification:

• Greenacre M.J. y P.J. Lewi. «Distributional Equivalence and Subcompositional Coherence in the Analysis of Contingency Tables, Ratio-Scale Measurements and Compositional Data». *Working paper* no. 908, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, 2005. Disponible en Internet: www.econ.upf.edu/en/research/onepaper.php?id=908.

La dimensión de conjuntos de datos categóricos multivariantes

Para finalizar este epílogo, vamos a plantear un problema sin resolver. Sabemos que en AC la dimensión de una tabla $I \times J$, es $(I-1, J-1)$. Para una matriz de Burt $J \times J$ obtenida a partir de Q variables categóricas, el número de dimensiones es $J-Q$. Sin embargo, sabemos que $J-Q$ dimensiones es mucho más de lo que necesitamos para reproducir de forma exacta las tablas que se hallan fuera de la diagonal. Podríamos definir la dimensionalidad de un conjunto de datos con Q variables como el número de dimensiones necesarias para reproducir exactamente la tabla de contingencia $\frac{1}{2}Q(Q-1)$. Es decir, el número de dimensiones necesarias en un AC conjunto para explicar el 100% de la inercia. La pregunta es: ¿podemos determinar las dimensiones de antemano? o, por el contrario, sólo podemos determinarlas empíricamente. Dar respuesta a esta cuestión sería muy útil. Por ejemplo, en el ACM ajustado en el que consideramos sólo las K^* dimensiones para las cuales $\sqrt{\lambda_k} > 1/Q$. En estudios empíricos, la inercia explicada utilizando este número (K^*) de dimensiones se acerca mucho al 100%, aunque no es una prueba suficiente de que la dimensionalidad sea K^*. ¡Quizá con el tiempo se llegue a publicar una tercera edición de este libro, en la que este problema ya esté resuelto!