La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 25

Consideraciones sobre estabilidad e inferencia

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

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

www.fbbva.es

CAPÍTULO 15

Consideraciones sobre estabilidad e inferencia

Hasta ahora, hemos centrado nuestra atención en las propiedades geométricas del AC y en su interpretación. De inferencia estadística sólo hemos visto la prueba χ^2 y algo sobre la significación de agrupaciones en el capítulo 15. En este capítulo final vamos a dar una visión general de cómo analizar la estabilidad de los resultados del AC y sobre las propiedades de estadísticos tales como inercia total, inercia principal y coordenadas principales. Vamos a distinguir entre (1) la estabilidad de la solución, con independencia de la fuente de datos, (2) la variabilidad de las muestras, suponiendo que éstas son el resultado de algún tipo de muestreo aleatorio de poblaciones grandes, y (3) el contraste de algunas hipótesis estadísticas.

Contenido

Transformación de la información *versus* inferencia estadística

Hemos explicado el AC como un método para la descripción de datos de forma gráfica que resulta fácil de interpretar, y así facilitar la exploración y la interpretación de información numérica. Saber si la información contenida en los mapas del AC refleja fenómenos reales o simplemente es resultado del azar es otro tema. Realizar afirmaciones sobre una población, es decir, hacer *inferencia*, es un ejercicio

distinto que exige consideraciones especiales que sólo son factibles cuando hemos obtenido correctamente los datos de una población. Para los datos categóricos considerados en este libro, existen muchos marcos de referencia que permiten contrastar hipótesis y hacer inferencia sobre algunas características de la población cuyos datos se han muestreado. Así, por ejemplo, la *modelización log-lineal* permite contrastar, de manera formal, interacciones entre variables y la *modelización de asociaciones* muy relacionada con el AC nos permite, por ejemplo, contrastar diferencias entre valores categóricos. El AC nos permite llevar a cabo inferencia estadística, así como explorar la variabilidad y la estabilidad de los mapas gracias a los modernos ordenadores de alta velocidad.

Estabilidad del AC

Cuando hablamos de *estabilidad* del resultado del AC (mapa, inercias, coordenadas en determinados ejes principales, etc.), hacemos referencia a unos determinados datos. No nos ocupamos de la población de origen de los datos. Por tanto, la estabililidad es un tema relevante en cualquier situación, para datos poblacionales o para datos obtenidos de un muestreo de conveniencia. Nuestra interpretación de un determinado mapa configurado por un conjunto específico de filas y columnas, ¿cómo puede verse afectada? Cuando eliminamos algunos puntos del mapa, ¿el mapa cambia sustancialmente (y, por tanto, nuestra interpretación)? Por ejemplo, si eliminamos una de las especies en los datos sobre biología marina, o uno de los autores en los datos sobre autores (cap. 10), ¿cambiarán sustancialmente los mapas? Cuando vimos el concepto de influencia y cuando analizamos la influencia de los puntos sobre la configuración de los ejes principales, ya entramos en la cuestión de la estabilidad del resultado del AC. En el capítulo 11, vimos que las *contribuciones a la inercia* de los puntos nos informan sobre su influencia. En los capítulos 11 y 12, vimos que si un punto contribuye mucho a la inercia de un eje, entonces éste puede tener una gran influencia sobre la configuración del mapa. Este hecho puede ser un problema cuando los puntos tienen poca masa. Por otro lado, hay puntos que contribuyen muy poco al resultado del AC y que, por tanto, podemos eliminar sin cambiar demasiado el mapa; es decir, el mapa es *estable* con respecto a la eliminación o a la inclusión de dichos puntos. Para poderlo valorar, la prueba de fuego consiste en llevar a cabo varios AC omitiendo puntos y ver cómo se ven afectados los resultados del AC.

Variabilidad muestral del resultado del AC

Supongamos ahora que hemos obtenido datos de una *población* siguiendo un determinado protocolo de muestreo. Por ejemplo, sabemos que los datos sobre los autores que mostramos en la tabla de la imagen 10.6, representan una pequeña parte de textos completos. Si repetimos el análisis con muestras distintas de los textos, seguro que el recuento de letras no será el mismo. Sería perfecto poder repetir muchas veces el muestreo, y en cada ocasión llevar a cabo el AC. De esta manera podríamos observar si cambian los mapas, si se mantienen más o menos

constantes o, por el contrario, cambian las posiciones de libros y letras. El mapa obtenido, ¿caracteriza realmente los 12 libros? o ¿es resultado del azar?

Automuestreo de datos

Sin embargo, dado que no podemos repetir el muestreo, para intentar comprender la variabilidad muestral de la matriz de datos, tenemos que basarnos en los datos que disponemos. En estadística, es habitual hacer algunas suposiciones sobre la población y luego obtener resultados sobre la incertidumbre de los parámetros estimados —en nuestro caso, las coordenadas de los puntos del mapa—. El *automuestreo*[*] es una manera menos formal de proceder, que evita tener que hacer suposiciones. Concretamente, consiste en contemplar los datos que disponemos como si fueran la población, ya que son la mejor estimación que tenemos de la misma y crear nuevos datos remuestreándolos como se muestrearon los datos originales. Consideremos, por ejemplo, los datos sobre los autores. Se muestrearon textos, no se muestrearon letras individuales. Por tanto, tenemos que remuestrear de esta manera. Así, en el primer libro, *Three Daughters*, se muestreó un texto de 7144 letras. Podemos imaginar estas 7144 letras alineadas en un largo vector en el que hay 550 `a`, 116 `b`, 147 `c`, ..., etc. A continuación obtenemos, de este vector, una muestra aleatoria de 7144 letras *con reemplazamiento*; por tanto, las frecuencias no serán exactamente iguales a las de la tabla original, sin embargo, éstas reflejarán la variabilidad de frecuencias existente en la muestra. Repetimos lo mismo con las restantes filas de la tabla de la imagen 10.6 hasta que obtengamos una nueva tabla, con los mismos totales de filas que la tabla original. Podemos repetir el procedimiento completo muchas veces, en general entre 100 y 1000 veces, para llegar a tener muchos automuestreos de la matriz de datos original.

Muestreo multinomial

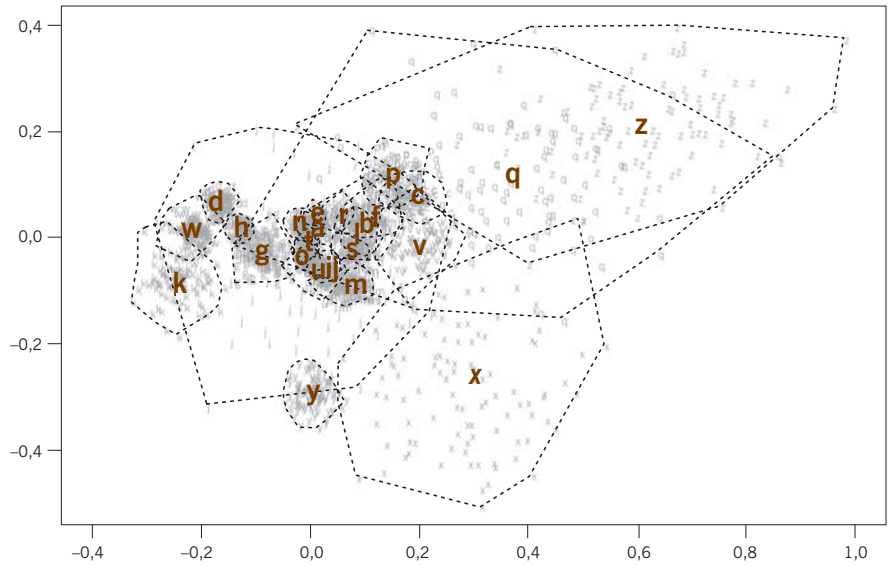
El *muestreo multinomial* es una manera equivalente de contemplar (y de llevar a cabo) el remuestreo. Se basa en que cada perfil fila define un conjunto de frecuencias que podemos considerar como las probabilidades de obtener, en cada texto, una `a`, una `b`, una `c`, etc. Por tanto, se trata de muestrear una población con estas mismas probabilidades. Lo podemos llevar a cabo con un simple algoritmo de cálculo, ya implementado en `R` (véase el apéndice de cálculo, B). Por tanto no tenemos la necesidad de crear un vector de 7144 letras, sólo necesitamos utilizar las probabilidades de las 26 letras en un procedimiento de muestreo multinomial.

Automuestreo parcial en un mapa del AC con perímetros convexos

Para ilustrar este procedimiento de automuestreo con los datos sobre los autores, en primer lugar calculamos 100 réplicas de la tabla con el procedimiento de cálculo que acabamos de ver. A continuación podemos seguir dos caminos. El más complicado consiste en llevar a cabo el AC en cada una de las réplicas y luego, de alguna manera, comparar los resultados con los obtenidos originalmente. El *au-*

[*] La expresión en inglés *pulling yourself up by your own bootstraps* significa salir de una situación difícil utilizando los propios recursos. Hemos traducido *bootstrap* por automuestreo.

Imagen 25.1:
Automuestreo (parcial) de 26 letras, después de 100 réplicas de la matriz de datos. Cuanto más frecuente sea una letra en los textos, más concentradas (menos variables) son las réplicas. Mostramos los perímetros convexos alrededor de cada conjunto de 100 réplicas



tomuestreo parcial es otra posibilidad más sencilla. Consiste en considerar cada una de las 100 tablas replicadas como un conjunto de perfiles fila o de perfiles columna, que proyectamos como puntos adicionales en el mapa original. En el mapa de la imagen 25.1 mostramos el resultado del automuestreo parcial de las 26 letras (se muestran en caracteres mayores las posiciones originales en coordenadas principales y en caracteres menores las 100 réplicas de cada letra). No suele ser habitual mostrar todos los puntos de cada réplica. Lo más frecuente es incluir en el mapa sólo las réplicas situadas en el *perímetro convexo*, es decir, los puntos exteriores que unimos mediante una línea discontinua, como si se tratara de una cinta elástica colocada alrededor de las réplicas de cada letra.

Recorte del perímetro convexo

Por *recorte* de perímetros convexos entendemos la eliminación de las observaciones atípicas que a menudo encontramos en los perímetros convexos (lo podemos ver, por ejemplo, para la letra z situada a la derecha del mapa de la imagen 25.1). Es habitual ir recortando el perímetro convexo hasta eliminar el 5% de los puntos más exteriores de las proyecciones de las subnubes de puntos. Los perímetros convexos de los puntos restantes constituyen una estimación, con un confianza del 95%, de la región de confianza de cada letra en el mapa. Para hacer más suave la estimación de las regiones convexas, podemos generar 1000 réplicas de cada letra y luego recortar los 50 puntos más exteriores de cada letra. En el mapa de la imagen 25.2 mostramos estos perímetros convexos recortados en esta última situación. Si dos perímetros convexos no se solapan, tenemos bastante seguridad de que en los textos, las letras son significativamente distintas. Dado que el procedimiento que utilizamos es bastante informal, y dado también el problema de

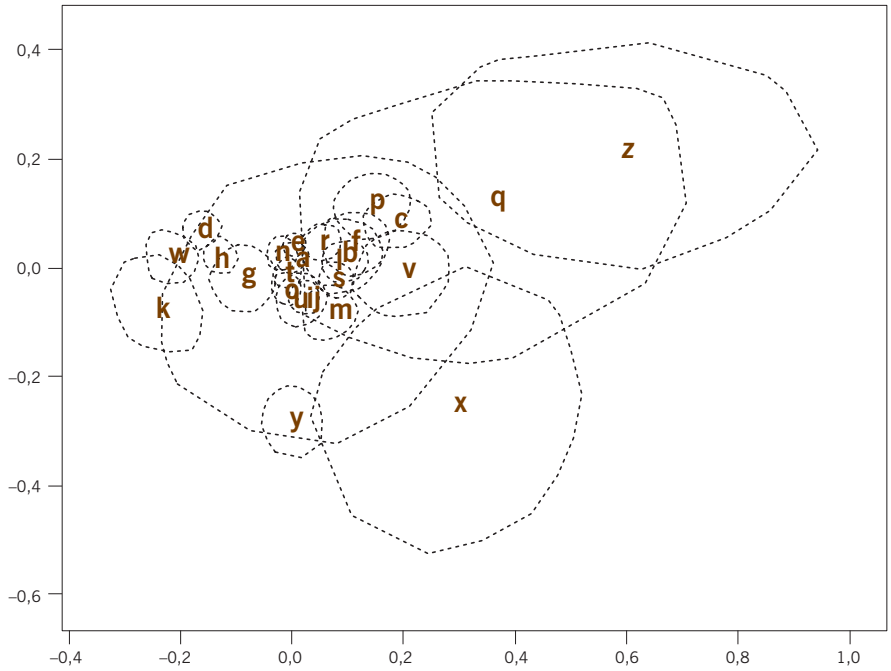


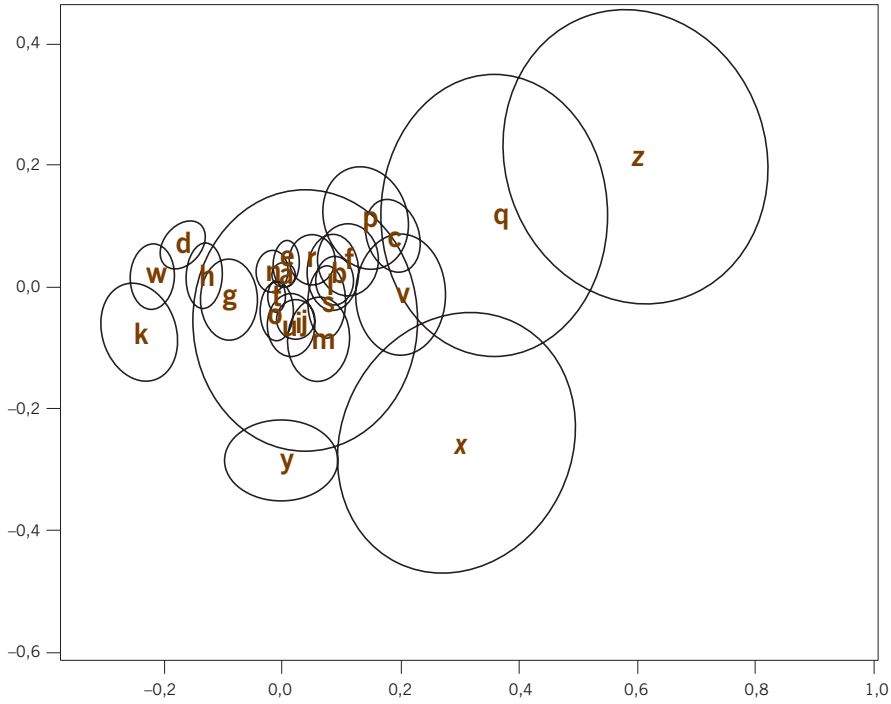
Imagen 25.2:
Recorte de perímetros convexos de puntos obtenidos de 1000 réplicas (10 veces más que en la imagen 25.1) que muestran, para sus distribuciones, regiones de confianza aproximadas al 95%

las comparaciones múltiples que vimos en el capítulo 15, es difícil calcular niveles de significación. Sin embargo, por suerte, como llevamos a cabo proyecciones de puntos sobre el mapa original, el procedimiento es conservador. Es decir, si dos perímetros convexos no se solapan en el mapa (como, por ejemplo, k e y), seguro que las nubes de puntos no se solaparán en el espacio completo. En cambio, aunque las proyecciones de dos nubes de puntos se solapen en el mapa (como por ejemplo x y q), desconocemos si las nubes de puntos se solapan o no en el espacio completo.

El método Delta

Un método alternativo para visualizar regiones de confianza de los puntos de un mapa de AC consiste en utilizar elipses de confianza. Podemos obtener estas elipses a partir de las réplicas del automuestreo parcial que vimos antes, o a partir de algunas suposiciones teóricas. Así, con el *método Delta* podemos calcular, de forma aproximada, las varianzas y las covarianzas de las coordenadas; utilizamos las derivadas parciales de los vectores propios con relación a las proporciones multinomiales. A continuación, suponiendo una distribución normal bivariante en el plano, podemos calcular elipses de confianza; estas elipses incluyen las verdaderas coordenadas con una confianza del 95%, de forma parecida a los intervalos de confianza de variables individuales. Esta aproximación se basa en el supuesto de un muestreo aleatorio independiente. Ello no se cumple completamente en el caso de los datos de los autores, ya que la presencia de una determinada letra no

Imagen 25.3:
Elipses de confianza obtenidas a partir del método Delta



es independiente de la concurrencia de otras (tenemos un problema similar con el muestreo en ecología, en el que las especies aparecen en las muestras en grupos). A pesar de ello, en el mapa de la imagen 25.3, mostramos las elipses de confianza de las letras en los datos sobre autores. Podemos observar que muestran un gran parecido con los perímetros convexos del mapa de la imagen 25.2, al menos en lo que se refiere al solapamiento.

Contraste de hipótesis: aproximación teórica

Hasta ahora, hemos presentado la prueba χ^2 como una prueba de independencia en una tabla de contingencia. Por ejemplo, la tabla de 5 × 3 de la imagen 4.1, que clasifica 312 personas según su nivel de lectura y su grupo de edad, tiene una inercia de 0,08326 y, por tanto, una χ^2 de 312 × 0,08326 = 25,98. Utilizando la aproximación habitual a la distribución χ^2, el valor p de esta prueba es igual a 0,0035, un valor altamente significativo. La prueba de la *distribución asintótica* permite contrastar la significación de la primera inercia principal de una tabla de contingencia. Los puntos críticos de esta prueba son exactamente los mismos que utilizamos en el capítulo 15 para contrastar la significación de agrupaciones. El valor de la primera inercia principal era de 0,07037, y su valor χ^2 es de 312 × 0,07037 = 21,96. Para contrastar la significación de este valor, tenemos que consultar la tabla del apéndice teórico, A. El punto crítico (a un nivel del 5%) para una tabla de 5 × 3 es de 12,68. Dado que 21,96 es mucho mayor que este

	Datos originales			*1.ª simulación*			*2.ª simulación*			…
Grupos educativos	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	*C1*	*C2*	*C3*	…
E1	5	7	2	2	9	5	4	5	7	…
E2	18	46	20	15	40	38	23	33	37	…
E3	19	29	39	13	36	27	17	34	25	…
E4	12	40	49	11	43	40	14	43	37	…
E5	3	7	16	8	12	13	5	12	16	…

Imagen 25.4:
Tabla de contingencia original mostrada en la imagen 4.1 y dos de las 9999 tablas simuladas según la hipótesis nula de que no existe asociación entre filas y columnas

valor, llegamos a la conclusión de que la primera dimensión del AC es significativa, es decir, no ha surgido del azar. Es más difícil contrastar la segunda inercia principal, especialmente, si suponemos que la primera inercia principal es significativa. Para superar este inconveniente recurrimos, de nuevo, a los métodos de cálculo con ordenador.

Contraste de hipótesis: simulación de Monte Carlo

La *simulación de Monte Carlo* nos permite, una vez planteada una hipótesis sobre una población, y una vez conocido cómo se muestrearon los datos, calcular la distribución del estadístico de contraste suponiendo que la hipótesis nula sea cierta. Por ejemplo, supongamos que queremos averiguar la significación de las dos inercias principales de los datos sobre el nivel de lectura. La hipótesis nula es que no existe asociación entre filas y columnas. En este caso, el muestreo no se realizó como con los datos sobre los autores, cuando dentro de cada libro se muestrearon textos (la analogía aquí podría ser un muestreo dentro de cada grupo educativo). Por el contrario, en este caso se obtuvo una muestra de 312 personas, y luego se averiguaron sus niveles de educación y de lectura. Por tanto, la distribución de los grupos educativos también es aleatoria. En consecuencia, debemos generar muestras de 312 personas a partir de una distribución multinomial que corresponda a toda la matriz, fila a fila, o columna a columna. En cada una de las 15 celdas de la tabla, las probabilidades esperadas son iguales a los productos de las masas. Si suponemos que la hipótesis nula es cierta, estas probabilidades esperadas definen un vector de 15 probabilidades que utilizaremos para generar muestras multinomiales simuladas de tamaño 312. En la tabla de la imagen 25.4 mostramos la tabla original y dos muestras simuladas (en total, generamos 9999 tablas). Tenemos, pues, un total 10000 conjuntos de datos (la tabla original y 9999 muestras simuladas). Llevamos a cabo el AC y calculamos las inercias principales de cada una de las tablas. En la imagen 25.5 mostramos un diagrama de dispersión con todos estos resultados, en la que hemos señalado el punto correspondiente al par de valores de la tabla de contingencia original. Observamos que solamente 12 valores de los 10000 son mayores que la primera inercia principal observada, por tanto, estimamos su valor p en 0,0012. Para la segunda inercia principal, hay 593 valores simulados mayores

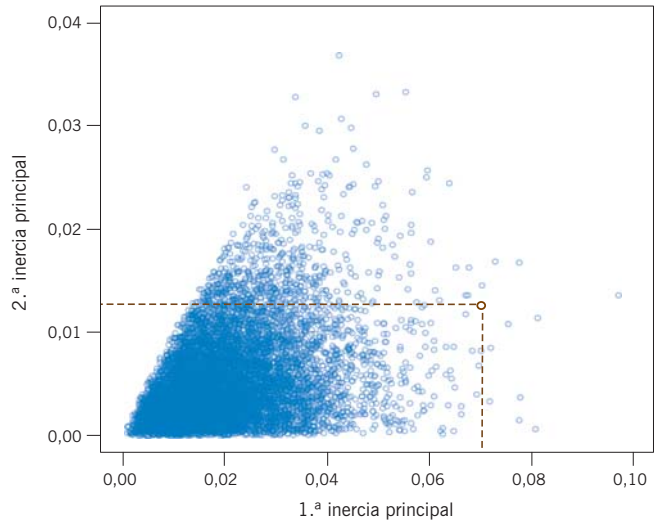


Imagen 25.5: *Diagrama de dispersión de las inercias principales del AC original y de las 9999 simulaciones de la tabla de contingencia de 5 × 3, bajo la hipótesis nula de que no existe asociación entre filas y columnas (en la imagen 25.4, se muestran dos de estas simulaciones). Las inercias principales observadas se han señalado con un círculo mayor (○) y líneas discontinuas*

que el observado, su valor *p* será de 0,0593. A un nivel del 5% solamente el primero es significativo. Al mismo tiempo, en cada simulación calculamos la inercia total: hay 19 valores simulados mayores que la inercia total observada de 0,08326. Por tanto, el valor *p* es de 0,0019, que es nuestra estimacion de Monte Carlo para la prueba χ^2, comparado con el valor *p* de 0,0035 calculado a partir de la distribución χ^2 habitual.

Pruebas de permutaciones

Las pruebas de permutaciones (o pruebas de aleatorización) son ligeramente distintas de los procedimientos de automuestreo y de Monte Carlo que acabamos de describir. Por ejemplo, en la ampliación de la parte central del mapa de la imagen 10.7 sobre autores, observamos que los pares de libros del mismo autor se hallaban próximos. Parece poco probable que ello se deba al azar, pero ¿cuál es el valor de probabilidad, o valor *p*, asociado con este resultado? Vamos a ver cómo responder esta pregunta. En primer lugar, calcularemos una medida de proximidad global entre los pares libro-autor. Una medida de proximidad inmediata es la suma de las seis distancias entre los pares libro-autor, lo que en nuestro caso da 0,4711. A continuación, generemos todas las posibles combinaciones de los seis pares libro-autor. Existen $11 \times 9 \times 7 \times 5 \times 3 = 10395$ maneras distintas de acomodar los pares libro-autor en grupos de seis. Las sumas de las seis distancias de los pares libro-autor de cada una de estas combinaciones definen la distribución del estadístico de contraste de la *prueba de permutación*. Tiene una media de 0,8400 y una desviación típica de 0,1246 (en el apéndice de cálculo B, se muestra el histograma de esta distribución). Resulta que no hay ninguna combinación de los seis pares libro-autor con una suma de distancias menor que el valor observado en el mapa del AC. Por tanto, el valor *p* de la prueba que afirma que los pares de

textos del mismo autor están próximos es $p = 1/10395$, es decir menor de 0,0001, ¡un valor altamente significativo! Realizamos pruebas de permutaciones similares de forma separada para consonantes y para vocales (imágenes 21.1 y 21.2) y obtuvimos valores p iguales a 0,0046 y a 0,0065, respectivamente. Por tanto, son las consonantes y las vocales las que explican las diferencias entre autores (aunque las vocales tengan menos inercia en total). En ACC es habitual llevar a cabo pruebas de permutaciones para contrastar la hipótesis de que el espacio restringido explica una parte significativa de la inercia. El estadístico de contraste consiste en el cociente entre la inercia restringida y la no restringida. Para ello llevamos a cabo un gran número de ACC en los que en cada análisis permutamos al azar las filas de la matriz de la variable explicativa. De esta manera obtenemos la distribución del estadístico de contraste según el supuesto de que la hipótesis nula es cierta (véase el apéndice de cálculo B).

RESUMEN:
Consideraciones sobre estabilidad e inferencia

1. Realizamos el análisis de *estabilidad* con los datos de que disponemos. Para ello, analizamos la influencia de cada fila o columna sobre el mapa. La estabilidad la valoramos (a) estudiando las contribuciones de filas y de columnas a la configuración de los ejes principales y (b) llevando a cabo AC en los que eliminamos determinados puntos o grupos de puntos de los datos, para ver así su efecto sobre la configuración del mapa.
2. Si conocemos cómo se muestrearon los datos, el *automuestreo* nos permite obtener réplicas de la muestra de datos. Si en el diseño del muestreo se fijaron los valores marginales de filas y columnas, las réplicas obtenidas por automuestreo deben tener los mismos valores marginales.
3. En el *automuestreo parcial*, proyectamos los perfiles de filas y/o columnas de las matrices replicadas en el mapa del AC original como puntos adicionales. Podemos sintetizar la distribución de estas proyecciones dibujando los perímetros convexos o las elipses de confianza.
4. También podemos analizar la configuración de los datos a partir de aproximaciones teóricas basadas en determinadas suposiciones sobre la distribución de la población. Por ejemplo, el método delta y la teoría asintótica se basan en la aproximación normal a la distribución multinomial.
5. Para contrastar hipótesis, podemos utilizar los métodos de *Monte Carlo* y las *pruebas de permutaciones*. Suponiendo que la hipótesis nula es cierta, utilizamos estos métodos para generar datos que nos permiten simular (o calcular exactamente) la distribución de los estadísticos de contraste elegidos. A partir de estas distribuciones podemos calcular los valores de p.