La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 24

Análisis de correspondencias canónico

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

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

www.fbbva.es

CAPÍTULO 13

Análisis de correspondencias canónico

El AC nos permite visualizar tablas de datos en subespacios de baja dimensionalidad que explican de forma óptima la inercia. Mediante puntos adicionales —que no tienen efecto alguno sobre la solución hallada (cap. 12)— podemos visualizar información externa suplementaria de filas o de columnas. Puede ocurrir que queramos que el resultado del AC esté directamente relacionado con variables externas, que queramos que tengan un papel activo en la definición del mapa del AC. Dicha situación se da con frecuencia en el contexto de la investigación medioambiental, en la que puede ocurrir, por ejemplo, que dispongamos al mismo tiempo, en distintas estaciones de muestreo, de información sobre la composición en determinadas especies biológicas y sobre parámetros ambientales. En estos casos, al llevar a cabo el AC, buscaríamos los subespacios que mejor expliquen los datos biológicos, pero con la condición de que éstos se hallen directamente relacionados con las variables ambientales. El *análisis de correspondencias canónico*, ACC, es una variante del AC en la que obtenemos las dimensiones del subespacio por regresión a partir de variables externas.

Contenido

Variables continuas adicionales

Para comprender el ACC, consideremos de nuevo los datos sobre biología marina que mostramos en la tabla de la imagen 10.4. Además de información sobre las especies presentes en el fondo marino de cada localidad de muestreo, se obtuvo información sobre algunas variables ambientales: concentración de metales (plomo, cadmio, bario, hierro, ...), composición de sedimentos (arcilla, arena, pelite, ...), así como el contenido en hidrocarburos y materia orgánica. Dado que estas variables están muy correlacionadas entre sí, escogimos como variables representativas, como mostramos en la tabla de la imagen 24.1, el contenido en bario, en hierro (expresados en partes por millón) y en pelite* (expresado como porcentaje). En el ACC, estas variables externas serán las variables explicativas de un modelo de regresión lineal que nos permitirá obtener las dimensiones del subespacio. Preferimos trabajar con los logaritmos de las mencionadas variables, una transformación habitual para pasar este tipo de medidas de una escala multiplicativa a una escala aditiva (en la tabla de la imagen 24.1 también mostramos estos valores). Esta transformación no sólo elimina el efecto de distintas escalas de medida de estas tres variables, sino que también reduce la influencia de los valores grandes.

Imagen 24.1: *Datos medioambientales medidos en 13 estaciones de muestreo (véase la tabla de la imagen 10.4); 11 estaciones próximas a una plataforma petrolífera y dos estaciones de referencia alejadas 10 km*

	Estaciones de muestreo (muestras)												
Variables	E4	E8	E9	E12	E13	E14	E15	E18	E19	E23	E24	R40	R42
Bario (Ba)	1656	1373	3680	2094	2813	4493	6466	1661	3580	2247	2034	40	85
Hierro (Fe)	2022	2398	2985	2535	2612	2515	3421	2381	3452	3457	2311	1804	1815
Pelite (PE)	2,9	14,9	3,8	5,3	4,1	9,1	5,3	4,1	7,4	3,1	6,5	2,5	2,0
log*(Ba)*	3,219	3,138	3,566	3,321	3,449	3,653	3,811	3,220	3,554	3,352	3,308	1,602	1,929
log*(Fe)*	3,306	3,380	3,475	3,404	3,417	3,401	3,534	3,377	3,538	3,539	3,364	3,256	3,259
log*(PE)*	0,462	1,173	0,580	0,724	0,623	0,959	0,724	0,613	0,869	0,491	0,813	0,398	0,301

Representación de variables explicativas como puntos adicionales

Antes de entrar en el ACC vamos a representar las tres variables externas en el mapa de la imagen 10.5 como puntos adicionales. Como vimos en el capítulo 14, para obtener las coordenadas de variables continuas en dos ejes principales, llevamos a cabo una regresión de mínimos cuadrados ponderada de las variables en la que las «variables explicativas» son las coordenadas estándares de las columnas γ_1 y γ_2 —en las dos primeras dimensiones— y los pesos son las masas de las columnas. Así, por ejemplo, a continuación mostramos parte de los datos para la regresión de log*(Ba)*:

Variable	log*(Ba)*	γ_1	γ_2	*Peso*
E4	3,219	1,113	0,417	0,0601
E8	3,138	–0,226	–1,327	0,0862
E9	3,566	1,267	0,411	0,0686
⋮	⋮	⋮	⋮	⋮
R42	1,929	2,300	0,7862	0,0326

* El pelite es un sedimento compuesto por finas partículas de textura arcillosa y limosa.

Los resultados de la regresión son:

Fuente	*Coeficiente*	*Coeficiente estandarizado*
Ordenada en el origen	3,322	—
γ_1	–0,301	–0,641
γ_2	–0,229	–0,488

$R^2 = 0{,}648$

Habitualmente, para la representación de variables adicionales utilizamos los coeficientes de regresión estandarizados. Como vimos en el capítulo 14, estos valores son idénticos a los coeficientes de correlación (ponderados) de log*(Ba)* con las coordenadas estándares de las dos columnas. Una vez realizada la regresión (o, de forma equivalente, calculando los coeficientes de correlación), podemos situar las tres variables ambientales en el mapa de la imagen 24.2, como se muestra en el mapa de la imagen 10.5, en la que hemos omitido los puntos correspondientes a las especies. El porcentaje de varianza de cada variable explicado (R^2) es igual a la suma de los coeficientes de correlación al cuadrado. Es decir, lo que llamamos *calidad* de representación de un punto. Para log*(Ba)* es bastante alto, 0,648 (64,8%); para log*(Fe)* es 0,326, y para log*(PE)*, solamente 0,126.

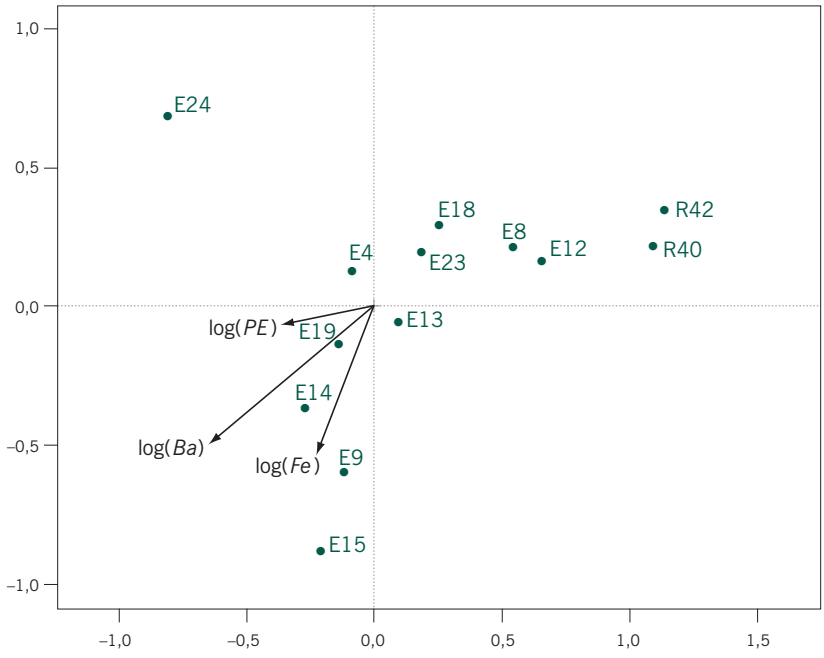


Imagen 24.2:
Mapa de las estaciones de la imagen 10.5, que muestra las posiciones de las tres variables ambientales externas como puntos adicionales, de acuerdo con sus correlaciones con los dos ejes principales

Imagen 24.3: *Regresión de las dos primeras dimensiones sobre las tres variables medioambientales*

Respuesta: dimensión 1 del AC

Fuente	*Coef.*	*Coef. estand.*
Ord. origen	–9,316	—
log*(Ba)*	–1,953	–0,918
log*(Fe)*	–4,602	0,398
log*(PE)*	0,068	0,014

$R^2 = 0{,}494$

Respuesta: dimensión 2 del AC

Fuente	*Coef.*	*Coef. estand.*
Ord. origen	14,465	—
log*(Ba)*	–0,696	–0,327
log*(Fe)*	–3,672	0,318
log*(PE)*	0,588	0,123

$R^2 = 0{,}319$

Dimensiones como funciones de las variables explicativas

Ahora vamos a dar la vuelta al problema, en vez de llevar a cabo la regresión de las variables continuas sobre las dimensiones, hagamos la regresión de las dimensiones sobre las variables explicativas, incorporando siempre en la regresión las masas como pesos. En las tablas de la imagen 24.3, mostramos los resultados de los dos análisis de regresión. Fijémonos en que, desafortunadamente, los coeficientes estándares ya no son los coeficientes de correlación que utilizamos en la imagen 24.2, para representar las variables. Por ejemplo, las correlaciones entre log*(Ba)* y las dos dimensiones son –0,641 y –0,488, mientras que en el análisis de regresión anterior los coeficientes de regresión estandarizados eran –0,918 y –0,327, respectivamente.

Restricción en las dimensiones del AC

El porcentaje de varianza (en realidad de inercia, ya que hemos ponderado las variables según los pesos de las estaciones) explicado por las regresiones de las dos dimensiones sobre las variables ambientales es del 49,4% y del 31,9%, respectivamente (véase la última línea de las tablas de la imagen 24.3). Vamos a tratar de aumentar la inercia explicada forzando que las dimensiones sean una función lineal de las tres variables explicativas. En el AC habitual optimizamos el ajuste de los perfiles de las especies halladas en los fondos marinos sin imponer restricción alguna sobre las dimensiones. Sin embargo, impongamos ahora la condición de que las dimensiones sean combinaciones lineales de las variables ambientales. De esta manera conseguiremos aumentar la inercia explicada de las dimensiones hasta el 100%. El inconveniente será que empeorará la explicación de los datos de las especies. La manera de proceder es la siguiente: proyectamos todos los datos sobre el subespacio definido por las tres variables ambientales, y a continuación llevamos a cabo el AC de la forma habitual. Esta es la idea del ACC: en vez de buscar la solución que mejor ajuste los ejes principales en el espacio completo de datos, lo que hacemos es ajustar los ejes principales, en una parte limitada o restringida del espacio (por tanto, podríamos considerar el ACC como un *análisis de correspondencias restringido*). En las tablas de la imagen 24.4 mostramos los resultados de las regresiones de las dos primeras dimensiones del ACC sobre las variables ambientales. Ahora, la varianza (inercia) explicada es del 100%, que es precisamente lo que buscábamos. Hemos impuesto que las dimensiones sean, ne-

Respuesta: dimensión 1 del AC		
Fuente	*Coef.*	*Coef. estand.*
Ord. origen	2,719	—
log*(Ba)*	–2,297	–1,080
log*(Fe)*	1,437	0,124
log*(PE)*	–0,008	–0,002
	$R^2 = 1$	

Respuesta: dimensión 2 del AC		
Fuente	*Coef.*	*Coef. estand.*
Ord. origen	14,465	—
log*(Ba)*	–0,877	–0,412
log*(Fe)*	12,217	1,058
log*(PE)*	–2,378	–0,497
	$R^2 = 1$	

Imagen 24.4:
Regresiones de las dos primeras dimensiones del ACC sobre las tres variables ambientales

cesariamente, combinaciones lineales de las variables ambientales (más adelante mostraremos los resultados completos).

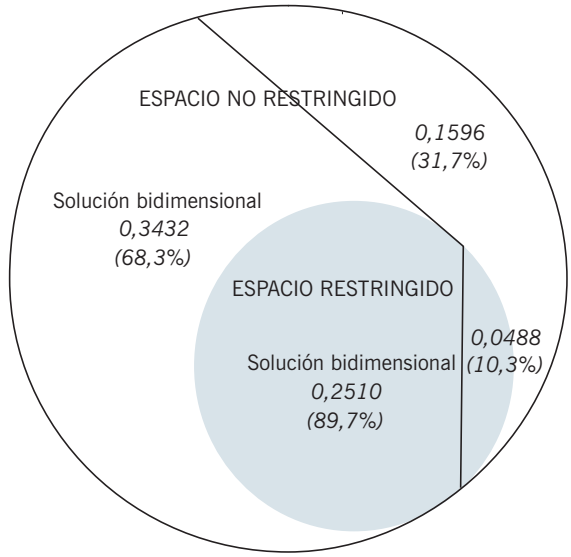
Espacios restringidos y no restringidos en ACC

En ACC, el *espacio restringido* (o *espacio canónico*) es la parte del espacio total en el que limitamos la búsqueda de los ejes principales óptimos, el *espacio no restringido* (o *espacio no canónico*) es el resto del espacio total. Dentro del espacio restringido, con el algoritmo habitual del AC, hallamos las dimensiones que mejor explican los datos de las especies. También podríamos buscar las mejores dimensiones dentro del espacio no restringido: el espacio que no se halla relacionado (correlacionado) con las variables ambientales. Podríamos estar interesados en llevar a cabo el ACC con determinadas variables ambientales, y luego buscar las dimensiones en la parte no restringida del espacio.

Descomposición de la inercia en ACC

En el ejemplo que nos ocupa, la inercia total de la tabla de especies por estaciones de muestreo es de 0,7826 (la inercia total de la tabla de la imagen 10.4). Los espacios restringido y no restringido nos permiten descomponer la inercia en dos partes, con valores de 0,2798 y 0,5028, respectivamente, el 35,8 y el 64,2% de la inercia total. Es decir, hay más inercia en el espacio no restringido que en el restringido. Esto nos proporciona una explicación de porqué las dimensiones del AC original daban correlaciones bajas con las variables ambientales. Efectivamente, el AC trata de explicar la máxima inercia, y hay más inercia en el espacio no restringido que en el espacio restringido. En la imagen 24.5 mostramos la descomposición de la inercia, incluyendo la descomposición en los ejes principales. Una vez limitamos la búsqueda de dimensiones dentro del espacio restringido (representada por el área sombreada de la imagen 24.5), las inercias de las dos primeras dimensiones son de 0,1895 y de 0,0615, respectivamente, un total de 0,2510, el 89,7% de la inercia restringida de 0,2798 y el 32,1% de la inercia total original de 0,7826 (en el mapa de la imagen 10.5, el AC bidimensional explica el 57,5%). Por otra parte, si estuviéramos interesados en el espacio no restringido (no sombreado en la imagen 24.5), veríamos que las dos primeras dimensiones tienen inercias principales de 0,1909 y de 0,1523, un total de 0,3432, el 68,3% de la inercia no restringida de 0,5028, y el 43,8% de la inercia total. Si lleváramos a

Imagen 24.5:
Diagrama esquemático de la descomposición de la inercia en espacio restringido (sombreado) y espacio no restringido, que muestra las partes de cada una de ellas explicadas por los respectivos mapas bidimensionales. Las partes situadas a la derecha de las líneas rectas (inercias de 0,0488 y 0,1596) permanecen inexplicadas por las respectivas soluciones bidimensionales



cabo la regresión de estas dos últimas dimensiones del espacio no restringido sobre las variables ambientales, veríamos que no existe relación (los coeficientes de correlación serían cero) y, por tanto, la inercia explicada sería cero.

Triplot del ACC

El ACC proporciona las mismas posibilidades de elección de coordenadas para filas y columnas que el AC habitual. El *triplot* es un diagrama en el cual, además, añadimos vectores correspondientes a las variables explicativas. Para la visualización de las variables explicativas tenemos dos posibilidades. Una es utilizar sus coeficientes de correlación con los ejes para definir sus posiciones; los representaríamos como si fueran puntos adicionales. La otra posibilidad es usar los coeficientes de regresión estandarizados derivados de su relación con los ejes. Nosotros consideramos que esta última posibilidad es mejor ya que refleja la idea de que en el ACC los ejes están relacionados linealmente con las variables explicativas. En el mapa de la imagen 24.6 mostramos un posible triplot del ACC correspondiente al ejemplo que nos ocupa. Para representar las variables ambientales hemos utilizado sus coeficientes de regresión. Hemos expresado las estaciones en coordenadas estándares y las especies en coordenadas principales; por tanto, el mapa básico es un mapa asimétrico de filas principales. Respecto a las estaciones de muestreo y especies, sigue siendo válida la interpretación del biplot. Efectivamente, dado que las localidades están en coordenadas estándares, éstas definen los ejes del biplot sobre los que podemos proyectar las especies y así determinar sus abundancias relativas en esa estación de muestreo (abundancia relativa respecto al total de todas las localidades). Asimismo, las posiciones de las estaciones de muestreo con relación a los ejes son, por construcción, combinacio-

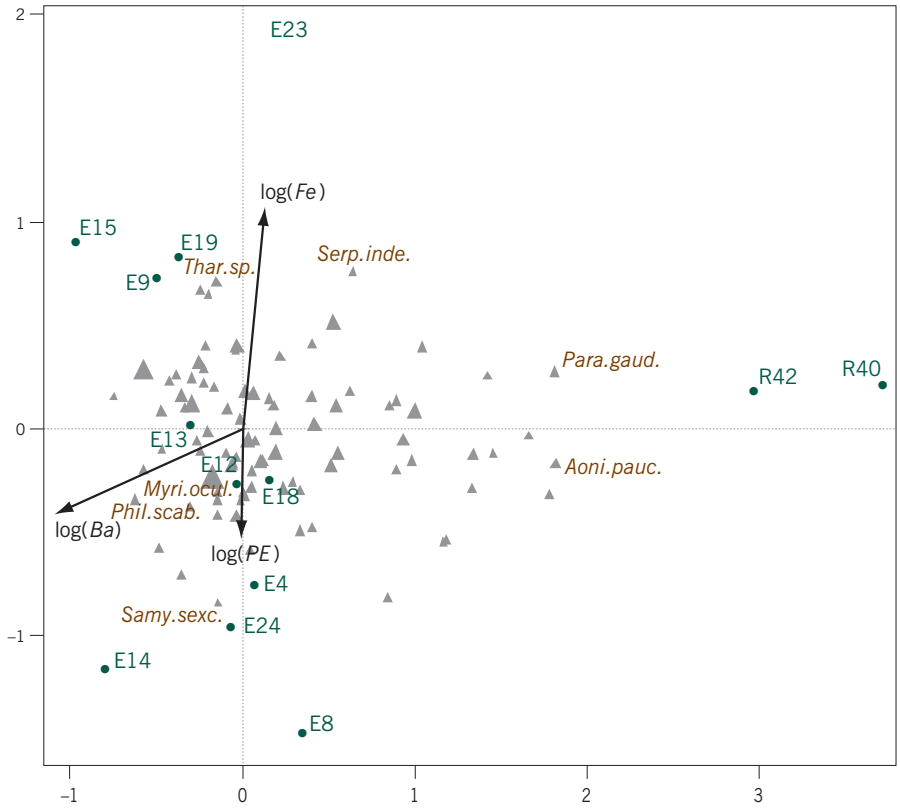


Imagen 24.6:
Triplot del ACC en el que hemos representado las especies (filas) y las localidades (columnas) en un mapa asimétrico de filas (es decir, las localidades en coordenadas estándares). Hemos situado las variables ambientales según los valores de sus coeficientes en las relaciones lineales con los dos ejes. El tamaño de los símbolos de las especies es proporcional a su abundancia total; sólo indicamos el nombre de algunas especies que citamos en el texto

nes lineales de los valores estandarizados de las tres variables ambientales. Si a una estación de muestreo le corresponde la media de una determinada variable ambiental, entonces la contribución de esta variable a su posición es cero. Por tanto, el hecho de que las estaciones de referencia R40 y R42 estén tan lejos, al otro lado de log*(Ba)*, indica que sus valores en bario deben ser bajos, lo cual es cierto. Asimismo, E23, E19, E15 y E9 deben tener valores altos en hierro (especialmente E23). En cambio, E8 y E14 deben tener valores altos en pelite. Podemos confirmarlo examinando los valores de la tabla de la imagen 24.1. Tenemos que investigar la relación entre especies y variables a través de las estaciones de muestreo. Así, especies como *Para.gaud.* y *Aoni.pauc.* están asociadas con las estaciones de referencia. Estas estaciones tienen poco bario. En cambio, especies como *Thar.sp.* y *Sep.inde.* están asociadas con estaciones que tienen mucho hierro y/o poco pelite, mientras que *Samy.sexc.*, en la parte baja, está asociada con estaciones que tienen mucho pelite y/o poco hierro. Las estaciones de referencia se hallan, más o menos, en la mitad del eje vertical. Es decir, tienen valores bajos tanto en hierro como en pelite, y este hecho ha equilibrado sus posiciones verticales.

Imagen 24.7: *Medias ponderadas de las tres variables ambientales de una selección de especies, calculadas a partir de los valores de las variables en cada estación de muestreo. Como pesos hemos utilizado frecuencias de las especies en cada estación de muestreo*

	Variables		
Especies	log*(Ba)*	log*(Fe)*	log*(PE)*
Myri.ocul.	3,393	3,416	0,747
⋮	⋮	⋮	⋮
Serp.inde.	3,053	3,437	0,559
Thar.sp.	3,422	3,477	0,651
Para.gaud.	2,491	3,352	0,534
Aoni.pauc.	2,543	3,331	0,537
Samy.sexc.	3,373	3,409	0,971
⋮	⋮	⋮	⋮
Media global	3,322	3,424	0,711

Variables explicativas categóricas

Si las variables explicativas fueran categóricas, como es el caso de *Región* (por ejemplo, con las categorías noreste/noroeste/sur) o como *Rocoso* (con las categorías sí/no), las incluiríamos en el ACC codificadas como variables binarias, al igual que en un análisis de regresión. En el mapa del ACC, no representamos las variables binarias mediante flechas, lo que hacemos es representar por puntos las medias ponderadas de las estaciones de muestreo que quedan en cada categoría (ponderando de la forma habitual).

Medias ponderadas de las variables explicativas de cada especie

Una manera alternativa de contemplar el ACC es verlo como un análisis de medias ponderadas de las variables explicativas para cada especie. En la tabla de la imagen 24.7 mostramos una pequeña parte de este conjunto de medias para algunas de las especies que hemos considerado anteriormente. Así, por ejemplo, en la tabla de la imagen 10.4, las frecuencias de *Myriochele oculata (Myri.ocul.)* de 193, 79, 150, etc., en las estaciones de muestreo son E4, E8, E9, etc. En estas mismas estaciones, los valores de log*(Ba)* son 3,219, 3,138, 3,566, etc. Por tanto, la media ponderada de *Myri.ocul.* para esta variable es:

$$\frac{193\times 3{,}219+79\times 3{,}138+150\times 3{,}566+\cdots}{193+79+150+\cdots}=3{,}393$$

es decir, el producto escalar de los perfiles de las especies y los valores de la variable. Hemos calculado de la misma manera la «media global» (ponderada) (última línea de la tabla de la imagen 24.7), pero con los totales de todas las especies. Podemos ver que *Myri.ocul.* se halla bastante cerca de la media global, y en consecuencia no ejerce un papel tan importante como el que tenía en el mapa del AC de la imagen 10.5. Para *Para.gaud.* y *Aoni.pauc.*, las medias ponderadas de la variable log*(Ba)* son bajas, debido a que sus frecuencias son relativamente elevadas en las localidades de referencia R40 y R42, en las que el bario es muy bajo. También podemos ver que la media ponderada de *Samy.sexc.* con relación a

log*(PE)* es alta. Finalmente, el hecho de que *Thar.sp.* y *Serp.inde.* se hallen en la parte superior, se debe más a sus bajas medias ponderadas en pelite que a sus valores altos en hierro.

ACC parcial

En el ACC *parcial* llevamos un poco más lejos la idea de separar la variación debida a algunas variables. Supongamos que dividimos las variables explicativas en dos grupos, grupos *A* y *B*. Supongamos también que el efecto de *A* no tiene demasiado interés, posiblemente porque es bien conocido como, por ejemplo, un gradiente geográfico norte-sur. Para llevar a cabo el ACC parcial, en primer lugar, eliminamos el efecto de las variables *A*, y llevamos a cabo el ACC con las variables *B*, en el espacio no relacionado con las variables *A*. Es decir, estamos llevando a cabo una descomposición de la inercia total original en tres partes: la parte debida a *A* que eliminamos, y la parte restante, en la que descomponemos la inercia en la parte restringida que debe estar relacionada con las variables *B* y la parte no restringida (que no está relacionada ni con las variables *A* ni con las variables *B*).

RESUMEN: Análisis de correspondencias canónico

1. En AC hallamos las dimensiones del subespacio que maximizan la inercia explicada.
2. En el *análisis de correspondencias canónico* (ACC) hallamos las dimensiones buscando el mismo objetivo que en el AC, pero con la restricción de que las dimensiones sean combinaciones lineales de un conjunto de variables explicativas.
3. Necesariamente, el ACC explica menos inercia que el AC, ya que el ACC busca la solución en un espacio restringido; sin embargo, puede ser que este espacio restringido tenga más interés para el investigador.
4. Podemos descomponer la inercia total en dos partes: la parte relacionada con el espacio restringido, en la que buscamos la solución del ACC, y la parte relacionada con el espacio no restringido, que no está relacionado linealmente con las variables explicativas. En ambos espacios, podemos identificar los ejes principales que expliquen el máximo de inercia; son las soluciones *restringida* y *no restringida*, respectivamente.
5. En el ACC *parcial*, antes de llevar a cabo el ACC, eliminamos el efecto de un grupo de variables, y realizamos el análisis con las otras variables explicativas.