La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 5

Representación gráfica de distancias ji-cuadrado

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

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

www.fbbva.es

CAPÍTULO 5

Representación gráfica de distancias ji-cuadrado

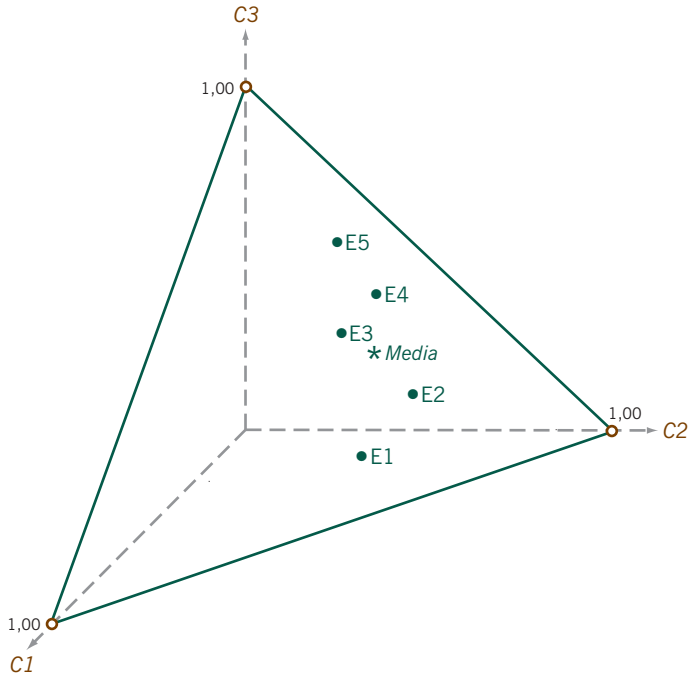
En el capítulo 3 interpretamos las posiciones de los perfiles bidimensionales en un sistema de coordenadas triangular mediante distancias euclídeas. En el capítulo 4 definimos la distancia ji-cuadrado (distancia χ^2) entre perfiles. Vimos la relación existente entre la distancia χ^2, el estadístico ji-cuadrado y la inercia de una matriz de datos. La distancia χ^2 es una distancia euclídea ponderada, en la que ponderamos los cuadrados de las diferencias entre coordenadas, con el inverso del correspondiente elemento del perfil medio. Hasta ahora, no hemos visualizado realmente las distancias χ^2 entre perfiles. Sólo lo hemos hecho en la imagen 4.2, en la que los elementos del perfil medio eran iguales y, por tanto, en este caso particular, las distancias χ^2 también eran distancias euclídeas. En este capítulo veremos cómo con una simple transformación del espacio de perfiles, las distancias que observamos en nuestras representaciones gráficas son distancias χ^2.

Contenido

Diferencia entre la distancia χ^2 y la distancia euclídea usual

En la imagen 5.1 hemos representado gráficamente los perfiles fila de la imagen 3.1 en unos ejes de coordenadas perpendiculares, en el espacio físico tridimensional habitual. Aquí las distancias entre perfiles no son distancias χ^2, son distancias euclídeas (sin ponderar) [fórmula (4.5)]. En este tipo de espacio, calculamos las distancias entre dos perfiles con elementos x_j e y_j, respectivamente

Imagen 5.1:
Espacio de perfiles, que muestra los perfiles de los niveles de educación en un triángulo equilátero en un espacio tridimensional; las distancias son euclídeas



(donde $j = 1, ..., J$), sumando los cuadrados de las diferencias de las coordenadas $(x_j - y_j)^2$, en todas las dimensiones j y calculando finalmente la raíz cuadrada de la suma resultante. Esta es la manera como usualmente calculamos las distancias «directamente» en el espacio físico con el que estamos familiarizados. Como hemos visto, el cálculo de la distancia χ^2 es distinto, ya que dividimos cada diferencia al cuadrado por el correspondiente elemento del perfil medio. Es decir, cada término es igual a $(x_j - y_j)^2/c_j$, donde c_j es el correspondiente elemento del perfil medio. Dado que solamente podemos interpretar y comparar distancias en nuestro espacio físico habitual, sería deseable algún tipo de modificación del mapa que hiciera que las distancias «directas» habituales se conviertieran en distancias χ^2. Afortunadamente, como veremos a continuación, esto es posible mediante transformaciones simples de los perfiles.

Transformación de las coordenadas antes de representarlas gráficamente

En el cálculo de la distancia χ^2, podemos reescribir cada término de la forma $(x_j - y_j)^2/c_j$, como $(x_j/\sqrt{c_j} - y_j/\sqrt{c_j})^2$. Esta forma equivalente de expresar el término general en el cálculo de la distancia es formalmente idéntica a la de la distancia euclídea usual; es decir, como una diferencia al cuadrado. El único cambio es que ahora las coordenadas no son los valores originales x_j e y_j, sino que las hemos transformado en $x_j/\sqrt{c_j}$ e $y_j/\sqrt{c_j}$. Ello sugiere que, en vez de utilizar como coordenadas los elementos originales de los perfiles, podríamos utilizar estos elementos divididos por las raíces cuadradas de los correspondientes elementos del per-

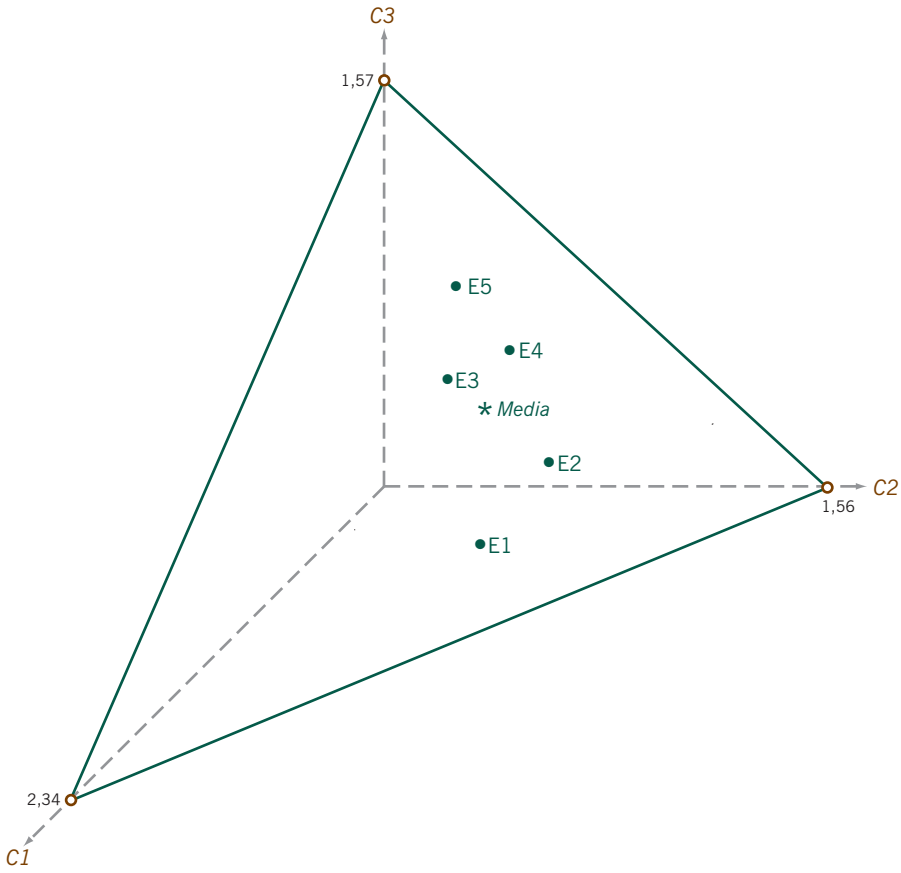


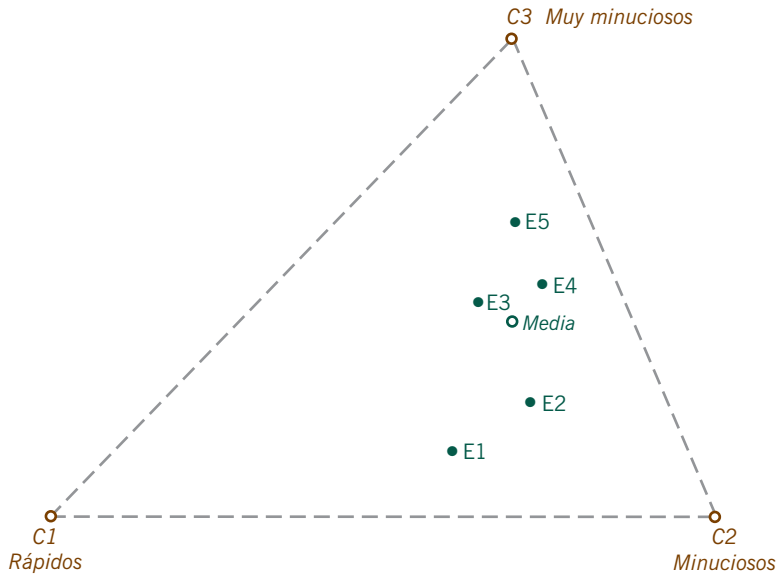
Imagen 5.2:
El espacio de perfiles muestra los ejes extendidos en distinta proporción, de manera que las distancias entre perfiles se convierten en distancias χ^2

fil medio. En tal caso, la distancia euclídea habitual entre estas coordenadas transformadas sería la distancia χ^2 que buscamos.

Efecto práctico de la transformación

Los valores c_j son los elementos del perfil medio, por tanto, todos son menores que 1. En consecuencia, la transformación consistente en dividir los elementos del perfil por $\sqrt{c_j}$ comportará un incremento del valor de todas las coordenadas. De todas maneras, unas aumentarán más que las otras. Si un determinado c_j es relativamente pequeño en comparación con los otros (es decir, la frecuencia de la *j*-ésima categoría de la columna es relativamente pequeña), entonces las correspondientes coordenadas $x_j/\sqrt{c_j}$ e $y_j/\sqrt{c_j}$ aumentarán de forma relativamente grande. A la inversa, una c_j grande, correspondiente a una categoría más frecuente, comportará un incremento relativamente menor de las coordenadas transformadas. Por tanto, la transformación aumenta los valores de las categorías con frecuencias bajas, relativamente más que los de las categorías con frecuencias altas. En el espacio sin transformar de la imagen 5.1, los vértices se hallan a una unidad

Imagen 5.3:
Espacio triangular de perfiles del espacio extendido de la imagen 5.2 situado en un «plano» (comparemos con la imagen 3.2). El triángulo se ha estirado más en la dirección de C1, la categoría menos frecuente



física del origen (es decir, del punto cero) de los tres ejes de coordenadas. El primer vértice, de coordenadas [1 0 0], se convertirá en la posición $[1/\sqrt{c_1}\ 0\ 0]$; es decir, su posición en el primer eje se estira hasta llegar al valor $1/\sqrt{0{,}183} = 2{,}34$. De forma similar, el segundo y el tercer vértice se estiran hasta los valores $1/\sqrt{c_2} = 1/\sqrt{0{,}413} = 1{,}56$ y $1/\sqrt{c_3} = 1/\sqrt{0{,}404} = 1{,}57$, respectivamente. En la imagen 5.2 mostramos estos valores junto a los vértices de los correspondientes ejes. Los perfiles transformados ocupan nuevas posiciones en el espacio, pero siguen dentro del triángulo definido por los vértices transformados. Fijémonos en que el estiramiento es mayor en la dirección de *C1*, la categoría con menor frecuencia marginal.

Interpretación alternativa en términos de ejes de coordenadas recalibrados

Geométricamente, podríamos ver la situación anterior de otra manera. En los tres ejes de los sistemas de coordenadas sin transformar de las imágenes 2.4 y 5.1, las marcas que indican las escalas (por ejemplo, los valores 0,1; 0,2; 0,3; etc.) se hallan separadas a intervalos iguales; sin embargo, como indica la imagen 5.2, la transformación provoca una extensión de los tres vértices. Aunque después de la transformación las escalas en los tres ejes son distintas, podríamos seguir considerando los tres vértices como si fueran perfiles unitarios, es decir, a una unidad del origen. En el eje *C1*, un intervalo de 0,1 entre dos marcas sería una longitud física de 0,234, mientras que, en los ejes *C2* y *C3*, estos intervalos serían de 0,156 y de 0,157, respectivamente. Por tanto, el intervalo unidad en el eje *C1* sería aproximadamente un 50% más largo que el mismo intervalo en los otros

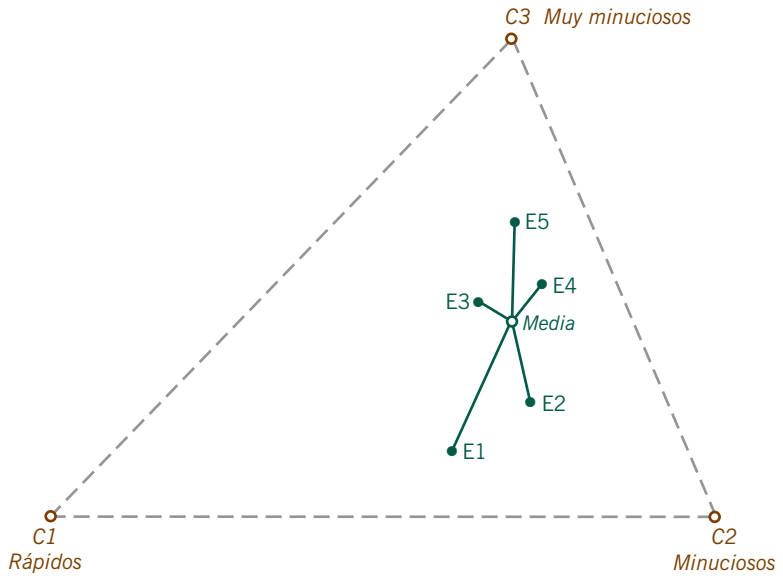


Imagen 5.4:
Espacio de perfiles «extendido» que muestra las distancias χ^2 de los perfiles a su centroide; la inercia es la media ponderada de la suma de los cuadrados de estas distancias y el estadístico χ^2 es la inercia multiplicada por el tamaño de la muestra (en este ejemplo, $n = 312$)

dos ejes. A pesar de ello, seguiríamos utilizando los valores originales de los perfiles para situarlos en el espacio tridimensional. En definitiva, sea cual sea la manera cómo veamos la transformación —ya sea como una transformación de los valores de los perfiles, o como un estiramiento y una recalibración de los ejes—, el resultado es el mismo: ahora situamos los perfiles en el espacio triangular estirado que mostramos en la imagen 5.2. En la imagen 5.3 hemos representado gráficamente el triángulo extendido en un plano. Queda claro que el vértice *C1*, correspondiente a la categoría más rara de lectores *rápidos*, la categoría que más se ha estirado.

Interpretación geométrica de la inercia y del estadístico χ^2

Ahora que en el espacio transformado las distancias son distancias χ^2, podemos trazar líneas de unión entre los perfiles y su media para mostrar, así, las distancias χ^2 entre los perfiles y su media (imagen 5.4). Por la fórmula (4.7), sabemos que la suma de las distancias de las filas a su media, ponderadas con sus respectivas masas, es igual a la inercia total de la tabla. Si en vez de ponderar con las masas, ponderamos con las frecuencias totales de las filas (la frecuencia total de una fila es n veces la masa de la fila, siendo n la suma total de la tabla), entonces la suma ponderada de los cuadrados de estas distancias es igual al estadístico χ^2. Obtenemos resultados equivalentes con los perfiles columna y el perfil columna medio. Por tanto, geométricamente, podemos interpretar la inercia y el estadístico χ^2 como medidas del grado de dispersión de los perfiles (de las filas o de las columnas) con relación a su media.

El principio de equivalencia distribucional

Para explicar este principio consideremos otra vez la imagen 3.1. Supongamos que podemos distinguir dos tipos de lectores *minuciosos*, los que se fijan más en la información política y los que se fijan más en la información cultural y deportiva. Simbolizaremos estas dos categorías por *C2a* y *C2b*, respectivamente. Supongamos, además, que en estas dos nuevas columnas, las frecuencias relativas de los niveles de educación son las mismas. Es decir, que no hay diferencias entre ambas subdivisiones del grupo de lectores *minuciosos* por lo que respecta a su educación. En el capítulo 3 dijimos que este tipo de columnas eran distribucionalmente equivalentes, en tanto que tienen los mismos perfiles. La subdivisión de la columna *C2* en *C2a* y *C2b* no aporta nueva información sobre las diferencias entre los niveles de educación. Por tanto, cualquier análisis de estos datos debería dar los mismos resultados, tanto si subdividimos *C2*, como si lo dejamos como una sola categoría. Decimos que un análisis que satisface esta propiedad cumple el *principio de equivalencia distribucional*. Si hubiéramos utilizado las distancias euclídeas habituales para medir las distancias entre los perfiles de los niveles de educación, no se cumpliría este principio ya que si hubiéramos hecho la mencionada subdivisión, hubiésemos obtenido resultados distintos. En cambio, la distancia χ^2 cumple siempre este principio, no se ve afectada por este tipo de subdivisiones de las categorías de la matriz de datos. Es decir, si unimos dos columnas distribucionalmente equivalentes, no cambian las distancias χ^2 entre las filas. En la práctica, esto significa que podemos unir columnas con perfiles similares sin que la geometría de las filas se vea afectada, y viceversa. El hecho de que en este tipo de análisis la introducción de arbitrariedades técnicas que modifiquen el número de categorías no afecte al resultado, y que éste sólo se vea modificado si introducimos modificaciones sustanciales, da ciertas garantías a los investigadores.

Las distancias χ^2 hacen que las contribuciones de las categorías sean más parecidas

Ya conocemos cómo organizar una representación gráfica para visualizar las distancias χ^2, pero ¿por qué tenemos que visualizar las distancias χ^2? ¿Por qué no utilizamos directamente distancias euclídeas? Podemos justificar la utilización de las distancias χ^2 de muchas maneras, unas más técnicas que otras. Existen razones más profundas que van más allá de las derivadas de la visualización del estadístico χ^2 que acabamos de presentar. Una de ellas se basa en la constatación de que existen diferencias importantes en las varianzas de los valores de las frecuencias de las distintas categorías. Así, por ejemplo, en la imagen 3.1 podemos ver el recorrido de los valores de los perfiles de la columna *C1* (de 0,115 a 0,357), una columna con frecuencias pequeñas, que es menor que el de la columna *C3* (de 0,143 a 0,615), una columna con frecuencias mayores. Esta observación ilustra una regla general sobre los datos de frecuencias: los conjuntos de frecuencias pequeñas presentan menor dispersión que los conjuntos de frecuencias grandes. Lo podemos ver calculando las contribuciones de las categorías de la imagen 3.1 a los cuadrados de las distancias euclídeas y χ^2, distancias entre los perfiles de los

	Euclídea			χ^2		
FILA	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*
E1	28,7	7,1	64,2	47,1	5,1	47,7
E2	2,1	38,7	59,1	4,7	37,2	58,1
E3	13,2	66,4	20,4	25,5	56,7	17,8
E4	37,1	2,8	60,1	56,6	1,9	41,5
E5	6,5	29,7	63,9	13,3	27,1	59,6
Global	*17,0*	*21,8*	*61,2*	*31,3*	*17,7*	*51,0*

Imagen 5.5:
Porcentajes de contribución de las categorías de las columnas a los cuadrados de las distancias euclídea y χ^2 de los perfiles fila a su centroide (datos de la imagen 3.1)

niveles de educación y su centroide (perfil medio). Por ejemplo, el cuadrado de la distancia euclídea entre el perfil del quinto nivel de educación E5 y el centroide es:

$$\begin{aligned}(\text{Distancia euclídea})^2 &= (0{,}115-0{,}183)^2+(0{,}269-0{,}413)^2+(0{,}615-0{,}404)^2\\ &= 0{,}00453+0{,}02080+0{,}04475\\ &= 0{,}07008\end{aligned}$$

mientras que el cuadrado de la distancia χ^2 es:

$$\begin{aligned}(\text{Distancia } \chi^2)^2 &= \frac{(0{,}115-0{,}183)^2}{0{,}183}+\frac{(0{,}269-0{,}413)^2}{0{,}413}+\frac{(0{,}615-0{,}404)^2}{0{,}404}\\ &= 0{,}02480+0{,}05031+0{,}11081\\ &= 0{,}18592\end{aligned}$$

[véanse ecuaciones (4.5) y (4.6)]. Cada una de estas distancias al cuadrado es la suma de tres valores correspondientes a las tres categorías de las columnas. Para valorar la contribución de cada tipo de lector, podemos expresar estos tres valores como porcentajes respecto de la distancia total. Por ejemplo, la contribución de la categoría *C1*, al cuadrado de la distancia euclídea es de 0,00453 sobre un total de 0,07008, es decir el 6,5%; mientras que la contribución de *C1* al cuadrado de la distancia χ^2 es de 0,02480 sobre un total de 0,18592, es decir el 13,3% (fila E5 de la imagen 5.5). En la imagen 5.5 mostramos las contribuciones de todos los términos, así como la contribución global de cada categoría, calculada considerando conjuntamente todos los términos de la misma (última fila de la imagen 5.5). Así, vemos que la contribución global de la categoría *C1* a la distancia euclídea es del 17,0%, mientras que la contribución global de esta categoría a la distancia χ^2 es del 31,3%. Este ejercicio ilustra el fenómeno general de que *C1*, la categoría con frecuencias más pequeñas, contribuye menos a la distancia euclídea que, por ejemplo, *C3*. Sin embargo, la contribución de *C1* a la distancia χ^2 se ve incrementada gracias a la división por las frecuencias medias.

Distancia euclídea ponderada

Como vimos en el capítulo 4, la distancia χ^2 es un ejemplo de distancia euclídea ponderada, su definición general es la siguiente:

$$\text{Distancia euclídea ponderada} = \sqrt{\sum_{j=1}^{p} w_j \left(x_j - y_j\right)^2} \tag{5.1}$$

donde w_j son los valores positivos de los pesos y x_j, con $j = 1, ..., p$ e y_j, con $j = 1, ..., p$ son dos puntos en un espacio p-dimensional. En el análisis de componentes principales (ACP), un método muy relacionado con el AC, las p dimensiones vienen definidas por variables continuas, a menudo en diferentes escalas de medida. En el ACP eliminamos el efecto de la escala sobre la varianza dividiendo los datos por las desviaciones estándar s_j de las respectivas variables. De esta manera, reemplazamos las observaciones x_j e y_j de la variable j por x_j/s_j e y_j/s_j. Podemos ver esta operación como la utilización de una distancia euclídea ponderada con pesos $w_j = 1/s_j^2$, los inversos de las varianzas. En la definición de la distancia χ^2 entre perfiles, los pesos son iguales a $w_j = 1/c_j$, es decir, son iguales a los inversos de los elementos del perfil medio.

Justificación teórica de la distancia χ^2

A pesar de que en el AC los perfiles se hallan en la misma escala de frecuencias relativas, seguimos teniendo la necesidad de compensar las diferencias entre varianzas, situación similar a la del ACP. En la *distribución de Poisson*, una de la distibuciones estadísticas estándar para recuentos, es inherente el hecho de que los conjuntos de frecuencias con medias más elevadas tienen varianzas mayores que los conjuntos de frecuencias con medias menores. Precisamente, una característica de la distribución de Poisson es que la varianza es igual a su media. En nuestro contexto, podemos interpretar la transformación de las frecuencias —consistente en dividir por la raíz cuadrada de la frecuencia esperada (media)— como una estandarización de los datos, ya que la raíz cuadrada de la frecuencia media es un equivalente a la desviación típica. De todas formas, existen otros procedimientos de estandarización. Pero, ¿por qué la distancia χ^2 es tan especial? Aparte de cumplir el principio de equivalencia distribucional y de hacer que el análisis de filas y el de columnas sean simétricos, otra ventaja de la utilización de la distancia χ^2 hay que buscarla en las propiedades de la *distribución multinomial*, una distribución estadística multivariante para recuentos. En el apéndice teórico (A) vemos este tema con más profundidad.

RESUMEN: Representación gráfica de distancias ji-cuadrado

1. Podemos visualizar las distancias χ^2 entre perfiles, en el espacio físico habitual (euclídeo) transformando los perfiles antes de representarlos gráficamente. Esta transformación consiste en dividir cada elemento del perfil por la raíz cuadrada del correspondiente elemento del perfil medio.

2. Otra posibilidad para visualizar las distancias χ^2 entre perfiles consiste, en vez de transformar los elementos del perfil antes de representarlos, en estirar los

ejes de manera que, en cada eje, la unidad tenga una longitud inversamente proporcional a la raíz cuadrada del correspondiente elemento del perfil medio.

3. La distancia χ^2 es un caso especial de distancia euclídea ponderada en la que los pesos son los inversos de los correspondientes elementos del perfil medio.
4. Cuando representamos gráficamente los perfiles fila, podemos ver el redimensionamiento de las coordenadas (o la extensión de los ejes) como una estandarización de las columnas de la tabla, que hace que las comparaciones entre los perfiles fila sean más equitativas.
5. Las distancias χ^2 cumplen el *principio de equivalencia distribucional*, que garantiza la estabilidad de las distancias entre las filas, cuando dividimos las columnas en componentes similares, o cuando unimos columnas similares.