La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 13

Biplots en análisis de correspondencias

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

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

www.fbbva.es

CAPÍTULO 13

Biplots en análisis de correspondencias

Hasta ahora hemos visto dos posibilidades de representación de filas y de columnas en AC. En los mapas asimétricos, por ejemplo en el análisis de filas, expresamos las filas en coordenadas principales y las columnas en coordenadas estándares. Las distancias χ^2 entre los perfiles fila del mapa son bastante exactas, y utilizamos los vértices de las columnas como referencias para la interpretación del mapa. En cambio, en los mapas simétricos, en los que que expresamos tanto las filas como las columnas en coordenadas principales, las distancias χ^2 entre los perfiles fila y los perfiles columna son sólo aproximadas. Los *biplots* son otra posibilidad para la representación conjunta de filas y de columnas que se basa en el producto escalar entre vectores fila y vectores columna —por tanto, depende más de las longitudes y de los ángulos formados por los vectores que de las distancias entre puntos—. En los biplots sólo representamos en coordendas principales las filas o las columnas. En este sentido, pues, los mapas asimétricos son biplots ya que en estos últimos también expresamos sólo las filas o las columas en coordenadas principales. La diferencia entre ambas representaciones radica en que, en los mapas asimétricos, siempre representamos los puntos de referencia en coordenadas estándares, mientras que en los biplots tenemos más posibilidades de elección.

Contenido

Definición de producto escalar

En geometría euclídea, el *producto escalar* entre dos vectores **x** e **y**, de coordenadas x_1, x_2, ... e y_1, y_2, ... es la suma de los productos de sus respectivos elementes $x_k y_k$, simbolizados como $\mathbf{x}^\top\mathbf{y} = \sum_k x_k y_k$ ($^\top$ indica la transposición de un vector o una matriz). Geométricamente, el producto escalar es igual al producto de las longitudes de dos vectores, multiplicado por el coseno del ángulo formado entre ellos.

$$\mathbf{x}^\top\mathbf{y} = \sum_k x_k y_k = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cdot \cos\theta \tag{13.1}$$

donde $\|\mathbf{x}\|$ simboliza la longitud del vector **x**, es decir, la distancia entre el punto **x** y el origen. En la imagen 13.1 hemos representado gráficamente este resultado en un espacio bidimensional (en un espacio multidimensional, siempre podemos representar dos vectores en un plano).

Relación entre el producto escalar y la proyección

Otro resultado geométrico bien conocido es que la proyección perpendicular de un vector **x** sobre una dirección definida por otro vector **y** tiene una longitud igual a la longitud de **x**, multiplicada por el coseno del ángulo entre **x** e **y**. Es decir el producto $\|\mathbf{x}\| \cdot \cos\theta$ es parte de la definición que vimos en (13.1). Por tanto, podemos contemplar el producto escalar de **x** e **y**, como la proyección de la longitud de **x** sobre **y**, multiplicada por la longitud de **y** (imagen 13.1). O de forma equivalente, como la proyección de la longitud de **y** sobre **x**, multiplicada por la longitud de **x**. Si la longitud de uno de los vectores es 1, por ejemplo el **y**, entonces el producto escalar es simplemente la longitud de la proyección del vector **x** sobre **y**.

Dado un determinado vector de referencia, los productos escalares son proporcionales a las proyecciones

Si consideramos que **y** es un determinado vector de referencia, y luego consideramos varios vectores $\mathbf{x}_1$, $\mathbf{x}_2$,... proyectados sobre **y**, entonces:

- Los productos escalares $\mathbf{x}_1^\top\mathbf{y}$, $\mathbf{x}_2^\top\mathbf{y}$,... tienen magnitudes proporcionales a las proyecciones, ya que son las proyecciones multiplicadas por la longitud del vector **y**.
- El signo del producto escalar es positivo si el vector forma un ángulo agudo (< 90º) con **y**, y es negativo si forma un ángulo obtuso (> 90º).

Estas propiedades son la base para la interpretación de los biplots en AC.

Un biplot simple y exacto

Un *biplot* es una representación en pocas dimensiones de una matriz rectangular de datos, en la que representamos las filas y las columnas como puntos con una interpretación específica en términos de productos escalares. La idea es recuperar, de forma aproximada, los elementos de la matriz a partir de estos productos escalares. Como ejemplo inicial de biplot que recupera exactamente los datos, consideremos la tabla **T** de 5 × 4:

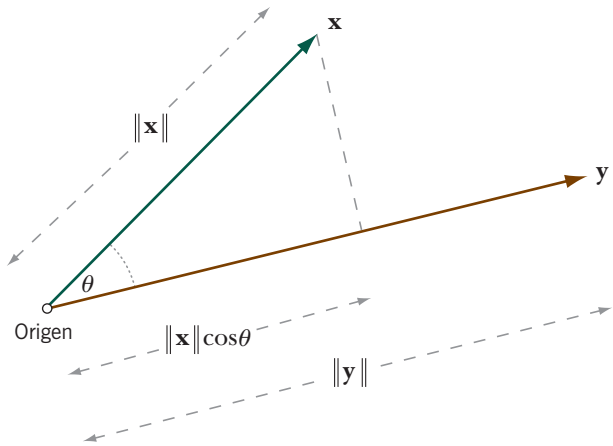
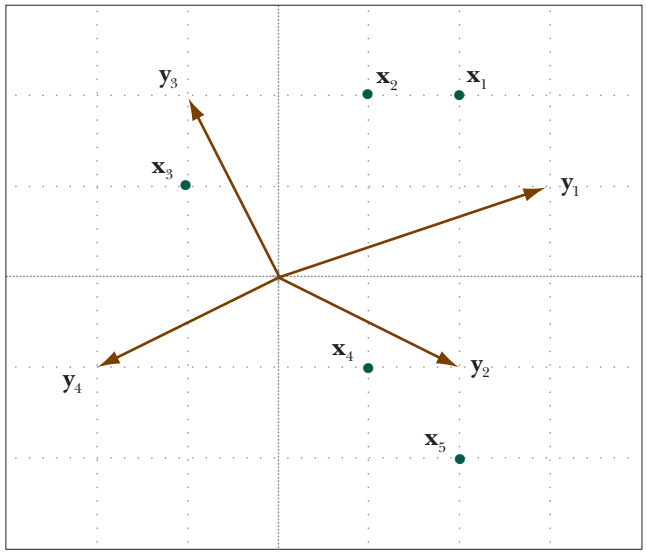


Imagen 13.1:
Ejemplo de dos puntos $\mathbf{x}$ *e* $\mathbf{y}$ *cuyos vectores forman un ángulo* θ *con relación al origen (en general, el centroide de un conjunto de puntos). El producto escalar entre los puntos es la longitud de la proyección de* $\mathbf{x}$ *sobre* $\mathbf{y}$*, multiplicada por la longitud de* $\mathbf{y}$

$$\mathbf{T} = \begin{bmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{bmatrix} \tag{13.2}$$

y luego comparémosla con el mapa de la imagen 13.2, que también proporciona las coordenadas de cada punto. (En álgebra de matrices, habitualmente representamos los vectores por columnas, de manera que una fila será un vector transpuesto.) Por ejemplo, el producto escalar entre $\mathbf{x}_1$ e $\mathbf{y}_1$, es igual a $2 \times 3 + 2 \times 1 = 8$, el



$\mathbf{x}_1 = [\ 2 \quad 2]^\top$
$\mathbf{x}_2 = [\ 1 \quad 2]^\top$
$\mathbf{x}_3 = [-1 \quad 1]^\top$
$\mathbf{x}_4 = [\ 1 \quad -1]^\top$
$\mathbf{x}_5 = [\ 2 \quad -2]^\top$

$\mathbf{y}_1 = [\ 3 \quad 1]^\top$
$\mathbf{y}_2 = [\ 2 \quad -1]^\top$
$\mathbf{y}_3 = [-1 \quad 2]^\top$
$\mathbf{y}_4 = [-2 \quad -1]^\top$

Imagen 13.2:
Mapa de cinco puntos fila $\mathbf{x}_i$ *y cuatro puntos columna* $\mathbf{y}_j$*. El producto escalar entre el punto correspondiente a la* i*-ésima fila y el correspondiente a la* j*-ésima columna proporciona el valor del* ij*-ésimo valor de la tabla (13.2). Hemos representado los puntos columna como vectores para facilitar la interpretación de los productos escalares como proyecciones de los puntos sobre los vectores, multiplicada por las longitudes de los respectivos vectores*

primer elemento de **T**. También podemos calcular el producto escalar, aunque de forma más laboriosa, como vimos en la ecuación (13.1). Es decir, en primer lugar, calculamos los ángulos formados por $\mathbf{x}_1$ e $\mathbf{y}_1$ con el eje horizontal. Por trigonometría básica: arctan(2/2) = 45º y arctan(1/3) = 18,43º, respectivamente. Por tanto, el ángulo entre $\mathbf{x}_1$ e $\mathbf{y}_1$ es igual a 45 – 18,43 = 26,57º. Finalmente utilizando la ecuación (13.1), vemos que el producto escalar es:

$$\mathbf{x}_1^\top \mathbf{y}_1 = \|\mathbf{x}_1\| \cdot \|\mathbf{y}_1\| \cdot \cos\theta = \sqrt{8} \cdot \sqrt{10} \cdot \cos(26{,}57°) = 8{,}00$$

con lo que este resultado coincide con el cálculo anterior. La proyección de $\mathbf{x}_1$ sobre $\mathbf{y}_1$ es igual a $\sqrt{8}\cos(26{,}57°) = 2{,}530$, y la longitud de $\mathbf{y}_1$ es $\sqrt{10} = 3{,}162$; por tanto, el producto escalar es 8,00.

Algunas características especiales de los biplots

En la palabra biplot, el prefijo «bi» indica que en el mapa representamos conjuntamente filas y columnas, y no indica que el mapa sea bidimensional, ya que los biplots pueden tener cualquier dimensionalidad. De todas formas, lo más frecuente es que los representemos en un plano. Los puntos de la imagen 13.2 ilustran algunas propiedades más de los biplots:

- $\mathbf{x}_2$ e $\mathbf{y}_2$ forman un ángulo recto, por tanto $\mathbf{x}_2$ se proyecta sobre el origen y, en consecuencia, en la tabla **T**, el valor de t_{22} es 0;
- $\mathbf{x}_2$ y $\mathbf{x}_3$ tienen la misma proyección sobre $\mathbf{y}_3$; por tanto, los valores t_{23} y t_{33} son iguales (3 en este caso);
- $\mathbf{x}_5$ es opuesto a $\mathbf{x}_3$ con respecto al origen y se halla dos veces más lejos, es decir $\mathbf{x}_5 = -2\mathbf{x}_3$; por tanto la quinta fila de la tabla **T** es igual a dos veces la tercera fila cambiada de signo;
- $\mathbf{x}_3$, $\mathbf{x}_4$ y $\mathbf{x}_5$ se hallan sobre una recta imaginaria (puede ser cualquier recta, no tiene porqué pasar por el origen), por lo que tienen una relación lineal, concretamente $\mathbf{x}_4 = \frac{1}{3}\mathbf{x}_3 + \frac{2}{3}\mathbf{x}_5$; esta expresión, tipo media ponderada, se transfiere a las correspondientes filas de **T**, por ejemplo, $t_{41} = \frac{1}{3}t_{31} + \frac{2}{3}t_{51} = \frac{1}{3}(-2) + \frac{2}{3}(4) = 2$.

Rango y dimensionalidad

Dado que podemos reconstruir perfectamente la tabla a partir de un biplot bidimensional, matemáticamente diríamos que el *rango* de la matriz **T** (13.2) es igual a 2. En nuestra aproximación geométrica, rango es equivalente a dimensión.

Los biplots proporcionan aproximaciones óptimas a los datos

En general, las matrices de datos tienen una alta dimensionalidad, por lo que no las podemos reconstruir exactamente a partir de un biplot de baja dimensionalidad. La idea que hay detrás del biplot es hallar una serie de puntos fila $\mathbf{x}_i$ y puntos columna $\mathbf{y}_j$, tales que los productos escalares entre los correspondientes vectores fila y los vectores columna se aproximen tan exactamente como sea posible a los respectivos elementos de la matriz de datos. Por tanto, podemos decir que

un biplot modeliza los datos t_{ij} como la suma de un producto escalar, en algún subespacio de baja dimensionalidad (por ejemplo de K^* dimensiones) y un término de «error» residual:

$$\begin{aligned} t_{ij} &= \mathbf{x}_i^{\mathsf{T}} \mathbf{y}_j + e_{ij} \\ &= \sum_{k=1}^{K^*} x_{ik} y_{jk} + e_{ij} \end{aligned} \tag{13.3}$$

El «modelo» de cálculo de los biplots se ajusta minimizando los errores —en general, por mínimos cuadrados—, cuya expresión minimizada es la siguiente $\sum_i \sum_j e_{ij}^2$. El modelo del biplot tiene el aspecto de una regresión lineal múltiple, salvo por el hecho de que hay dos conjuntos de parámetros desconocidos, las coordenadas de las filas $\{x_{ik}\}$ y las coordenadas de las columnas $\{y_{jk}\}$. En el capítulo 14 veremos con más profundidad esta relación con el análisis de la regresión.

El modelo del AC

Para comprender el vínculo entre el AC y el biplot, tenemos que introducir una fórmula matemática que exprese los datos originales n_{ij} en términos de las masas de las filas, las masas de las columnas y las coordenadas. Una versión de esta fórmula, que llamamos *fórmula de reconstitución* (véase el apéndice teórico, A), es:

$$p_{ij} = r_i c_j \left(1 + \sum_{k=1}^{K} \sqrt{\lambda_k}\, \phi_{ik} \gamma_{jk}\right) \tag{13.4}$$

donde

- p_{ij} son las proporciones relativas n_{ij}/n, siendo n la suma total $\sum_i \sum_j n_{ij}$;
- r_i y c_j son las masas de las filas y de las columnas, respectivamente;
- λ_k es la k-ésima inercia principal;
- ϕ_{ik} y γ_{jk} son las coordenadas estándares de las filas y de las columnas, respectivamente.

En el sumatorio de la ecuación (13.4), el número de sumandos es igual a K, la dimensión de la matriz de datos, que vimos que era igual al menor del número de filas menos uno y del número de columnas menos uno. La representación gráfica del AC en K^* dimensiones en el mapa (en general, K^* es igual a 2), es óptima en el sentido de que, a partir de $K^* + 1$, minimizamos los términos de la ecuación (13.4): estos términos constituyen el «error» o residuo.

Biplot de cocientes de contingencia

Podemos reacomodar ligeramente la ecuación (13.4), de manera que el término de la derecha aparezca como un producto escalar en un espacio de dimensión K^*, más un término de error, como en la ecuación (13.3):

$$\frac{p_{ij}}{r_i c_j} - 1 = \sum_{k=1}^{K^*} f_{ik}\gamma_{jk} + e_{ij} \tag{13.5}$$

siendo $f_{ik} = \sqrt{\lambda_k}\phi_{ik}$, la coordenada principal de la i-ésima fila en el k-ésimo eje. Esta ecuación demuestra que el mapa asimétrico de filas, en el que expresamos las filas en coordenadas principales f_{ik} y las columnas en coordenadas estándares γ_{jk}, es un biplot aproximado de los valores situados a la izquierda de la ecuación (13.5). Llamamos *cocientes de contingencia* a los cocientes entre las proporciones observadas y las proporciones esperadas, $p_{ij}/(r_i c_j)$, y cuanto más cerca se hallen estos cocientes a 1, más cerca de hallan los datos al modelo de independencia (o supuesto de homogeneidad).

El biplot desde el punto de vista de los perfiles fila

También podemos expresar la ecuación (13.5) como:

$$\left(\frac{p_{ij}}{r_i} - c_j\right) / c_j = \sum_{k=1}^{K^*} f_{ik}\gamma_{jk} + e_{ij} \tag{13.6}$$

es decir, el mapa asimétrico de filas es un biplot aproximado que nos permite recuperar las desviaciones de los perfiles fila de su media con relación a su media (en la imagen 10.2 podemos ver una representación gráfica de un mapa asimétrico de filas). Como ya se ha comentado, un inconveniente de los mapas asimétricos es que, cuando la inercia es pequeña, el mapa puede ser poco satisfactorio, ya que los perfiles de las filas (las coordenadas f_{ik}) se concentran en un espacio pequeño en el centro del mapa, mientras que los vértices de las columnas (coordenadas γ_{jk}) se hallan muy lejos.

El biplot estándar del AC

En los biplots tienen especial interés las direcciones de los vértices ya que éstas definen los ejes sobre los que podemos proyectar los perfiles fila. Se han propuesto diferentes modificaciones del biplot que acabamos de ver para redefinir las longitudes de los vectores definidos por los vértices. La opción más oportuna consiste en reexpresar (13.6) de la siguiente manera:

$$\left(\frac{p_{ij}}{r_i} - c_j\right) / c_j^{1/2} = \sum_{k=1}^{K^*} f_{ik}\left(c_j^{1/2}\gamma_{jk}\right) + e_{ij} \tag{13.7}$$

(fijémonos en que los residuos e_{ij} en (13.7) tienen una definición y estandarización distinta a la que posee en (13.6), aunque estemos usando la misma notación en ambos casos). Efectivamente, en el lado izquierdo hemos estandarizado las desviaciones de los perfiles fila de su media, de manera que hemos pasado el factor $c_j^{1/2}$ remanente a la derecha multiplicando. Al multiplicar los vértices de las columnas por las correspondientes raíces cuadradas de las masas, éstos se acercan al origen. De esta forma, las categorías poco frecuentes se acercarán más, justo lo que queríamos para mejorar la legibilidad del mapa asimétrico. Dado que en este

tipo de biplot representamos los valores originales estandarizados, lo llamamos *biplot estándar* del AC. En la imagen 13.3 mostramos el biplot estándar del AC correspondiente a los datos del ejemplo sobre la financiación de la investigación científica; comparemos este mapa con los mapas de las imágenes 10.2 y 10.3. En todos ellos, las posiciones de las filas son las mismas, siendo las posiciones de las columnas las que cambian (comparemos las escalas de cada mapa).

Interpretación de los biplots

En el bipot de la imagen 13.3 no podemos interpretar las distancias entre columnas, estos puntos solamente indican las direcciones de los ejes del biplot. En cambio, las proyecciones de las filas sobre estos ejes del biplot estiman los valores estandarizados que aparecen en el lado izquierdo de la ecuación (13.7). Es decir, tomamos una determinada dirección de referencia, por ejemplo la *D*, y luego proyectamos todas las filas sobre dicho eje, con lo que aparecen alineadas. Así vemos que Zoología es la fila que tiene el mayor elemento perfil en esta categoría, le siguen Botánica, Geología, y así sucesivamente, Física y Bioquímica tienen los menores valores de perfil en *D*. (Los valores que aparecen en la tabla de la imagen 10.1 muestran que esto es correcto, con algunas pequeñas excepciones; resulta lógico ya que se trata de un biplot aproximado, y representa el 84% de la inercia total de la tabla.)

Imagen 13.3:
Biplot estándar del AC de los datos sobre la financiación de la investigación científica de la imagen 10.1. Hemos expresado las filas en coordenadas principales, y las columnas, que indican las direcciones de los vértices, en coordenadas estándares, pero multiplicadas por la raíz cuadrada de la masa de cada columna. Así, por ejemplo, la posición de A la hemos obtenido multiplicando la posición de A de la imagen 10.2, por $\sqrt{0{,}0389} = 0{,}197$

Calibración de ejes de los biplots

Dado que las proyecciones de las filas sobre los ejes del biplot son proporcionales a los valores del lado izquierdo de (13.7), podemos calibrar los ejes del biplot en las mismas unidades que los perfiles para leer directamente sus valores. Así, por ejemplo, para estimar los valores estandarizados del perfil en el eje *A* del biplot, tenemos que multiplicar las proyecciones de las filas por la longitud del vector *A*, que es igual a 0,484. Para desestandarizar y así recuperar los valores originales de los perfiles, multiplicaremos esta longitud por la raíz cuadrada de la masa de la columna ($\sqrt{0{,}0389} = 0{,}197$), y así obtenemos el factor de escala 0,0955. La longitud de una unidad en el eje *A* del biplot será igual a $1/0{,}0955 = 10{,}47$. Por tanto, la longitud de un intervalo del 1% (es decir 0,01) en este eje del biplot de la imagen 13.4, será una centésima de esta longitud, es decir 0,1047. Por tanto, conocemos los tres elementos necesarios para calibrar el eje *A*: a) el origen del mapa se halla en el valor 0,039 (3,9%) del eje *A*; b) una longitud de 0,01 (1%) es igual a 0,1047; y c) el vector de la imagen 13.3 apunta hacia la dirección positiva del eje. En la imagen 13.4 podemos ver la calibración del eje *A*, así como la del eje *D*, que hemos efectuado de forma similar.

Calidad global de la representación

Anteriormente determinábamos la calidad global de un mapa bidimensional del AC, como el valor de la inercia explicada por los dos primeros ejes principales. El biplot proporciona una manera alternativa de determinar la calidad de los mapas, concretamente la capacidad de recuperar los valores de los perfiles a par-

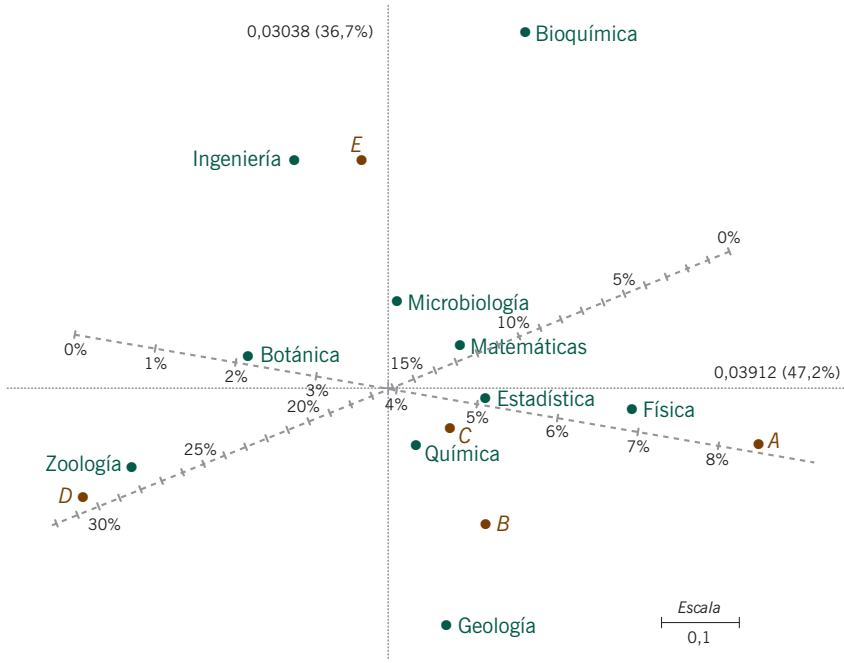


Imagen 13.4: *Mapa simétrico de la tabla 10.1 (datos sobre la financiación de la investigación científica) incluye los ejes de las categorías A y D calibrados. Fijémonos en que los ejes calibrados se hallan en la dirección de los vértices y en que no pasan exactamente por los puntos correspondientes a los perfiles de la categoría (en este ejemplo pasan muy cerca de los puntos en coordenadas principales debido a que las diferencias entre las inercias de los dos ejes es pequeña)*

tir del mapa. Por ejemplo, proyectando todos los puntos fila sobre los ejes del biplot que mostramos en la imagen 13.3, convenientemente calibrados, podemos recuperar de forma aproximada los valores de la tabla que mostramos en la imagen 10.1. Cuanto más próximos estén los valores estimados de los perfiles a los reales, mejor será la calidad del mapa. A la inversa, para obtener una medida global de error, podemos ir acumulando las diferencias entre los valores verdaderos de los elementos del perfil y los estimados. Cuando calculamos estas diferencias en forma de ji-cuadrado, es decir, calculando los cuadrados de las diferencias divididas por los valores esperados, obtenemos exactamente la misma medida de error que obtuvimos anteriormente. En este ejemplo en concreto, el porcentaje de inercia explicada en el mapa bidimensional es el 84%; por tanto el error es del 16%.

RESUMEN:
Biplots en análisis de correspondencias

1. El *producto escalar* entre dos vectores es igual al producto de sus longitudes multiplicado por el coseno del ángulo que forman.
2. Dado que la proyección perpendicular de un vector **x** sobre la dirección definida por un segundo vector **y**, tiene una longitud igual a la de **x** multiplicada por el coseno del ángulo formado por **x** e **y**, podemos ver el producto escalar como el producto de la longitud de la proyección de **x** y la longitud de **y**.
3. El *biplot* es un mapa que representa conjuntamente las filas y las columnas de una matriz de datos, de manera que los productos escalares entre los vectores fila y los vectores columna se aproximen tanto como sea posible a los correspondientes valores de la matriz.
4. En AC, los mapas asimétricos son biplots; en cambio, en sentido estricto, los mapas simétricos no lo son, a pesar de que en la práctica las direcciones definidas por los perfiles del mapa simétrico y los correspondientes vértices del mapa asimétrico, a menudo, no son muy distintas, de modo que la interpretación del biplot sigue siendo válida.
5. Multiplicando las posiciones de los vértices de los mapas asimétricos por la raíz cuadrada de la masa de las correspondientes columnas acercamos las posiciones de los vértices al origen. A esta interesante variación del mapa asimétrico le llamaremos *biplot estándar* del AC.
6. Podemos calibrar los ejes del biplot en las unidades de los perfiles (como proporciones o en porcentajes). De esta manera, las proyecciones de los perfiles nos darán directamente sus valores aproximados.