La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del apéndice D

Glosario de términos

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

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

www.fbbva.es

Glosario de términos

En este apéndice presentamos por orden alfabético una lista de los términos más comúnmente utilizados en este libro, junto con definiciones abreviadas de cada uno de ellos. Las palabras en cursiva corresponden a los términos incluidos en el glosario.

— *agrupación de Ward:* algoritmo de agrupación jerárquica que minimiza la inercia dentro de los grupos en cada paso de agrupación. Equivalente a maximizar la inercia entre grupos.

— *análisis de correspondencias (AC):* método de representación de filas y columnas de una tabla como puntos en un mapa, con una interpretación geométrica específica de sus posiciones, que nos permite interpretar las similitudes y las diferencias entre filas y entre columnas, así como la asociación entre filas y columnas.

— *análisis de correspondencias canónico (ACC):* ampliación del AC que incluye variables explicativas externas. Restringimos la solución del AC para que las dimensiones estén relacionadas linealmente con estas variables explicativas.

— *análisis de correspondencias conjunto (ACCo):* variante del *análisis de correspondencias múltiples* para el análisis de todas las tablas de contingencia derivadas del cruzamiento de un conjunto de Q variables categóricas en el que ignoramos los cruzamientos de cada variable con ella misma.

— *análisis de correspondencias múltiples (ACM):* AC de la *matriz binaria* o de la *matriz de Burt* formadas a partir de más de dos variables categóricas.

— *análisis de correspondencias de subgrupos:* variante del AC en la que aunque analizamos parte de las filas o de las columnas de una tabla mantenemos la geometría de la tabla completa.

— *automuestreo* [ingl. *bootstrap*]*:* método para la investigación de la variabilidad de un estadístico. Consiste en generar, mediante un ordenador, un gran número de réplicas de una muestra a partir de la muestra observada.

— *automuestreo parcial* [ingl. *partial bootstrap*]*:* en AC, representación de muchas muestras replicadas obtenidas por *automuestreo*, como puntos adicionales en el mapa de la tabla original.

— *biplot:* mapa conjunto de puntos que representa las filas y las columnas de una tabla de manera que los *productos escalares* entre filas y columnas se aproximen de forma óptima a los elementos de la tabla.

— *bootstrap* [véase *automuestreo*].

— *calibración:* en *biplots*, es el proceso de configuración de una escala en un *eje de un biplot* mediante marcas y valores. En la representación de *perfiles* en AC se trata de una escala de razón o de porcentajes.

— *centroide:* punto medio ponderado.

— *cociente de contingencia:* para una *tabla de contingencia*, frecuencia observada dividida por la frecuencia esperada de acuerdo con el *modelo de independencia*.

— *codificación interactiva:* creación de una sola variable categórica a partir de todas las combinaciones de categorías de dos variables categóricas.

— *condición de identificación:* condición que debe imponerse en un problema de optimización para obtener una sola solución.

— *contribución a la inercia:* componente de la *inercia* explicada por un determinado punto en un *eje principal*. En general la expresamos en relación con la *inercia principal* correspondiente (que nos informa sobre cómo se han construido los ejes) o en relación a la inercia del punto (que nos informa sobre cómo queda explicado el punto en el eje).

— *coordenadas estándares:* coordenadas de un conjunto de puntos en un eje que cumplen que la suma ponderada de sus cuadrados es igual a 1.

— *coordenadas principales:* coordenadas de un conjunto de puntos proyectados sobre un *eje principal*, que cumple que la suma ponderada de sus cuadrados en dicho eje es igual a la *inercia principal* del eje.

— *descomposición en valores singulares (DVS):* descompomposición de una matriz similar a la descomposición en vectores y *valores propios*, pero aplicado a matrices rectangulares. Los cuadrados de los valores singulares son *valores propios* de las matrices cuadradas, y los vectores singulares de la izquierda y de la derecha son también vectores propios.

— *dimensión:* número de dimensiones geométricas inherentes de una tabla necesarias para reproducir exactamente sus elementos en un *mapa* de AC.

— *distancia euclídea:* distancia entre puntos que calculamos como la raíz cuadrada de la suma de las diferencias al cuadrado entre los correspondientes elementos de los vectores.

— *distancia euclídea ponderada:* similar a la *distancia euclídea*, pero con un factor de ponderación positivo para cada diferencia al cuadrado.

— *distancia ji-cuadrado:* distancia euclídea ponderada entre *perfiles*, en la que hemos dividido cada diferencia al cuadrado entre los elementos de los perfiles por el correspondiente elemento del perfil medio.

— *doblado:* procedimiento por el que recodificamos filas (o columnas) como pares de filas (o de columnas) con el objetivo de dibujar en un mapa los extremos, o polos, de una escala. Lo utilizamos en AC para ordenaciones, preferencias o comparaciones por pares.

— *dummy variable* [Véase *variable binaria*].

— *efecto arco:* la tendencia de los puntos en un mapa de AC a formar una curva debido a la particular geometría del AC por la que los perfiles se hallan dentro de un símplex. También conocido como «efecto herradura».

— *eje de un biplot:* una dirección marcada por un vector de un *biplot* sobre la cual podemos proyectar puntos con el objetivo de estimar los valores de la tabla que analizamos.

— *eje principal:* dirección de dispersión de puntos de un espacio multidimensional que optimiza la *inercia* o, de forma equivalente, eje que mejor se ajusta a los puntos en el sentido de distancias mínimo-cuadráticas ponderadas.

— *escala óptima:* conjunto de valores asignados a las categorías de variables categóricas, que optimizan algún criterio como, por ejemplo, la correlación máxima (con otra variable) o la discriminación máxima (entre grupos).

— *estadístico ji-cuadrado:* estadístico utilizado habitualmente para contrastar el modelo de independencia de una *tabla de contingencia*; calculado como la suma de diferencias al cuadrado entre frecuencias observadas y esperadas de acuerdo con el modelo. Dividimos cada diferencia al cuadrado por la correspondiente frecuencia esperada.

— *indicator matrix* [Véase *matriz binaria*].

— *inercia:* suma ponderada de distancias al cuadrado de un conjunto de puntos con relación a su *centroide*. En AC los puntos son *perfiles*, los pesos son *masas* de los perfiles y las distancias son *distancias ji-cuadrado*.

— *inercia principal:* la correspondiente de un *eje principal*; también llamada *valor propio*.

— *inercias principales ajustadas:* una modificación de los resultados del *análisis de correspondencias múltiples,* que proporciona una estimación más realista de la inercia explicada por el AC.

— *mapa:* representación en el espacio de puntos (perfiles fila o perfiles columnas en AC) en la que podemos interpretar distancias o productos escalares *(biplot)*.

— *mapa asimétrico:* una representación conjunta en la que hemos normalizado (escalado) de forma distinta los puntos de filas y de columnas. En general unos en *coordenadas principales* y los otros en *coordenadas estándares*. A menudo los mapas asimétricos son *biplots*.

— *masa:* suma marginal total de una fila o una columna de una tabla dividida por la suma total de la tabla. La utilizamos como pesos en AC.

— *matriz antisimétrica:* matriz cuadrada con ceros en la diagonal y que cumple la propiedad de que los elementos por encima de la diagonal tienen el mismo valor absoluto que los elementos opuestos situados por debajo de la diagonal, pero con signo opuesto.

— *matriz binaria* [ingl. *indicator matrix*]: codificación de datos multivariantes categóricos en forma de *variables binarias*.

— *matriz de Burt:* un tipo de *matriz compuesta*, que consiste en todas los cruzamientos de Q variables categóricas, incluyendo los cruzamientos de las variables con ellas mismas.

— *modelo de independencia* (o «hipótesis de homogeneidad»): modelo para los recuentos de una *tabla de contingencia*, que supone que hemos muestreado la filas (o las columnas) al azar de la misma población. Es decir, que las frecuencias relativas esperadas (proporciones) de filas, o de columnas, son las mismas.

— *observación atípica:* punto situado en la periferia de una representación gráfica que se halla bien separado de la dispersión general de puntos.

— *perfil:* valores de una fila o columna de una tabla de contingencia dividida por su total. Los puntos que visualizamos en AC son perfiles.

— *producto escalar:* de dos vectores definidos por dos puntos. Es el producto de sus longitudes multiplicado por el coseno del ángulo entre ellos. Directamente proporcional a la proyección de uno de los puntos sobre el vector definido por el otro punto.

— *pruebas de permutaciones:* obtención de permutaciones de datos; todos las posibles o una gran muestra aleatoria de ellas, con el objetivo de obtener la distribución de un determinado estadístico de contraste suponiendo cierta la hipótesis nula y así poder estimar el valor p asociado del estadístico.

— *punto adicional* o *punto pasivo* o *punto suplementario:* punto del mapa (*perfil* en AC) con masa cero. Es decir, punto que representamos en el mapa, pero que no interviene en su configuración.

— *razón de escalas:* en una representación gráfica, el cociente entre una unidad de longitud en el eje horizontal y una unidad de longitud en el eje vertical. En un mapa de AC debe ser 1.

— *relación de transición:* relación entre las coordenadas de filas y de columnas de un mapa.

— *símplex:* en dos dimensiones, un triángulo, en tres dimensiones un tetraedro, y la generalización de estas figuras geométricas en más dimensiones. En AC, los *perfiles* de *J*-dimensiones se hallan dentro de un simplex definido por *J vértices* en un espacio de $(J - 1)$ dimensiones.

— *tabla concatenada:* tabla formada concatenando horizontal o verticalmente o en ambas direcciones tablas de contingencia, que hemos obtenido clasificando los mismos individuos cruzando variables categóricas.

— *tabla de contingencia:* clasificación de un conjunto de individuos de acuerdo con el cruce de dos variables categóricas. Por tanto, el total de la tabla es el número total de individuos.

— *valor propio:* valor inherente de una matriz cuadrada. Forma parte de la descomposición de una matriz como el producto de matrices más simples. En general, las matrices cuadradas tienen tantos valores propios y vectores propios asociados como su rango. En AC, valor propio es sinónimo de *inercia principal*.

— *variable binaria* [ingl. *dummy variable*]: variable que sólo toma los valores 0 o 1. Las utilizamos en una variante del *análisis de correspondencias múltiples* para codificar datos multivariantes categóricos.

— *vértice:* perfil unitario, es decir, perfil con todos sus elementos iguales a cero excepto uno que toma el valor 1.