La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 6

Reducción de la dimensionalidad

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

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

www.fbbva.es

CAPÍTULO 6

Reducción de la dimensionalidad

Hasta ahora hemos trabajado con conjuntos pequeños de datos (imágenes 2.1 y 3.1). Estos datos tienen pocas dimensiones y los podemos visualizar de forma exacta. Las tablas con tres columnas conllevan perfiles tridimensionales que, en realidad, y como vimos en el capítulo 2, son bidimensionales. Las podemos representar en un sistema de coordenadas triangular situado en un plano. Sin embargo, en la mayoría de aplicaciones del AC, las tablas de interés tienen muchas más filas y columnas y, por tanto, los perfiles se sitúan en un espacio de mayor dimensionalidad. Dado que no podemos ni observar ni imaginar fácilmente puntos en un espacio de más de tres dimensiones, es necesario reducir la dimensionalidad de los puntos. La reducción de la dimensionalidad es un aspecto analítico crucial del AC, por lo que llevarlo a cabo implica una cierta pérdida de información. Debemos restringir en lo posible esta pérdida y así conservar la máxima información.

Contenido

Conjunto de datos 3: Encuesta Nacional de Salud 66
Comparación de los perfiles de los grupos de edad (filas) 66
Identificación de subespacios de baja dimensionalidad 67
Proyección de los perfiles en subespacios 67
Determinación de la calidad de la representación 68
Una aproximación a la distancia entre los perfiles 68
Representación de las proyecciones de los vértices 69
Interpretación conjunta de perfiles y vértices 70
Definición de proximidad de los puntos a un subespacio 70
Definición formal del criterio de proximidad en el AC 71
Descomposición en valores singulares (DVS) 71
Hallar el subespacio óptimo no es una regresión 72
RESUMEN: Reducción de la dimensionalidad 72

Conjunto de datos 3: Encuesta Nacional de Salud

En la imagen 6.1 podemos ver un ejemplo de tabla multidimensional. Se trata de una tabla de contingencia obtenida a partir de la base de datos de la Encuesta Nacional de Salud de España de 1997. Una de las preguntas de esta encuesta trataba sobre la autopercepción de la salud de los encuestados, quienes la podían considerar *muy buena*, *buena*, *regular*, *mala* o *muy mala*. La mencionada tabla cruza estas respuestas de los encuestados con sus grupos de edad. En el momento de la encuesta, la tabla de contingencia, que contiene datos de 6371 encuestados, incluía siete grupos de edad (las filas de la imagen 6.1) y cinco categorías de salud (las columnas), proporcionando una instantánea de cómo veían los españoles su salud. Pero, ¿cómo cambia esta percepción de la salud con la edad? Utilizando el AC podremos interpretar de forma rápida la relación entre la edad y la autopercepción de la salud.

Comparación de los perfiles de los grupos de edad (filas)

Supongamos por el momento que estamos interesados en los perfiles de los grupos de edad (perfiles fila) con relación a las categorías de salud. En la tabla de la imagen 6.2, hemos expresado los perfiles fila como porcentajes. La última fila es el perfil fila medio, o el perfil fila resultante de considerar conjuntamente todos los grupos de edad de la muestra, es decir, sin distinguir entre grupos de edad. Así, por ejemplo, podemos ver que de los 6371 encuestados de la muestra, el 12,8% se ven a sí mismos con *muy buena* salud, el 55,6% con *buena* salud, etc. Fijándonos en grupos de edad específicos, vemos que hay diferencias esperables; por ejemplo, el grupo de edad más joven tiene porcentajes más altos de estas categorías (el 19,9% *muy buena* y el 64,5% *buena*) que el grupo de mayor edad, que tiene porcentajes más bajos (el 5,1% y el 34,4%, respectivamente). Examinada con detalle esta tabla pronto llegamos a la conclusión de que la autopercepción de la salud empeora con la edad, lo que no constituye una sorpresa. Sin embargo, solamente con los valores numéricos, no es fácil que nos demos cuenta de la intensidad con la que ocurren estos cambios, o de entre qué grupos de edad son mayores (o menores) los cambios en la autopercepción de la salud.

Imagen 6.1: *Cruce del grupo de edad con la autopercepción de la salud*

Grupo de edad	*Muy buena*	*Buena*	*Regular*	*Mala*	*Muy mala*	*Suma*
16–24	243	789	167	18	6	*1223*
25–34	220	809	164	35	6	*1234*
35–44	147	658	181	41	8	*1035*
45–54	90	469	236	50	16	*861*
55–64	53	414	306	106	30	*909*
65–74	44	267	284	98	20	*713*
75+	20	136	157	66	17	*396*
Suma	*817*	*3542*	*1495*	*414*	*103*	*6371*

Fuente de datos: Encuesta Nacional de Salud de España, 1997.

Identificación de subespacios de baja dimensionalidad

El AC nos permite visualizar los grupos de edad y nos proporciona más agudeza en el análisis de los datos. En este ejemplo, no podemos visualizar de forma exacta los perfiles de los grupos de edad porque los perfiles son puntos que se sitúan en un espacio de cinco dimensiones. En realidad, como vimos en los anteriores ejemplos tridimensionales, al tener los perfiles de los grupos de edad cinco elementos y ser su suma igual a 1, éstos se sitúan en un espacio de una dimensión menos. Sin embargo, incluso la visualización directa de un espacio de cuatro dimensiones es imposible. Por tanto, sería interesante poder visualizar los perfiles aunque fuera de forma aproximada en un espacio de pocas dimensiones. Así pues, dado que no podemos visualizar el espacio de cuatro dimensiones, podríamos visualizar los perfiles de forma aproximada en un subespacio de una, dos o tres dimensiones. Precisamente ésta es la esencia del AC: la identificación de subespacios de pocas dimensiones que contengan los perfiles, aunque sea de forma aproximada. También podríamos decir que el AC identifica dimensiones para las cuales existe muy poca dispersión de los perfiles, y que elimina las direcciones de dispersión que aportan poca información. Reduciendo la dimensionalidad de la nube de puntos visualizaremos más fácilmente las posiciones relativas de los perfiles.

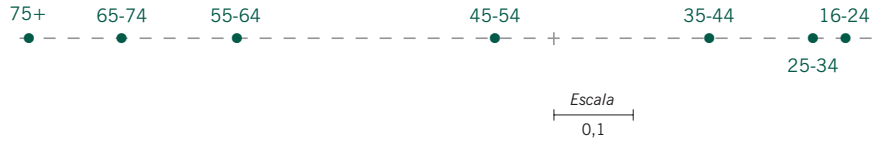
Proyección de los perfiles en subespacios

En este ejemplo, los perfiles se sitúan, en realidad, muy cerca de una recta. Es decir, podemos imaginar los perfiles formando una nube de puntos alargada en forma de cigarro situado en un espacio de perfiles de cuatro dimensiones. Si identificamos la recta «más próxima» a la nube de puntos (pronto definiremos cómo medir la proximidad), podemos dejar caer *(proyectar)* los puntos perpendicularmente sobre esta recta, sacarla del espacio multidimensional y representar las proyecciones de izquierda a derecha de forma que su interpretación sea mucho más fácil. En el mapa de la imagen 6.3 mostramos esta representación unidimensional de los perfiles de los grupos de edad. Podemos ver que, a pesar de que el método desconoce el orden natural de las categorías, éstas se sitúan de forma natural de la de mayor edad (a la izquierda) a la de menor edad (a la derecha). En

Grupo de edad	*Muy buena*	*Buena*	*Regular*	*Mala*	*Muy mala*	*Suma*
16–24	19,9	64,5	13,7	1,5	0,5	*100,0*
25–34	17,8	65,6	13,3	2,8	0,5	*100,0*
35–44	14,2	63,6	17,5	4,0	0,8	*100,0*
45–54	10,5	54,5	27,4	5,8	1,9	*100,0*
55–64	5,8	45,5	33,7	11,7	3,3	*100,0*
65–74	6,2	37,4	39,8	13,7	2,8	*100,0*
75+	5,1	34,3	39,6	16,7	4,3	*100,0*
Media	*12,8*	*55,6*	*23,5*	*6,5*	*1,6*	*100,0*

Imagen 6.2:
Perfiles de los grupos de edad, con relación a las categorías de salud, expresados como porcentajes

Imagen 6.3:
Mapa unidimensional óptimo de los perfiles de los grupos de edad



esta representación podemos ver fácilmente que las diferencias menores se hallan entre los grupos de edad más jóvenes y que las diferencias mayores se hallan entre los grupos de mediana edad.

Determinación de la calidad de la representación

Dado que las proyecciones de los perfiles en subespacios de pocas dimensiones no son sus verdaderas posiciones, deberíamos conocer cuál es la magnitud de la discrepancia entre las posiciones exactas y las aproximadas. Para hacerlo utilizaremos la inercia total de los perfiles como una medida de la variabilidad total; es decir, como una medida de la dispersión geométrica de los puntos en sus verdaderas posiciones tetradimensionales. Expresaremos tanto la calidad de la representación, como su contrapartida, la pérdida de calidad, o error de representación, como porcentajes de la inercia total; por tanto, su suma debe ser 100%. Cuanto menor sea la pérdida de inercia, mayor será la calidad, y cuanto mayor sea su pérdida, menor la calidad. En este ejemplo, la pérdida de inercia que se produce al proyectar los puntos sobre la recta del mapa de la imagen 6.3 es sólo del 2,7%. Así, la calidad de la aproximación unidimensional de los perfiles es del 97,3%. Se trata de un resultado muy favorable: empezamos con una tabla de contingencia de 7×5 con una dimensionalidad inherente de 4, y —eliminadas tres dimensiones sacrificando solamente el 2,7% de la dispersión de puntos— el 97,3% restante corresponde a su dispersión en una sola dimensión. Podemos interpretar este porcentaje exactamente igual que, en regresión, explicamos «el porcentaje de varianza explicada». En la imagen 6.3, la dimensión que muestra las proyecciones de los siete perfiles, explica el 97,3% de la inercia de los verdaderos perfiles (el 97,3% de la inercia total de la tabla de la imagen 6.1).

Una aproximación a la distancia entre los perfiles

En el mapa de la imagen 6.3, las distancias entre las proyecciones de los perfiles fila son aproximaciones de las verdaderas distancias χ^2 en el espacio tetradimensional completo. Podemos comparar las distancias χ^2 exactas, calculadas a partir de los datos de la tabla de la imagen 6.2, con las distancias que representamos gráficamente en el mapa de la imagen 6.3. En la figura de la imagen 6.4, ilustramos gráficamente esta comparación —dado que tenemos 7 puntos, existen $\frac{1}{2} \times 7 \times 6 = 21$ distancias posibles entre estos puntos—. Podemos ver que la concordancia es excelente, lo que era de esperar debido a que, al reducir los perfiles a una sola dimensión, en términos relativos, la pérdida de precisión es pequeña: el 2,7%. En la figura de la imagen 6.4, se aprecia que las distancias observadas son siempre menores o iguales que las verdaderas distancias (decimos entonces que las distancias

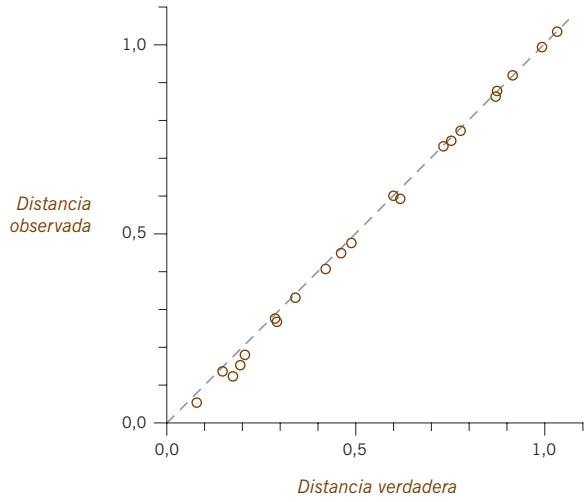


Imagen 6.4:
Distancias observadas entre todos los pares de puntos del mapa de la imagen 6.3, representadas con relación a las correspondientes distancias χ^2 verdaderas entre los perfiles fila del mapa de la imagen 6.3

se han aproximado «desde abajo»). Es decir, el cuadrado de la verdadera distancia es la suma de una serie de componentes al cuadrado —una por cada dimensión del espacio de perfiles—, mientras que el cuadrado de la distancia observada es la suma de un número reducido de estas componentes —en este ejemplo unidimensional, una sola componente—. En la figura de la imagen 6.4, la parte «no explicada» de la distancia aparece como desviaciones de los puntos con relación a la bisectriz.

Representación de las proyecciones de los vértices

En el espacio de perfiles de los siete grupos de edad existen cinco vértices que representan las cinco categorías de la salud. Recordemos, una vez más, que cada uno de estos vértices representa un perfil ficticio totalmente concentrado en una sola categoría de la salud; por ejemplo, el vértice [1 0 0 0 0] representa un grupo con una autopercepción de la salud *muy buena*. Igual que los perfiles, los vértices también los podemos proyectar sobre la dimensión que representamos gráficamente en el mapa de la imagen 6.3, que como vimos es la dimensión que mejor explica los perfiles de los grupos de edad (imagen 6.5). Fijémonos en el cambio de escala en comparación con el mapa de la imagen 6.3 —en ambos mapas, los perfiles de los grupos de edad se hallan exactamente en las mismas posiciones—. No obstante, la dispersión de los vértices es mucho mayor, lo que se puede explicar por el hecho de que éstos representan los perfiles más extremos.

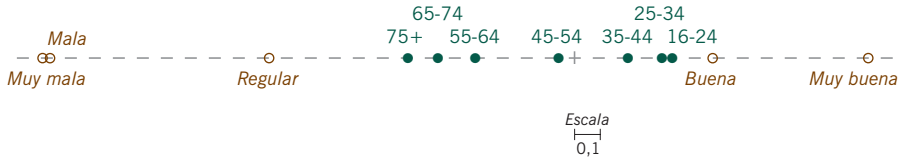


Imagen 6.5:
Mapa óptimo del mapa de la imagen 6.3, que muestra las proyecciones de los vértices de las categorías de salud

Interpretación conjunta de perfiles y vértices

En la representación conjunta de perfiles y vértices del mapa de la imagen 6.5, las categorías de salud también se hallan dispuestas en su orden natural; en el extremo izquierdo hallamos la categoría *muy mala* salud y en el extremo derecho la categoría *muy buena* salud. Las posiciones de estos puntos de referencia en la dimensión nos proporcionan la llave para la interpretación de la asociación entre las filas (grupos de edad) y las columnas (categorías de salud). Así, vemos que el grupo de edad más joven está alejado, pero cerca de la buena salud; en cambio, el grupo de más edad se halla hacia la mala salud. El origen (o punto cero, que hemos indicado por + en los mapas de las imágenes 6.3 y 6.5) representa el perfil medio. Así pues, deducimos que, con relación a la media, los grupos de edad de hasta 44 años se hallan en el lado «bueno», mientras que los de más de 45 años se hallan en el lado «malo». El hecho de que el vértice *muy mala* se halle tan lejos de los perfiles de los grupos de edad indica que ningún grupo de edad se halla cerca de este extremo (en la tabla de la imagen 6.2 podemos ver porcentajes del 0,5 al 4,3%, para esta categoría, cuya media es sólo del 1,6%, el valor medio que se halla en el origen). La categoría *mala* se halla casi en la misma posición, pero con porcentajes del 1,5 al 16,7% y una media del 6,5% en el origen (en los capítulos 8 y 13, veremos más detalles sobre la interpretación conjunta de filas y columnas). En esta proyección unidimensional, la relación entre los perfiles fila y los vértices columna es la misma que describimos en el espacio triangular de los capítulos 2 y 3, es decir, el perfil de cada grupo de edad se halla en la media ponderada de los vértices de las categorías de salud, ponderados con los elementos del perfil. El grupo de edad más joven (de 16-24 años) es el que se encuentra más a la derecha porque su perfil es el que tiene mayores valores para las categorías situadas a la derecha, o sea, las categorías *muy buena* y *buena*.

Definición de proximidad de los puntos a un subespacio

El ejemplo que hemos visto es más simple de lo habitual ya que una sola dimensión describe adecuadamente los datos. En realidad, en la mayoría de casos necesitamos al menos un plano de dos dimensiones para «aproximarnos» o «ajustarnos» al espacio multidimensional de la nube de perfiles. Sobre dicho plano proyectaremos los perfiles y los vértices del espacio de perfiles. En la imagen 6.6 hemos representado gráficamente algunos perfiles en un espacio multidimensional imaginario, así como sus proyecciones sobre un plano que corta dicho espacio. Tanto si proyectamos los perfiles sobre la recta que mejor se ajusta (un subespacio unidimensional), como sobre un plano (un subespacio bidimensional) o incluso sobre un subespacio de mayor dimensionalidad, tenemos que definir lo que entendemos por «proximidad» de los puntos a los mencionados subespacios. Imaginemos una recta en un espacio multidimensional de perfiles, ¿cómo podemos calcular la distancia más corta de los puntos a la mencionada recta? (en este contexto entendemos por distancia, la distancia χ^2). Una posibilidad intuitiva obvia, para llegar a una sola medida de proximidad de todos los puntos a la recta, podría ser la suma de las distancias de todos los perfiles a la mencionada recta

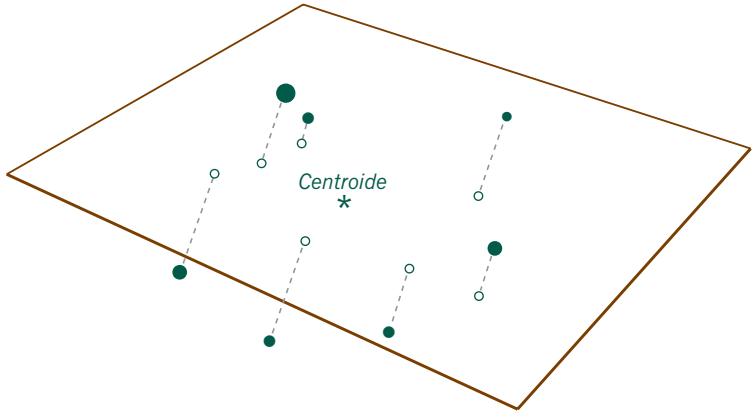


Imagen 6.6:
Perfiles en un espacio multidimensional y un plano que corta dicho espacio; el plano que mejor se ajuste en el sentido mínimo-cuadrático debe pasar por el centroide de los puntos (Los perfiles tienen masas diferentes tal como indican los tamaños de los puntos.)

imaginaria. Tendríamos que hallar la recta para la cual la suma de estas distancias sea menor. En principio no hay nada que nos impida hacer exactamente esto, sin embargo resulta matemáticamente bastante complicado minimizar esta suma de distancias. Igual que en muchas otras áreas de la estadística, el problema se simplifica mucho si definimos un criterio basado en sumas de distancias al cuadrado y no uno basado directamente en la suma de distancias. De esta forma llegamos al llamado problema de la *suma mínimo-cuadrática*. Sin embargo, en nuestro caso tenemos también una masa asociada a cada perfil, masa que cuantifica la importancia del perfil en el análisis. El criterio que utilizaremos en el análisis de correspondencias será, pues, una suma ponderada de distancias al cuadrado.

Definición formal del criterio de proximidad en el AC

Supongamos que, en un espacio multidimensional, tengamos I perfiles y que S sea un candidato a subespacio de pocas dimensiones en el espacio original. Simbolicemos como $d_i(S)$ la distancia χ^2 entre el i-ésimo perfil de masa m_i y S. Calcularemos la proximidad de este perfil al subespacio como $m_i[d_i(S)]^2$, es decir, el cuadrado de la distancia ponderada con la masa. Calcularemos la proximidad de todos los perfiles a S, como la suma de estos valores:

$$\text{proximidad a } S = \sum_i m_i[d_i(S)]^2 \qquad (6.1)$$

El objetivo del AC es identificar el subespacio S que minimice el criterio anterior. Se puede demostrar que, necesariamente, el subespacio S buscado tiene que pasar por el centroide de los puntos (imagen 6.6), por lo tanto, sólo debemos considerar los subespacios que contienen el centroide.

Descomposición en valores singulares (DVS)

No es necesario que entremos en las operaciones matemáticas implicadas en la minimización anterior. Es suficiente con que indiquemos que la manera más elegante de definir la teoría del AC, así como de calcular la solución de la mini-

mización anterior, es utilizar lo que en matemáticas se llama *descomposición en valores singulares* (DVS de forma abreviada). La DVS es uno de los resultados más útiles de la teoría de matrices. En estadística, la DVS es especialmente relevante en todos los métodos de reducción de la dimensionalidad. La DVS es a las matrices rectangulares lo que la descomposición en vectores y valores propios es a las matrices cuadradas. Es decir, una manera de descomponer una matriz en sus componentes, de los más a los menos importantes. El concepto algebraico de *rango* de una matriz es equivalente a nuestro concepto geométrico de dimensión. La DVS proporciona un mecanismo directo para aproximar una matriz rectangular a otra matriz de menor rango por mínimos cuadrados. Los resultados que obtenemos de la DVS nos llevan directamente a la teoría del AC, y a todos los elementos que necesitamos (coordenadas, inercias principales, etc.). Dado que la DVS se halla implementada en muchos lenguajes informáticos, es fácil llevar a cabo la parte analítica del AC. En el apéndice de cálculo (B) se muestra lo fácil que es llevar a cabo el AC utilizando la función DVS del lenguaje de programación R.

Hallar el subespacio óptimo no es una regresión

Acabamos de describir cómo hallar subespacios de pocas dimensiones (por ejemplo, rectas y planos) por mínimos cuadrados. Parece como si fuera lo mismo que hace el análisis de la regresión, que también ajusta rectas y planos a puntos que podemos imaginar en un espacio multidimensional. Sin embargo, existe una gran diferencia entre la regresión y lo que nosotros hacemos aquí. En el análisis de la regresión, consideramos una de las variables como variable respuesta. Además, en regresión, las distancias se minimizan en la dirección del eje de esta variable respuesta. En cambio, en nuestro caso no existe ninguna variable respuesta, hacemos el ajuste minimizando distancias perpendiculares al subespacio que estamos ajustando (en la imagen 6.6 podemos ver que las proyecciones son perpendiculares al plano; son las menores distancias entre los puntos y el plano). De todas formas, como las dimensiones identificadas en el AC se pueden contemplar como variables explicativas de los datos, ajustar subespacios de pocas dimensiones a puntos, a veces, se le llama «regresión ortogonal».

RESUMEN: Reducción de la dimensionalidad

1. Los perfiles constituidos por m elementos se sitúan, exactamente, en espacios de dimensionalidad $m - 1$. Por tanto, los perfiles con más de cuatro elementos se sitúan en espacios de dimensionalidad mayor de tres, que no podemos observar directamente.
2. Si identificamos un espacio de poca dimensionalidad, preferentemente con no más de dos o tres dimensiones, que se halle cerca de los perfiles, podremos proyectar dichos perfiles sobre el mencionado subespacio y observar las posiciones de sus proyecciones como una aproximación a sus verdaderas posiciones en el espacio original de mayor dimensionalidad.

3. En el proceso de reducción de la dimensionalidad perdemos información sobre las verdaderas posiciones de los perfiles con relación al subespacio (dirección y separación del subespacio). Sin embargo, ganamos la posibilidad de ver conjuntamente todos los perfiles, lo que de otra forma sería imposible.
4. Expresamos la precisión de una representación como *porcentaje de inercia*. Por ejemplo, si el 85% de la inercia de los perfiles queda representada en el subespacio, la inercia residual, o error, que queda fuera del subespacio es del 15%.
5. También podemos proyectar los vértices, o los perfiles unitarios, sobre el subespacio óptimo. En este caso, el objetivo no es representar de forma exacta los vértices, sino utilizarlos como puntos de referencia para la interpretación de los perfiles representados.
6. El cálculo del espacio de poca dimensionalidad se basa en la determinación de la proximidad entre un conjunto de puntos y un subespacio. Calculamos dicha proximidad como la suma ponderada de los cuadrados de las distancias χ^2 entre los puntos y el subespacio, y ponderamos los puntos con sus respectivas masas.